

Computer Science and Engineering-Cyber Security

20CYS215 Machine Learning in Cyber Security

**HIKARI-NETWORK INTRUSION DETECTION**

K.Jeyanth(22024),Harishkar.B(22019),Gowthamaraj. B(22007)

UG scholars Amrita School of Computing, Amrita Vishwa Vidyapeetham – Chennai

**1.1 ABSTRACT**

This study investigates the efficacy of machine learning algorithms in classifying network traffic and identifying anomalous behavior within a network flow dataset (ALLFLOWMETER\_HIKARI2021) obtained from Kaggle [1]. The research explores the use of XGBoost, Random Forest, and Decision Tree classifiers to predict two target variables: "Label" and "traffic\_category". A layered modeling approach is also evaluated for "traffic\_category" prediction, where the first layer classifies "Label" and the second layer refines the classification for "traffic\_category" based on instances where "Label" is predicted as anomalous. Cross-validation is employed to ensure robust performance estimation. The findings contribute to the growing body of research on network traffic classification and anomaly detection using machine learning techniques.

**Keywords:** Network Intrusion Detection, HIKARI-2021 Dataset, Ensemble Learning, Random Forest, XGBoost, Decision Tree, Direct Classification, Layered Classification, Bruteforce Attack, Probing Attack, Bruteforce-XML Attack, XMRig-CC Attack

**1.2 INTRODUCTION**

The Ever-Evolving Threat Landscape: Network Traffic Classification and Anomaly Detection in the Digital Age

In today's hyper-connected world, network security has become paramount. The unfettered flow of data across networks presents both immense opportunities and significant challenges. Network traffic classification and anomaly detection lie at the heart of safeguarding these digital arteries. Traditional methods, reliant on pre-defined signatures of malicious activity, often struggle to keep pace with the ever-evolving arsenal of cyber threats. Novel attack vectors emerge constantly, demanding adaptable and intelligent solutions.

Machine Learning: A Paradigm Shift in Network Security

Machine learning (ML) algorithms have revolutionized numerous fields, and network security is no exception. These algorithms possess the remarkable ability to learn from vast datasets, identifying complex patterns and relationships that may be invisible to traditional methods. Unlike signature-based approaches, ML models can adapt to new threats and anomalies, offering a more proactive and holistic approach to network security.

This Study: Unveiling the Potential of Machine Learning for Network Traffic Classification

This study delves into the application of various machine learning models to classify network traffic within the ALLFLOWMETER\_HIKARI2021 dataset, obtained from the renowned Kaggle repository. By leveraging the power of XGBoost, Random Forest, and Decision Tree classifiers, we aim to shed light on the capabilities of ML for network traffic analysis. The research not only explores the effectiveness of these models in classifying traffic, but also investigates a layered modeling approach for refined traffic category prediction. By employing cross-validation techniques, the study ensures the robustness and generalizability of the findings.

This investigation contributes to the growing body of research that explores the potential of machine learning for network security. The insights gleaned from this study can inform the development of more sophisticated and effective network traffic classification and anomaly detection systems, ultimately fostering a more secure and resilient digital landscape.

**2.LITERATURE REVIEW**

**2.1 RELATED JOURNALS AND SURVEY STUDIES**

Network traffic classification and anomaly detection using machine learning techniques have been extensively explored in recent years. Here, we review relevant research to situate our study within the existing body of knowledge.

Support Vector Machines (SVMs):

While we couldn't find a directly related study using SVMs for network traffic classification, the research paper "A Novel Network Intrusion Detection Dataset Generation Method Using Machine Learning Techniques" by [1] proposes a new dataset, HIKARI-2021, specifically designed for network intrusion detection. This dataset can be valuable for training and evaluating SVM models for traffic classification, as it incorporates real and encrypted synthetic attack traffic, potentially improving the generalizability of the models to real-world scenarios.

