Machine Learning Based Adaptive Cybersecurity Incident Detection

---

Progress report: 70%

Thanawat Tejapijaya
Krittawat Thongnoppakao
Punyawat Jaroensiripong

Karin Sumongkayothin, Ph.D (Main-Advisor)
Suratose Tritilanunt, Ph.D (Co-Advisor)
Konglit Hunchangsith, Ph.D (Co-Advisor)
Prarinya Siritanawan, Ph.D (JAIST-Advisor)

Content

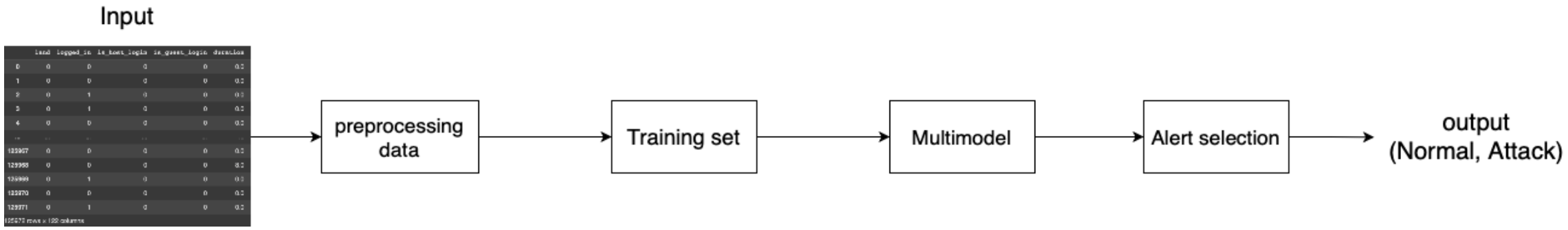
- Introduction
- Proposed Method Review
- Dataset
- Data preprocessing
- Experiments
  - Normal training
  - Mixed training
- Result
- On progress
- Future Work

Introduction

- In the Security Operations Center (SOC), the responsibility is to monitor systems to prevent cyber-attacks. Also, Security information and event management (SIEM) has been used in SOC and sometimes gets a lot of False Alarms.
- We will propose a method that uses Machine Learning (ML) not only to detect the cyber attack but also to adapt behavior following the expert of the SOC.

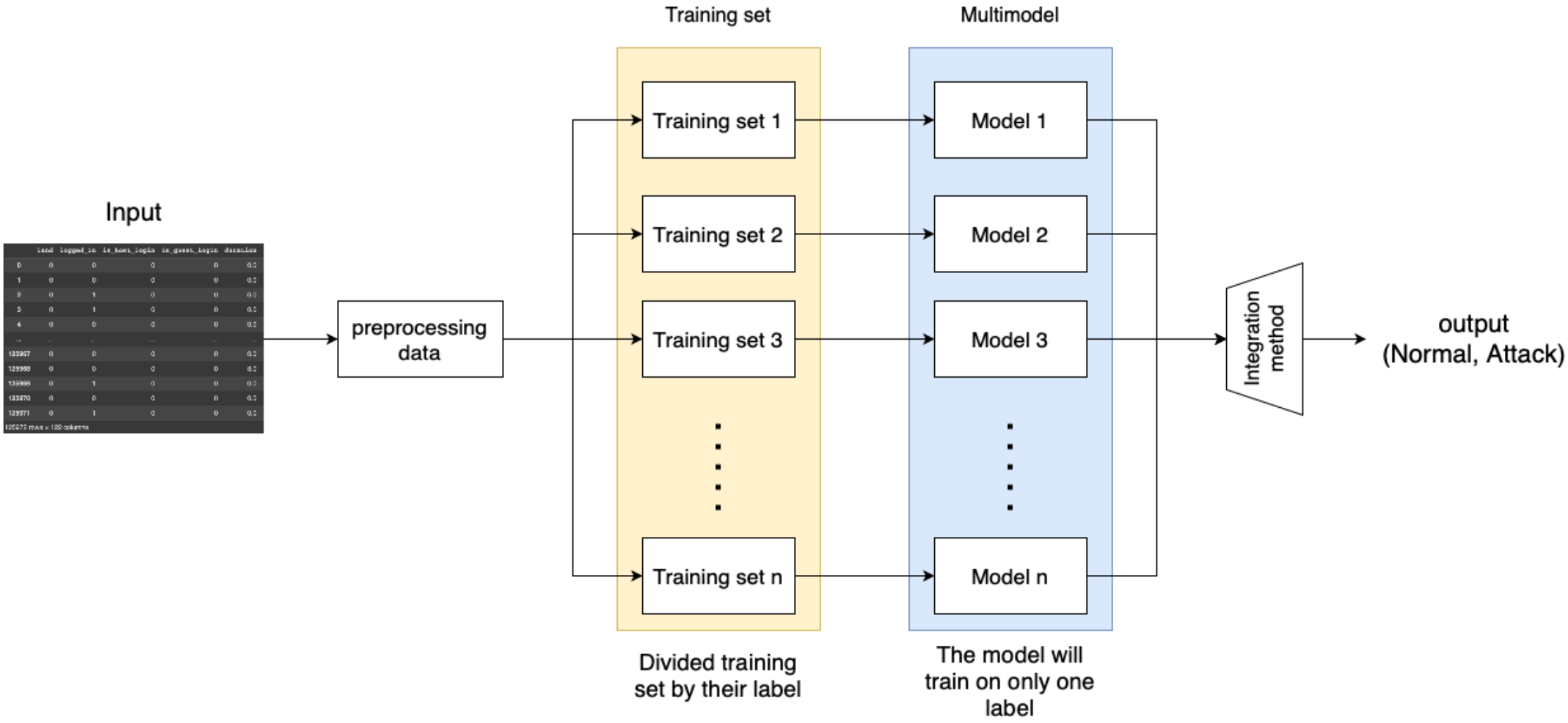
Proposed Anomaly detection Method Review

