Traffic Analysis of UNSW-NB15 Dataset

Abstract

The significant increase in technology development over the internet makes network security a crucial issue. An intrusion detection system (IDS) shall be introduced to protect the networks from various attacks. Even with the increased amount of works in the IDS research, there is a lack of studies that analyze the available IDS datasets. Therefore, this study presents a comprehensive analysis of how the features in UNSW-NB15 dataset can be used in an efficient way so that they result in more accurate and precise predictions when used in real world Intrusion Detetction Systems (IDS). The features were also frequently selected and drop during various process. The findings of this study are anticipated to help the cybersecurity academics in creating a lightweight and accurate IDS model with a smaller number of features for the developing technologies.

1. Introduction

The Australian Center for Cyber Security (ACCS), Cyber Range Lab has made an IXIA PerfectStorm tool for creating the hybrid of synthetic contemporary attack behavior and real modern activities. To find the 100 GB of raw traffic (for example, Pcap files), a Tcpd ump tool is used. By utilizing the tools like Argus and Bro-IDS, and by developing 12 algorithms collectively 45 features with the class label of total 2,540,044 records. Dataset Features are broadly categorized into 6 subsets as Labeled Features, Time Features, Content Features, Flow Features, Basic and features which are additionally generated. More attacks on UNSW-NB15 are further categorized as 9 various types, namely Worms, Fuzzers, DoS, Exploit, Reconnaissance, Backdoor, Analysis, Shellcode, and Generic. By further categorizing additional generated features, two sub-groups formed namely

Connection and General Purpose Features. The numbering of features from 36-40 and from 41-47 are called as General Purpose and Connection Features respectively.

Keywords : Analysis, Intrusion Detection, UNSW-NB15 Dataset, Attack Category

2. Literature Review

i) UNSW-NB15 Dataset Feature Selection and Network Intrusion Detection using Deep Learning [1]

They applied a combination fusion of Random Forest Algorithm with Decision Tree Classifier in which 45 features have been decreased to the strongest four features. The proposed system detects normal and attacks with a better accuracy using Deep Learning technique.

ii) A network forensic scheme using correntropy-variation for attack detection [2]

They used the UNSW-NB15 Dataset for packet-feature analysis; however, their model is reasonably weaker. The selected features used in the model for network forensic analysis are inadequate for accurately predicting attacks on the system. Hence, unlike our model, it may leave loopholes of breaches in the NIDS system due to information loss.

iii) A new feature selection IDS based on genetic algorithm and SVM [3]

In this paper, for each attack in the UNSW-NB15 dataset, they introduced a hybrid model for IDS based on a Genetic Algorithm (GA) and Support Vector Machine (SVM). They converted the features into chromosomes and selected the highest accuracy from them. Then, as a detection method, they proposed the Least Squares Support Vector Machine (LSSVM). The results were tested for accuracy, true positive rate, and false-positive rate.

iv) A new feature selection IDS based on genetic algorithm and SVM [4]

The machine learning (ML) algorithms that have been used here are Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Artificial Neural Network (ANN). The evaluation metrics used in the comparison of performance are accuracy, precision, recall, f1 score, and area under the Receiver Operating Characteristic Curve. The system obtained 99.4% test accuracy for Decision Tree, Random Forest, and ANN.

v) Experimental evaluation of a multi-layer feedforward artificial neural network classifier for network intrusion detection system [5]

In this paper, a deep learning binomial classifier for Network Intrusion Detection System is proposed and experimentally evaluated using the UNSW-NB15 dataset. Three different experiments were executed in order to determine the optimal activation function, then to select the most important features and finally to test the proposed model on unseen data.

vi) Attack Detection in IoT using Machine Learning [6]

In this paper, a framework is recommended for the detection of malicious network traffic. The framework uses three popular classification-based malicious network traffic detection methods, namely Support Vector Machine (SVM), Gradient Boosted Decision Trees (GBDT), and Random Forest (RF), with RF supervised machine learning algorithm achieving far better accuracy

vii) An Experimental Analysis of Attack Classification using Machine Learning in IoT Networks [7]