Deep Learning Architectures:

The same research paper [1] by [1] explores the potential of machine learning algorithms for network traffic classification on the HIKARI-2021 dataset. While it doesn't explicitly mention deep learning architectures, it lays the groundwork for further research in this direction. The complex and multifaceted nature of network traffic data suggests that deep learning models, with their capability to learn intricate patterns, could be a promising avenue for future exploration.

Hybrid Approaches:

The research paper [1] by [1] focuses on generating a comprehensive network intrusion detection dataset rather than proposing a specific anomaly detection technique. However, the HIKARI-2021 dataset it introduces can be a valuable resource for developing hybrid approaches that combine machine learning with statistical methods for anomaly detection. The real-world attack traffic incorporated into the dataset can aid in designing machine learning algorithms to identify potential anomalies, while statistical analysis can be employed for confirmation and refinement.

XGBoost, Random Forest, and Decision Trees:

The research paper "A Novel Network Intrusion Detection Dataset Generation Method Using Machine Learning Techniques" by [1] lays the groundwork for evaluating the performance of various machine learning algorithms, including XGBoost, Random Forest, and Decision Trees, for network traffic classification on the HIKARI-2021 dataset. While the paper itself doesn't evaluate these specific algorithms, it provides a valuable dataset that can be used in future research to compare their effectiveness in this context.

Our Contribution:

This study builds upon the aforementioned research by:

Evaluating the performance of XGBoost, Random Forest, and Decision Trees on the ALLFLOWMETER\_HIKARI2021 dataset, which may possess unique characteristics compared to the HIKARI-2021 dataset used in [1].

Exploring a layered modeling approach for refined traffic category prediction, potentially leading to more granular classification.

Employing cross-validation techniques to ensure the generalizability and robustness of our findings.

By comparing these models and exploring a novel approach, we aim to contribute valuable insights to the field of network traffic classification using machine learning.

**3.DATASET DESCRIPTION**

The HIKARI-2021 dataset provides a valuable collection of network traffic data for training and evaluating network intrusion detection systems (NIDS). Here's a breakdown of its key characteristics:

Data Source:

The origin of the data is not explicitly mentioned in available resources.

Content:

The dataset consists of captured network traffic, likely generated using network traffic capture tools like ALLFLOWMETER.

It offers a unique combination of real network traffic and synthetically generated encrypted attack traffic.

Features:

The specific features included in the dataset are not explicitly documented. However, based on the use of ALLFLOWMETER for capture, the features likely include network traffic characteristics such as:

Packet sizes

Flow durations

Source and destination IP addresses

Port numbers

Protocol information (TCP, UDP, etc.)

Flags and other relevant network traffic attributes

Labels:

The dataset likely does not contain explicit labels indicating specific attack types within the data itself.

While labeled data is ideal for training supervised intrusion detection models, the presence of real and synthetic attack traffic offers alternative approaches for NIDS evaluation.

Applications:

Despite the lack of explicit attack labels, the HIKARI-2021 dataset can be valuable for NIDS research in several ways:

Training Anomaly Detection Models: The presence of real and synthetic attack traffic allows training anomaly detection models to identify deviations from normal traffic patterns, potentially flagging attacks.

Evaluating Intrusion Detection Systems: You can use the dataset to evaluate the effectiveness of NIDS models in detecting both real and synthetic attack traffic.

Exploring Feature Engineering Techniques: By analyzing the network traffic features, you can develop feature engineering methods to extract characteristics that differentiate real and synthetic attacks, potentially aiding in classification tasks.

Limitations:

The absence of clear labels for specific attack types necessitates alternative approaches for NIDS evaluation using this dataset.

