Characterizing Prediction Model Responses to Attack Inputs: A Study with Time-Series Power Consumption Data

Srinidhi Madabhushi and Rinku Dewri

University of Denver, Denver CO 80208
nidhi.madabhushi@du.edu, rdewri@cs.du.edu

Abstract. In time series anomaly detection systems, a prediction model plays a pivotal role in identifying anomalies. This is because the gap between the observed and predicted values contributes to determining whether a time instance is anomalous or not. Following the traditional approach of choosing the best prediction model in terms of accuracy may lead to unexpected results in the detection system. A prediction model performing well on test data due to its generalization capability might inadvertently adapt to attack data as well. Therefore, it becomes important to understand the prediction model's properties to determine if it accurately identifies a target attack or adapts to it. To compare the behavior of these prediction models under attacks, we propose a framework for analyzing prediction responses to specific inputs, similar to control system analysis. We identify four model responses that can be used to assess five prediction models- MLP, LSTM, GRU, TCN and CNN-LSTM, and gain insights into their behavior. Our results show that each prediction model has distinct behavior and choosing a model by the prediction accuracy alone is insufficient. It emphasizes the need for nuanced model selection and highlights the potential for developing more effective anomaly detection systems.

Keywords: Power consumption · Anomaly detection · Time series prediction · Neural networks · Attacks

1 Introduction

Power grids have become more complex and interconnected with the advancement of Industry 4.0, which involves the integration of smart devices, IoT devices, renewable energy sources and other control technologies. There is a higher risk of various anomalies and cyberattacks with the increasing complexity of the grid. Anomaly detection systems play an important role in ensuring the stability, security and reliability of modern power grids by identifying anomalous instances, thus allowing operators to promptly enforce the necessary contingency measures. The application of anomaly detection varies by the target detection task to be performed in one of the four power grid processes which are generation, transmission, distribution and consumption [12]. Assessing the robustness

and performance of these detection systems is important to ensure the detection accuracy and safety of the grid. Measuring detection performance for power grid applications is challenging due to the lack of anomalous data to train the detection systems [12, 18]. This makes it more difficult to understand whether a system can detect new types of attacks and anomalies. While modern machine learning approaches can generalize better with new data, this adjustment can lead to the inability to detect attack instances and be categorized as the new norm. Prior research indicates that deep learning models often exhibit a tendency to generalize excessively to new data such that it fails to identify anomalies [3, 7, 14]. It has also been shown that certain configurations of undetected demand manipulation attacks in power grids are feasible with detection systems that use state-of-the-art neural networks as the forecasting model, due to their generalization property [11]. Since the prediction model is crucial in a detection system, it is necessary to understand the model behavior during undetected attacks.

As shown in Figure 1, there are various components in an anomaly detection system for power consumption, such as the prediction model, scoring mechanism and thresholding mechanism. Our research targets the prediction model, specifically neural networks that are commonly used for time series forecasting. We discuss some of the current prediction models that are used in anomaly detection systems, in addition to pointing out some of the limitations and robustness of the models to set up this work's motivation in Section 2. We craft attack inputs that represent different types of attacks on power consumption in Section 3.2. We construct four properties that the prediction values undergo during an attack by using the principles of control systems and discuss this in Sections 3.3 and 3.4. We provide insights into each model's adaptability and generalization characteristics, allowing us to compare and assess the models in terms of their responses in Section 4.1. Our results reveal that popular time series neural network models, such as the Multi-Layer Perceptron (MLP), Long Short-Term Memory Network (LSTM) and Gated Recurrent Unit (GRU), tend to either generalize too extensively or memorize training data patterns, both of which can pose challenges in detecting the target attack. Interestingly, when combining a Convolutional Neural Network with LSTM (CNN-LSTM) , the model does not readily adjust to certain attacks while exhibiting stable behavior. We discuss the potential use of this assessment in Section 4.2. We provide future directions for this framework and conclude in Section 5.

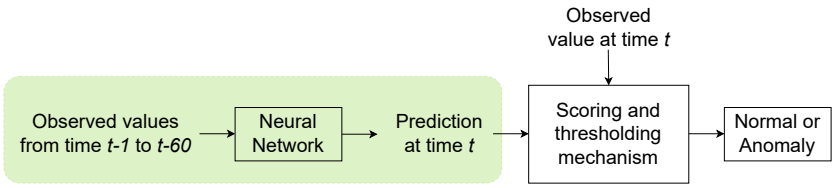

Fig. 1: This paper focuses on an anomaly detection system's prediction process

2 Background and Motivation

Anomaly detection for power consumption time series involves a prediction model which is crucial to accurate detection. As illustrated in Figure 1, a prediction model is used to forecast the power consumption at time t using the historical consumption values from say, the last 60 minutes ($t-1$ to $t-60$). The difference between the predicted and observed values is quantified as an anomaly score by a scoring mechanism. If the score appears to be less than a threshold, then it is tagged as normal and otherwise, it is flagged as an anomaly. In this section, we discuss the prediction models that are used in anomaly detection for power consumption, their limitations, robustness against attacks and present the design goals for this study.