Authors in used ML techniques such as KNN, SVM, DT, Naïve Bayes, neural networks, and RF which can be applied in IDS. The authors compared ML models for multi and binary class combinations on the data set of Bot-IoT

viii) Detection of unauthorized iot devices using machine learning techniques [8]

In this paper, Random Forest algorithm, was applied to features extracted from network traffic data.

ix) Towards the development of realistic botnet dataset in the internet of things for network forensic analytics [9]

They used LSTM, SVM, and RNN machine learning models to evaluate the IoT dataset, but in their analysis they did not determine the adversarial robustness of their models.

x) Survey of Random Forest Based Network Anomaly Detection Systems [10]

In this paper, the authors propose an approach to Anomaly Detection of IoT system based on Random Forest which can effectively provide protection to all types of attacks and to reduce false positive rate in the system.

3. Description of UNSW-NB15 dataset

There are nine attackstypes discovered in UNSW-NB15 Dataset.

i. Fuzzers: An attack in which the attacker tries to discover security loopholes in the Operating System, program or network and make these resources suspended for some time period and can even crash them.

ii. Analysis: A type intrusions that penetrate the web applications through port scanning, malicious web scripting and dispatching spam emails etc.

iii. Backdoor: A technique in which attacker can bypass the usual authentication and can get unauthorized remote access to a system.

iv. DoS: An intrusion in which attacker tries to disrupt the computing resources, by making them extremely busy in order to prevent the authorized access to the resources.

v. Exploit: The intrusions which utilize the software vulnerabilities, error or glitch within the operating systems(OS) or software.

vi. Generic: This attack act against a cryptographical system and it tries to break the key of the security system.

vii. Reconnaissance: It can be defined as a probe; an attack that gathers information about the target computer network in order to bypass its security control.

viii. Shellcode: A malware attack in which the attacker penetrates a slight piece of code starting from a shell to control the compromised machine.

ix. Worms: Malware that replicate themselves and spread to other computers by using the network to spread the attack, depending on the security failures on the target computer which it want to access.

The UNSW-NB15 data set features are classified into six groups as follows:

1) Flow features: These features have the identifier attributes between hosts, such as client-to-serve or server-to-client.

2) Basic features: These features include the attributes that represent protocols connections.

3) Content features: These features contain the attributes of TCP/IP; also they contain some attributes of http services.

4) Time features: This group contains the attributes of time, for example, arrival time between packets, start/end packet time and round trip time of TCP protocol.

5) Additional generated features: This group can be further divided into two groups: (1) General purpose features which each feature has its own purpose, in order to protect the service of protocols. (2) Connection features are built from the flow of 100 record connections based on the sequential order of the last time feature.

