“Are Encrypted Client Hello and Encrypted Server Name Indication challenge to Traffic Classification”

Problem Statement Provided by Samsung R&D Bangalore Team

An M. Tech project report submitted
in partial fulfilment of the requirements
for the degree of

Master of Technology

by

Ujjwal Chaudhary
(2311CS30)

under the guidance of
Dr. Samrat Mondal

to the

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING INDIAN INSTITUTE OF TECHNOLOGY PATNA

ACKNOWLEDGEMENT

I am incredibly happy and proud to acknowledge various people for their insightful counsel and helpful direction throughout the process, as well as to convey my sentiments of gratitude. I truly appreciate the tight supervision I received from my supervisor, **Dr. Samrat Mondal**, as I worked on my thesis. Being able to work under his guidance has been an honour since he has supported me at every turn and enabled me to reach my objectives. Ever since I joined the Indian Institute of Technology, Patna research team led by **Dr. Samrat Mondal**, I have received constant encouragement from him. I would like to acknowledge my friends and my family for emotionally supporting me till the end of the project. Without their endless love and encouragement, I would never have been able to complete my thesis, staying miles away from them.

ABSTRACT

TLS has undergone several enhancements like the Encrypted Client Hello (ECH) and the Encrypted Server Name Indication (ESNI) which effectively secured the handshake information boosting privacy and making it hard for classic traffic classification techniques to be employed. In this project which tackled a problem statement from the Samsung Team and guided by an industrial team, novel methods aiming at real-time traffic classification relying on the unencrypted parts of the TLS handshake were put into practice. Scapy and tshark were used to develop scripts that transformed the datasets of TLSv1.1, TLSv1.2, and TLSv1.3 into ECH and ESNI versions. Above mentioned models, ALIGNED BYTES RANDOM FOREST (AB-RF), RECOMPOSED BYTES RANDOM FOREST (RB-RF), BI-DIRECTIONAL GATED RECURRENT UNIT WITH SELF ATTENTION (BGRUA), and MULTI HEAD ATTENTION ENCODER WITH CONVOLUTION LAYERS (MATEC) were tested on datasets ISCXVPN2016, VNAT and WNL. Also LightGBM had been used because of quicker training time and greater accuracy than Random Forest algorithm. The study findings reveal the availability of high and very effective classification accuracy along with low error rates in order to make the encrypted traffic classification efficient.

Contents

1.Objective

The advancements in the Transport Layer Security (TLS) protocol have greatly enhanced the confidentiality of communications on the internet through data encryption, however, there remain unencrypted parts of its handshake such as the Server Name Indication (SNI), which in the past, facilitated traffic classification frameworks to be able to find applications.. Such developments are essential for applications like Quality of Service (QoS) management, network optimization, and encrypted threat detection.The introduction of the Encrypted ClientHello (ECH) amendment and Encrypted Server Name Indication amendment to TLS addresses this privacy gap, making traffic classification much more challenging and necessitating the development of innovative solutions for real-time classification.

Following are the objective :

1- Implementing classification models using unencrypted parts of the TLS handshake(Client Hello and Server Hello) as features, that can be use in real-time classification.

2- Making the algorithms fast and accurate, which can perform better than random forest algorithm

3- Development of scripts for converting TLSv1.1, TLSv1.2 and TLSv1.3 Dataset into Encrypted Client Hello.

4- Scripts for converting TLS v1.1, TLS v1.2 and TLS v1.3 Dataset into Encrypted Server Name Indication

5-Evaluation of implemented models on different Datasets(ISCXVPN2016, VNAT and WNL)

2 Introduction

The Transport Layer Security (TLS) protocol plays a key role in internet security. It protects almost 90% of global web traffic as of 2021. TLS encrypts user data, but it leaves some metadata unencrypted.

This includes the Server Name Indication (SNI) in the ClientHello (CH) message. The SNI shows the domain name of the server in plain text. This creates a big privacy risk .

Bad actors take advantage of this weakness in several ways. Middleboxes group traffic by finding the type of data linked to domain names. This allows mobile operators to handle subscription-based services . Internet providers use it to study network operations . Some governments use it to censor content by blacklisting SNIs.

2.1 **Transport Layer Security**

The Transport Layer Security (TLS) protocol occupies a very prominent position in internet security, as 90% of all web traffic in the world in 2021 is said to pass through it . However, TLS is not totally effective in providing privacy; it encrypts user traffic while certain information such as the Server Name Indication (SNI) remains unencrypted in the ClientHello (CH) message.

The SNI, which indicates a plain-text domain name of the server, constitutes a major privacy flaw . This defect is exploited in different ways: middleboxes classify traffic by matching the data types associated with domain names, so mobile operators maintain their subscriptions-based services [4], internet providers analyze their networks' operations , and certain governments would use SNI blacklisting to impose censorship.

2.2 Encrypted Client Hello

The Encrypted ClientHello (ECH) protocol is the one that is meant to deal with the privacy concerns such as the unencrypted transmission of fields and extensions during the handshake for TLS 1.3. Unencrypted, by default, fields such as Server Name Indication (SNI) are sent in requests, and hence expose connection details to middleboxes as well as path attackers.

ECH is aimed at keeping these parameters in TLS private via encryption so that they do not end up in the hands of unauthorized persons, but in consideration of backward compatibility and correct functioning, it is necessary to keep some few fields, for example, the protocol version and cipher suites, extensions like Key Share, Pre-Shared Key and Supported Versions unencrypted.

This will take care of the client's and server's critical need to have information shared on their secret during handshake.

2.3 Encrypted Server Name Indication

Encrypted Server Name Indication (ESNI) is a function that was brought in to improve the security directly relating to someone's privacy with regard to the Transport Layer Security (TLS) protocol. This it particularly does by encrypting the field which contains the Server Name Indication. The SNI, which is part of the ClientHello message during the TLS handshake, specifies the server hostname to which the user wants to connect, telling the servers which host multiple domains on one IP address to which domain the call should be directed..

ESNI addresses this problem by encrypting this key in the SNI field with the cryptographic keys provided by the server so that sensitive information such as the accessed domain names is kept private and not visible to any middleboxes or network observers. Thus, this also prevents traffic analysis, censorship, and even unauthorized classification.

3.Models

3.1 BGRUA (Bi-Directional Gated Recurrent Unit with Self Attention)

The Bi-Directional Gated Recurrent Unit with Self-Attention (BGRUA) Model has been built to traffic classification in the right way, especially when dealing with encrypted environments. GRU has been the requirement for this model as it has a simpler structure and lower computational cost, more preferably than LSTM.

This model makes use of bidirectional GRU to the processing of sequences in forward and reverse directions. This helps the hidden state capture the information from the entire sequence.

Other important aspects of this model are induction of self-attention to focus on the most relevant features from the sequential outputs by tightening its scopes. To make an encrypted traffic flow labeling,

Server Name Indication is used, whereby the using SNI's bytes are replaced with zeros during preprocessing to further help in privacy. Out of the Open HTTPS dataset, made up of TLS 1.2 traffic, BGRUA proves its better performance over CNN-LSTM regarding classification accuracy and computational efficiency.

The BGRUA has a good trade-off between understanding capabilities, computational efficiency, and privacy protection, making it robust to encrypted traffic analyses.

3.2 MATEC(Multi-Head Attention Encoder with Convolutional Layers)

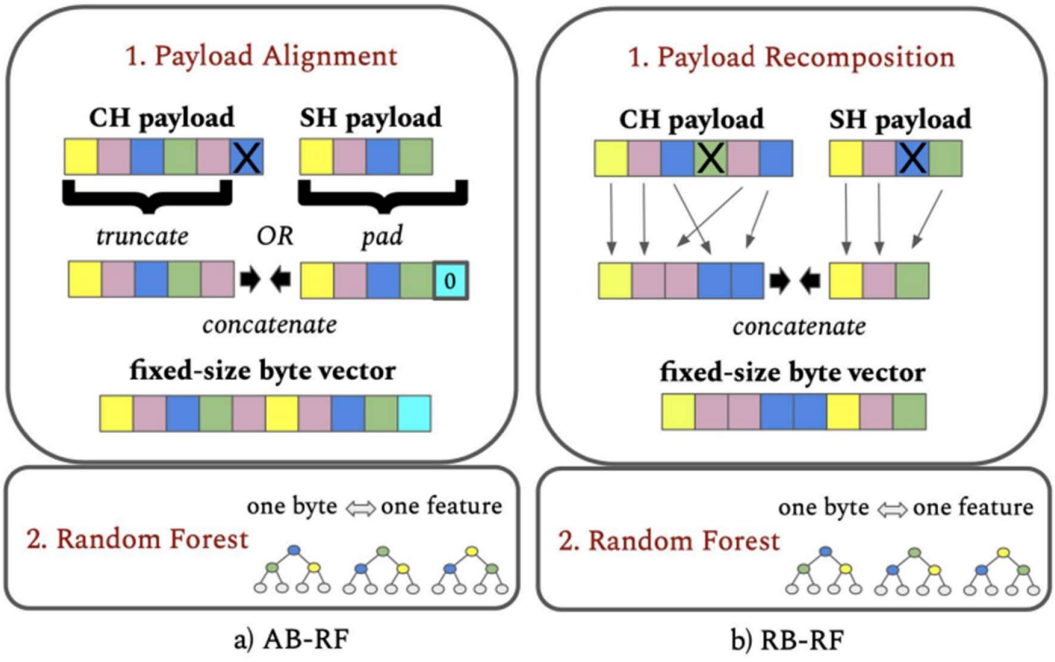
The Multi-Head Attention Encoder with Convolutional Layers (MATEC) is a lean model intended for the efficient processing of high-dimensional input data such as encrypted network traffic for classification purposes.

