EFFECTIVE AND EFFICIENT INTRUSION DETECTION

VIA MACHINE LEARNING ON CICIOT2023 DATASET

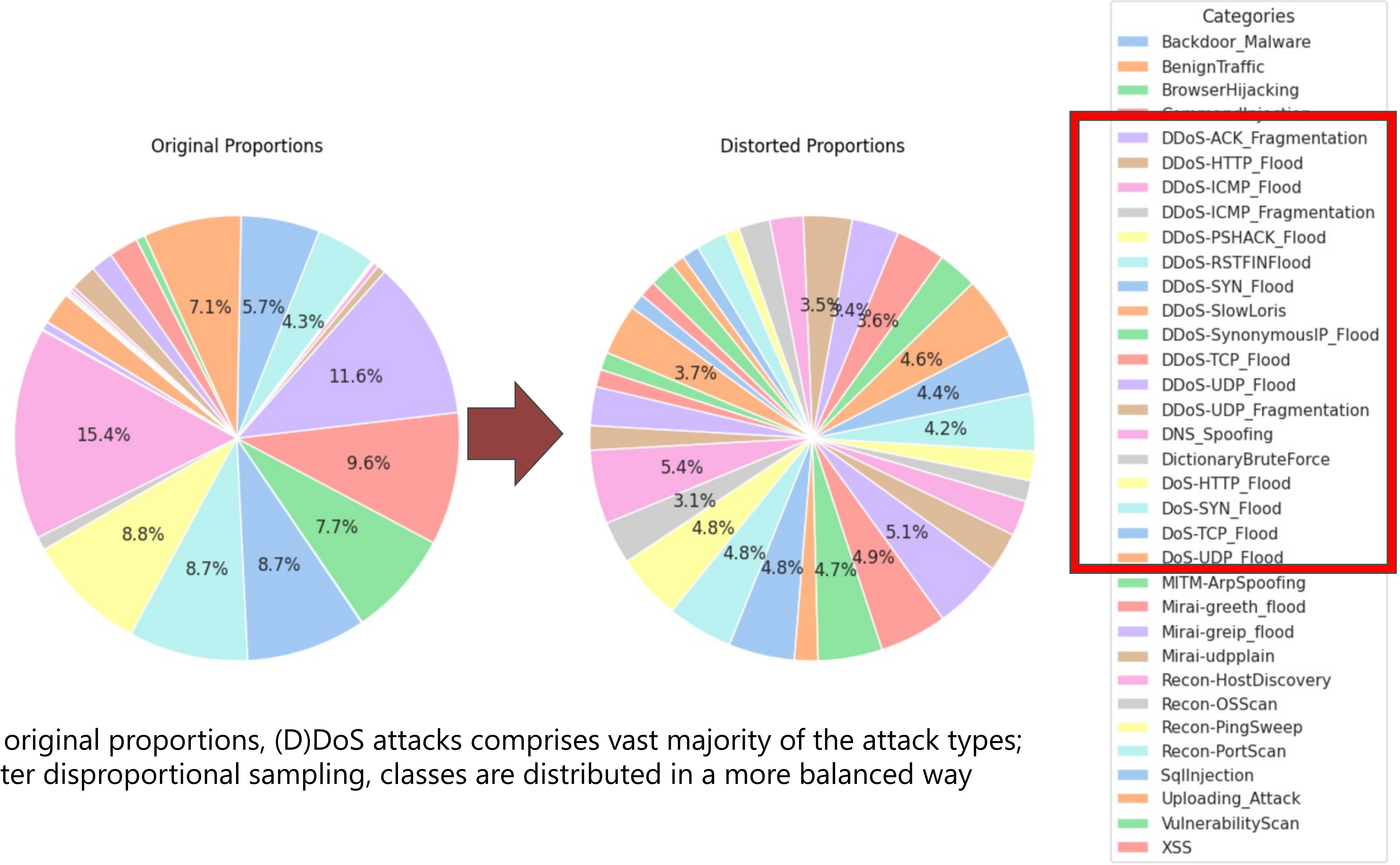
DAVID DAI 235821890

TABLE OF CONTENTS

- Effective Machine Learning
  - EDA & Preprocessing
  - Model Selection & Evaluation
- Efficient Machine Learning
  - Model performance on different scale of sampling data
  - Potential integration with IPS (Intrusion Prevention System)