- Each model will be trained by different type of cyber attack.
- Integrate every trained model to predict one result.



Model training methodology

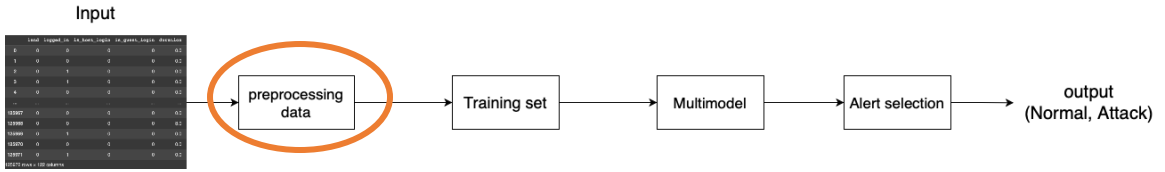
Overall Model Review



Dataset

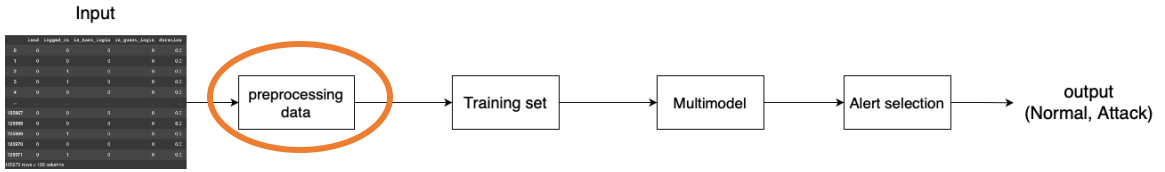
- CIC-IDS
  - There are 82 features with 16,601,424 data of network flow.
  - Containing 14 types of network attack.

- NSL-KDD
  - There are 42 features with 148,517 data of network flow.
  - Contain 4 major categories, that is Probing, Denial-of-Service (DoS), Remote to Local (R2L), and User to Root (U2R).
    - There are 40 attack subcategories.