2.1 Prediction Models in Anomaly Detection

Prediction-based anomaly detection systems use a forecasting model to predict consumption using recent historical power consumption values. The choice of the prediction model is made by evaluating either the accuracy of the prediction model itself or the overall detection performance using the prediction model. Different types of models can be used for prediction, such as statistical methods, regression-based techniques and neural networks. Due to the complex nature of the time series data, which often involves non-linear relationships between variables, traditional models cannot efficiently capture this intricate behavior. Neural networks on the other hand can automatically learn and extract relevant features from raw data instead of requiring manual feature engineering and transformation. Neural network architectures like recurrent neural networks and long short-term memory networks can capture sequential dependencies and memory effects in time series data, making these models outperform traditional time series models in terms of capturing temporal dynamics. In this work, we perform the prediction assessment for five neural network models from the literature that are predominantly used for prediction and detection tasks in the power grid – Multi-layer perceptron (MLP) [5], long short-term memory network (LSTM) [4], gated recurrent unit (GRU) [19], temporal convolutional network (TCN) [15] and convolutional neural network with LSTM (CNN-LSTM) [9].

2.2 Limitations and Robustness of Prediction Models

When considering neural networks for power consumption anomaly detection systems, they offer significant advantages along with certain limitations. Training a deep learning model can be computationally intensive. Additionally, the black-box nature of neural networks makes their internal workings not easily interpretable, hindering the transparency of anomaly detection systems overall. This makes it challenging to understand how and why certain anomalies are detected or undetected. Overfitting is a concern that arises when neural networks learn noise in training data rather than genuine patterns. This can lead to poor generalization on unseen data, especially since a dataset with anomalies in power

consumption is difficult to acquire. On the other hand, choosing to generalize a model to new patterns comes with its own disadvantage of its inability to detect anomalies and consider them as normal. Therefore, careful attention is required to balance the network's ability to memorize and generalize.

Neural networks can tolerate adversarial inputs and give predictions by adapting to malicious input data [8]. The current literature demonstrates this problem for the vision and classification tasks and provides techniques to handle such adversarial data. But when it comes to time series data or any sequential numerical data, we have a research gap in addressing and solving this problem for a prediction task. By quantifying the impact of an undetected attack from a detection system, we can assess its robustness to attack inputs [11]. If an undetected attack has an impact on the target application, which in our case is the power grid, then we must improve the performance of the detection system and find which sub-process within the detection system can be improved. Since the scoring and thresholding mechanisms are some form of a statistical value over the predictions, understanding and improving the performance of the prediction model becomes the first step towards robustness against adversarial inputs. In this work, we observe the response of prediction models to attack inputs to compare five different neural networks and discuss their characteristics.

2.3 Our Design Targets

With the goal of understanding and comparing the prediction responses and behavior of different neural network models under adversarial conditions, we come up with the following desirable targets for this work.

- Craft attack examples for capturing differences between models.
- List properties (or states) that align with control system analysis to study prediction model responses to attack inputs.
- Discuss how these properties can be used for comparison when choosing a prediction model for a power consumption anomaly detection system.

3 Prediction Model Responses in Adversarial Context

In the realm of predictive modeling, the assessment of model properties becomes particularly critical when deployed within adversarial contexts. The dynamic and complex nature of attack scenarios necessitates a comprehensive understanding of how prediction models respond to external perturbations, ensuring their reliability and robustness. This section delves into the definition and identification of prediction properties within adversarial contexts. We adopt a closed-system approach to establish the prediction properties. A closed system refers to a controlled and isolated environment that resists influences to change without being affected by external factors [17]. We assume that the prediction model is in a closed system as it does not update its internal parameters during the prediction phase. Since we are focused on the prediction behavior only under adversarial settings, the input to the model is crafted to represent a specific attack and the prediction output is used for the analysis.

3.1 Dataset and Model Training

We use the UMass Smart* apartment dataset from the UMass Trace Repository[1], which consists of minute-level power consumption data for 114 single-family apartments for the year 2016. These apartments are located within the same geographical area. To obtain grid-level power consumption, we sum the power consumption values of all apartments for each minute and use only the aggregated power consumption for detecting demand manipulation attacks. There are 60 missing values in the dataset that are interpolated using the linear interpolation method. The first occurrence of any duplicate values is dropped. For the prediction model training, data from the first six months (January to June) is used as the training data, July to September as validation data, and October to December as testing data.

