Abstract

Over the last decade, the number of cyberattacks targeting power systems and causing physical and economic damages has increased rapidly. Among them, False Data Injection Attacks (FDIAs) is a class of cyberattacks against power grid monitoring systems. Adversaries can successfully perform FDIAs to manipulate the power system State Estimation (SE) by compromising sensors or modifying system data. SE is an essential process performed by the Energy Management System (EMS) towards estimating unknown state variables based on system redundant measurements and network topology. SE routines include Bad Data Detection (BDD) algorithms to eliminate errors from the acquired measurements, e.g., in case of sensor failures. FDIAs can bypass BDD modules to inject malicious data vectors into a subset of measurements without being detected; and thus, manipulate the results of the SE process.

To overcome the limitations of traditional residual-based BDD approaches, data-driven solutions based on machine learning algorithms have been widely adopted for detecting malicious manipulation of sensor data due to their fast execution times and accurate results. Machine learning algorithms have been proposed as a promising solution for detecting FDIAs, as they can automatically learn patterns and anomalies in the data that are indicative of an attack. However, these algorithms are also vulnerable to adversarial examples, which are maliciously crafted inputs that are designed to mislead the model into making a wrong decision.

In this dissertation, we focus on evaluating the vulnerability of machine learning algorithms against adversarial examples in the context of FDIAs. Specifically, we study six different cases of adversarial attacks, including Adversarial Label Flipped Attack on SVM, Targeted Fast Gradient Sign Method Attack on MLP, Limited-memory BFGS Attack on MLP, Jacobian-based Saliency Attack on MLP, Carlini and Wagner Adversarial Attack, and Zeroth Order Optimization-based Attack. We implement these attacks on a simulated power system, and evaluate the performance of the machine learning algorithms in detecting them. The results of this study provide insights into the strengths and weaknesses of different machine learning algorithms in detecting FDIAs and adversarial examples. We also provide recommendations on how to improve the robustness of these algorithms against adversarial examples. The findings of this research are useful for practitioners in the field of power systems and machine learning, as well as for researchers working on the security of cyber-physical systems. This dissertation is organized into several chapters, starting with background, literature review, objective, adversarial examples, adversarial examples on power systems state estimation, evasion attacks with adversarial deep learning against power system state estimation, adversarial machine learning designs against learning-based attack detection algorithms in power systems and a summary of the work and future work.