Adversarial Attacks on Deep Learning-Based False Data Injection Detection in Differential Relays

Ahmad Abdelsamie[1], Aditi Maheshwari[1], Amr Youssef[2], Deepa Kundur[1]

[1]University of Toronto, [2]Concordia University

The application of Deep Learning-based Schemes (DLSs) for detecting False Data Injection Attacks (FDIAs) in smart grids has attracted significant attention. This paper demonstrates that adversarial attacks–carefully crafted FDIAs can evade existing DLSs used for FDIA detection in Line Current Differential Relays (LCDRs). We propose a novel adversarial attack framework, utilizing the Fast Gradient Sign Method, which exploits DLS vulnerabilities by introducing small perturbations to LCDR remote measurements, leading to misclassification of the FDIA as a legitimate fault while also triggering the LCDR to trip. We evaluate the robustness of multiple deep learning models–including multi-layer perceptrons, convolutional neural networks, long short-term memory networks, and residual networks–under adversarial conditions. Our experimental results demonstrate that while these models perform well, they exhibit high degrees of vulnerability to adversarial attacks. For some models, the adversarial attack success rate exceeds 99.7%. To address this threat, we introduce adversarial training as a proactive defense mechanism, significantly enhancing the models' ability to withstand adversarial FDIAs without compromising fault detection accuracy. Our results highlight the significant threat posed by adversarial attacks to DLS-based FDIA detection, underscore the necessity for robust cybersecurity measures in smart grids, and demonstrate the effectiveness of adversarial training in enhancing model robustness against adversarial FDIAs.

979-8-3315-8138-1