For prediction model training, the five neural network models chosen are implemented with the same number of layers as used by the authors in the literature – MLP [5], LSTM [4], GRU [19], TCN [15], and CNN-LSTM [9]. However, the number of nodes, epochs, and batch size are selected by searching for the best combination of parameters based on the least mean squared error. To facilitate reproducibility, we have made the code used for the model search using different parameters and training publicly available[2].

3.2 Adversarial Settings

Crafting attacks for power consumption prediction models is crucial due to the vulnerabilities that these examples expose within the neural network, giving output that directly impacts the attack detection rate of an anomaly detection system [11]. In this work, we focus on creating examples that represent a demand manipulation attack. Demand manipulation attacks (MAD) occur when an adversary manipulates the demand of the power grid from the utility side using consumer devices. These devices are assumed to be in a residential setting and are manipulated either directly by the attacker or by the consumer. Such attacks can be performed by controlling a botnet of devices that can manipulate the power demand much faster than the power plants can react [6]. When an attacker has access to various high-wattage IoT devices, they can synchronously switch them on and off, which leads to the disruption of the power grid [16]. The attacker can also influence the behavior of the consumers by sending false messages to fake maintenance shutdown alerts, suggesting the consumers to use appliances during peak consumption periods [13].

Threat Model. We assume that an adversary has unauthorized access to IoT devices across different apartments. These devices could include smart thermostats, water heaters, and electric vehicles, all of which consume significant

[1] https://traces.cs.umass.edu/index.php/Smart/Smart
[2] https://github.com/NidhiMadab/prediction-model-training/

amounts of electricity. By exploiting vulnerabilities in these devices or the mobile applications used to control them, the adversary can increase their power usage simultaneously [6, 16]. Historical attacks like the Mirai botnet have demonstrated the feasibility of compromising large numbers of IoT devices. The Mirai botnet infected approximately 600,000 devices, primarily using default or weak passwords to gain control [1]. More recent attacks, such as the Hajime worm [2] and VPNFilter [10], have shown the continued vulnerability of IoT devices to large-scale exploitation. Assuming an attacker gains control over 600,000 smart thermostats, each controlling two 1 kW air conditioners, the adversary could manipulate up to 35 GW of power. Scaling this up to include other high-wattage devices, such as electric water heaters (typically 4.5 kW) and electric vehicles (charging at 7 kW), the potential power under control could easily exceed 35 GW [16]. When the adversary creates a coordinated surge in power demand using these high-wattage devices, it could destabilize the grid, leading to significant operational, financial, and infrastructural impacts.

Types of Attacks. To study the properties of the prediction model under adversarial settings, we consider three types of attacks on the power consumption data, namely *point attack*, *shift attack* and *ramp attack*. Figure 2 shows visual examples of each type of attack. These attacks are used to achieve demand manipulation by making alterations to the consumption data of multiple apartments such that there is an increase or decrease in the power grid demand. In this work, we consider only increasing the consumption values to avoid negative values during the experiments. Since we assume that the adversary has access to various high-wattage IoT devices in each apartment, each type of attack is performed by considering the average increase of wattage for every apartment. For example, if we consider the average increase of wattage per apartment to be 4 kW, then the total increase in the grid-level wattage across 114 apartments will be 456 kW.

Point Attack. A point attack is performed by choosing a specific data point (or timestamp) in the power consumption data where we want to introduce the point anomaly. We then decide on the magnitude of the spike, which we alternatively call an *adder*. This attack represents an abrupt surge in consumption across all apartments, resulting in a sudden increase in power grid demand.

Shift Attack. A shift attack involves introducing intentional shifts or offsets, usually by uniformly adding a constant value (adder) to all the power consumption readings in a time period, thus creating collective anomalies. This can trigger grid frequency instability or even lead to line failures when the demand reaches a critical value [16].

Ramp Attack. A ramp attack involves gradually increasing or decreasing the power consumption values over a specific time period to simulate a continuous change in consumption behavior. This also comes under the category of collective

