The Use of Machine Learning in The Detection of Anomaly Network Intrusion

Akshat Behera
A20516439
abehera2@hawk.iit.edu

Nagarjuna Bolla
A20524548
nbolla@hawk.iit.edu

Project Proposal

1 Problem Description

Cybersecurity is one of the major concerns for organizations worldwide. Network Intrusion Detection is a crucial component of cybersecurity, which detects and mitigates the attempts of attackers to exploit vulnerabilities in a network. The detection of these attacks is a challenging task as the attackers are continuously modifying their techniques to bypass the detection methods. Traditional rule-based approaches for detecting network intrusions are not effective as they have limitations in identifying novel and sophisticated attacks. Machine Learning approaches can play a vital role in detecting these attacks as they can learn from large datasets and identify patterns in network traffic.

The proposed project aims to develop an effective Machine Learning-based approach for Network Intrusion Detection using the UNSW-NB15 dataset. The dataset contains both normal and attack traffic, and the goal is to classify the traffic into different attack categories accurately.

2 Brief survey of the Literature and Proposed Work

Cybersecurity is an area of great concern for organizations worldwide, and one of its key components is Network Intrusion Detection, which aims to detect and prevent attackers from exploiting network vulnerabilities. To this end, Machine Learning (ML) approaches have been proposed, with various researchers exploring their effectiveness for Network Intrusion Detection.

For example, Moustafa et al. (2015) [3] proposed the UNSW-NB15 dataset, which contains both normal and attack traffic. They also suggested a feature selection method that utilizes Principal Component Analysis (PCA) to select the most relevant features. Similarly, Alazab et al. (2016) [1] conducted a survey of anomaly-based Network Intrusion Detection techniques, identifying the challenges in this field, and proposing a framework that employs multiple ML algorithms.

In a more recent survey, Islam and Ahmed (2019) [2] provided a comprehensive overview of ML approaches for Network Intrusion Detection, classifying them into three categories: supervised, unsupervised, and hybrid. They also highlighted some of the limitations of these methods and suggested directions for future research. Meanwhile, Mukherjee et al. (2020) [5] compared the performance of different Deep Learning (DL) approaches, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Autoencoder, finding that CNNs were the most accurate and had the highest F1-Score.

Moustafa and Slay (2016) [4] proposed the UNSW-NB15 dataset, which contains both normal and attack traffic, and conducted a statistical analysis of Network Anomaly Detection Systems using this dataset. They compared the performance of various machine learning algorithms, including Decision Trees, Random Forest, Naive Bayes, K-Nearest Neighbor, and Support Vector Machine, with the traditional KDD99 dataset. They found that machine learning approaches achieved higher accuracy and detection rates on the UNSW-NB15 dataset. This study supports the use of machine learning approaches for Network Intrusion Detection and provides a benchmark dataset for evaluating their performance.

This proposed project aims to develop a Machine Learning-based approach to Network Intrusion Detection, utilizing the UNSW-NB15 dataset. It seeks to build on existing research by exploring the potential of Neural Networks (NNs), such as Neural Network (NN) using MLP Classifier, Convolutional Neural Networks (CNNs) and Recurrent Convolutional Neural Networks (RCNNs), for Network Intrusion Detection. Additionally, the project will investigate the benefits of feature selection methods in improving the performance of the classifiers. Finally, the project aims to compare the effectiveness of NNs with traditional classifiers, including XGBoost, SVM, Random Forest, Decision Tree, and Logistic Regression.

Overall, this proposed project aims to advance the field of Network Intrusion Detection by exploring the use of Neural Networks and Feature Selection methods to improve classifier accuracy. This could lead to more effective and efficient Network Intrusion Detection systems capable of detecting even the most advanced and novel attacks.

3 Preliminary Plan (Milestones):

1. Preprocessing of the Dataset: Here, We begin with Extraction and Cleaning of the Data. The UNSW-NB15 dataset was obtained from UNSWResearchSite, which is also available at Kaggle. The UNSW-NB15 dataset contains missing values and categorical variables, which need to be addressed before feeding the data into Machine Learning models. The plan is to preprocess the data by dropping unwanted columns, encoding categorical variables, and standardizing numerical variables.

2. Exploratory Data Analysis (EDA): EDA is essential to understand the distribution of the features, correlation among the features, and to identify any outliers in the data. The plan is to perform EDA using visualizations such as histograms, heatmaps, and boxplots.

3. Selection of the Features: Feature selection aims to select the most relevant features that can improve the performance of the classifiers. The plan is to use Mutual Information Score to select the top k features.

4. Selection of the Models: The plan is to select different types of Neural Networks, such as CNNs, RCNNs, and NNs (MLP), and compare their performance with traditional classifiers such as Random Forest, Decision Tree, XGBoost, SVM and Logistic Regression.

5. Evaluation of the Models: The performance of the classifiers will be evaluated using metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC. The plan is to use the confusion matrix to identify the misclassified samples and analyze the results.

References

[1] Mamoun Alazab, Michael Hobbs, and Jemal Abawajy. Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security*, 60:84–102, 2016.

[2] Md Rafiul Islam and Kazi Mohammed Ahmed. Machine learning approaches for network intrusion detection: A comprehensive survey. *IEEE Access*, 7:27459–27484, 2019.

[3] Ahmed Moustafa, Jill Slay, and Gregory Creech. Unsw-nb15: A comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In *Military Communications and Information Systems Conference (MilCIS)*, pages 1–6. IEEE, 2015.

[4] Nour Moustafa and Jill Slay. The evaluation of network anomaly detection systems: Statistical analysis of the unsw-nb15 dataset and the comparison with the kdd99 dataset. *Information Security Journal: A Global Perspective*, 25(1-3):1–14, 2016.

[5] Sourav Mukherjee, Ananda Roy Chowdhury, Shukla Das, Sayan Chakraborty, and Mita Nasipuri. A comparative study of deep learning approaches for network intrusion detection. *Future Generation Computer Systems*, 107:1063–1077, 2020.