Its employs an embedding layer that extracts shallow-level features, while the multi-head attention encoder captures global dependencies of the data. Multiple heads, in turn, will operate on this input simultaneously for enhanced pattern recognition. To further refine the features,

MATEC uses the 1D Convolutional Neural Network (1D-CNN) as a feed-forward layer with multiple convolutional kernels for local capturing of patterns. These kernels, alongside the attention heads, operate independently and parallelly, which boosts the model's computational efficiency.

MATEC compared better to other byte-based classifiers on the Open HTTPS dataset (with plastered SNI), inferring greater f scores and prediction throughput out of it. Its lighter architecture, which harmoniously intertwines the local and global extraction of features, promises high accuracy and scalability, thus being suitable for real-time as well as resource-limited scenarios.

3.3 Aligned-Bytes-Based Random Forest



The Aligned-Bytes-Based Random Forest (AB-RF) algorithm is a two-phase and two-stage available effective solution for classifying traffic in Transport Layer Security (TLS) messages. It adopts the best features of payload alignment as well as the great prediction strength of Random Forests. The following is a stepwise explanation.

First step: Payload Alignment By Byte Extraction:

The algorithm extracts exactly B bytes from every payload of the particular message.

In case a message is shorter than B bytes, the algorithm pads the payload with a number of zeros.

If the message has more than B bytes, the algorithm truncates the payload.

-Concatenate Bytes:

The aligned bytes extracted from multiple messages-alike ClientHello (CH) and ServerHello (SH) will be concatenated in order to form a single vector.

This feature vector will then be a fixed one for uniformity of input into the machine learning model.

Step 2: **Classification by Random Forest (RF)**

-**Feature Inpu**t:

The aligned byte vector will be an input to a Random Forest (RF) classifier.

-**Training Phase**:

During training, the RF builds up a set of D decision trees.

K attributes are randomly selected from the input vector for each tree.

The best split is computed for these attributes at each node of the tree for classification.

-**Prediction Phase**:

Here, the already trained RF model can predict the traffic class by aggregating the outputs from all decision trees.

Every tree takes an individual decision, and the final outcome is usually decided by voting majority across all trees

3.4 RECOMPOSED-BYTES-BASED RF (RB-RF)

Record Version	Record Len	Message Len	Message Version	SID Len	Cipher Suites Len	Cipher Suites	Extensions Len
2 bytes	2 bytes	3 bytes	2 bytes	1 byte	2 bytes	70 bytes	2 bytes

Ext 1 Type	Ext 2 Type	...	Ext 20 Type	Padding (21) Len	Session Ticket (35) Len	PSK (41) Len	Cookie (44) Len	SNI (0) Len
2 bytes	2 bytes	34 bytes	2 bytes	2 bytes	2 bytes	2 bytes	2 bytes	2 bytes

Cached info (25) Len	Key Share (51) Len	ALPN (16) Len	Trusted CA keys (3) Data	Heartbeat (15) Data	PSK KE modes (45) Data
2 bytes	2 bytes	2 bytes	2 bytes	2 bytes	2 bytes

Compress Certificate (27) Data	Record size limit (28) Data	user mapping (6) Data	EC point formats (11) Data	Client Cert type (19) Data
4 bytes	4 bytes	4 bytes	4 bytes	4 bytes

Server Cert type (20) Data	Ticket Request (58) Data	Supported Versions (43) Data	Supported Groups (10) Data	SA (13) Data	ALPN (16) Data
4 bytes	4 bytes	12 bytes	26 bytes	26 bytes	4 bytes

(a)

Record Version	Record Len	Message Len	Message Version	SID Len	Cipher Suite	Extensions Len
2 bytes	2 bytes	3 bytes	2 bytes	1 byte	2 bytes	2 bytes

Ext 1 Type	...	Ext 10 Type	PSK (41) Len	Key Share (51) Len	Key Share (51) Data	Supported Versions (43) Data
2 bytes	16 bytes	2 bytes	2 bytes	2 bytes	2 bytes	2 bytes

(a) ClientHello and (b) ServerHello recomposed payload

The Recomposed-Bytes Random Forest (RB-RF) model enhances the Aligned-Bytes Random Forest (AB-RF) method as it replaces the payload alignment procedure with more sophisticated payload recomposition.

Payload parameters will be rearranged in the recomposition approach so that they each have positions and lengths assigned to them. With fixed positions, the parameters remain consistent in different TLS handshake implementations, which, by their very nature, vary in structure, making it possible for the algorithm to identify and, hence, match the same ones across different handshakes. As a result, classification accuracy increases.

The recomposition process has two crucial stages: First, payload decomposition, in which the handshake is split into individual payload parameters. Second, new composition, where these parameters are rearranged into a predetermined, fixed-size byte structure. This structured arrangement ensures an improvement in performance in classification and also makes the model more interpretable since the byte arrangement is predictable and consistent with the former purity of the Random Forest (RF) by nature. Thus, RB-RF stands as a strong and clear option for traffic classification.

3.5 LightGBM Classifier

The LightGBM classifier is a fast, efficient gradient boosting framework designed for classification tasks, offering superior performance on large, high-dimensional datasets. It uses a histogram-based algorithm for faster training and lower memory usage, supports categorical features natively, and excels at handling imbalanced datasets with parameters like scale_pos_weight . Compared to Random Forest, which builds multiple decision trees independently and averages their predictions, LightGBM builds trees sequentially, optimizing each based on previous results to minimize errors. While Random Forest is robust and less sensitive to hyperparameters, LightGBM often achieves higher accuracy due to its boosting approach but requires careful tuning to avoid overfitting. Additionally, LightGBM is significantly faster and more memory-efficient than Random Forest, especially on large-scale datasets.

4.Results

4.1 WNL Dataset

The dataset contains download traces of TLS-encrypted flows of four traffic types: buffered video, buffered audio, uplink live video streaming, and web. Overall, the dataset consists of 12 classes and 3547 flows

Value Counts

```
target
ww                1045
Netflix            433
YandexMusic        376
AppleMusic         292
SoundCloud         281
Kinopoisk          268
Spotify            255
YouTube_PC         249
PrimeVideo         191
Live_YouTube       108
Live_Facebook      106
Vimeo              106
Name: count, dtype: int64
```

4.2 Bi-Directional Gated Recurrent Unit Attention Model-(BGRUA)

Encrypted Server Name Indication-

```
92/92 ━━━━━━━━━━━━━━━━━━━━ 5s 50ms/step - accuracy: 1.0000 - loss: 9.5666e-04 - val_accuracy: 0.9576 - val_loss: 0.2963 - learnin
Epoch 39/40
92/92 ━━━━━━━━━━━━━━━━━━━━ 5s 50ms/step - accuracy: 1.0000 - loss: 0.0019 - val_accuracy: 0.9603 - val_loss: 0.2858 - learning_ra
Epoch 40/40
92/92 ━━━━━━━━━━━━━━━━━━━━ 4s 48ms/step - accuracy: 1.0000 - loss: 8.2782e-04 - val_accuracy: 0.9590 - val_loss: 0.2907 - learnin
23/23 ━━━━━━━━━━━━━━━━━━━━ 0s 16ms/step - accuracy: 0.9677 - loss: 0.2435
Test Loss: 0.2906722128391266, Test Accuracy: 0.9589603543281555
23/23 ━━━━━━━━━━━━━━━━━━━━ 0s 18ms/step - accuracy: 0.9677 - loss: 0.2435
23/23 ━━━━━━━━━━━━━━━━━━━━ 2s 42ms/step
```

```
Class-wise Error Rate Table:
            Class  Error Rate
0      AppleMusic    0.015873
1       Kinopoisk    0.046154
2   Live_Facebook    0.000000
3    Live_Youtube    0.000000
4         Netflix    0.011364
5      PrimeVideo    0.018182
6      SoundCloud    0.109375
7         Spotify    0.043478
8           Vimeo    0.142857
9     YandexMusic    0.075949
10     YouTube_PC    0.020000
11             ww    0.067164
```

BGRUA classifier uses a combination of two bidirectional GRU layers (the hidden state size is 256), a selfattention layer, and a fully connected layer with a softmax activation function that generates a probability over class labels. The packet lengths are aligned to 900 bytes and reshaped to six vectors with 150 bytes. Each vector is normalized to [0,1] by dividing by 255.