6) Labelled Features: This category represents the label of each record

4. Methodology

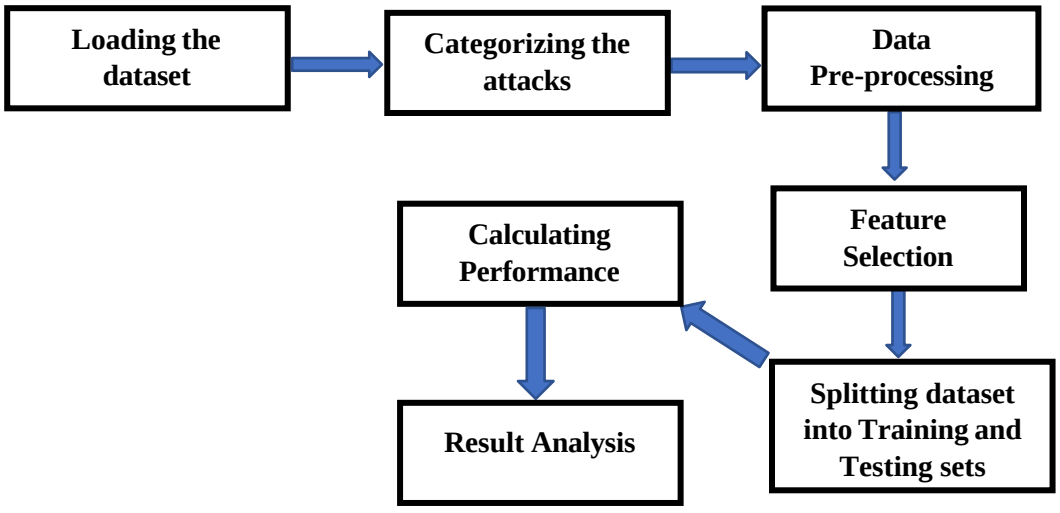


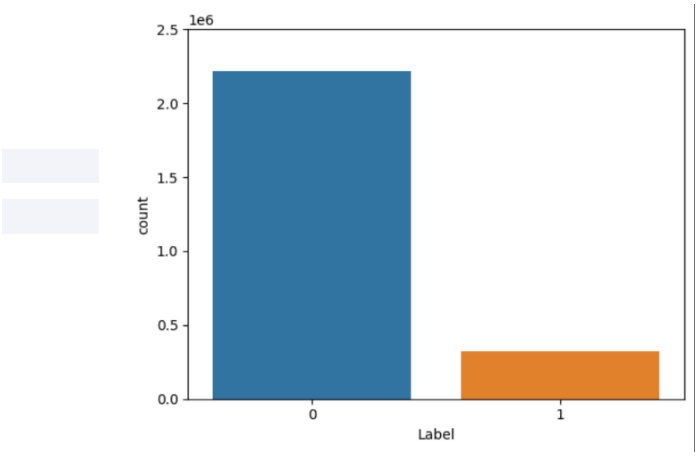
Figure 1: Analysis Architecture

4.1. Loading the dataset

The dataset consists of 4 partitions(UNSW_NB15_1.csv, UNSW_NB15_2.csv, UNSW_NB15_3.csv, UNSW_NB15_4.csv) which are loaded and concatenated to form a single dataframe. The loaded dataset consists of 49 features in which 1-47 are considered as independent features and 48,49 are considered as dependant features. There are totally 25,40,047 records in all the four partitions of the dataset.

4.2. Categorizing the attacks

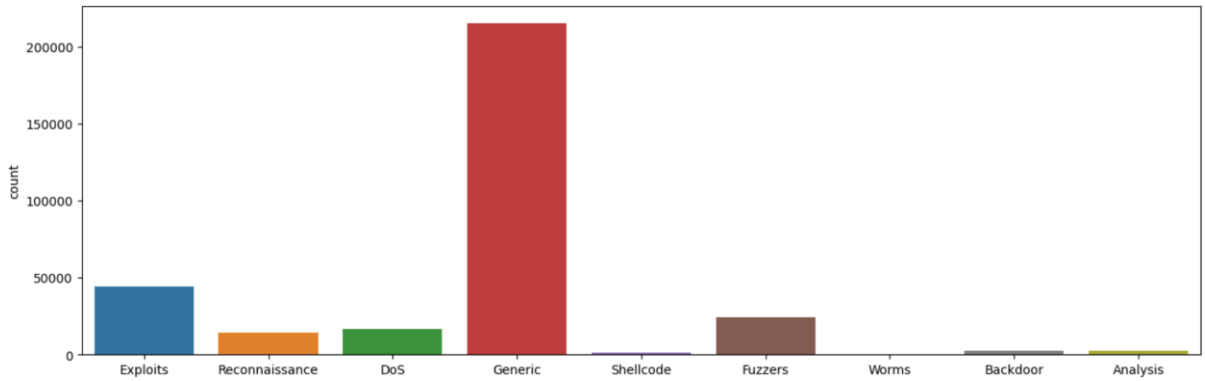
The dataset totally consists of 3,21,283 attack records inside it, which is visualized in below given graph.