Data preprocessing

- Using One-Hot Encoding to preprocess categorical data.
  - In NSL-KDD dataset, there are 3 features that contain categorical data (protocol_type, service, flag)
  - In CIC-IDS2017 dataset, we preprocess the 'port' feature by categorize the range of port into 3 categories following Internet Assigned Numbers Authority (IANA) port number.
    - port 0-1023 -> Well-known ports (1)
    - port 1024-49151 -> Registered ports (2)
    - port 49152 - 65535 -> Dynamic/Private ports (3)

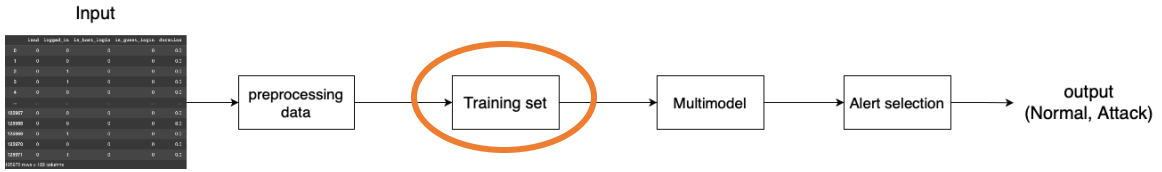


Data preprocessing

- Sorting the data by timestamp
- Dropping the non-numerical data (the data that cannot calculated)
  - In CIC-IDS2017 dataset, there are 4 non-numerical data ('flow_id', 'source_ip', 'destination_ip', 'timestamp')
- Use Min-Max scaler to scale the data in dataset into range (0, 1)

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

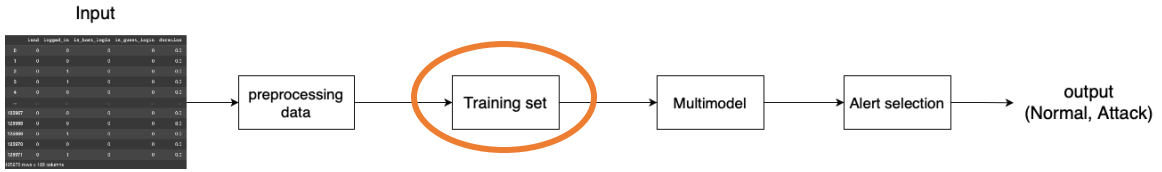
Min-Max scaler formular



Experiments

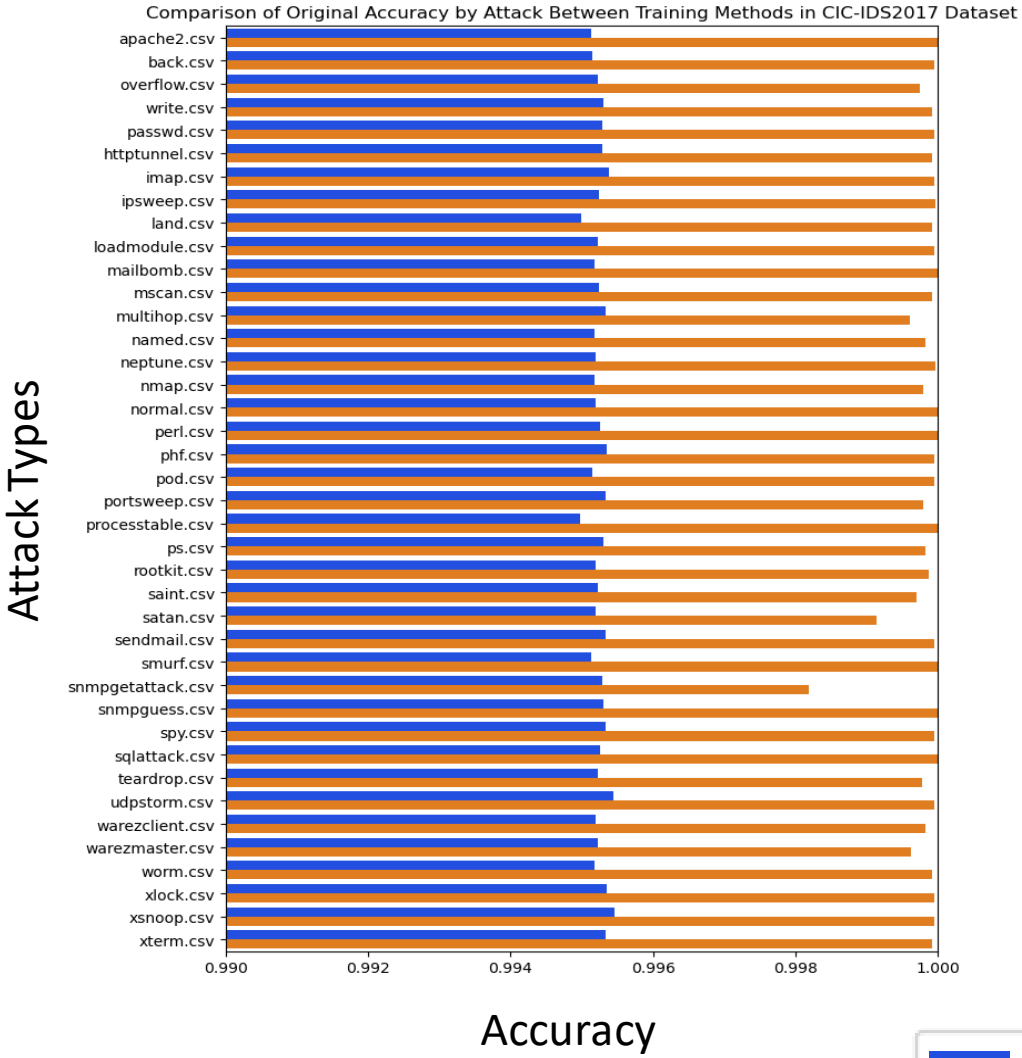
- Normal training set
  - Train the model using each attack categories and normal flow.
  - Consequently, the model will only learn on the specific attack with normal flow.
- Mixed training set
  - Train the model each attack categories along with the other attack that label as normal flow (Aiming to add noise) and genuine normal flow.
  - Therefore, the model will learn the characteristic of the specific attack and ignores the other attacks.