Accuracy : 0.9581

BGRUA is taking less training time compare to MATEC and accuracy is better than MATEC,

Accuracy is better in case of Encrypted Server Name Indication compare to Encrypted Server Hello (Accuracy 0.7647)

Encrypted CIlent Hello Dataset-(BGRUA)

```
92/92 ━━━━━━━━━━━━━━━━━━━━ 4s 43ms/step - accuracy: 0.9357 - loss: 0.1788 - val_accuracy: 0.7661 - val_loss: 0.9724 - learning_rate:
Epoch 36/40
92/92 ━━━━━━━━━━━━━━━━━━━━ 4s 44ms/step - accuracy: 0.9401 - loss: 0.1558 - val_accuracy: 0.7592 - val_loss: 1.0058 - learning_rate:
Epoch 37/40
92/92 ━━━━━━━━━━━━━━━━━━━━ 4s 44ms/step - accuracy: 0.9485 - loss: 0.1448 - val_accuracy: 0.7620 - val_loss: 1.0696 - learning_rate:
Epoch 38/40
92/92 ━━━━━━━━━━━━━━━━━━━━ 4s 43ms/step - accuracy: 0.9471 - loss: 0.1425 - val_accuracy: 0.7524 - val_loss: 1.0539 - learning_rate:
Epoch 39/40
92/92 ━━━━━━━━━━━━━━━━━━━━ 4s 45ms/step - accuracy: 0.9463 - loss: 0.1448 - val_accuracy: 0.7606 - val_loss: 1.1544 - learning_rate:
Epoch 40/40
92/92 ━━━━━━━━━━━━━━━━━━━━ 4s 44ms/step - accuracy: 0.9595 - loss: 0.1259 - val_accuracy: 0.7647 - val_loss: 1.1296 - learning_rate:
23/23 ━━━━━━━━━━━━━━━━━━━━ 0s 13ms/step - accuracy: 0.7656 - loss: 1.1121
Test Loss: 1.1296416521072388, Test Accuracy: 0.7647058963775635
23/23 ━━━━━━━━━━━━━━━━━━━━ 0s 13ms/step - accuracy: 0.7656 - loss: 1.1121
23/23 ━━━━━━━━━━━━━━━━━━━━ 1s 36ms/step
```

```
Class-wise Error Rate Table:
           Class  Error Rate
0      AppleMusic    0.079365
1       Kinopoisk    0.546875
2   Live_Facebook    0.000000
3    Live_Youtube    0.000000
4         Netflix    0.068182
5      PrimeVideo    0.181818
6      SoundCloud    0.046875
7         Spotify    0.571429
8           Vimeo    0.619048
9     YandexMusic    0.405063
10     YouTube_PC    0.020000
11             ww    0.201493
```

BGRUA classifier uses a combination of two bidirectional GRU layers (the hidden state size is 256), a selfattention layer, and a fully connected layer with a softmax activation function that generates a probability over class labels. The packet lengths are aligned to 900 bytes and reshaped to six vectors with 150 bytes. Each vector is normalized to [0,1] by dividing by 255.

Accuracy-0.7647

Incase of Encrypted Client Hello prediction are not that consistent (due to more extensions are encrypted compare to ESNI)

4.3-MATEC MODEL-

Encrypted CIlent Hello Dataset-

```
Epoch 37/40
79/79 ━━━━━━━━━━━━━━━━━━━━ 43s 423ms/step - accuracy: 0.8445 - loss: 0.3682 - val_accuracy: 0.7893 - val_loss: 0.5577
Epoch 38/40
79/79 ━━━━━━━━━━━━━━━━━━━━ 39s 400ms/step - accuracy: 0.8539 - loss: 0.3655 - val_accuracy: 0.7571 - val_loss: 0.7067
Epoch 39/40
79/79 ━━━━━━━━━━━━━━━━━━━━ 41s 397ms/step - accuracy: 0.8261 - loss: 0.4542 - val_accuracy: 0.7750 - val_loss: 0.7650
Epoch 40/40
79/79 ━━━━━━━━━━━━━━━━━━━━ 42s 407ms/step - accuracy: 0.8522 - loss: 0.3546 - val_accuracy: 0.7607 - val_loss: 0.7353
22/22 ━━━━━━━━━━━━━━━━━━━━ 1s 38ms/step - accuracy: 0.7485 - loss: 0.7485
Test Loss: 0.6776
Test Accuracy: 0.7650
22/22 ━━━━━━━━━━━━━━━━━━━━ 1s 54ms/step - accuracy: 0.7485 - loss: 0.7485
```

```
Class-wise Error Rate Table:
             Class  Error Rate (%)
0     AppleMusic22       78.723404
1      Kinopoisk22       46.969697
2   LiveFacebook22        0.000000
3    LiveYouTube22       52.941176
4        Netflix22        0.000000
5     PrimeVideo22        2.941176
6     SoundCloud22        1.754386
7        Spotify22      100.000000
8          Vimeo22      100.000000
9            Web22       28.643216
10   YandexMusic22       13.333333
11       YouTube22        9.615385
```

The MATEC classifier passes each input packet normalized to [0,1] to an embedding layer (embedding size is 432), which extracts low-level features. Then MATEC extracts high-level features with an encoder. The encoder is composed of T (T = 2) multi-head attention layers (the head number is 3) and 1D-CNN feed-forward layers (the kernel number is 432, the kernel size is 1). A dense layer with a softmax activation function generates the probabilities of various class labels.

Accuracy-0.7485

MATEC is taking more training time compare to BGRUA due to use of Multi Head Attention.

Encrypted CIlent Server Name Indication Dataset-

```
Epoch 40/40
80/80 ━━━━━━━━━━━━━━━━━━━━ 40s 249ms/step - accuracy: 0.9679 - loss: 0.1279 - val_accuracy: 0.9718 - val_loss: 0.1275
23/23 ━━━━━━━━━━━━━━━━━━━━ 1s 32ms/step - accuracy: 0.9445 - loss: 0.3126
Test Loss: 0.3161
Test Accuracy: 0.9451
Total Evaluation Time: 0.82 seconds
23/23 ━━━━━━━━━━━━━━━━━━━━ 1s 45ms/step

Class-wise Error Rate Table:
          Class  Error Rate (%)
0     AppleMusic        5.882353
1      Kinopoisk        2.469136
2   Live_Facebook        0.000000
3    Live_Youtube        0.000000
4         Netflix        2.127660
5      PrimeVideo       14.285714
6      SoundCloud        7.547170
7         Spotify        7.692308
8           Vimeo       16.666667
9     YandexMusic       11.111111
10     YouTube_PC        0.000000
11             ww        3.649635
```

The MATEC classifier passes each input packet normalized to [0,1] to an embedding layer (embedding size is 432), which extracts low-level features. Then MATEC extracts high-level features with an encoder. The encoder is composed of T (T = 2) multi-head attention layers (the head number is 3) and 1D-CNN feed-forward layers (the kernel number is 432, the kernel size is 1). A dense layer with a softmax activation function generates the probabilities of various class labels.

Accuracy: 0.9451

In case of Encrypted Server Name Indication accurarcy is similar to BGRUA but training time is more in case of MATEC.