0 – Non Attacks

1 - Attacks

The attacks are further classified into different categories and scattered all over the dataset in variable counts which can be visualized in the following graph.



The scattering of records of different attack categories are shown in the below table.

Exploits	44525
Reconnaissance	13987
Dos	16353
Generic	215481
Shellcode	1836
Fuzzers	24246
Worms	174
Backdoor	2328
Analysis	2677

4.3. Data Pre-processing and Feature Extraction

The dataset is checked for NULL values and if found in the columns of independent variables the corresponding features are dropped. The dependent and independent features are separated. The features that will serve good for training the model are identified and the remaining features are dropped. In the end 37 independent features are identified and extracted to train the model along with the independent features.

4.4. Splitting data into training and testing sets and transforming.

It is important to split dependent and independent features as training set and testing set. Here, we split the original data as 70% training set and 30% testing set. Scikit-learn provides a library of transformers. Like other estimators, these are represented by classes with a fit method, which learns model parameters (e.g. mean and standard deviation for normalization) from a training set, and a transform method which applies this transformation model to unseen data.

In addition, it is very common to want to perform different data transformation techniques on different columns in your input data. The ColumnTransformer is a class in the scikit-learn library that allows you to selectively apply data preparation transforms. For example, it allows you to apply a specific transform or sequence of

transforms to just the numerical columns, and a separate sequence of transforms to just the categorical columns. In our case we need to perform OneHotEncoder on categorical columns and StandardScaler on numerical columns. Again we will apply LabelEncoder to training and testing set of the dependant features.

5. Experimental Results and Analysis

5.1. Model Selection

We will train several machine learning models for the training set and evaluate their performance on both training and testing set. Before doing this, let's first go through a standard procedure of training a certain classifier. The classifiers we selected includes the following

- **Logistic Regression :** Logistic regression is one of the most popular Machine learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

- **Decision Tree :** Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the

given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

- **Random Forest :** RF is a machine learning approach that uses decision trees. In this method, a “forest” is created by assembling a large number of different decision tree structures that are formed in different ways. This algorithm has many advantages, such as the ability run on huge datasets efficiently, its light weight compared to other methods, and robustness against noise and outliers when compared to single classifiers.

