iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

IOT NETWORK TRAFFIC CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

Ch. Shanthi Priya
Assistant Professor, Hyderabad Institute of Technology and Management,
Hyderabad, India

Gopi Somjiyani
B. Thulasi
V. Yashaswini
P. Shivani
UG Students, Hyderabad Institute of Technology and Management,
Hyderabad, India

ABSTRACT

In past years there is been a substantial rise in the number of IoT devices and the volume of data they generate. These IoT devices, often limited in terms of computational and power resources, are typically deployed without robust security mechanisms, making them prime targets for cyberattacks. As the number of attack attempts on these devices grows, conventional intrusion detection systems (IDS) face increasing challenges in promptly detecting and responding to threats. Accurate classification of IoT-generated data is crucial for various applications including anomaly detection, intrusion prevention, resource management, and ensuring Quality of Service (QoS). This paper explores the application of machine learning (ML) techniques for classifying network traffic in IoT environments. A range of ML models—such as Random Forest, Support Vector Machines, and deep learning methods—are analyzed for their effectiveness in classifying IoT traffic based on network flow characteristics. The study focuses on key elements like feature selection, the datasets utilized, and the evaluation metrics applied. The results highlight the potential of AI-driven approaches to enhance the accuracy of IoT traffic classification, revealing the promise of intelligent systems in strengthening IoT network security.

Keywords:
Attack, Dataset, IoT, Intrusion detection, Machine learning.

1. INTRODUCTION

The IoT refers to a vast network of smart interconnected devices capable of sensing their environment and exchanging data autonomously without the need for direct human involvement. These technologies are revolutionizing a wide range of industries. In retail, IoT facilitates efficient inventory management through automation. In healthcare, it enables remote patient monitoring, supporting continuous care outside traditional clinical settings. In the manufacturing sector, intelligent systems can track worker movements and initiate automated operations in unmanned warehouses.

According to industry projections, the number of IoT-connected devices worldwide is expected to surpass 75 billion by 2025, with the trend continuing upward. However, this rapid growth raises serious security concerns. IoT devices are often deployed in the field for long durations with minimal human oversight, making them attractive targets for cyberattacks. Their limited processing power, memory, and bandwidth pose challenges to implementing conventional security solutions. As a result, IoT networks are increasingly vulnerable to various types of cyber threats, including Denial-of-Service (DoS) attacks, which can disrupt normal operations by flooding systems with malicious traffic.

To counter these threats, machine learning (ML) has emerged as a promising approach for detecting and responding to anomalous or malicious activities within IoT environments. Given the dynamic nature of both IoT technologies and cyber threats, there is a growing demand for intelligent, adaptive defense mechanisms. This study investigates the effectiveness of various machine learning algorithms in detecting malicious behavior in IoT systems using a benchmark dataset. The research evaluates the performance of these models in both binary and multiclass classification tasks, aiming to enhance the reliability and security of IoT networks.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

2. OBJECTIVES

The use of machine learning for IoT network traffic classification has gained substantial attention in recent years due to its potential to address the limitations of traditional methods. Several researchers have explored various ML algorithms and datasets to improve the accuracy and efficiency of traffic classification and anomaly detection in IoT environments. These studies collectively indicate that ML techniques, when trained on appropriate datasets and optimized with the right features, can significantly enhance traffic classification and threat detection in IoT networks. However, gaps remain in terms of real-time deployment, generalization across diverse devices, and handling encrypted or obfuscated traffic.

3. METHODOLOGY

3.1 Data Acquisition and Preprocessing

3.1.1 Dataset Collection

The first step involves obtaining network traffic data that accurately reflects real-world IoT scenarios. This can be achieved by capturing live traffic from IoT networks using tools such as Wireshark or TCPDump, which allow for packet-level data collection. Choosing datasets that incorporate a variety of IoT device types and application domains, ensuring comprehensive coverage of realistic communication patterns and behaviors. Leveraging open-access datasets (e.g., those containing botnet-infected IoT traffic or normal device traffic traces) to support reproducible experimentation and benchmark evaluation.

