ABSTRACT

Deep Neural Networks (DNNs) in machine learning are vulnerable to backdoor attacks, where attackers insert unauthorized watermarks into the model to conceal specific patterns that can be exploited to manipulate predictions. These attacks can reduce model accuracy and cause misclassification when triggered by specially crafted inputs containing backdoors.

To address this issue, this research proposes a watermarking method based on Wet Paper Coding (WPC), embedding watermarks directly into DNN model parameters, and using a Siamese Neural Network (SNN) for verification and backdoor attack detection. The watermark used is an email identifier, ensuring dataset integrity and ownership verification.The embedding process utilizes the Optimized Probabilistic Selection Strategy (OPSS) to select model parameters that have minimal impact on accuracy. WPC embeds the watermark by modifying ”dry blocks” while preserving ”wet blocks” to maintain model stability. After embedding, the watermark is extracted and verified using SNN, which compares it against registered email identifiers. If the extracted watermark does not match the authorized list, the model is flagged as potentially compromised by a backdoor attack.

Experiments on the MNIST dataset show optimal results with a Layer Ratio of 0.5, Prob Alpha of 0.5, and Dry Block Ratio of 0.2, achieving a Detection Rate (DR) of 0.9922 and a Bit Error Rate (BER) of 0.0078. The model remains stable, with accuracy improving from 90.37% to 92.09% after watermarking. However, the extraction process still encounters errors due to high watermark capacity, and SNN achieves only 56.66% accuracy in detecting backdoor attacks, indicating challenges in distinguishing authorized watermarks from unauthorized ones.

Keywords: : Watermark, Wet-Paper-Coding, deep Neural Network, OPSS, Siamese Neural Network, backdoor attack