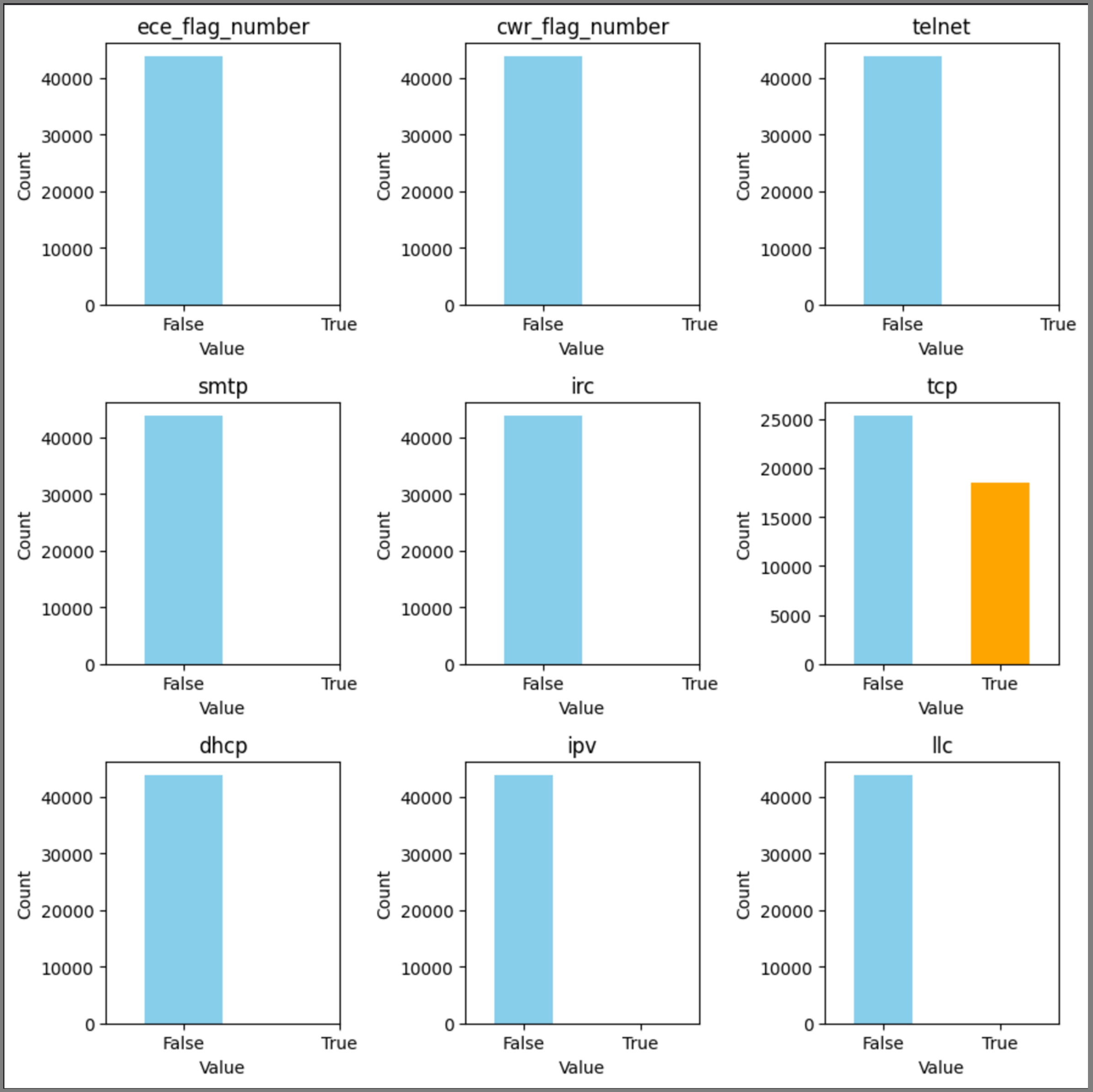
EDA & PREPROCESSING

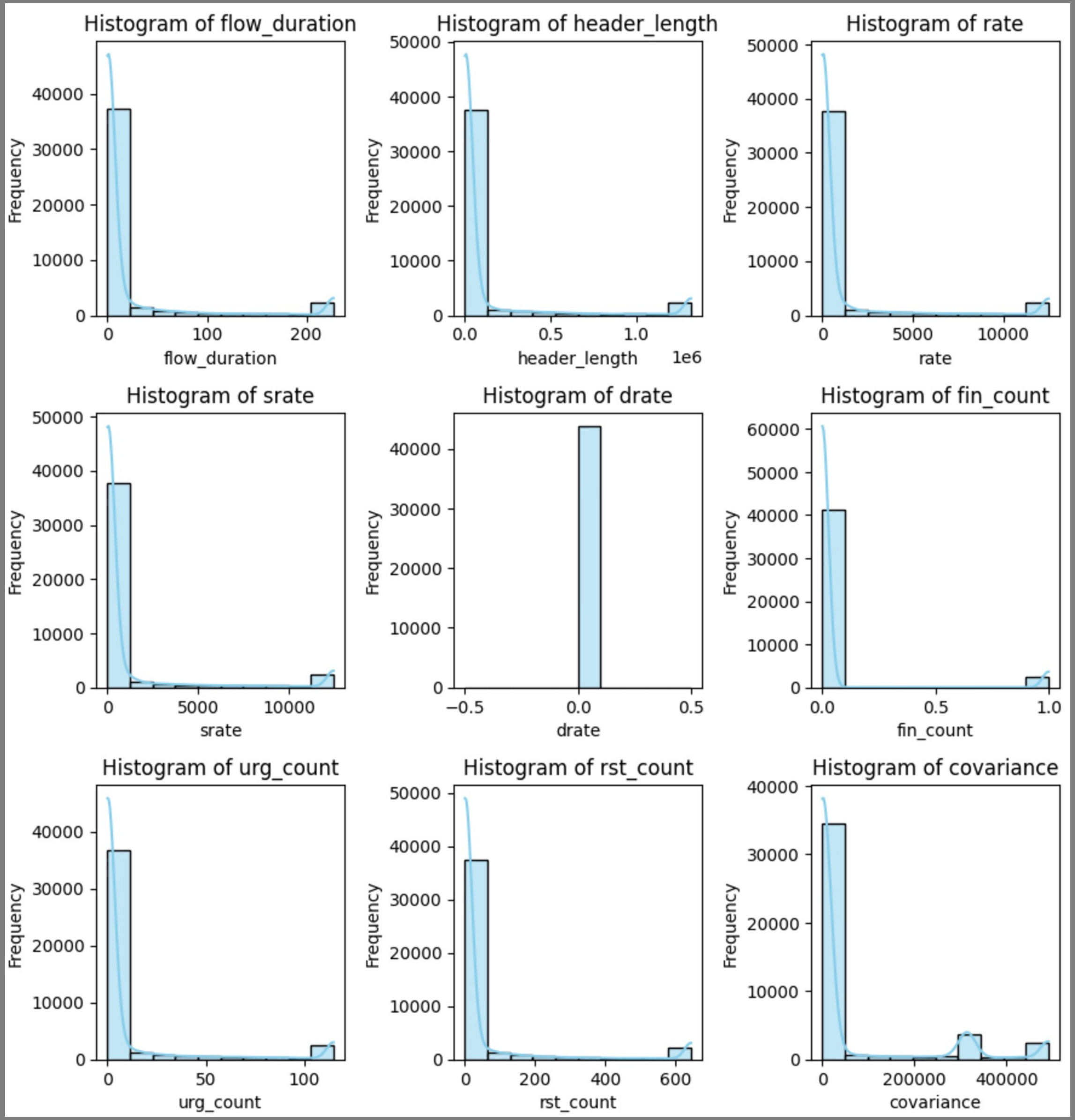
EDA Method	Typical Preprocessing/ Feature Engineering	Relevancy to current probelm
Data Cleaning	Handle missing values, duplicates, and outliers.	outlier is the challenge of the dataset, capping techniques are applied to extreme data
Univariate Analysis	Normalize/encode features, transform skewed data.	show skewness of the data and visualize with histogram, boxplot, drop column with limited info
Multivariate Analysis	Normalize, reduce multicollinearity, apply dimensionality reduction.	show Correlation table and reduce dimension with PCA/t-CNS
Target Label Analysis	Balance classes, transform skewed distributions.	the label class is highly biased , especially for (D)DOS. Disproportional Sampling is used to balance class
Time Series Analysis	Interpolate missing data, extract date/time features, apply smoothing.	N/A, the temporal information dataset has been compacted into each flow as row of data. We only have statistical info of each flow.