4.4.ALIGNED-BYTES-BASED RANDOM FOREST

-ENCRYPTED SERVER NAME INDICATION

```
Training Time: 4.5907 seconds

Class-wise Performance Table:
            Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0      AppleMusic    100.000000        0.000000     100.000000  100.000000
1       Kinopoisk    100.000000        0.000000      98.148148  100.000000
2   Live_Facebook    100.000000        0.000000     100.000000  100.000000
3    Live_Youtube    100.000000        0.000000     100.000000  100.000000
4         Netflix    100.000000        0.000000     100.000000  100.000000
5      PrimeVideo    100.000000        0.000000     100.000000  100.000000
6      SoundCloud    100.000000        0.000000     100.000000  100.000000
7         Spotify    100.000000        0.000000     100.000000  100.000000
8           Vimeo    100.000000        0.000000      95.000000  100.000000
9     YandexMusic    100.000000        0.000000      97.402597  100.000000
10     YouTube_PC    100.000000        0.000000     100.000000  100.000000
11             ww     98.039216        1.960784     100.000000   98.039216

Overall Accuracy: 0.9945
F1 Score: 0.9945
```

CONFUSION MATRIX

```
Confusion Matrix:
[[ 58   0   0   0   0   0   0   0   0   0   0   0]
 [  0  53   0   0   0   0   0   0   0   0   0   0]
 [  0   0  21   0   0   0   0   0   0   0   0   0]
 [  0   0   0  22   0   0   0   0   0   0   0   0]
 [  0   0   0   0  85   0   0   0   0   0   0   0]
 [  0   0   0   0   0  38   0   0   0   0   0   0]
 [  0   0   0   0   0   0  56   0   0   0   0   0]
 [  0   0   0   0   0   0   0  50   0   0   0   0]
 [  0   0   0   0   0   0   0   0  19   0   0   0]
 [  0   0   0   0   0   0   0   0   0  75   0   0]
 [  0   0   0   0   0   0   0   0   0   0  50   0]
 [  0   1   0   0   0   0   0   0   1   2   0 200]]
Prediction Time: 0.0220 seconds
```

Hyperparameters of Random Forest used

1- **n_estimators=150,**
2- **min_samples_split=10,**
3- **max_features=70,**
4- **random_state=42**

From Client Hello and Server Hello we are extracting 185 bytes each.

Training Time is 4.5seconds , which is significantly less than Training Time taken by base classifiers like BGRUA and MATEC

For 11 classes this model giving 100 percent accuracy and for other classes accuracy near to 100 percent.

Prediction Time is .022 seconds ,due to which it can be use real-time predictions.

Average Accuracy on Subset of Recomposed ESNI Preprocessed Dataset,total number of subset is 3 each containing 70 percent of the total Dataset(Resampling) is **0.9922 .**

-ENCRYPTED CLIENT HELLO

```
Training Time: 6.4665 seconds

Class-wise Performance Table:
            Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0      AppleMusic    100.000000        0.000000      98.305085  100.000000
1       Kinopoisk     64.150943       35.849057      97.142857   64.150943
2   Live_Facebook    100.000000        0.000000     100.000000  100.000000
3    Live_Youtube    100.000000        0.000000     100.000000  100.000000
4         Netflix    100.000000        0.000000      78.703704  100.000000
5      PrimeVideo     84.210526       15.789474      94.117647   84.210526
6      SoundCloud    100.000000        0.000000      98.245614  100.000000
7         Spotify     98.000000        2.000000      87.500000   98.000000
8           Vimeo     73.684211       26.315789      87.500000   73.684211
9     YandexMusic    100.000000        0.000000      69.444444  100.000000
10     YouTube_PC    100.000000        0.000000     100.000000  100.000000
11             ww     77.941176       22.058824      96.363636   77.941176

Overall Accuracy: 0.8960
F1 Score: 0.8951
```

CONFUSION MATRIX

```
Confusion Matrix:
[[ 58   0   0   0   0   0   0   0   0   0   0   0]
 [  0  34   0   0   0   0   0   0   0  19   0   0]
 [  0   0  21   0   0   0   0   0   0   0   0   0]
 [  0   0   0  22   0   0   0   0   0   0   0   0]
 [  0   0   0   0  85   0   0   0   0   0   0   0]
 [  0   0   0   0   0  32   0   0   0   0   0   6]
 [  0   0   0   0   0   0  56   0   0   0   0   0]
 [  0   1   0   0   0   0   0  49   0   0   0   0]
 [  0   0   0   0   0   0   0   5  14   0   0   0]
 [  0   0   0   0   0   0   0   0   0  75   0   0]
 [  0   0   0   0   0   0   0   0   0   0  50   0]
 [  1   0   0   0  23   2   1   2   2  14   0 159]]
Prediction Time: 0.0252 seconds
```

Hyperparameters of Random Forest used

1- **n_estimators=150,**
2- **min_samples_split=10,**
3- **max_features=70,**
4- **random_state=42**

Training Time is 4.6 seconds , which is significantly less than Training Time taken by base classifiers like BGRUA and MATEC

For 7 classes this model giving 100 percent accuracy and for www,Vimeo,Kinopoisk accuracy getting less than 80 percent.

Prediction Time is .025 seconds ,due to which it can be use real-time predictions.

Average Accuracy on Subset of Recomposed ESNI Preprocessed Dataset,total number of subset is 3 each containing 70 percent of the total Dataset(Resampling) is **0.89.**

4.5.RECOMPOSED-BYTES-BASED RANDOM FOREST

-ENCRYPTED SERVER NAME INDICATION

```
Training Time: 1.3431 seconds

Class-wise Performance Table:
           Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0      AppleMusic    100.000000        0.000000     100.000000  100.000000
1       Kinopoisk     98.148148        1.851852     100.000000   98.148148
2   Live_Facebook    100.000000        0.000000     100.000000  100.000000
3    Live_YouTube    100.000000        0.000000     100.000000  100.000000
4         Netflix    100.000000        0.000000     100.000000  100.000000
5      PrimeVideo     94.736842        5.263158     100.000000   94.736842
6      SoundCloud    100.000000        0.000000     100.000000  100.000000
7         Spotify    100.000000        0.000000      96.226415  100.000000
8           Vimeo    100.000000        0.000000     100.000000  100.000000
9     YandexMusic    100.000000        0.000000      98.684211  100.000000
10     YouTube_PC    100.000000        0.000000     100.000000  100.000000
11             ww     99.043062        0.956938      99.043062   99.043062

Overall Accuracy: 0.9933
F1 Score: 0.9932
```

CONFUSION MATRIX

```
Confusion Matrix:
[[ 58   0   0   0   0   0   0   0   0   0   0   0]
 [  0  53   0   0   0   0   0   1   0   0   0   0]
 [  0   0  21   0   0   0   0   0   0   0   0   0]
 [  0   0   0  22   0   0   0   0   0   0   0   0]
 [  0   0   0   0  87   0   0   0   0   0   0   0]
 [  0   0   0   0   0  36   0   0   0   0   0   2]
 [  0   0   0   0   0   0  56   0   0   0   0   0]
 [  0   0   0   0   0   0   0  51   0   0   0   0]
 [  0   0   0   0   0   0   0   0  21   0   0   0]
 [  0   0   0   0   0   0   0   0   0  75   0   0]
 [  0   0   0   0   0   0   0   0   0   0  50   0]
 [  0   0   0   0   0   0   0   1   0   1   0 207]]
Prediction Time: 0.0593 seconds
```

Hyperparameters of Random Forest used

1. **n_estimators=150**
2. **min_samples_split=10,**
3. **max_features=70,**
4. **random_state=42**

Training Time is 1.34 seconds , which is significantly less than Training Time taken by Aligned Byte Random Forest on same hyperparameters.

For 9 classes this model giving 100 percent accuracy and for other classes accuracy near to 100 percent.

Prediction Time is .0593 seconds ,due to which it can be use real-time predictions.

Average Accuracy on Subset of Recomposed ESNI Preprocessed Dataset,total number of subset is 3 each containing 70 percent of the total Dataset(Resampling) is **0.9929 .**

-ENCRYPTED CLIENT HELLO

```
Training Time: 0.6305 seconds

Class-wise Performance Table:
           Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0      AppleMusic    100.000000        0.000000     100.000000  100.000000
1       Kinopoisk     62.962963       37.037037      85.000000   62.962963
2   Live_Facebook    100.000000        0.000000     100.000000  100.000000
3    Live_YouTube    100.000000        0.000000     100.000000  100.000000
4         Netflix    100.000000        0.000000      85.294118  100.000000
5      PrimeVideo     78.947368       21.052632      85.714286   78.947368
6      SoundCloud    100.000000        0.000000      96.551724  100.000000
7         Spotify    100.000000        0.000000      87.931034  100.000000
8           Vimeo     76.190476       23.809524     100.000000   76.190476
9     YandexMusic     90.666667        9.333333      70.833333   90.666667
10     YouTube_PC    100.000000        0.000000      98.039216  100.000000
11             ww     83.732057       16.267943      94.594595   83.732057

Overall Accuracy: 0.9003
F1 Score: 0.8991
```

CONFUSION MATRIX

```
Confusion Matrix:
[[ 58   0   0   0   0   0   0   0   0   0   0   0]
 [  0  34   0   0   0   0   0   1   0  19   0   0]
 [  0   0  21   0   0   0   0   0   0   0   0   0]
 [  0   0   0  22   0   0   0   0   0   0   0   0]
 [  0   0   0   0  87   0   0   0   0   0   0   0]
 [  0   0   0   0   0  30   0   0   0   0   0   8]
 [  0   0   0   0   0   0  56   0   0   0   0   0]
 [  0   0   0   0   0   0   0  51   0   0   0   0]
 [  0   0   0   0   0   0   0   5  16   0   0   0]
 [  0   5   0   0   0   0   0   0   0  68   0   2]
 [  0   0   0   0   0   0   0   0   0   0  50   0]
 [  0   1   0   0  15   5   2   1   0   9   1 175]]
Prediction Time: 0.0371 seconds
```

Hyperparameters of Random Forest used

1- **n_estimators=150,**
2- **min_samples_split=10,**
3- **max_features=70,**
4- **random_state=42**

Training Time is 0.63 seconds , which is significantly less than Training Time taken by Aligned Byte Random Forest on same hyperparameters.