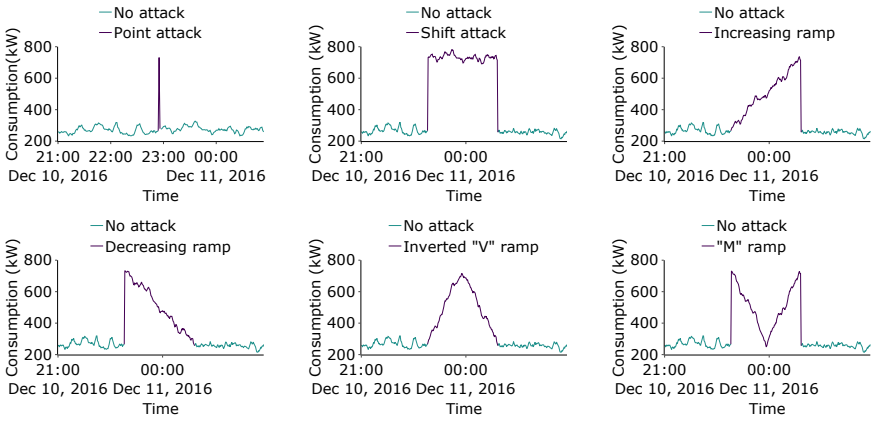

Fig. 2: Examples of point, shift and different ramp attacks

anomalies. This attack is introduced by determining the overall desired gain for a chosen time window and then computing how much to add or subtract from the original consumption values at each time stamp. However, if we want to mimic a decreasing ramp, we perform a point attack first and then subtract from the consumption to make a ramp. To incorporate more complexity, we introduce a blend of actions beyond solely increasing or decreasing the values. This involves a combination of increasing followed by decreasing consumption (an inverted "V" attack), as well as the reverse scenario ("M" attack), resulting in four variations of the ramp attack shown in Figure 2. These attacks are particularly crafted to gradually turn on or off devices from multiple apartments to mimic a slow increase or decrease in demand as smaller changes in demand are difficult for a detection system to capture. When multiple apartments simultaneously reach a high demand value, it can lead to any of the adverse outcomes associated with MAD attacks or even incur increased operating costs.

3.3 Types of Model Responses

The behavior of the prediction models to adversarial events can be characterized by four distinct responses or states: *steady state*, *transient response*, *conditional translation* and *post-translation response*. These states play a crucial role in understanding the dynamic nature of a prediction model and how different each model is in terms of their response to perturbations. Each state represents a specific phase of an ongoing attack, offering insights into how a model adapts to perturbations and eventually stabilizes. In this section, we will delve into each state's definition and significance as the predicted values transition from their initial equilibrium to a new state and ultimately return to their original balance.

State 1: Steady State. The steady state is observed when the differences between the observed and predicted consumption values remain constant over

time. While in this state, the predictions closely follow the shape of the observed time series. This state does not necessarily indicate that the differences will be small or closer to zero, but suggests that the change in the differences will be approximately zero. We can observe in Figure 3 that the steady state occurs during both non-attack and attack periods. The steady state is represented as follows, where dt represents a time window.

$$\frac{d(predicted - observed)}{dt} \approx 0 \tag{1}$$

State 2: Transient Response. A transient response pertains to the behavior of the prediction values immediately after experiencing an abrupt change or disturbance in the observed consumption values. It is a temporary and non-steady state behavior that occurs during the prediction model's adjustment period. The prediction values are often characterized by fluctuations and oscillations before eventually reaching a new steady state. The transient response can be observed in Figure 3 taking place from 3:00 AM to 4:00 AM.

State 3: Conditional Translation. When the observed values are perturbed for a prolonged time period, the prediction model generalizes the predicted values to the observed values. However, some models like the convolutional neural networks (CNN) come with inherent properties such as translation invariance, which does not let them scale to any arbitrary perturbation. They are invariant to certain translations in the observed values, thus saturating the predictions to some constant value. This property can be observed in Figure 3, where the predictions during the entire attack duration from 3:00 AM to 5:00 AM do not adjust to the translation applied after a certain point. Mathematically, conditional translation can be represented as follows.

$$prediction = \begin{cases} observed + f(x), x < T. \\ c, x \geq T. \end{cases} \tag{2}$$

where x is the translation applied to the observed values, $f(x)$ suggests how much the observed values will change, T is a large translation value beyond which the model is invariant to translations and c is a saturation value reached by the prediction model due to translation invariance.

State 4: Post-translation Response. The state when the prediction model is adjusting to the original steady state after the translation or perturbation to the original values has ended is called the post-translation response. This state also shows a transient behavior as the prediction values gradually return to the original state. Figure 3 shows the post-translation response after the attack ended at 5:00 AM.

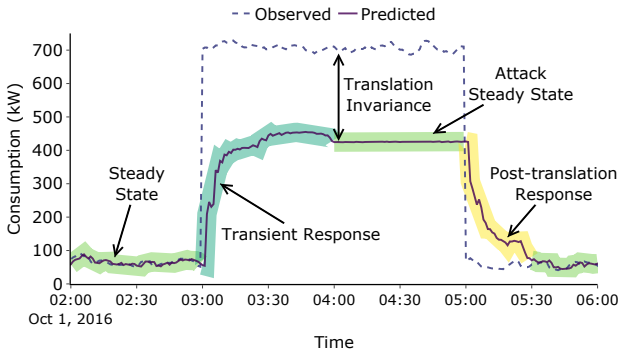

Fig. 3: Different states of the prediction values during a shift attack from 3:00 AM to 5:00 AM with a 650 kW (5.7 kW average wattage increase per apartment)

