

# Response to reviewer comments on paper: “@PAD: Adversarial Training of Power Systems Against Denial-of-Service Attacks”

by A. I. Ozdagli; C. Barreto; X. Koutsoukos

*The authors greatly appreciate the comments and revision suggestions by the reviewer. We have addressed each of the review comments and are providing a detailed description of the specific changes that have been made.*

---

---

## Comments from Reviewer #1:

**{1-1}:** Aren't there other mechanisms in the power system that can detect corrupted sensor measurements? It would be useful to look more into the current state of the art power system protection schemes that are relevant to this problem.

**{1-1 Response}:** *Authors thank the reviewer for this comment. In this paper, we focus on Denial of Service attacks where the adversarial actor disables a specific set of sensors rather than corrupting the sensor measurements. Our aim is not to detect whether a sensor measurement is corrupted or not. Instead, in our attack model, the adversary wants to cause misclassification of an event. Likewise, the defender wants to defend against this misclassification.*

*Inherently, the defender has a mechanism to detect which sensors are disabled. The nature of the attack is relevant to how the defender reacts to it. We assume that the measurement from a disabled sensor reading will be zero. In this context, action upon detecting corrupt sensor measurement is therefore out of scope.*

**{1-2}:** How is the solution specific to power systems? This seems to be an adversarial machine learning problem and the power grid was just used as the context, but nothing in the proposed approach would change if the context changes.

**{1-2 Response}:** *Thanks for this comment. We agree with the fact that the adversarial attack/defense model proposed in this manuscript is generalizable and applicable to any problem. We focus deliberately on power systems from the application point of view. We addressed this comment by adding the following text to the introduction:*

*It should be noted that the proposed attack and defense model is generalizable for many machine learning-based classifiers in various domains. In this paper, we specifically focus on the security of power grid systems from the application point of view. Since the sensors of grid systems are susceptible to DoS attacks, the methodology discussed in this paper aligns well with the research needs.*

**{1-3}:** More details on how this affects the power system besides the wrong labeling would be useful. Would like to see more quantitative and qualitative analysis on how this attack impacts the power grid – how much does it affect the operation performance?

**{1-3 Response}:** *Thanks for this comment. In this paper, to demonstrate the effectiveness of the attack and defense models, we utilize a dataset (Hink et al. 2014 – [19]) that is generated only for testing detection algorithms against cyber-attacks. Thus, we cannot demonstrate the direct impact of the wrong labeling on the power grid. We are aware that false positives may incur costs. On the other hand, false negatives may cause the grid to break apart. To address this comment, we included the confusion matrices before the attack, after the attack and the defense. The following modification is included in the Evaluation section:*

*The confusion matrices for the Original Model before and after the DoS attack over the Testing Dataset (K=1) are given in Tables 3 and 4, respectively. Similarly, Table 5 presents the confusion matrix for the Resilient Model against DoS attacks over Testing Dataset. Accordingly, the Original Model predicts 708 out of 774 attack events as attack which corresponds to 90 percent correct*

labeling. When the Original Model is attacked by DoS, the class for 506 attack events switched to normal label summing to 572 false positives. As a result, only 26 percent of the attack events are predicted correctly. After retraining using the defense scheme, the Resilient Model was able to predict 266 out of 774 attack events correctly and the recall value for the attack events increased to 34 percent. The increase in recall demonstrates that the defense mechanism may improve the prediction accuracy and reduce the impact of cyber-attacks on the performance of the grid systems, albeit the defense is limited.

**{1-4}:** How does this compare to false data injection attacks? More details about false data injection attacks would be useful (there are several entries in the references that cover FDIs)

**{1-4 Response}:** *We thank the reviewer for the comments. In this work, we propose a Denial of Service attack scheme, and we assume the sensors sending the measurements to the control room can be disabled rather than corrupted. Hence, false data injection is somewhat out of scope for this paper. We are willing to modify the manuscript if the reviewer is certain that FDI is relevant to our attack scheme. We addressed this comment by adding the following change to the conclusion as future research:*

*Last but not least, we want to compare the DoS attacks to more complicated attack schemes such as false data injection attacks in terms of the cost of devising the attack and its effectiveness.*

**{1-5}:** Assuming that the attacker has access to the machine learning algorithm and other details used to detect disturbances may be an unrealistic assumption.

**{1-5 Response}:** *We thank the reviewer for the comments. We agree that the attacker may not have full access to the model. However, we also think that every system is eventually prone to white-box attacks. We addressed the limitations of the attack scheme in the conclusion by adding the following section to the manuscript:*

*It should be noted that the attack scheme discussed here assumes the adversarial actor has access to the modeling parameters of the machine learning classifier. While, in reality, the attacker may not have full access to the system all the time, we assume that the system is eventually prone to white-box attack.*

**{1-6}:** Spelling mistake on the title ("Adverserial" should be "Adversarial")

*"Scada systems are vulnerabilities because ....." (grammatical error)*

**{1-6 Response}:** *We thank the reviewer for the corrections. We have fixed the typos and grammatical errors as the reviewer suggested.*