5.2 Training and Testing ML Models

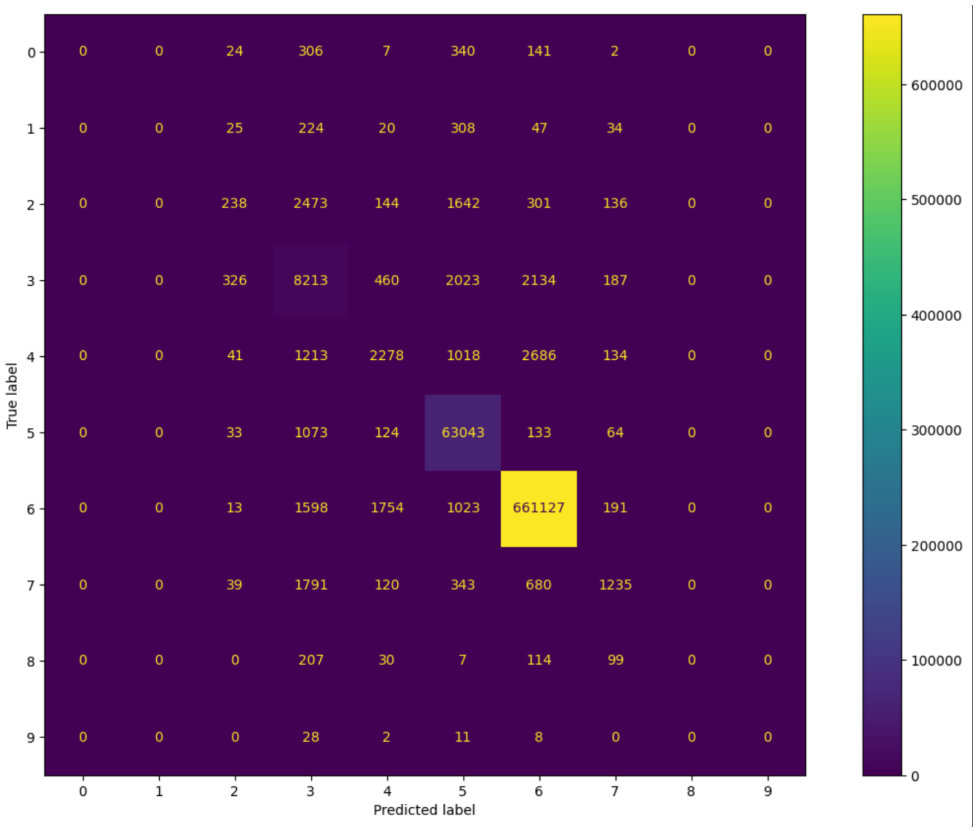
The training dataset is fit to the each classifier to train the model. The testing dataset is again fed to the model to make predictions. Based on the predictions on the testing set performance metrics such Accuracy, Precision and Recall are calculated which helps us in identifying the best classifier for designing the model.

5.3 Comparing Models with performance metrics

The performance metrics of various classifiers are listed below which shows us that Logistic Regression gives us the best accuracy.

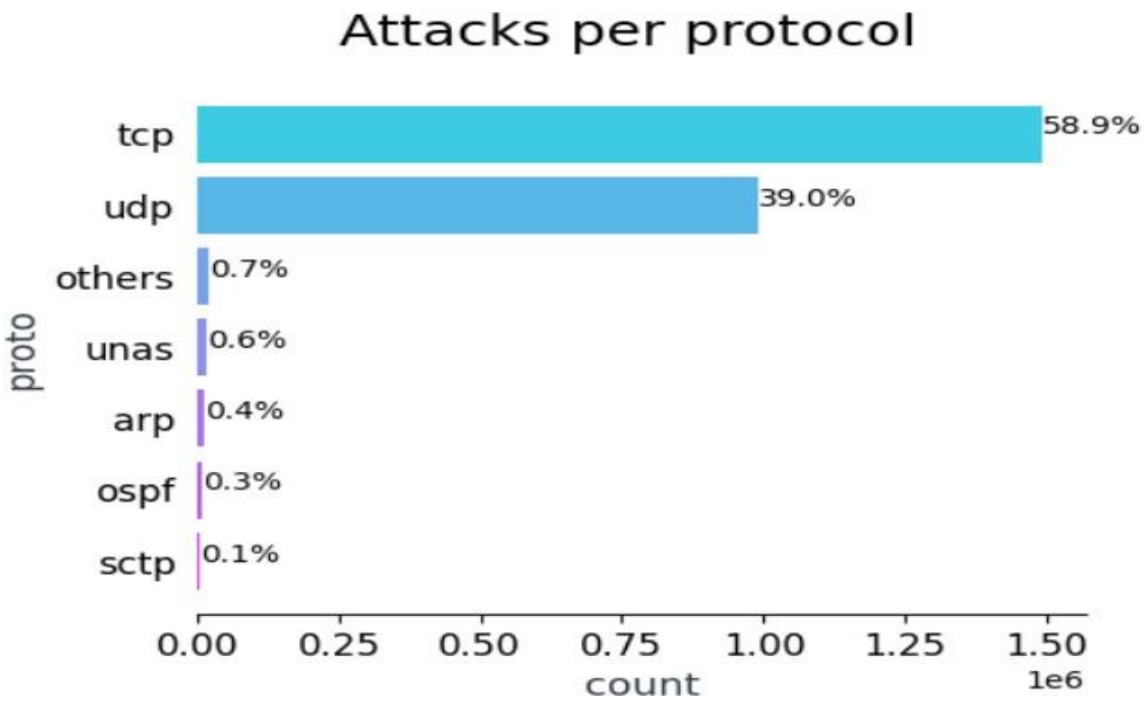
Sl.No	Classifier	Accuracy	Precision	Recall	F1 score
1	Logistic Regression	0.966	0.966	0.966	0.966
2	Decision Tree	0.886	0.886	0.886	0.866
3	Random Forest	0.880	0.880	0.880	0.880