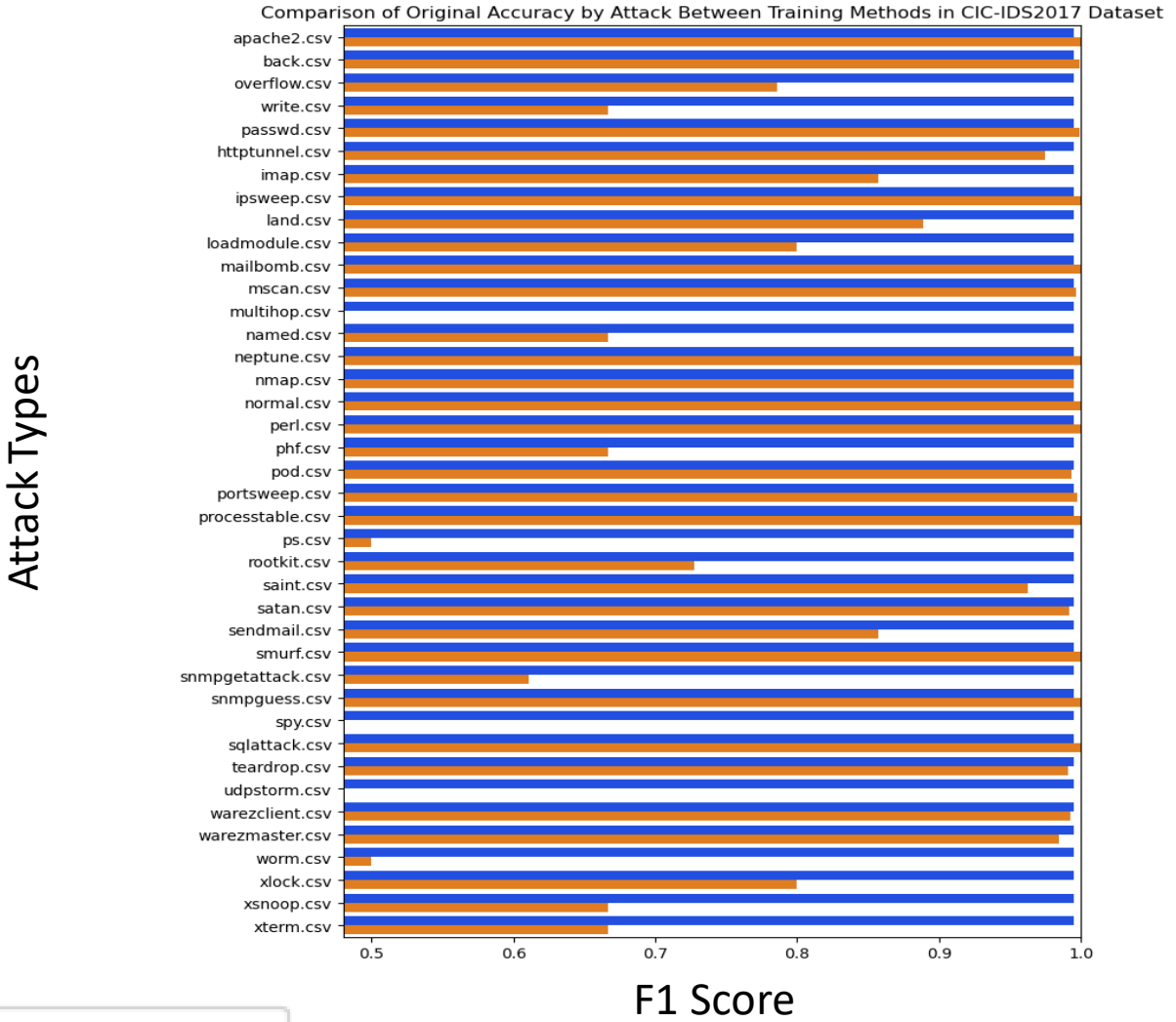
Experiments



- Selected Classical ML Model
  - Logistic Regression
  - Extra Trees Classifier
  - Bagging Classifier (using Decision tree as estimator)
  - Linear Discriminant Analysis (LDA)
  - Quadratic Discriminant Analysis (QDA)
  - Decision Tree Classifier
  - Random Forest Classifier
  - Gradient Boosting Classifier
  - K-Neighbors Classifier
  - Gaussian Naive Bayes
  - Linear perceptron classifier
  - Ada Boost Classifier