Justification of Selected States. The proposed four states encapsulate the dynamics of prediction models within an adversarial setting. The reason we have identified these states is because they connect with well-established ideas in control system analysis. This helps us understand how prediction models adjust and react to changes. The steady state, recognized as the equilibrium point, is a fundamental concept in various domains, and its application to prediction models provides a logical basis for assessing model stability and performance. The transient response state captures the immediate adjustments following a perturbation, aligning with the behavior of dynamic systems transitioning from one state to another. The introduction of a state such as the translation invariance, signifies the model's generalization capability to new inputs of varying magnitudes. Lastly, the return to the original steady state acknowledges the cyclic nature of the model's ability to revert to its initial behavior once the attack subsides. This rationale is further fortified by the application of control system principles, which brings out the similarities between prediction models and dynamic systems. By connecting these states with well-known theories, we build a strong foundation for understanding how predictions react and adapt in adversarial circumstances.

3.4 State Identification

The beginning of a state can be detected by looking for point-based or short-term changes that occur during the state transition. When a sudden increment in the observed values occurs, the predictions try to immediately adjust to the new values due to the model's generalization property. Therefore, this change can be captured by using the variance of the differences between the observed and predicted values. During the non-attack steady state, differences tend to approach zero as prediction values align closely with observed values. In Figure 4, the graph illustrates the differences and the five-minute rolling window

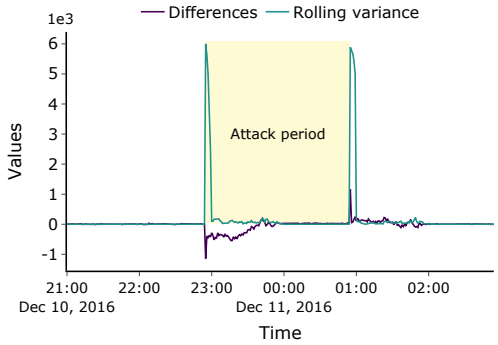

Fig. 4: Variance of the differences between the predicted and observed values scaled down to 2.7% of the original magnitude

variances of these differences. We can recognize the start and end of the attack by identifying two distinct peaks on December 11 at 22:55 and on December 12 at 00:54 respectively. Conversely, for the time span outside the attack window, which denotes the non-attack phase, the variances are closer to zero. Therefore, by implementing a threshold over the variances (discussed in Section 4), we have the capability to identify the transitions between the non-attack phase to the attack phase as well as the subsequent return to the non-attack phase. We rely on using a hierarchical approach when identifying the states. First, we capture the transient and post-translation responses by using variances. Once the onset of a response is identified, we also use visual analysis to understand its behavior. Each distinct behavior becomes apparent in a graphical representation (as seen in Figure 3) and allows the user to interact with the prediction values directly allowing for the state duration to be identified accurately. By using statistical and visual analysis in conjunction, we are able to give more insights into the behavior, duration of the states and its relation with other parameters like input window size and attack window size.

4 Evaluation Settings and Results

In this section, we delve into the evaluation settings, results of the proposed framework and compare the prediction models using the four states. We perform all six types of attacks (discussed in Section 3.2) at 100 distinct time points in the test data. We apply six different adders for each apartment, representing various high-wattage devices switching on at those chosen points going from an average of 4 kW to 14 kW per apartment in increments of 2 kW. Except for the point attacks, the other attack types extend over a two-hour duration. These experiments are designed to analyze the prediction behavior at different times

of the day and on various days. For each experiment, attack type and adder combination, we perform the prediction using the trained model configuration.

For identifying the beginning of transient and post-translation responses, we apply a threshold to the differences between predicted and observed values, as discussed in Section 3.4. During the threshold selection process, we calculate the number of experiments in which state transitions are identified outside of the attack window (false positives) and the number of experiments where state transitions remain unidentified within the attack window (false negatives). The optimal threshold is the one that minimizes the number of experiments with false positives and false negatives. After we perform this search, we finalize the thresholds to be 778 for MLP, 22724 for LSTM, 2298 for GRU, 486 for TCN and 830 for CNN-LSTM.

4.1 Model Comparison

Steady State. All the models follow a steady state when there is no attack, and during an attack, they show similar behavior in terms of reaching a new steady state. During a point attack, the attack duration is only one minute for a single spike, making the occurrence of a new steady state during the attack not applicable. For an increasing ramp and an inverted "V" ramp attack, the prediction values remain in a steady state throughout the entire attack because smaller increments in the adder do not trigger a transient response between the normal and attack scenarios. In contrast, for the decreasing ramp, "M" ramp, and shift attacks, a new steady state is observed after the transient response occurs. In conclusion, the occurrence of the steady state does not differ across the models and behaves similarly among them.