The confusion matrix for the predictions made can be given as follows

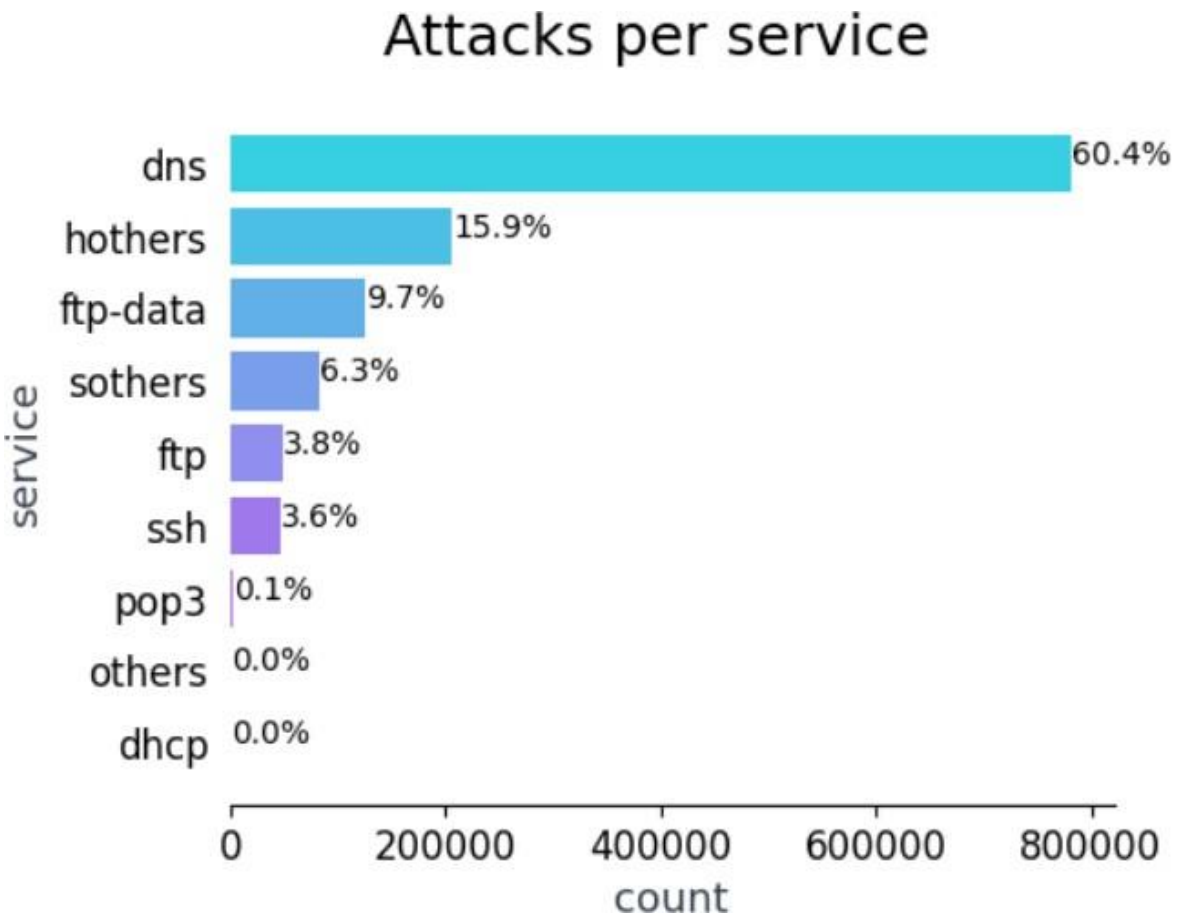


5.4 Data Analysis

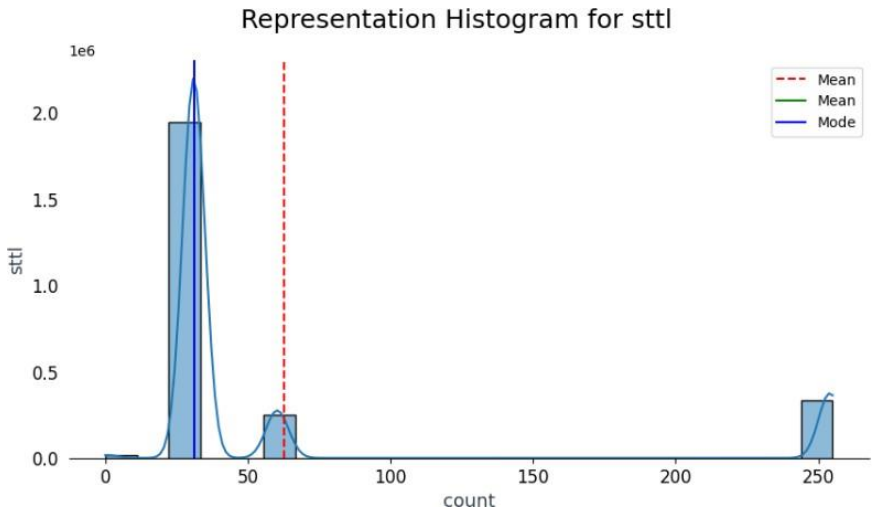
The following the section gives us the distribution