3.1.2 Data Cleaning

Once raw traffic data is collected, it must be cleaned to ensure quality and consistency. Discard corrupted or malformed packets that cannot be parsed correctly. Address missing or incomplete data fields either by imputing values or removing affected entries. Eliminate irrelevant traffic, duplicate entries, and noise to streamline the dataset for meaningful analysis.

3.1.3 Feature Extraction

Feature extraction is the process of deriving quantifiable attributes from raw packet data, enabling machine learning models to learn from structured inputs. Common features include:

- Statistical metrics on packet lengths (e.g., average, max, standard deviation).
- Inter-arrival times between packets in a flow.
- Protocol details such as whether the traffic uses TCP, UDP, MQTT, etc.
- Network identifiers including source and destination IP addresses and port numbers.
- Flow-based attributes such as session duration and the number of packets exchanged.
- Payload characteristics such as byte frequency or entropy.

Tools like TShark and CICFlowMeter can be used to automate the extraction of flow-level features from raw packet captures.

3.1.4 Data Transformation

Before feeding the data into a machine learning model, transformation is necessary to prepare the features. Normalization or standardization ensures all numerical features share a similar scale, avoiding bias during model training. Encoding categorical features, such as protocol types, is done using techniques like one-hot encoding or label encoding to convert them into numerical format.

3.1.5 Feature Selection and Dimensionality Reduction

To enhance model performance and reduce overfitting, only the most informative features should be retained. Methods include:

- **Information gain**: measures how much a feature contributes to reducing uncertainty.
- **Chi-square test**: evaluates the independence between features and target labels.
- **Recursive Feature Elimination (RFE)**: iteratively removes the least important features based on model performance.
- **Correlation-Based Feature Selection (CFS)**: selects subsets of features that are highly correlated with the output but uncorrelated with each other.

If the dataset contains a high number of features, dimensionality reduction techniques may be applied:

- **Principal Component Analysis (PCA)**: projects data into a lower-dimensional space while preserving variance.
- **Linear Discriminant Analysis (LDA)**: reduces dimensions while maximizing class separability.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

- **T-distributed Stochastic Neighbor Embedding (t-SNE)**: used for visualizing high-dimensional data in 2D or 3D space.

3.3 Model Selection and Training

3.3.1 Algorithm Selection

A variety of machine learning algorithms can be used for traffic classification, each offering different benefits depending on the complexity and size of the dataset:

- **Random Forest**: a robust ensemble learning method suitable for high-dimensional data.
- **Support Vector Machines (SVM)**: effective for linearly and nonlinearly separable data.
- **K-Nearest Neighbors (KNN)**: a simple instance-based learning method.
- **Naive Bayes**: fast and efficient for problems with probabilistic class distributions.
- **Convolutional Neural Networks (CNNs)**: suitable for spatial data representations.
- **Recurrent Neural Networks (RNNs)** and **LSTMs**: ideal for sequential or time-series traffic data.

3.3.2 Model Training

To train and validate the models:

- Split the dataset into training and testing sets, commonly 70:30 or 80:20.
- Train each model using the training portion of the dataset.
- Optimize hyperparameters through techniques such as grid search, random search, or cross-validation to find the best performing model configuration.

3.4 Model Evaluation

Evaluating the effectiveness of the trained models involves applying a range of performance metrics:

- **Accuracy**: proportion of correctly classified instances over the total.
- **Precision**: ability of the model to return only relevant results (i.e., true positives vs. false positives).
- **Recall**: measures the model's sensitivity (i.e., true positives vs. actual positives).
- **F1-Score**: harmonic mean of precision and recall; useful in imbalanced datasets.
- **Confusion Matrix**: provides a detailed breakdown of classification outcomes.
- **ROC-AUC (Receiver Operating Characteristic – Area Under Curve)**: indicates the model's performance across different threshold settings.