Transient and Post-Translation Responses. Since the transient and post-translation responses gradually adjust to the sudden changes in the observed data, the initial phase of both states can be detected using the selected threshold over the variance of differences. We analyze how many experiments out of the 100 for each model have the entire duration of the transient response detected and summarize the observations below:

- For point, shift, and "M" ramp attacks, we are able to observe both responses in all models due to the sudden increase and decrease in consumption at the beginning and the end of the attack.
- When an attack begins with an increasing ramp, the transient response is missing in all models, but a post-translation response is observed. When an attack begins with a decreasing ramp, the post-translation response is missing, but the transient response is observed.
- Since an inverted "V" ramp attack is a combination of increasing and decreasing ramps, both responses are missing across all models.

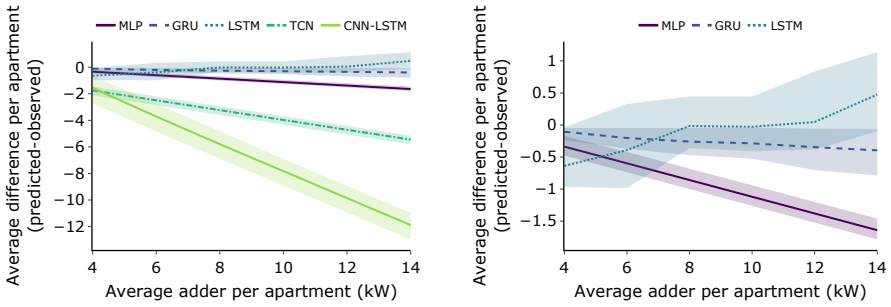

Fig. 5: The range (shaded) and mean (line) of the average differences between predicted and observed values during the attack for each adder across 100 experiments. The left figure shows all models, while the right figure provides a zoomed-in view of the MLP, GRU, and LSTM models for clarity.

Translation Invariance. Figure 5 illustrates the range of differences between predicted and observed consumption values, averaged over the attack window, for each experiment and average apartment adder value. A negative trend in the differences indicates that the predicted values are lower than the observed ones. This suggests that as the adder increases, the model fails to generalize predictions closely aligned with the observed consumption values. Conversely, an increasing trend suggests that the model over-predicts consumption compared to observed values. For MLP, LSTM, and GRU, all experiments are around zero, indicating strong generalization. Additionally, LSTM has a subtle increasing trend, indicating over-prediction of consumption values as the adder increases. The most notable behavior in the plot is exhibited by CNN-LSTM, which displays invariance for increasing adders. CNN-LSTM gradually resists adjusting to higher observed consumption values, maintaining relatively stable predictions even when faced with unusually high consumption in the observed values. TCN also exhibits invariance to some extent, but not as aggressively as CNN-LSTM.

State Duration. The duration of each state is directly influenced by the model's input size. The transient and post-translation responses are observed for 60 minutes if they are present, depending on the attack type. When the input size is reduced from 60 minutes to 30 minutes, the duration of both responses also decreases from 60 to 30 minutes. Similarly, the duration of the attack steady state is affected, which impacts the time duration between the end of the transient response and the start of the post-translation response. This pattern holds when the input duration is extended from 60 to 120 minutes. In this case, the transient and post-translation responses extend over a duration of 120 minutes, leaving the remaining attack period as the attack steady state.

4.2 Impact in Practice

In this section, we discuss how state characterization aids in early attack detection and provide guidelines on how to assess models using the framework.

Early Detection of Attacks. Each state occurs during different attack phases – the transient response indicates the beginning of the attack, the post-translation response marks the end, and the attack steady state exists between the beginning and end. Early detection relies on observing the unstable but gradual adjustment to attack data. We perform six attack variations on the WSCC 9-bus system using PowerWorld simulator to assess frequency instability in the worst-case scenario from [16] (low inertia and 30% load increase). The WSCC system operates in a low inertia setting, with generators 2 and 3 having inertia constants (H) of 5 and 10 seconds, respectively. The minimum time unit in the apartment dataset is minute with 120-minute attack duration, but we use a 120-second duration for attack simulations to align both the time units for consistency. Therefore, a total of 30 MW is increased over 120 seconds starting at 5 seconds to match the dataset's time unit. Figure 6 shows the frequency in all load buses during the attack. Point and inverted "V" ramp attacks are the only variations within the safe frequency interval of 58.2 Hz to 61.2 Hz. All other attacks lead to frequency deviations beyond this interval. In shift, decreasing ramp, and "M" ramp attacks, the frequency drops below 58.2 Hz five seconds after the attack begins. The transient response helps detect these attacks as they begin (within one time unit) due to a spike in the variance of differences. In cases where attacks are obscured without frequency instability (point, increasing ramp, inverted "V" ramp), we detect them using the translation invariance property of convolutional neural network models (TCN and CNN-LSTM). This is particularly useful for the increasing ramp case, as we can detect the attack before the frequency exceeds 61.2 Hz. Therefore, state characterization is crucial for early detection, enabling necessary contingency actions before grid failures and identifying attacks that may not cause physical impacts but increase operating costs.