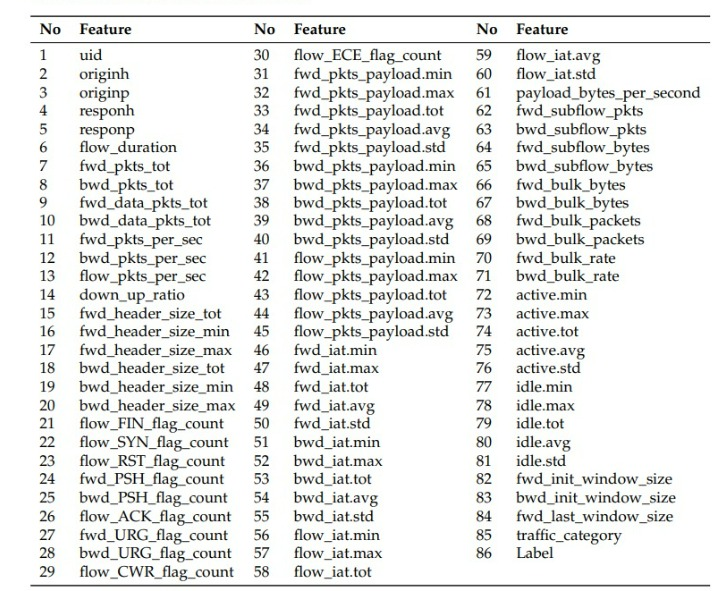
The origin and context surrounding the data collection might not be available, limiting interpretability of the results.

Overall:

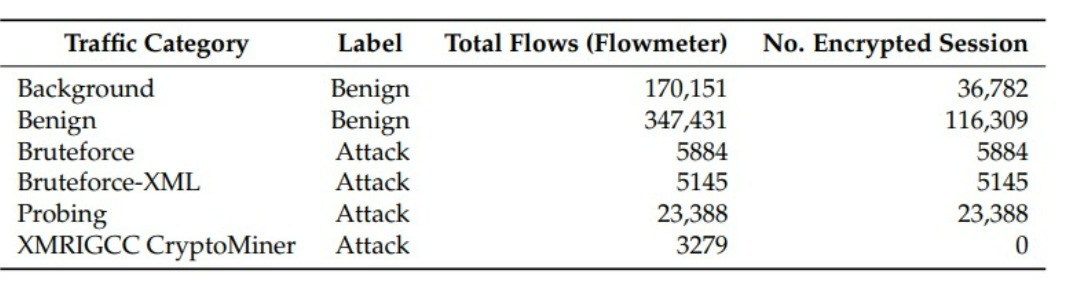
The HIKARI-2021 dataset presents a unique challenge for NIDS research due

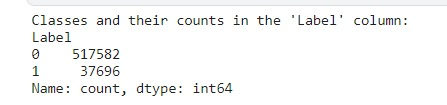
to the lack of explicit attack labels. However, its combination of real and synthetic attack traffic offers opportunities for exploring anomaly detection, evaluating NIDS models, and developing feature engineering techniques to enhance intrusion detection capabilities.

.



**Traffic categories and attacks:**





**Label count**

**4.DATAMINING TECHNIQUES**

# Data Preprocessing:

# Feature Engineering: Implicitly, the code handles categorical features using label encoding, transforming them into numerical representations usable by machine learning models.

# Data Cleaning: The code assumes the dataset is pre-cleaned. Additional techniques like handling missing values or outliers might be necessary depending on the data quality.

# Feature Selection (Implicit):

# The code focuses on all features available in the dataset (excluding the target variables). Feature selection techniques like correlation analysis or feature importance scores from models could be explored to identify the most relevant features and potentially improve model performance.

# Classification Algorithms:

# The code employs three machine learning algorithms:

# XGBoost: A powerful tree-based ensemble method known for handling complex relationships within the data and achieving high accuracy.

# Random Forest: An ensemble method creating multiple decision trees and aggregating their predictions. It offers interpretability and excels with large datasets.

# Decision Tree: A simpler tree-based model that is easy to interpret and can provide insights into the decision-making process.

