ABSTRACT

Network security has become a global challenge that requires effective and innovative solutions. Intrusion Detection Systems (IDS) play a crucial role in protecting network infrastructures from evolving cyberattacks. The use of Machine Learning (ML) techniques in IDS offers high accuracy in detecting and identifying threats. However, challenges arise when dealing with imbalanced and high-dimensional datasets. This paper introduces a novel approach for ML-based network intrusion detection by employing Random Oversampling (RO) to handle data imbalance and K-fold validation, along with Feature Selection and Extraction using Random Forest and Principal Component Analysis (PCA) to address dimensionality reduction and K-vold validation to ensures that the feature selection process (Random Forest + PCA) and model training are optimized to avoid overfitting. Additionally, each model undergoes Maximum Optimization using Optuna to enhance accuracy, precision, recall, F1-score, NIDS traffic parameters, and ROC Curve performance. The approach was evaluated on three benchmark datasets: UNSW-NB15, CIC-IDS-2017, and CIC-IDS-2018. Each dataset was modeled using KNN, Logistic Regression, Decision Tree, Random Forest, GBM, XGBM, Adaboost, Light GBM, CatBoost, and Extra Tree algorithms to achieve a high accuracy of 99%. Notably, this method proves effective for large and imbalanced datasets, as evidenced by the CIC-IDS-2018 dataset, which contains over one million records. The results outperform state-of-the-art models, marking a significant advancement in network intrusion detection. This flexible framework paves the way for further exploration of ML algorithms to enhance IDS effectiveness.

Keywords: Machine Learning, Network Intrusion Detection System, Feature Extraction, Random Oversampling, Principal Component Analysis.