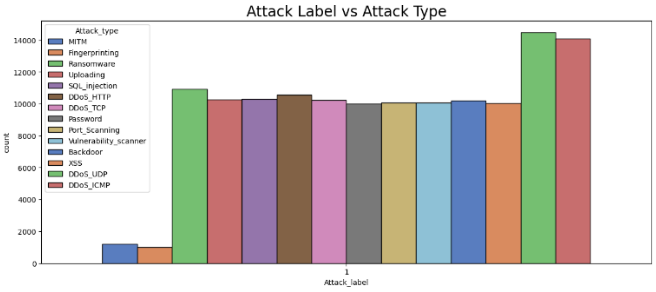


Figure 1 Attack label vs attack type

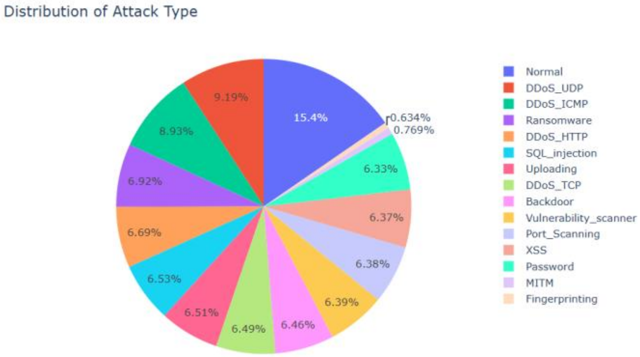
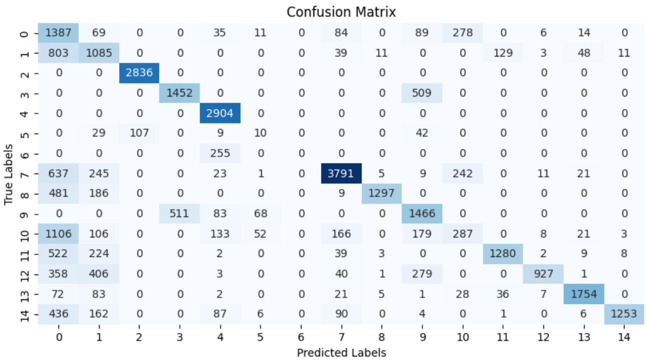


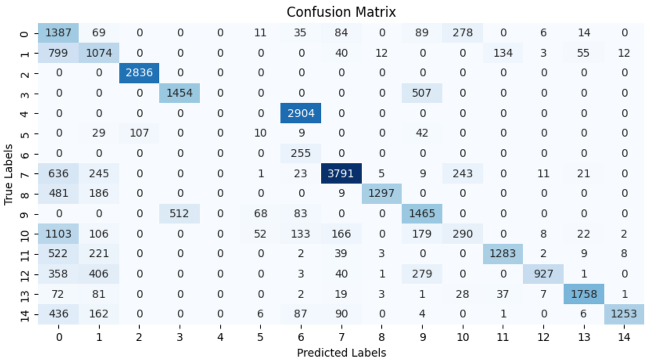
Figure 2 Distribution of attack type

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1387	69	0	0	35	11	0	84	0	89	278	0	6	14	0
1	803	1085	0	0	0	0	0	39	11	0	0	129	3	48	11
2	0	0	2836	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1452	0	0	0	0	0	509	0	0	0	0	0
4	0	0	0	0	2904	0	0	0	0	0	0	0	0	0	0
5	0	29	107	0	9	10	0	0	0	42	0	0	0	0	0
6	0	0	0	0	255	0	0	0	0	0	0	0	0	0	0
7	637	245	0	0	23	1	0	3791	5	9	242	0	11	21	0
8	481	186	0	0	0	0	0	9	1297	0	0	0	0	0	0
9	0	0	0	511	83	68	0	0	0	1466	0	0	0	0	0
10	1106	106	0	0	133	52	0	166	0	179	287	0	8	21	3
11	522	224	0	0	2	0	0	39	3	0	0	1280	2	9	8
12	358	406	0	0	3	0	0	40	1	279	0	0	927	1	0
13	72	83	0	0	2	0	0	21	5	1	28	36	7	1754	0
14	436	162	0	0	87	6	0	90	0	4	0	1	0	6	1253