Result : NSL-KDD dataset



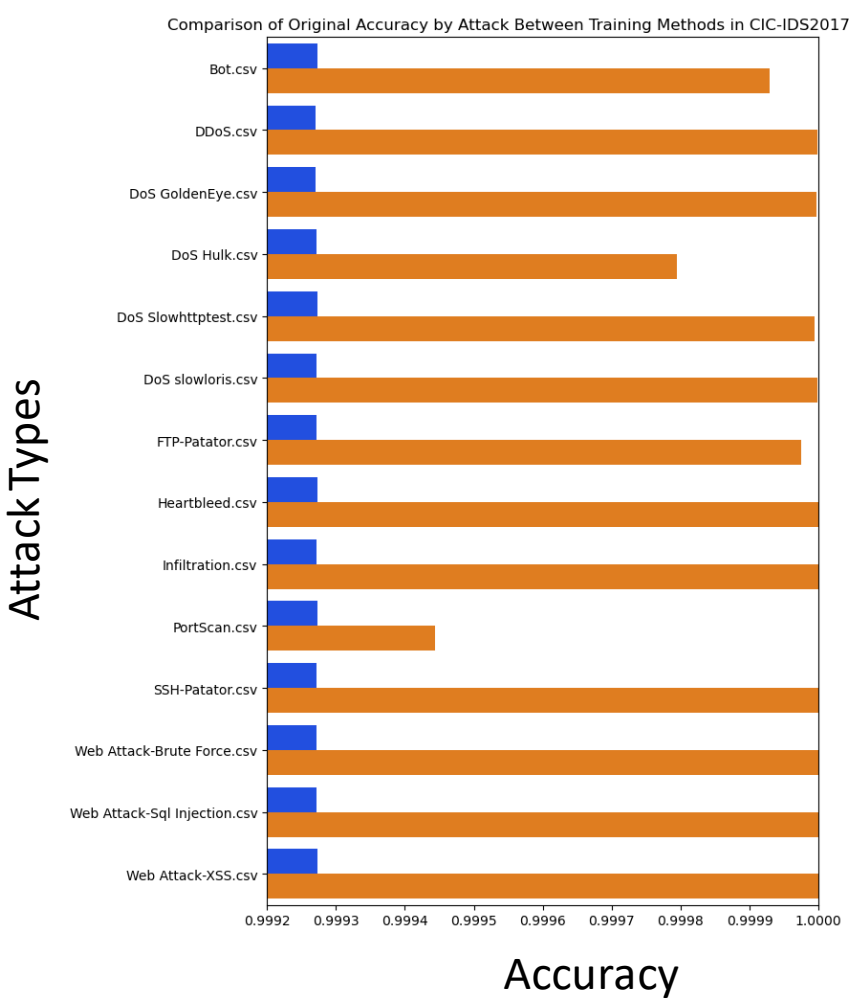


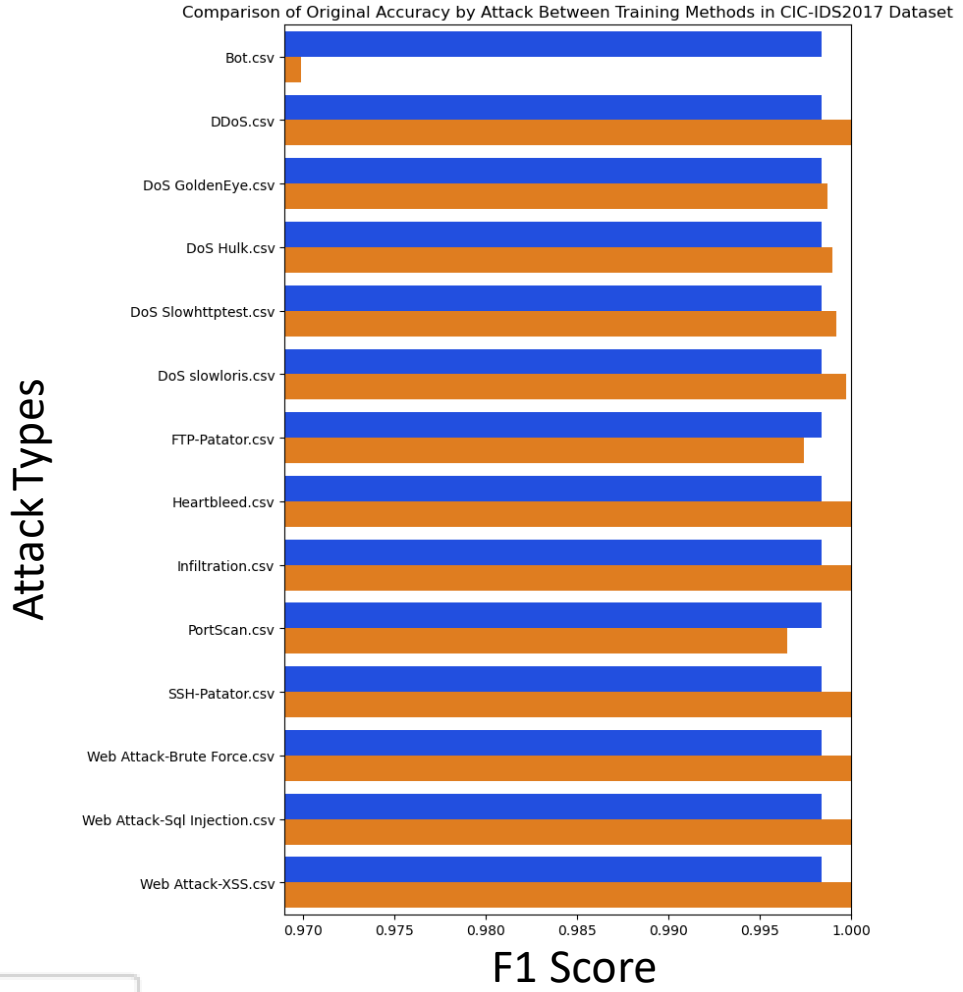
Result : NSL-KDD dataset

Training type	Mean Accuracy	Mean F1-Score
Normal KDD Training	0.9998	0.8251
Mixed KDD Training	0.9952	0.9951

The mean of overall of normal training and mixed training

Result : CIC-IDS2017





Result : CIC-IDS2017

Training type	Mean Accuracy	Mean F1-Score
Normal CIC Training	0.9999	0.9972
Mixed CIC Training	0.9993	0.9984

The mean of overall of normal training and mixed training

On progress

- Training the data on the sequential Deep learning models.

Future Work

- Implement integration method to integrate output from Multi-model into one output.
- Writing thesis

Q&A

Thank You