For 7 classes this model giving 100 percent accuracy and for Vimeo ,Kinopoisk and PrimeVideo is less than 90 percent accuracy.

Prediction Time is .0371 seconds ,due to which it can be use real-time predictions.

Average Accuracy on Subset of Recomposed ECH Preprocessed Dataset,total number of subset is 3 each containing 70 percent of the total Dataset(Resampling) is **0.8985 .**

4.6 VNAT and ISCXVPN2016 Dataset

Both Datasets provided by Samsung Team for evaluation of implemented model and proposed model.

These Dataset Contains VPN and Non-VPN pcap files of different Application (containing TLS v1.2 and TLS v1.3) .

Both Datasets contain many different Server Name Indication,SNI which contain atleast more than 20 value counts we took that class/label for evaluation.

All classes which have value counts more than 20 then converted into Encrypted Client Hello and Encrypted Server Name Indication

VALUE COUNTS

```
target
Vimeo              200
Gmail              180
Youtube            169
Google Services     73
Netflix             67
FileTransfer        52
ssl.gstatic         45
Hangout             22
Name: count, dtype: int64
```

ALIGNED BYTES-RANDOM FOREST

ENCRYPTED SERVER NAME INDICATION

```
Training Time: 1.4272 seconds

Class-wise Performance Table:
            Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0     FileTransfer    100.000000        0.000000          100.0  100.000000
1            Gmail    100.000000        0.000000          100.0  100.000000
2  Google Services    100.000000        0.000000          100.0  100.000000
3          Hangout    100.000000        0.000000           80.0  100.000000
4          Netflix    100.000000        0.000000          100.0  100.000000
5            Vimeo    100.000000        0.000000          100.0  100.000000
6          Youtube    100.000000        0.000000          100.0  100.000000
7      ssl.gstatic     88.888889       11.111111          100.0   88.888889

Overall Accuracy: 0.9938
F1 Score: 0.9940
```

CONFUSION MATRIX

```
Confusion Matrix:
[[10  0  0  0  0  0  0  0]
 [ 0 36  0  0  0  0  0  0]
 [ 0  0 15  0  0  0  0  0]
 [ 0  0  0  4  0  0  0  0]
 [ 0  0  0  0 14  0  0  0]
 [ 0  0  0  0  0 40  0  0]
 [ 0  0  0  0  0  0 34  0]
 [ 0  0  0  1  0  0  0  8]]
Prediction Time: 0.0583 seconds
```

Hyperparameters of Random Forest used

1- **n_estimators=150,**
2- **min_samples_split=10,**
3- **max_features=70,**
4- **random_state=42**

From Client Hello and Server Hello we are extracting 185 bytes each.

Training Time is 1.4 seconds, which is significantly less than Training Time taken by base classifiers like BGRUA and MATEC

For 7 classes this model giving 100 percent accuracy and for other class accuracy near to 90 percent.

Prediction Time is .0583 seconds ,due to which it can be use real-time predictions.

Average Accuracy on Subset of Recomposed ESNI Preprocessed Dataset,total number of subset is 3 each containing 70 percent of the total Dataset(Resampling) is **0.9766 .**

-ENCRYPTED CLIENT HELLO

```
Training Time: 0.6245 seconds

Class-wise Performance Table:
            Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0     FileTransfer    100.000000        0.000000     100.000000  100.000000
1            Gmail    100.000000        0.000000     100.000000  100.000000
2  Google Services     80.000000       20.000000      60.000000   80.000000
3          Hangout      0.000000      100.000000       0.000000    0.000000
4          Netflix    100.000000        0.000000     100.000000  100.000000
5            Vimeo    100.000000        0.000000     100.000000  100.000000
6          Youtube    100.000000        0.000000      97.142857  100.000000
7      ssl.gstatic     55.555556       44.444444      71.428571   55.555556

Overall Accuracy: 0.9321
F1 Score: 0.9223

Confusion Matrix:
[[10  0  0  0  0  0  0  0]
 [ 0 36  0  0  0  0  0  0]
 [ 0  0 12  0  0  0  1  2]
 [ 0  0  4  0  0  0  0  0]
 [ 0  0  0  0 14  0  0  0]
 [ 0  0  0  0  0 40  0  0]
 [ 0  0  0  0  0  0 34  0]
 [ 0  0  4  0  0  0  0  5]]
Prediction Time: 0.0296 seconds
```

Hyperparameters of Random Forest used

1- **n_estimators=150,**
2- **min_samples_split=10,**
3- **max_features=70,**
4- **random_state=42**

From Client Hello and Server Hello we are extracting 185 bytes each.

Training Time is 0.62 seconds, which is significantly less than Training Time taken by base classifiers like BGRUA and MATEC

For 5 classes this model giving 100 percent accuracy but for Hangout error rate is high ,it is due to very less value count for Hangout.

Prediction Time is .0296 seconds ,due to which it can be use real-time predictions.

Average Accuracy on Subset of Recomposed ECH Preprocessed Dataset,total number of subset is 3 each containing 70 percent of the total Dataset(Resampling) is **0.89 .**

Recomposed-Bytes-Random Forest

-ENCRYPTED SERVER NAME INDICATION

```
Training Time: 0.2038 seconds

Class-wise Performance Table:
            Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0     FileTransfer    100.000000        0.000000          100.0  100.000000
1            Gmail    100.000000        0.000000          100.0  100.000000
2  Google Services    100.000000        0.000000          100.0  100.000000
3          Hangout    100.000000        0.000000           80.0  100.000000
4          Netflix    100.000000        0.000000          100.0  100.000000
5            Vimeo    100.000000        0.000000          100.0  100.000000
6          Youtube    100.000000        0.000000          100.0  100.000000
7      ssl.gstatic     88.888889       11.111111          100.0   88.888889

Overall Accuracy: 0.9938
F1 Score: 0.9940

Confusion Matrix:
[[11  0  0  0  0  0  0  0]
 [ 0 36  0  0  0  0  0  0]
 [ 0  0 14  0  0  0  0  0]
 [ 0  0  0  4  0  0  0  0]
 [ 0  0  0  0 14  0  0  0]
 [ 0  0  0  0  0 40  0  0]
 [ 0  0  0  0  0  0 34  0]
 [ 0  0  0  1  0  0  0  8]]
Prediction Time: 0.0164 seconds
```

Hyperparameters of Random Forest used

1- **n_estimators=150,**
2- **min_samples_split=10,**
3- **max_features=70,**
4- **random_state=42**

Training Time is 0.2 seconds, which is significantly less than Training Time taken by base classifiers like BGRUA , MATEC and Aligned-Bytes (1.4 sec)

For 7 classes this model giving 100 percent,one other class ssl.gstatic accuracy is about 90 percent.

Prediction Time is 0.0164 seconds ,due to which it can be use real-time predictions.