In original proportions, (D)DoS attacks comprises vast majority of the attack types;
After disproportional sampling, classes are distributed in a more balanced way

HIGH SKEWNESS – EXTREME DATA



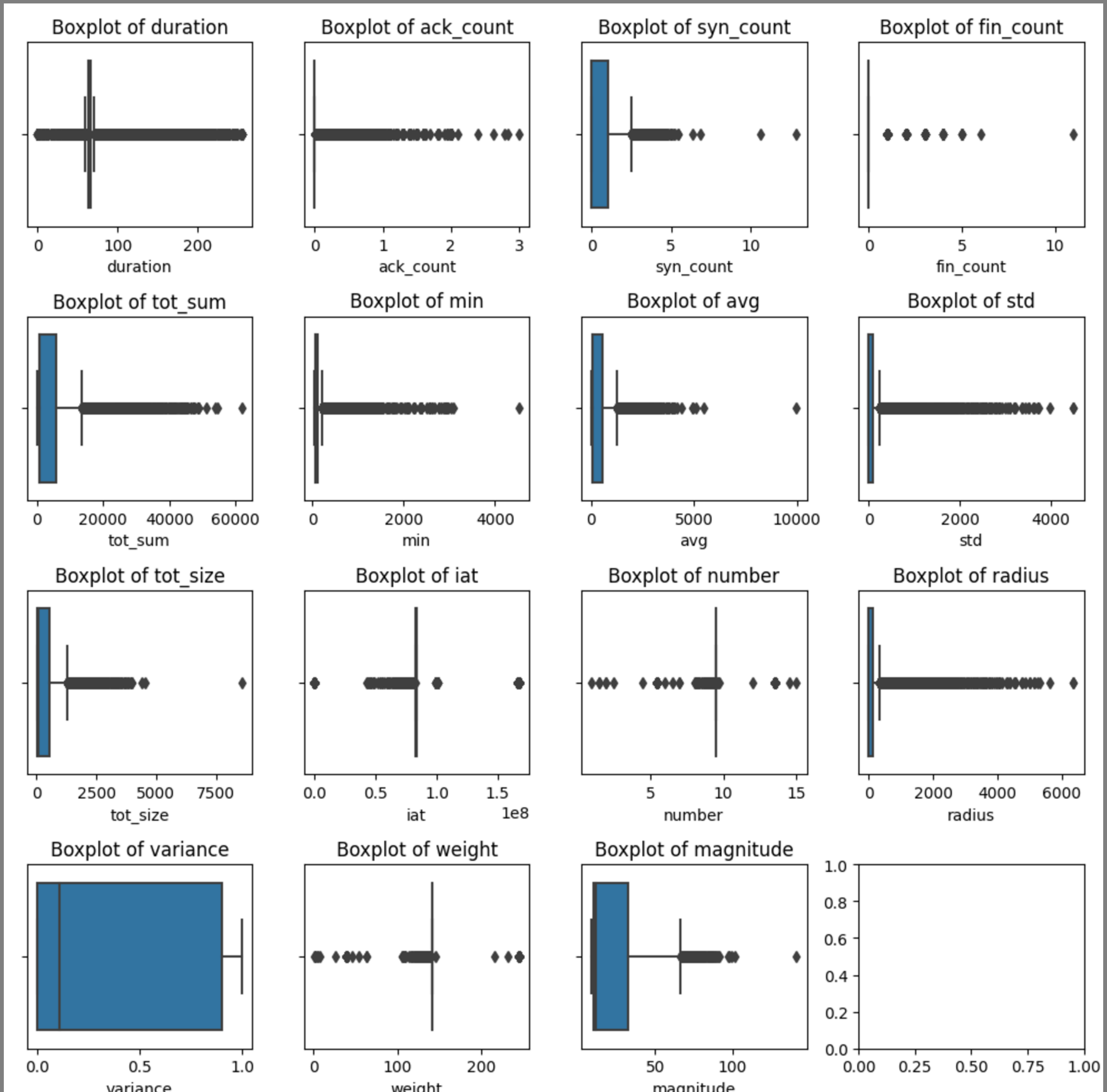


zero skewness for bool value => all single value except for tcp indicator

Extreme high skewness => capping outliers exceeding (0.05, 0.95)

HIGH SKEWNESS – "NORMAL" DATA

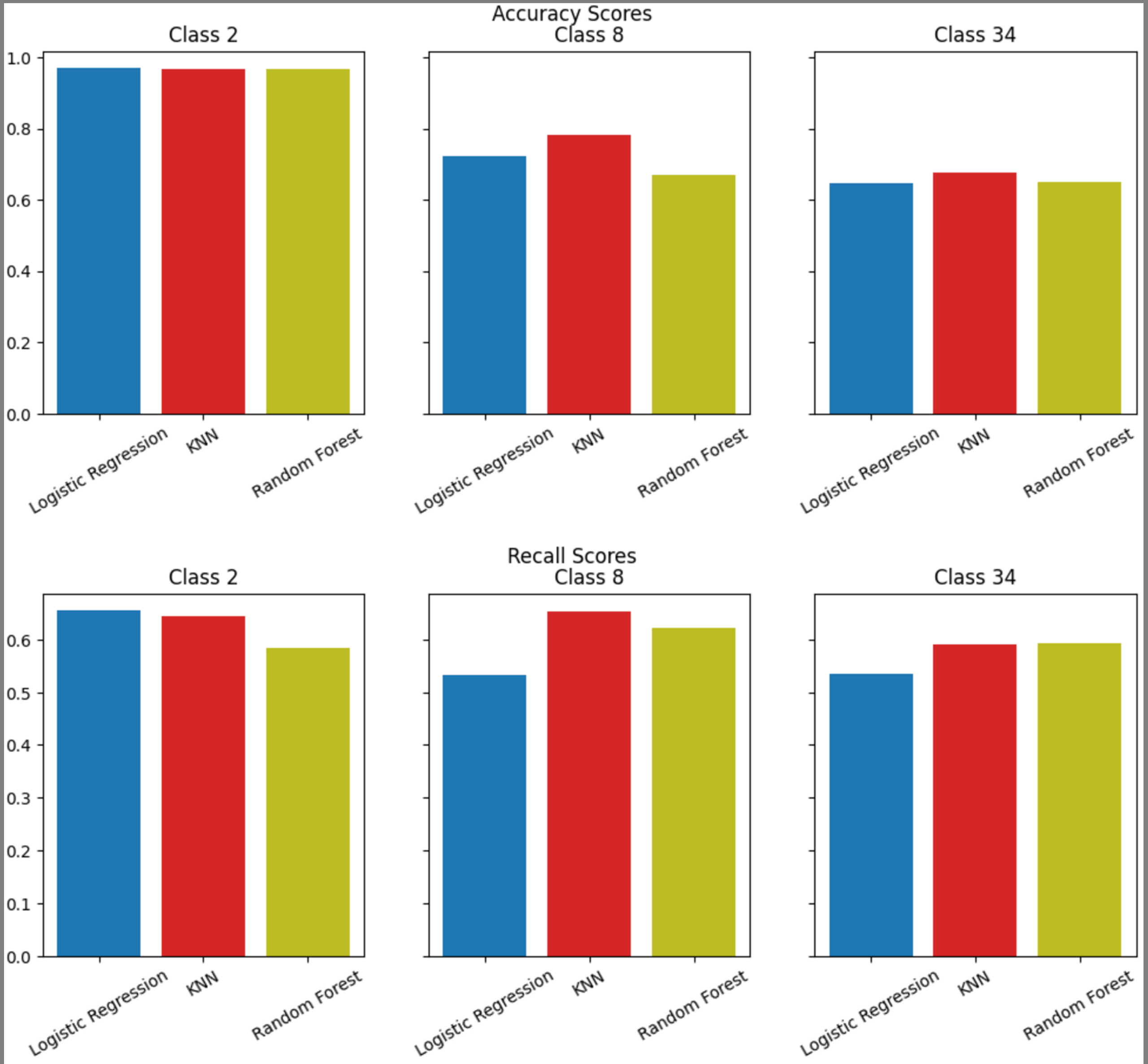
- Evident outliers
  - Almost all features, except for variance
- High skewness in feature with less outliers
  - The median value lean to one end of the box
  - scale transform(learning transform) doesn't work well
- How to improve?
  - Capping outliers – double-edged sword when the goal is to detect abnormality…
  - Transform skewness of the features – together after evaluating the effect of multivariate analysis

