ABSTRACT

Most NIDS rely on predefined signatures patterns of known attacks allowing zero-day exploits to slip through. Anomaly-based NIDS profile “normal” activity but incur high false-positive rates. Traditional classifiers (Naïve Bayes, SVM) need extensive feature engineering; LSTMs struggle with long-range dependencies and tuning; autoencoders overfit benign patterns and lack adversarial robustness. We introduce an unsupervised Transformer-Encoder that reconstructs only benign flows via multi-head self-attention and positional encoding, using a fixed reconstruction-error threshold calibrated on held-out data to flag zero-day intrusions. On UNSW-NB15 it achieves Accuracy 95.09%, Precision 97.31%, Recall 91.63%, and F1 Score 94.38%, on TON-IoT it achieves Accuracy 88.75%, Precision 99.25%, Recall 88.73%, F1 Score 93.69%, demonstrating strong detection with reduced false alarms.

Keywords — Zero-day attack detection, Network intrusion detection system (NIDS), Anomaly-based detection, Transformer Encoder, Unsupervised learning, UNSW-NB15, TON-IoT