Average Accuracy on Subset of Recomposed ESNI Preprocessed Dataset,total number of subset is 3 each containing 70 percent of the total Dataset(Resampling) is **0.9971 .**

-Encrypted Client Hello

```
Training Time: 0.1810 seconds

Class-wise Performance Table:
             Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0     FileTransfer    100.000000        0.000000     100.000000  100.000000
1            Gmail    100.000000        0.000000     100.000000  100.000000
2  Google Services     42.857143       57.142857      66.666667   42.857143
3          Hangout    100.000000        0.000000      57.142857  100.000000
4          Netflix    100.000000        0.000000     100.000000  100.000000
5            Vimeo    100.000000        0.000000     100.000000  100.000000
6          Youtube    100.000000        0.000000     100.000000  100.000000
7      ssl.gstatic     55.555556       44.444444      45.454545   55.555556

Overall Accuracy: 0.9259
F1 Score: 0.9242

Confusion Matrix:
[[11  0  0  0  0  0  0  0]
 [ 0 36  0  0  0  0  0  0]
 [ 0  0  6  2  0  0  0  6]
 [ 0  0  0  4  0  0  0  0]
 [ 0  0  0  0 14  0  0  0]
 [ 0  0  0  0  0 40  0  0]
 [ 0  0  0  0  0  0 34  0]
 [ 0  0  3  1  0  0  0  5]]
Prediction Time: 0.0145 seconds
```

Hyperparameters of Random Forest used

1- **n_estimators=150,**
2- **min_samples_split=10,**
3- **max_features=70,**
4- **random_state=42**

Training Time is 0.18 seconds, which is significantly less than Training Time taken by base classifiers like BGRUA and MATEC ,training time similar to Aligned Bytes RF

For 6 classes this model giving 100 percent,prediction for ssl.gstatic and Google Services is below 60 percent

Prediction Time is 0.03 seconds ,due to which it can be use real-time predictions.

Average Accuracy on Subset of Recomposed ESNI Preprocessed Dataset,total number of subset is 3 each containing 70 percent of the total Dataset(Resampling) is **0.947 .**

4.7 Proposed Models(Using LightGBM classifier)

WNL Dataset

Aligned-Bytes-LightGBM Classifier

Encrypted Server Name Indication

```
Training Time: 0.4570 seconds

Class-wise Performance Table:
          Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0      AppleMusic    100.000000        0.000000     100.000000  100.000000
1       Kinopoisk    100.000000        0.000000      98.148148  100.000000
2   Live_Facebook    100.000000        0.000000     100.000000  100.000000
3    Live_Youtube    100.000000        0.000000     100.000000  100.000000
4         Netflix    100.000000        0.000000     100.000000  100.000000
5      PrimeVideo    100.000000        0.000000     100.000000  100.000000
6      SoundCloud    100.000000        0.000000     100.000000  100.000000
7         Spotify    100.000000        0.000000     100.000000  100.000000
8           Vimeo    100.000000        0.000000     100.000000  100.000000
9     YandexMusic    100.000000        0.000000      97.402597  100.000000
10     YouTube_PC    100.000000        0.000000     100.000000  100.000000
11             ww     98.529412        1.470588     100.000000   98.529412

Overall Accuracy: 0.9959
F1 Score: 0.9959
```

```
Confusion Matrix:
[[ 58   0   0   0   0   0   0   0   0   0   0   0]
 [  0  53   0   0   0   0   0   0   0   0   0   0]
 [  0   0  21   0   0   0   0   0   0   0   0   0]
 [  0   0   0  22   0   0   0   0   0   0   0   0]
 [  0   0   0   0  85   0   0   0   0   0   0   0]
 [  0   0   0   0   0  38   0   0   0   0   0   0]
 [  0   0   0   0   0   0  56   0   0   0   0   0]
 [  0   0   0   0   0   0   0  50   0   0   0   0]
 [  0   0   0   0   0   0   0   0  19   0   0   0]
 [  0   0   0   0   0   0   0   0   0  75   0   0]
 [  0   0   0   0   0   0   0   0   0   0  50   0]
 [  0   1   0   0   0   0   0   0   0   2   0 201]]
Prediction Time: 0.0510 seconds
```

Following are the hyperparameters for LightGBM which are used:

Boosting_Type: 'Gradient Boosting Decision Tree',

learning_rate:0.1,

num_leaves: 4

Metric: ‘multi_logloss’,

Max_dept: 3

Training Time is 0.45 seconds , which is significantly less than Training Time taken by Aligned Byte Random Forest (4.5 seconds).

For 11 classes this model giving 100 percent accuracy and for other class accuracy near to 100 percent.

Prediction Time is .05 seconds ,due to which it can be use real-time predictions.

Accuracy of Aligned Bytes LightGBM is slightly less than Aligned Bytes Random Forest.

Overall Aligned Bytes LightGBM is performing better than Aligned Bytes LightGBM incase of Encrypted Server Hello

2-ENCRYPTED CLIENT HELLO

```
Training Time: 0.5445 seconds

Class-wise Performance Table:
      Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0      AppleMusic    100.000000        0.000000      98.305085  100.000000
1       Kinopoisk     64.150943       35.849057      87.179487   64.150943
2   Live_Facebook    100.000000        0.000000     100.000000  100.000000
3    Live_Youtube    100.000000        0.000000     100.000000  100.000000
4         Netflix     98.823529        1.176471      80.000000   98.823529
5      PrimeVideo     89.473684       10.526316      91.891892   89.473684
6      SoundCloud    100.000000        0.000000      98.245614  100.000000
7         Spotify     98.000000        2.000000      87.500000   98.000000
8           Vimeo     63.157895       36.842105     100.000000   63.157895
9     YandexMusic     94.666667        5.333333      68.269231   94.666667
10     YouTube_PC    100.000000        0.000000     100.000000  100.000000
11             ww     79.411765       20.588235      95.857988   79.411765

Overall Accuracy: 0.8933
F1 Score: 0.8925
```

```
Confusion Matrix:
[[ 58   0   0   0   0   0   0   0   0   0   0   0]
 [  0  34   0   0   0   0   0   0   0  19   0   0]
 [  0   0  21   0   0   0   0   0   0   0   0   0]
 [  0   0   0  22   0   0   0   0   0   0   0   0]
 [  0   0   0   0  84   0   0   0   0   0   0   1]
 [  0   0   0   0   0  34   0   0   0   0   0   4]
 [  0   0   0   0   0   0  56   0   0   0   0   0]
 [  0   0   0   0   0   0   0  49   0   1   0   0]
 [  0   0   0   0   0   0   0   5  12   0   0   2]
 [  0   4   0   0   0   0   0   0   0  71   0   0]
 [  0   0   0   0   0   0   0   0   0   0  50   0]
 [  1   1   0   0  21   3   1   2   0  13   0 162]]
Prediction Time: 0.0480 seconds
```

Training Time is 0.54 seconds , which is more compare to Training Time taken by Aligned Byte Random Forest (4.5 seconds).

For 6 classes this model giving 100 percent accuracy but for

Prediction Time is .05 seconds ,due to which it can be use real-time predictions.

Accuracy of Aligned Bytes LightGBM is slightly less than Aligned Bytes Random Forest

Recomposed-Bytes- LightGBM Classifier

-ENCRYPTED SERVER NAME INDICATION

```
Training Time: 0.1662 seconds

Class-wise Performance Table:
          Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0      AppleMusic    100.000000        0.000000     100.000000  100.000000
1       Kinopoisk     98.148148        1.851852     100.000000   98.148148
2   Live_Facebook    100.000000        0.000000     100.000000  100.000000
3    Live_YouTube    100.000000        0.000000     100.000000  100.000000
4         Netflix    100.000000        0.000000     100.000000  100.000000
5      PrimeVideo     97.368421        2.631579     100.000000   97.368421
6      SoundCloud    100.000000        0.000000     100.000000  100.000000
7         Spotify    100.000000        0.000000      96.226415  100.000000
8           Vimeo    100.000000        0.000000     100.000000  100.000000
9     YandexMusic    100.000000        0.000000     100.000000  100.000000
10     YouTube_PC    100.000000        0.000000     100.000000  100.000000
11             ww     99.521531        0.478469      99.521531   99.521531

Overall Accuracy: 0.9960
F1 Score: 0.9960
```

```
Confusion Matrix:
[[ 58   0   0   0   0   0   0   0   0   0   0   0]
 [  0  53   0   0   0   0   0   1   0   0   0   0]
 [  0   0  21   0   0   0   0   0   0   0   0   0]
 [  0   0   0  22   0   0   0   0   0   0   0   0]
 [  0   0   0   0  87   0   0   0   0   0   0   0]
 [  0   0   0   0   0  37   0   0   0   0   0   1]
 [  0   0   0   0   0   0  56   0   0   0   0   0]
 [  0   0   0   0   0   0   0  51   0   0   0   0]
 [  0   0   0   0   0   0   0   0  21   0   0   0]
 [  0   0   0   0   0   0   0   0   0  75   0   0]
 [  0   0   0   0   0   0   0   0   0   0  50   0]
 [  0   0   0   0   0   0   0   1   0   0   0 208]]
Prediction Time: 0.0450 seconds
```

Following are the hyperparameters for LightGBM which are used:

Boosting_Type: 'Gradient Boosting Decision Tree',

learning_rate:0.1,

num_leaves: 4

Metric: 'multi_logloss',

Max_dept: 3

Training Time is 0.16 seconds , which is significantly less than Training Time taken by Aligned Byte Random Forest (1.34 seconds).