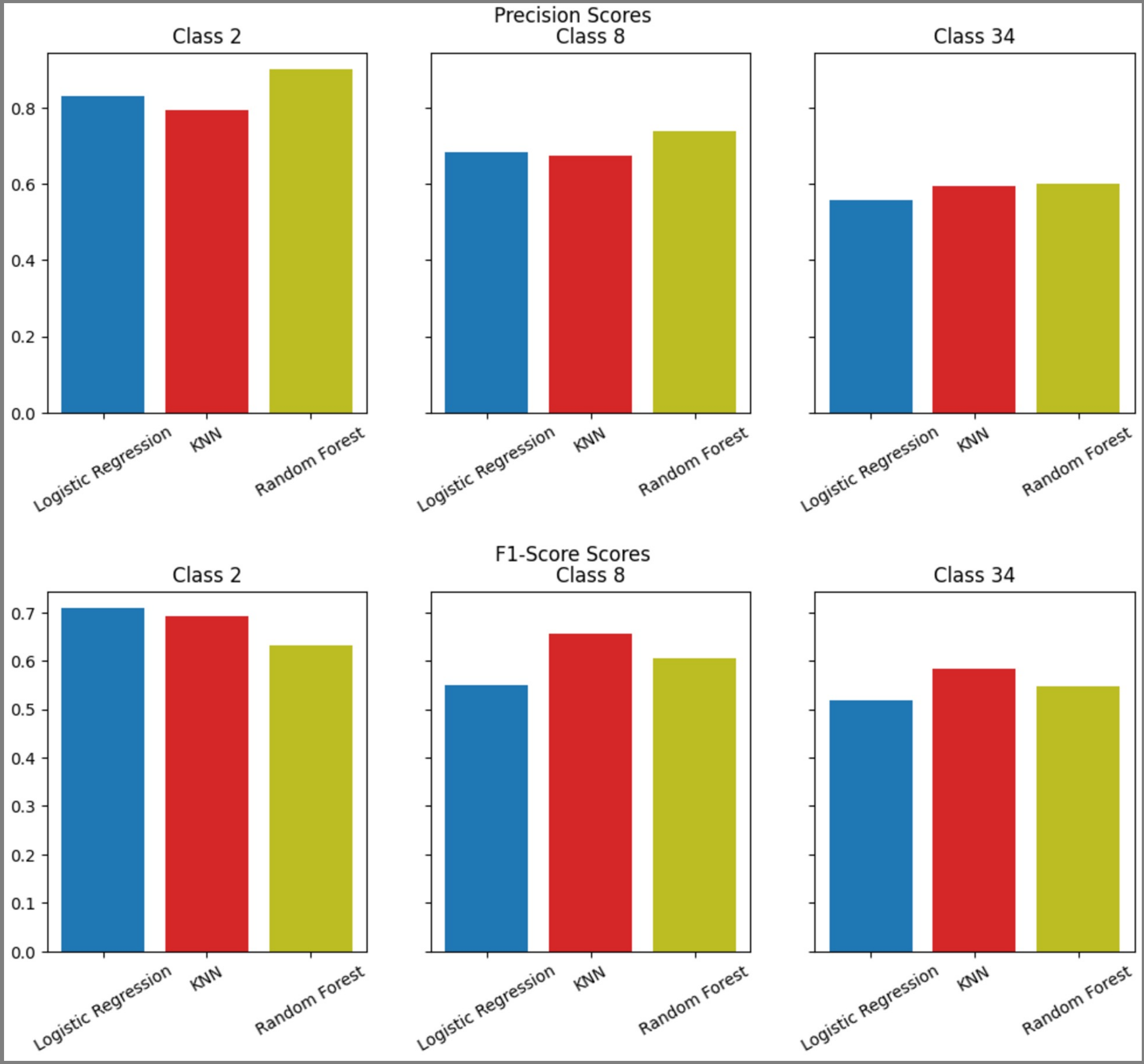


MODEL SELECTION

Model	**Type**	**Advantage**
Logistic Regression	Supervised	Easy to implement, computational efficient
KNN	Supervised	widely-used for classification of known number(k) of clusters, can upgrade to outlier robust variant: DBSCAN
Random Forest	unsupervised	Ensemble decision trees, robust to skewness of data
k-means	Unsupervised	similar to KNN, can be used as comparison with KNN
Auto-encoder	Unsupervised	Deep Learning model for abnormal detection