Guideline for Prediction Model Selection. In this section, we discuss the steps on how state characterization can be applied to select better prediction models for anomaly detection in power consumption.

1. Crafting different types of attacks is important to assess each prediction model in terms of their responses to such inputs. Therefore, the first step is to create attack profiles that are feasible in power grids and evaluate their impact on the grid with different attack intensities (i.e. injected wattage).
2. After tuning the hyperparameters of the prediction models, assess their performance on the test data to ensure that the models can generalize to data outside the training process. Then, perform state characterization by providing attack inputs to the models and observing their prediction responses.

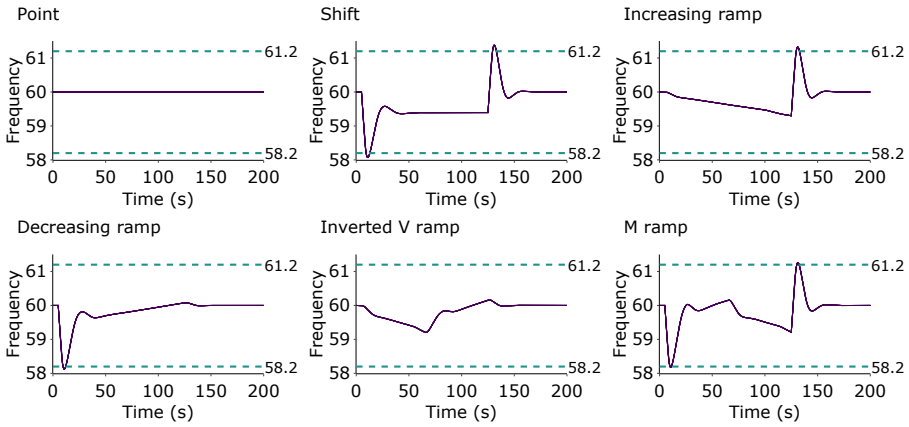

Fig. 6: Frequency of all load buses in the WSCC system for attack types, ignoring generators' frequency cut-off limits (dashed lines)

3. A model with a slow transient response to a sustained increase in consumption will enable early detection of attacks, as this state shows the maximum gap between the attack and predicted responses at the onset of an attack.
4. The post-translation response identifies the end of an attack, signaling when to revert contingency measures, if applied. This period is critical as it reveals the duration of an unstable response that could be exploited for subsequent attacks. Hence, a model that exhibits a gradual decrease in the predicted consumption can aid in better identification of this state.
5. Translation invariance supports anomaly detection by increasing the gap between predicted and observed responses. For example, convolutional neural networks, being inherently translation-invariant, exhibit better detection and false alarm rates compared to other time series models [11]. Therefore, it's important to choose models that demonstrate this characteristic.

Evaluating models using the four states provides insights into their behavior under various attacks. This systematic analysis reveals how models adapt to perturbations, distinguishing stable models from those prone to overfitting during attacks. In machine learning, generalization is critical for a model to perform accurately on unfamiliar data. However, in anomaly detection, models must strike a balance between generalization and avoiding overfitting. Overgeneralization can lead to overlooked anomalies, while excessive memorization can result in unreliable predictions [11]. Using state transitions as indicators for detecting potential anomalies and understanding the duration of each state aids in an early intuition of anomaly detection performance just by using the prediction model responses, rather than evaluating the detection performance itself, which is done later in the anomaly detection system.

Prediction Performance Improvement. There can be cases where we are required to use only a specific model and cannot necessarily choose between multiple models. For example, in edge detection, it is preferred to use lightweight models compared to models that require more memory or take more time to make predictions. Let's say we want an MLP to resist generalization to higher wattage values. An easy way to limit generalization would be to include an element that constrains the predictions coming from the neural network to be under a cut-off value. Another technique is to use adversarial training [8]. By training the model on a small proportion of crafted inputs where the input has an ongoing attack while the output is restricted to a smaller wattage value, the model will stop generalizing easily to attack data. We can additionally control the loss during the adversarial training process by penalizing the model if it generalizes to high-wattage increases. We can also control generalization by limiting the number of training epochs or changing other hyperparameters. However, careful analysis is required to ensure that limiting generalization does not affect the prediction performance on non-attack data.