For 7 classes this model giving 100 percent accuracy and for other classes accuracy near to 100 percent.

Prediction Time is .045 seconds ,due to which it can be use real-time predictions.

Accuracy of Aligned Bytes LightGBM is slightly better than Aligned Bytes Random Forest.

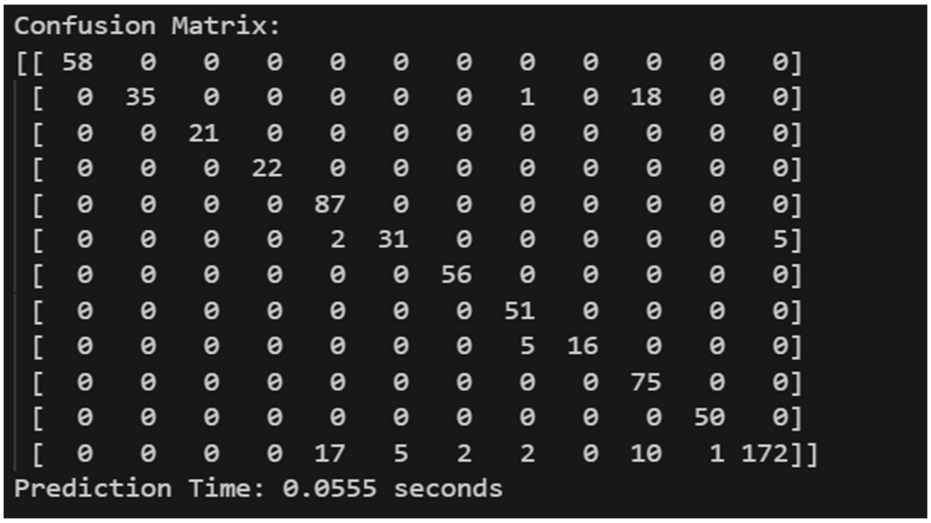
Overall Aligned Bytes LightGBM is performing better than Aligned Bytes LightGBM incase of Encrypted Server Hello

-ENCRYPTED CLIENT HELLO

```
Training Time: 0.1313 seconds

Class-wise Performance Table:
          Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0      AppleMusic    100.000000        0.000000     100.000000  100.000000
1       Kinopoisk     64.814815       35.185185     100.000000   64.814815
2   Live_Facebook    100.000000        0.000000     100.000000  100.000000
3    Live_YouTube    100.000000        0.000000     100.000000  100.000000
4         Netflix    100.000000        0.000000      82.075472  100.000000
5      PrimeVideo     81.578947       18.421053      86.111111   81.578947
6      SoundCloud    100.000000        0.000000      96.551724  100.000000
7         Spotify    100.000000        0.000000      86.440678  100.000000
8           Vimeo     76.190476       23.809524     100.000000   76.190476
9     YandexMusic    100.000000        0.000000      72.815534  100.000000
10     YouTube_PC    100.000000        0.000000      98.039216  100.000000
11             ww     82.296651       17.703349      97.175141   82.296651

Overall Accuracy: 0.9084
F1 Score: 0.9073
```



Boosting_Type: 'Gradient Boosting Decision Tree',

learning_rate:0.1,

num_leaves: 4

Metric: 'multi_logloss',

Max_dept: 3

Training Time is 0.13 seconds , which is significantly less than Training Time taken by Aligned Byte Random Forest (0.63 seconds).

For 8 classes this model giving 100 percent accuracy and for other class accuracy near to 80 percent except Kionpoisk accuracy 65 percent .

Prediction Time is .05 seconds ,due to which it can be use real-time predictions.

Accuracy of Aligned Bytes LightGBM is slightly better than Aligned Bytes Random Forest.

Overall Aligned Bytes LightGBM is performing better than Aligned Bytes LightGBM incase of Encrypted Server Hello

VNAT and ISCXVPN2016 Dataset

Aligned-Bytes-LightGBM Classifier

1-ENCRYPTED SERVER NAME INDICATION

```
Training Time: 0.1483 seconds

Class-wise Performance Table:
          Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0      FileTransfer    100.000000        0.000000          100.0  100.000000
1             Gmail    100.000000        0.000000          100.0  100.000000
2   Google Services    100.000000        0.000000          100.0  100.000000
3           Hangout    100.000000        0.000000           80.0  100.000000
4           Netflix    100.000000        0.000000          100.0  100.000000
5             Vimeo    100.000000        0.000000          100.0  100.000000
6           Youtube    100.000000        0.000000          100.0  100.000000
7       ssl.gstatic     88.888889       11.111111          100.0   88.888889

Overall Accuracy: 0.9938
F1 Score: 0.9940

Confusion Matrix:
[[10  0  0  0  0  0  0  0]
 [ 0 36  0  0  0  0  0  0]
 [ 0  0 15  0  0  0  0  0]
 [ 0  0  0  4  0  0  0  0]
 [ 0  0  0  0 14  0  0  0]
 [ 0  0  0  0  0 40  0  0]
 [ 0  0  0  0  0  0 34  0]
 [ 0  0  0  1  0  0  0  8]]
Prediction Time: 0.0340 seconds
```

Following are the hyperparameters for LightGBM which are used:

Boosting_Type: 'Gradient Boosting Decision Tree',

learning_rate:0.1,

num_leaves: 4

Metric: 'multi_logloss',

Max_dept: 3

Training Time is 0.15 seconds , which is significantly less than Training Time taken by Aligned Byte Random Forest (0.69 seconds).

For 7 classes this model giving 100 percent accuracy and for other class accuracy near to 90 percent.

Prediction Time is 0.034 seconds ,due to which it can be use real-time predictions.

Accuracy of Aligned Bytes LightGBM is similar to Aligned Bytes Random Forest.

Overall Aligned Bytes LightGBM is performing better than Aligned Bytes LightGBM incase of Encrypted Server Hello in terms of training time

-ENCRYPTED CLIENT HELLO

```
Training Time: 0.1180 seconds
```

```
Class-wise Performance Table:
          Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0     FileTransfer    100.000000        0.000000     100.000000  100.000000
1            Gmail    100.000000        0.000000     100.000000  100.000000
2  Google Services     86.666667       13.333333      59.090909   86.666667
3          Hangout      0.000000      100.000000       0.000000    0.000000
4          Netflix    100.000000        0.000000     100.000000  100.000000
5            Vimeo    100.000000        0.000000     100.000000  100.000000
6          Youtube    100.000000        0.000000     100.000000  100.000000
7      ssl.gstatic     44.444444       55.555556      66.666667   44.444444

Overall Accuracy: 0.9321
F1 Score: 0.9219

Confusion Matrix:
[[10  0  0  0  0  0  0  0]
 [ 0 36  0  0  0  0  0  0]
 [ 0  0 13  0  0  0  0  2]
 [ 0  0  4  0  0  0  0  0]
 [ 0  0  0  0 14  0  0  0]
 [ 0  0  0  0  0 40  0  0]
 [ 0  0  0  0  0  0 34  0]
 [ 0  0  5  0  0  0  0  4]]
Prediction Time: 0.0310 seconds
```

Following are the hyperparameters for LightGBM which are used:

Boosting_Type: 'Gradient Boosting Decision Tree',

learning_rate:0.1,

num_leaves: 4

Metric: 'multi_logloss',

Max_dept: 3

Training Time is 0.11 seconds , which is significantly less than Training Time taken by Aligned Byte Random Forest (0.63 seconds).

For 5 classes this model giving 100 percent accuracy and accuracy for Hangout is very less.

Prediction Time is 0.031 seconds ,due to which it can be use real-time predictions.

Accuracy of Aligned Bytes LightGBM is similar to Aligned Bytes Random Forest.

Overall Aligned Bytes LightGBM is performing better than Aligned Bytes LightGBM incase of Encrypted Server Hello in terms of training time.