Figure 3 Confusion matrix for decision tree



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1387	69	0	0	0	11	35	84	0	89	278	0	6	14	0
1	799	1074	0	0	0	0	0	40	12	0	0	134	3	55	12
2	0	0	2836	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1454	0	0	0	0	0	507	0	0	0	0	0
4	0	0	0	0	0	0	2904	0	0	0	0	0	0	0	0
5	0	29	107	0	0	10	9	0	0	42	0	0	0	0	0
6	0	0	0	0	0	0	255	0	0	0	0	0	0	0	0
7	636	245	0	0	0	1	23	3791	5	9	243	0	11	21	0
8	481	186	0	0	0	0	0	9	1297	0	0	0	0	0	0
9	0	0	0	512	0	68	83	0	0	1465	0	0	0	0	0
10	1103	106	0	0	0	52	133	166	0	179	290	0	8	22	2
11	522	221	0	0	0	0	2	39	3	0	0	1283	2	9	8
12	358	406	0	0	0	0	3	40	1	279	0	0	927	1	0
13	72	81	0	0	0	0	2	19	3	1	28	37	7	1758	1
14	436	162	0	0	0	6	87	90	0	4	0	1	0	6	1253

Figure 4 Confusion matrix for random forest

Evaluation Parameter	**Result**
accuracy	88%
precision	99%
recall	88.76%
F1 score	93.95%

RESULTS AND DISCUSSION

The implementation of the machine learning algorithms successfully classify the IoT network traffic based on various parameters into normal or malicious behavior. Aim was to attain better precision and accuracy scores which makes this model stand out among the others.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Hyderabad Institute of Technology and Management for providing the necessary resources and support to carry out this research. I extend my heartfelt appreciation to my mentor, Ch. Shanthi, for their invaluable guidance, insightful suggestions, and continuous encouragement throughout this project. Lastly, I am grateful to Google Colab, Kaggle, and jupyter notebook for providing robust platforms that enabled the successful classification of traffic using the machine learning algorithms. Their integration has made it possible to create a system that enhances efficiency in the model performance. Thank you all for your support and guidance.

CONCLUSION

IoT network traffic classification using machine learning algorithms has emerged as a critical solution to address the limitations of traditional port-based and payload-based methods, particularly in handling encrypted traffic and dynamic network environments.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

In this research, we investigated the application of machine learning techniques for the accurate and efficient classification of IoT network traffic. The results demonstrate the viability of utilizing statistical features and behavioral patterns, extracted from network flows, to effectively distinguish between various IoT device types and applications. By employing specific ML models, decision trees, random forest classifier, ensemble methods, deep learning architectures, we achieved significant improvements in classification accuracy compared to traditional methods. Furthermore, our analysis highlighted the importance of key findings, e.g., feature selection, model optimization, real-time processing in achieving robust and scalable performance in dynamic IoT environments.

The proposed framework offers a valuable tool for network administrators to enhance security monitoring, optimize resource allocation, and improve overall network management in the face of the ever-increasing volume and complexity of IoT traffic. Future work will focus on addressing challenges such as handling encrypted traffic, adapting to concept drift, deploying lightweight models on resource-constrained devices. Exploring federated learning approaches to preserve data privacy and developing adaptive models capable of real-time classification in highly dynamic environments are also crucial directions for future research. Ultimately, the advancement of ML-based IoT traffic classification is essential for ensuring the secure and efficient operation of future IoT ecosystems.

REFERENCES

[1] Moustafa, N., & Slay, J. (2015). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). MILCOM 2015.
[2] García, S., et al. (2020). An empirical evaluation of botnet detection methods using the IoT-23 dataset.
[3] A. Tavallaee et al., "A detailed analysis of the KDD CUP 99 data set," IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009.
[4] Scikit-learn documentation. https://scikit-learn.org
[5] TensorFlow documentation. https://www.tensorflow.org
[6] https://colab.research.google.com/
[7] https://www.kaggle.com/
[8] https://www.python.org/