METRICS ANALYSIS

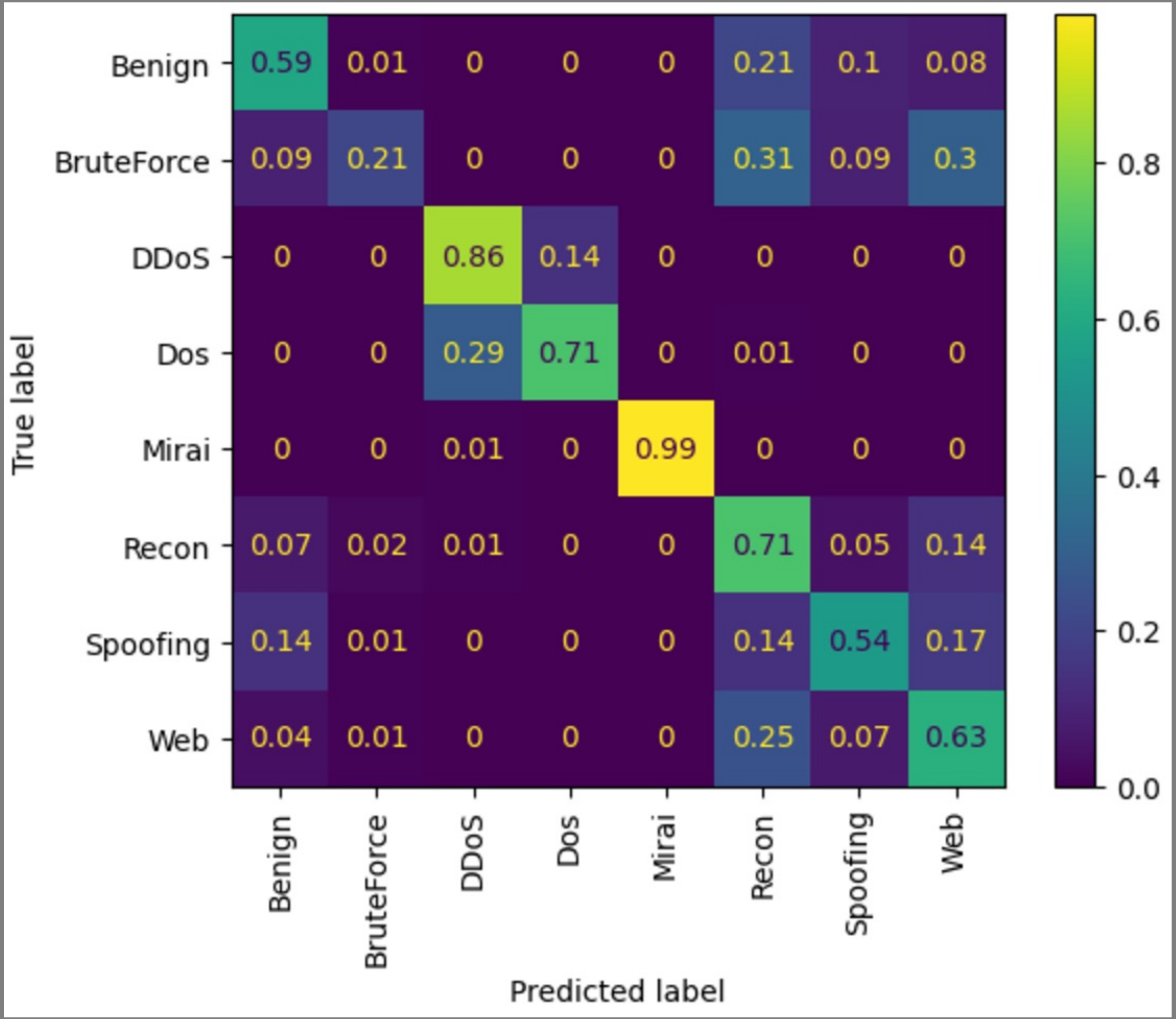




Accuracy/Recall/Precision/F1-score on smallest percentage(0.001) of original dataset

CLASSIFICATION ANALYSIS

- Two large group of attack type
  - (D)Dos-like: DDoS, DoS, Mirai
  - Recon-like: Brute Force, Recon, Spoofing, Web + Benign
- Two distinct characteristics
  - (D)Dos-like: very distinctive among others
  - Recon-like: easy to be mistaken as others
- How to improve?
  - Multi-variant analysis and dimension reduction
  - granular feature engineering on tangled categories



	Benign	BruteForce	DDoS	Dos	Mirai	Recon	Spoofing	Web
Benign	0.59	0.01	0	0	0	0.21	0.1	0.08
BruteForce	0.09	0.21	0	0	0	0.31	0.09	0.3
DDoS	0	0	0.86	0.14	0	0	0	0
Dos	0	0	0.29	0.71	0	0.01	0	0
Mirai	0	0	0.01	0	0.99	0	0	0
Recon	0.07	0.02	0.01	0	0	0.71	0.05	0.14
Spoofing	0.14	0.01	0	0	0	0.14	0.54	0.17
Web	0.04	0.01	0	0	0	0.25	0.07	0.63

Confusion Matrix : KNN in detecting 8 categories including Benign traffic vs. 7 types of attack traffic

SUMMARY & OUTLOOK

- Effective Machine Learning
  - Get some result, still room to improve
    - Multi-variant analysis
    - transform highly skewed data
    - More models
- Efficient Machine Learning
  - Have the potential to use small data to predict large data
  - Analysis on attribute importance together with optimal tuning library has the potential to transform ML result to rule settings for existing IPS

Suricata is a high-performance open-source Network IDS, IPS and Network Security Monitoring engine.
It uses **meerkat** (the meaning of "Suricata" in Latin) to resemble vigilance, adaptability, speed and efficiency and teamwork.