Recomposed-Bytes-LightGBM classifier

ENCRYPTED SERVER NAME INDICATION

```
Training Time: 0.1022 seconds

Class-wise Performance Table:
                Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0         FileTransfer    100.000000        0.000000          100.0  100.000000
1                Gmail    100.000000        0.000000          100.0  100.000000
2       Google Services    100.000000        0.000000          100.0  100.000000
3              Hangout    100.000000        0.000000           80.0  100.000000
4              Netflix    100.000000        0.000000          100.0  100.000000
5                Vimeo    100.000000        0.000000          100.0  100.000000
6              Youtube    100.000000        0.000000          100.0  100.000000
7          ssl.gstatic     88.888889       11.111111          100.0   88.888889

Overall Accuracy: 0.9938
F1 Score: 0.9940

Confusion Matrix:
[[11  0  0  0  0  0  0  0]
 [ 0 36  0  0  0  0  0  0]
 [ 0  0 14  0  0  0  0  0]
 [ 0  0  0  4  0  0  0  0]
 [ 0  0  0  0 14  0  0  0]
 [ 0  0  0  0  0 40  0  0]
 [ 0  0  0  0  0  0 34  0]
 [ 0  0  0  1  0  0  0  8]]
Prediction Time: 0.0375 seconds
```

Following are the hyperparameters for LightGBM which are used:

Boosting_Type: 'Gradient Boosting Decision Tree',

learning_rate:0.1,

num_leaves: 4

Metric: 'multi_logloss',

Max_dept: 3

Training Time is 0.10 seconds , which is less than Training Time taken by Recomposed Bytes Random Forest (0.16 seconds).

For 7 classes this model giving 100 percent accuracy and for other class accuracy near to 90 percent.s

Prediction Time is .0375 seconds ,due to which it can be use real-time predictions.

Accuracy of Aligned Bytes LightGBM is slightly better than Recomposed Bytes Random Forest.

Overall Recomposed Bytes LightGBM is performing better Recomposed Bytes Random Forest incase of Encrypted Server Hello.

2-ENCRYPTED CLIENT HELLO

```
Training Time: 0.0952 seconds

Class-wise Performance Table:
                Class  Accuracy (%)  Error Rate (%)  Precision (%)  Recall (%)
0         FileTransfer    100.000000        0.000000     100.000000  100.000000
1                Gmail    100.000000        0.000000     100.000000  100.000000
2      Google Services     42.857143       57.142857      75.000000   42.857143
3              Hangout    100.000000        0.000000      57.142857  100.000000
4              Netflix    100.000000        0.000000     100.000000  100.000000
5                Vimeo    100.000000        0.000000     100.000000  100.000000
6              Youtube    100.000000        0.000000     100.000000  100.000000
7          ssl.gstatic     66.666667       33.333333      50.000000   66.666667

Overall Accuracy: 0.9321
F1 Score: 0.9302

Confusion Matrix:
[[11  0  0  0  0  0  0  0]
 [ 0 36  0  0  0  0  0  0]
 [ 0  0  6  2  0  0  0  6]
 [ 0  0  0  4  0  0  0  0]
 [ 0  0  0  0 14  0  0  0]
 [ 0  0  0  0  0 40  0  0]
 [ 0  0  0  0  0  0 34  0]
 [ 0  0  2  1  0  0  0  6]]
Prediction Time: 0.0270 seconds
```

Following are the hyperparameters for LightGBM which are used:

Boosting_Type: ‘Gradient Boosting Decision Tree’,

learning_rate:0.1,

num_leaves: 4

Metric: ‘multi_logloss’,

Max_dept: 3

Training Time is 0.09 seconds , which is less than Training Time taken by Recomposed Byte Random Forest (0.13 seconds).

For 6 classes this model giving 100 percent accuracy and for other class accuracy is less than 60 percent

Prediction Time is .0375 seconds ,due to which it can be use real-time predictions.

Accuracy of Aligned Bytes LightGBM is slightly better than Aligned Bytes Random Forest.

Overall Aligned Bytes LightGBM is performing better than Aligned Bytes LightGBM incase of Encrypted Server Hello

5. Conclusion

Based on the analysis and comparison of the various models used for classifying the Encrypted Server Name Indication (ESNI) and Encrypted Client Hello (ECH) datasets, the following conclusions can be drawn:

1. BGRUA Model (Bi-Directional GRU Attention Model):

- ESNI Dataset: BGRUA performs well, with higher accuracy than MATEC, while requiring significantly less training time than MATEC. However, BGRUA takes significantly more time compared to Aligned Byte Random Forest and LightGBM models.
- ECH Dataset: BGRUA's accuracy drops for ECH due to the greater complexity of the encrypted extensions. Its training time is still shorter than MATEC, but it takes considerably more time compared to the Random Forest and LightGBM models, making it less efficient for real-time predictions.

2. MATEC Model:

- ESNI Dataset: MATEC achieves slightly lower accuracy than BGRUA but provides good performance, though it has the longest training time among the models.
- ECH Dataset: MATEC’s performance on ECH is also lower, with accuracy dropping due to the encryption complexity. MATEC's training time is the slowest, which makes it less suitable for applications requiring fast responses.

3. Random Forest Models (Aligned-Bytes and Recomposed-Bytes):

- ESNI Dataset: Both Aligned-Bytes Random Forest and Recomposed-Bytes Random Forest deliver excellent accuracy and significantly faster training times compared to BGRUA and MATEC. The aligned-byte model takes the longest to train among the Random Forest models, but both types are fast enough for real-time predictions.
- ECH Dataset: The Random Forest models perform well across most classes but show slight drops in accuracy for specific classes. They are still highly effective, with real-time prediction capability. In terms of training time, Recomposed-Bytes Random Forest outperforms the Aligned-Bytes Random Forest and is much faster than both BGRUA and MATEC.

4. LightGBM Models (Aligned-Bytes and Recomposed-Bytes):

- ESNI Dataset: Both Aligned-Bytes LightGBM and Recomposed-Bytes LightGBM offer very fast training times, with the Recomposed-Bytes LightGBM model being the fastest. These models also maintain high accuracy, making them highly efficient for real-time predictions. The LightGBM models are faster than Random Forest and

perform similarly in terms of accuracy, with a slight edge in Recomposed-Bytes LightGBM over Aligned-Bytes Random Forest.

- ECH Dataset: For the ECH dataset, the LightGBM models perform better than the Random Forest models, with faster training times and more consistent accuracy across classes. The Aligned-Bytes LightGBM shows improved accuracy over Random Forest for most classes, and the Recomposed-Bytes LightGBM performs similarly well, especially in terms of real-time prediction capability.

Key Insights:

- BGRUA takes less training time compared to MATEC, but significantly more time compared to Aligned-Bytes Random Forest and LightGBM models, particularly the Recomposed-Bytes Random Forest and Recomposed-Bytes LightGBM models, which are much faster.
- Proposed LightGBM models, particularly the Recomposed-Bytes LightGBM, perform extremely well, delivering fast training times, high accuracy, and real-time prediction capabilities. These models outperform both BGRUA and MATEC in terms of speed and efficiency.
- Random Forest models are also highly efficient, with both the aligned- and recomposed-byte versions showing excellent accuracy and fast training times. The Recomposed-Bytes Random Forest is the fastest of the Random Forest models.
- MATEC is the slowest in terms of both training time and prediction time, making it less ideal for real-time prediction applications.

For real-time applications where speed and efficiency are critical:

- Proposed LightGBM models (especially Recomposed-Bytes LightGBM) are the best choice, offering the fastest training and prediction times while maintaining high accuracy.
- Random Forest models, particularly Recomposed-Bytes Random Forest, also perform well with fast processing times, making them a solid choice for real-time predictions.
- BGRUA offers a good balance of accuracy and speed, but it still takes significantly longer to train and predict compared to the Random Forest and LightGBM models.
- MATEC is best suited for scenarios where accuracy is prioritized over speed, but it is not ideal for real-time applications due to its slower processing times.

6. Contribution

1-Developed Script for converting TLS v1.2,TLS v1.3 datasets into Encrypted Client Hello Dataset with the help of Scapy and tshark.

2-Developed Script for converting TLS v1.2,TLS v1.3 datasets into Encrypted Server Name Indication with the help of Scapy and tshark.

3-Implemented Aligned Bytes Random Forest (AB-RF) for Traffic classification of ECH and ESNI Dataset

4-Implemented Recomposed Bytes Random Forest (RB-RF) on ECH and ESNI

5-Implemented Base Classifiers like BGRUA (Bi-Directional Gated Recurrent Unit with Self Attention) and MATEC(Multi-Head Attention Encoder with Convolutional Layers)

6-Analysis and Result Evaluation of WNL Dataset

7-Analysis of Models and Result Evaluation on different Datasets(VNAT and ISCXVPN2016 Dataset

8-Creation of Encrypted Client Hello Dataset from VNAT and ISCXVPN2016 Dataset

9-Creation of Encrypted Server Name Indication Dataset from VNAT and ISCXVPN2016 Dataset

10-Implemented a new approach of using LightGBM classifier instead of Random Forest for Recomposed Bytes and Aligned Bytes