5 Conclusion and Future Work

This study assesses the prediction behavior of different neural networks for anomaly detection in power consumption data using a control system approach. Assuming a closed system where the prediction model remains unchanged during forecasting, we observe four proposed states during anomalies or attacks. In control system analysis, steady state and transient responses are analyzed for various inputs. Similarly, we construct adversarial examples to study and compare prediction model responses, providing insights into their performance in a detection system. The proposed states offer a framework to enhance detection performance by identifying model behaviors during different attack phases. Our results show that MLP, LSTM and GRU adapt to new data, including attack instances, while CNN-LSTM and TCN exhibit all states throughout the attack and resist to wattage injections. Recognizing these properties allows us to customize model architectures for better attack identification. Future work can focus on defining desired properties for prediction models and evaluating different neural network architectures based on those criteria. In summary, our study enhances the understanding of prediction model behavior and offers insights for designing more effective anomaly detection systems, contributing to the evolving field of artificial intelligence.

References

1. Aloqaily, M., Kantarci, B., Mouftah, H.T.: Trusted third party for service management in vehicular clouds. In: 2017 13th International Wireless Communications and Mobile Computing Conference. pp. 928–933 (2017)
2. Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., Durumeric, Z., Halderman, J.A., Invernizzi, L., Kallitsis, M., Kumar, D., Lever, C.,

Ma, Z., Mason, J., Menscher, D., Seaman, C., Sullivan, N., Thomas, K., Zhou, Y.: Understanding the mirai botnet. In: Proceedings of the 26th USENIX Conference on Security Symposium. p. 1093–1110. SEC'17, USENIX Association (2017)
3. Bergman, L., Cohen, N., Hoshen, Y.: Deep nearest neighbor anomaly detection. arXiv preprint:2002.10445 (2020)
4. Chahla, C., Snoussi, H., Merghem, L., Esseghir, M.: A deep learning approach for anomaly detection and prediction in power consumption data. Energy Efficiency **13**(8), 1633–1651 (2020)
5. Chou, J.S., Telaga, A.S.: Real-time detection of anomalous power consumption. Renewable and Sustainable Energy Reviews **33**, 400–411 (2014)
6. Dabrowski, A., Ullrich, J., Weippl, E.R.: Grid shock: Coordinated load-changing attacks on power grids: The non-smart power grid is vulnerable to cyber attacks as well. In: 33rd Annual Computer Security Applications Conference. pp. 303–314. Association for Computing Machinery (2017)
7. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: International Conference on Computer Vision. pp. 1705–1714. IEEE Computer Society (2019)
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint:1412.6572 (2014)
9. Kim, T.Y., Cho, S.B.: Predicting the household power consumption using cnn-lstm hybrid networks. In: International Conference on Intelligent Data Engineering and Automated Learning. pp. 481–490. Springer (2018)
10. Largent, W.: New vpnfilter malware targets at least 500k networking devices worldwide. Cisco (2018), https://blog.talosintelligence.com/vpnfilter/
11. Madabhushi, S., Dewri, R.: On the impact of model tolerance in power grid anomaly detection systems. In: International Conference on Information Systems Security. pp. 220–234. Springer (2022)
12. Madabhushi, S., Dewri, R.: A survey of anomaly detection methods for power grids. International Journal of Information Security pp. 1–34 (2023)
13. Raman, G., Peng, J.C.H., Rahwan, T.: Manipulating residents' behavior to attack the urban power distribution system. IEEE Transactions on Industrial Informatics **15**(10), 5575–5587 (2019)
14. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Conference on Computer Vision and Pattern Recognition. pp. 14318–14328. IEEE (2022)
15. Shaikh, A.K., Nazir, A., Khalique, N., Shah, A.S., Adhikari, N.: A new approach to seasonal energy consumption forecasting using temporal convolutional networks. Results in Engineering **19**, 101296 (2023)
16. Soltan, S., Mittal, P., Poor, H.V.: BlackIoT: IoT Botnet of High Wattage Devices Can Disrupt the Power Grid. In: 27th USENIX Security Symposium. pp. 15–32. USENIX Association (2018)
17. Wirick, D.M., Teufel-Prida, L.A.: Closed systems in family systems theory. In: Encyclopedia of couple and family therapy, pp. 466–469. Springer (2019)
18. Wolsing, K., Kus, D., Wagner, E., Pennekamp, J., Wehrle, K., Henze, M.: One IDS is not Enough! Exploring Ensemble Learning for Industrial Intrusion Detection. In: 28th European Symposium on Research in Computer Security. Springer (2023)
19. Wu, W., Liao, W., Miao, J., Du, G.: Using gated recurrent unit network to forecast short-term load considering impact of electricity price. Energy Procedia **158**, 3369–3374 (2019)