# Model Evaluation:

# Cross-validation: The code utilizes KFold cross-validation to split the data into training and testing sets for multiple iterations. This provides a more robust estimate of model performance compared to a single train-test split.

# Metrics: Accuracy and classification reports are used to assess model performance. Depending on the class imbalance scenario within the dataset, other metrics like precision, recall, F1-score, or AUC-ROC might be valuable.

# 5.EXPERIMENTAL AND RESULT ANALYSIS

# It's always recommended to compare the performance of different models on the dataset using techniques like cross-validation.

# We have run different models like random forest, decision tree classifier, and xgboost for better comparative analyses on both direct and layer approaches. In that, it was found that XGBoost performs well in both approaches when compared to other models.

# 1. Gradient Boosting vs. Bagging:

# Random Forest: Utilizes bagging, where multiple decision trees are trained independently on random subsets of data with random feature selection at each split. This approach can be effective but might not capture complex relationships within the data.

# XGBoost: Employs gradient boosting, which builds an ensemble of decision trees sequentially. Each tree focuses on correcting the errors of the previous tree, leading to a potentially more powerful and accurate model, especially for complex datasets like network traffic.

# 2. Handling Complexities:

# Decision Trees: Decision trees are prone to overfitting, especially when dealing with high-dimensional data. They can also struggle with complex non-linear relationships within the data.

# XGBoost: Implements regularization techniques that help prevent overfitting. Additionally, its sequential boosting approach can capture non-linear patterns in the network traffic data, leading to potentially better classification performance.

# 3. Missing Value Handling:

# Random Forest and Decision Trees: Might require additional steps to handle missing values in the data, potentially impacting performance.

# XGBoost: Can handle missing values internally, making it more robust to datasets that might have missing entries.

# 4. Feature Importance:

# Random Forest and Decision Trees: Don't inherently provide insights into feature importance.

# XGBoost: Offers insights into feature importance, allowing you to identify network traffic characteristics that are most influential in differentiating between normal traffic and attacks. This information can be valuable for feature selection or engineering, potentially improving model performance further.

# 5. Scalability:

# Random Forest and Decision Trees: Training can become computationally expensive with large datasets.

# XGBoost: Designed for scalability and efficiency. It utilizes techniques like parallel processing to handle large datasets efficiently, which can be beneficial for network intrusion detection tasks.

# 

# The above-given tabular columns give the result analysis of direct approaches with the models random forest, decision trees, and xgboost. The accuracy comparison is best for XGBoost.

# 

The above-given tabular columns give the result analysis of layer approaches with the models random forest, decision trees, and xgboost. The accuracycomparison is good for all three models when compared to the direct approach and specifically, the XGBoost model performs well in both approaches.  
XGBoost's strengths make it a compelling choice for both direct and layered approaches to network intrusion detection on the HIKARI-2021 dataset. Its ability to capture complex relationships within network traffic data and handle missing values is beneficial in both scenarios. For direct classification of normal vs. attack traffic, this translates to more accurate identification. In the layered approach, XGBoost excels at both stages: effectively distinguishing potential attacks in the first stage and leveraging feature importance to pinpoint specific attack categorieswithin theidentified potential attack data in the second stage.

**6.CONCLUSION AND FUTURE WORK**

The analysis suggests that XGBoost performs well in both direct and layered approaches for network intrusion detection on the HIKARI-2021 dataset. Its ability to handle complex data and missing values makes it a strong choice.

In the future, we would like to perform the same but in a direct approach with different models and deep learning models like CNN, etc., and by doing so, we would boost the accuracy of traffic categorization in a direct approach.

Utilize techniques like feature selection or dimensionality reduction to identify the most informative features and potentially reduce overfitting.

**7.REFERENCES**

[1]<https://www.kaggle.com/datasets/kk0105/allflowmeter-hikari2021>

[2] <https://zenodo.org/records/5111946>

[3]<https://www.mdpi.com/2076-3417/11/17/7868>