of attacks for different values of important features in the dataset.



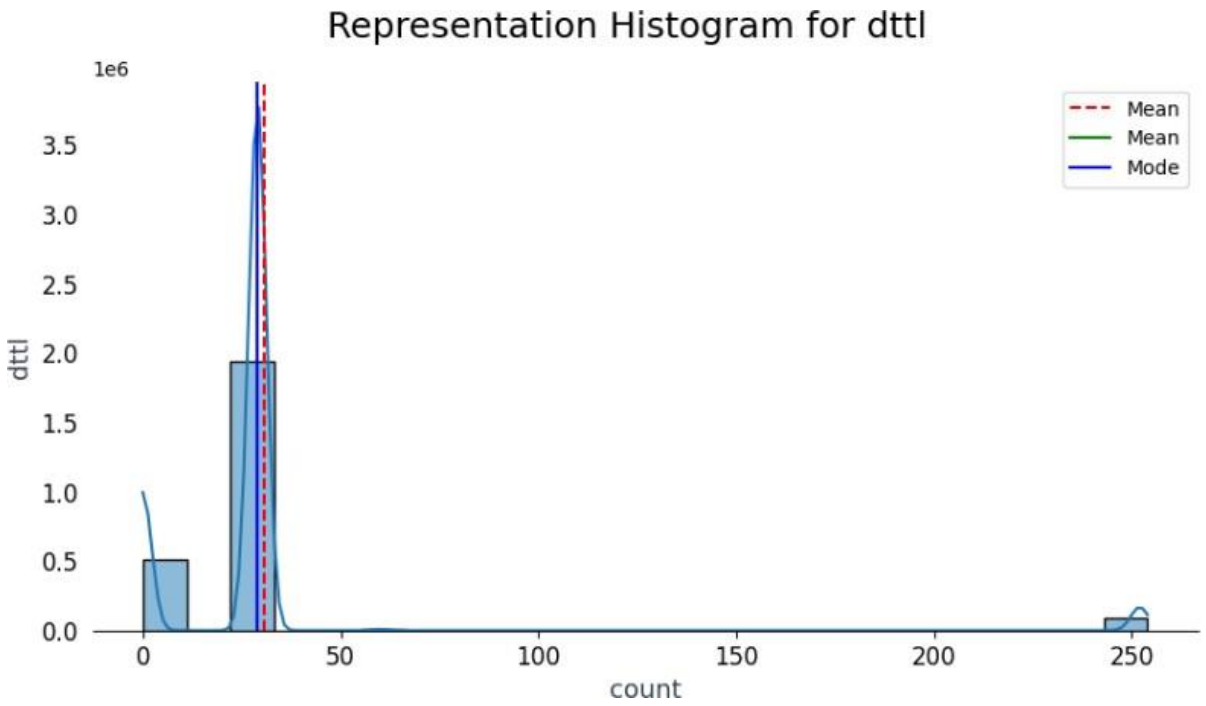
The above bar graph represents the percentage of attacks that happened on various types of protocols.



The above bar graph represents the percentage of attacks that happened on various types of services.



The above histogram represents the distribution of attack and non-attack for Source Time to Live and the red dotted (Mean) which separate attack and non-attack as Attack lies below 60 and Non-Attack lies above 60.



The above histogram represents the distribution of attack and non-attack for Destination Time to Live and the red dotted (Mean) which separates attack and non-attack as Attack lies below 35 and Non-Attack lies above 35.

6. Conclusion

In today's evolution of cyberthreats, IDS and IPS systems are playing an inevitable role in detecting and preventing those attacks. Implementing machine learning models in such systems efficiently helps in classifying and identifying those threads. In our research on this UNSW-NB15 dataset we had shown how this dataset can be used to train those machine learning model with high performance so that they could help in detecting cyberattacks.

7. References

[1] V. Kanimozhi, Prem Jacob. UNSW-NB15 Dataset Feature Selection and Network Intrusion Detection using Deep Learning. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-5S2, January 2019.

[2] Moustafa N., Slay J. A network forensic scheme using correntropy-variation for attack detection. IFIP Adv. Inf. Commun. Technol. 2018

[3] H. Gharaee and H. Hamid, “A new feature selection IDS based on genetic algorithm and SVM,” in Proceedings of the 2016 8th International Symposium on Telecommunications (IST), IEEE, Tehran, Iran, June 2016.

[4] Mahmudul Hasan. Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. Internet of Things 7 (2019) 100059

[5] Al-Zewairi M., Almajali S., Awajan A. Experimental evaluation of a multi-layer feedforward artificial neural network classifier for network intrusion detection system; Proceedings of the 2017 International Conference on New Trends in Computing Sciences (ICTCS); Amman, Jordan. 11–13 October 2017

[6] Maryam Anwer, Shariq Mahmood Khan, Muhammed Umar Farooq, Waseemullah. Attack Detection in IoT using Machine Learning Engineering, Technology & Applied Science Research Vol. 11, No. 3, 2021, 7273-7278.

[7] A. Churcher et al., "An Experimental Analysis of Attack Classification using Machine Learning in IoT Networks," Sensors, vol. 21, no. 2, Jan2021, Art. no. 446

[8] Y. Meidan, M. Bohadana, A. Shabtai, M. Ochoa, N. O. Tippenhauer, J. D. Guarnizo, and Y. Elovici, “Detection of unauthorized iot devices using machine learning techniques,” arXiv preprint arXiv:1709.04647,2017

[9] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, “Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset,” Future Generation Computer Systems, vol. 100, pp. 779–796, 2019.

[10] Rashmi H Roplekar and N V Buradkar, “Survey of Random Forest Based Network Anomaly Detection Systems”, Vol. 6, Issue 12, December 2017