RIT Rochester Institute of Technology

Guarding Against Cyber Threats: A Scalable Solution for Intrusion Detection in Imbalanced Network Traffic

CHARITHA MADHAMSETTY cm7513@rit.edu

INTRODUCTION

Motivation: As cyber attacks continue to evolve, detecting intrusions in the network traffic is very important.

Aim: The main aim of the project is to classify the network traffic as normal or anomalous using LSTM and to perform the efficient multi-class classification where the random forest model is used to classify the detected malicious network data into specific attack types. Isolation Forest is used for Unsupervised learning based approach.

Research Questions:

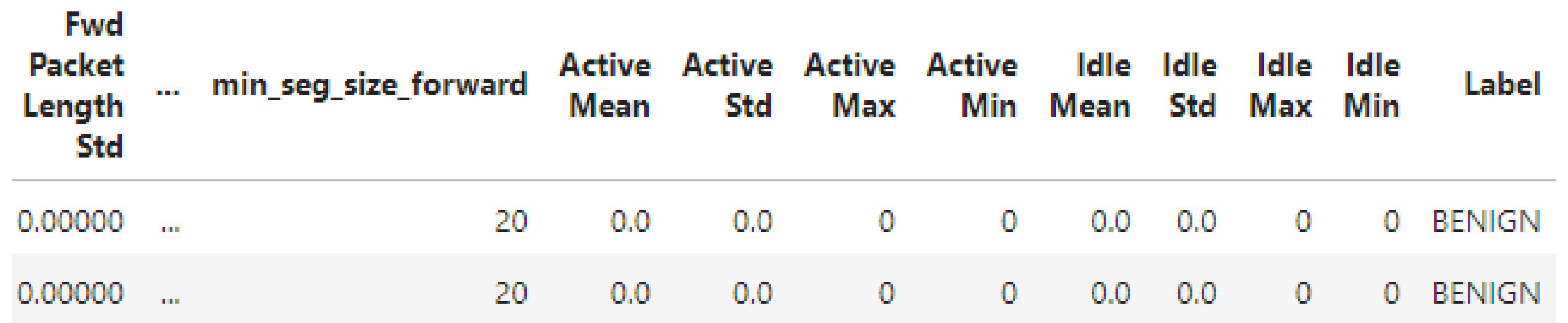
RQ1: Is it important to consider temporal dependencies for this data? If yes, how are they captured in the proposed model?

RQ2: Can the model process data in real-time quickly (real-time intrusion detection) and efficiently (without more false negatives).

DATASET

Dataset: CICIDS 2017 Dataset

- 2830743 samples with 79 features each.
- 78 numerical features, 1 categorical feature - target.
- DoS, PortScan, DDoS, Brute Force, Web Attack, Bot, Infiltration, Heartbleed.



Fwd Packet Length Std	...	min_seg_size_forward	Active Mean	Active Std	Active Max	Active Min	Idle Mean	Idle Std	Idle Max	Idle Min	Label
0.00000	...	20	0.0	0.0	0	0	0.0	0.0	0	0	BENIGN
0.00000	...	20	0.0	0.0	0	0	0.0	0.0	0	0	BENIGN

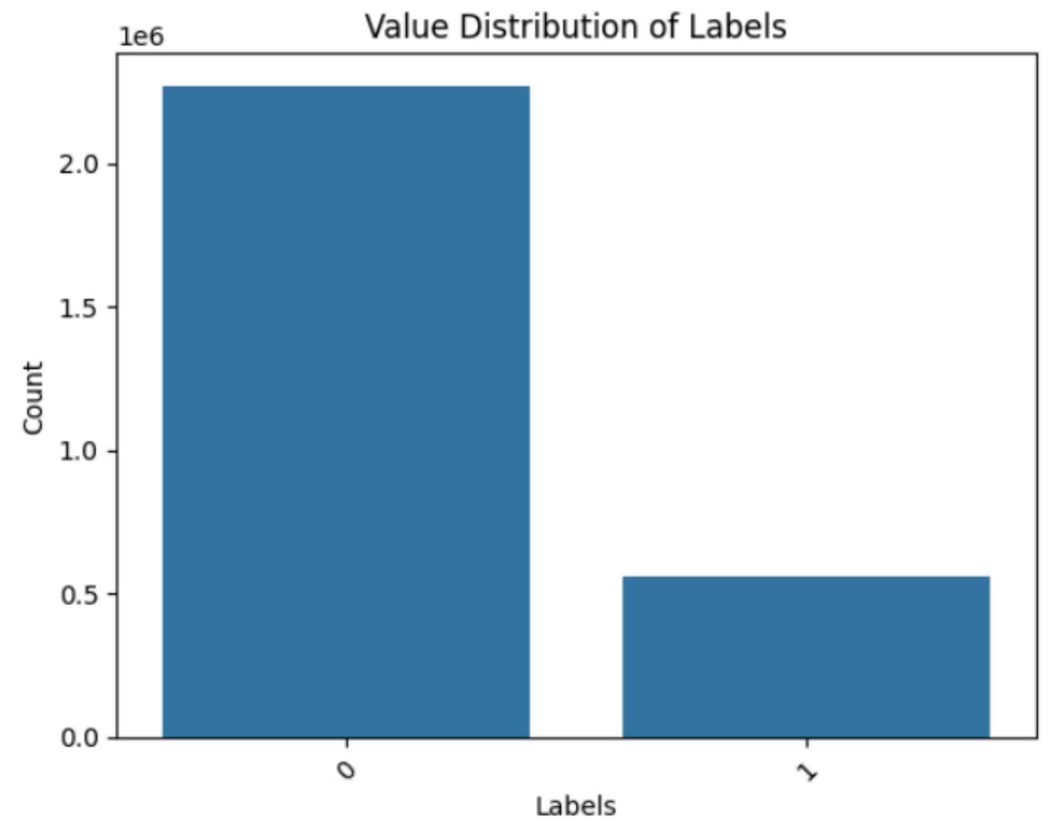


Figure 1: Bar graph representing the imbalance in the class data

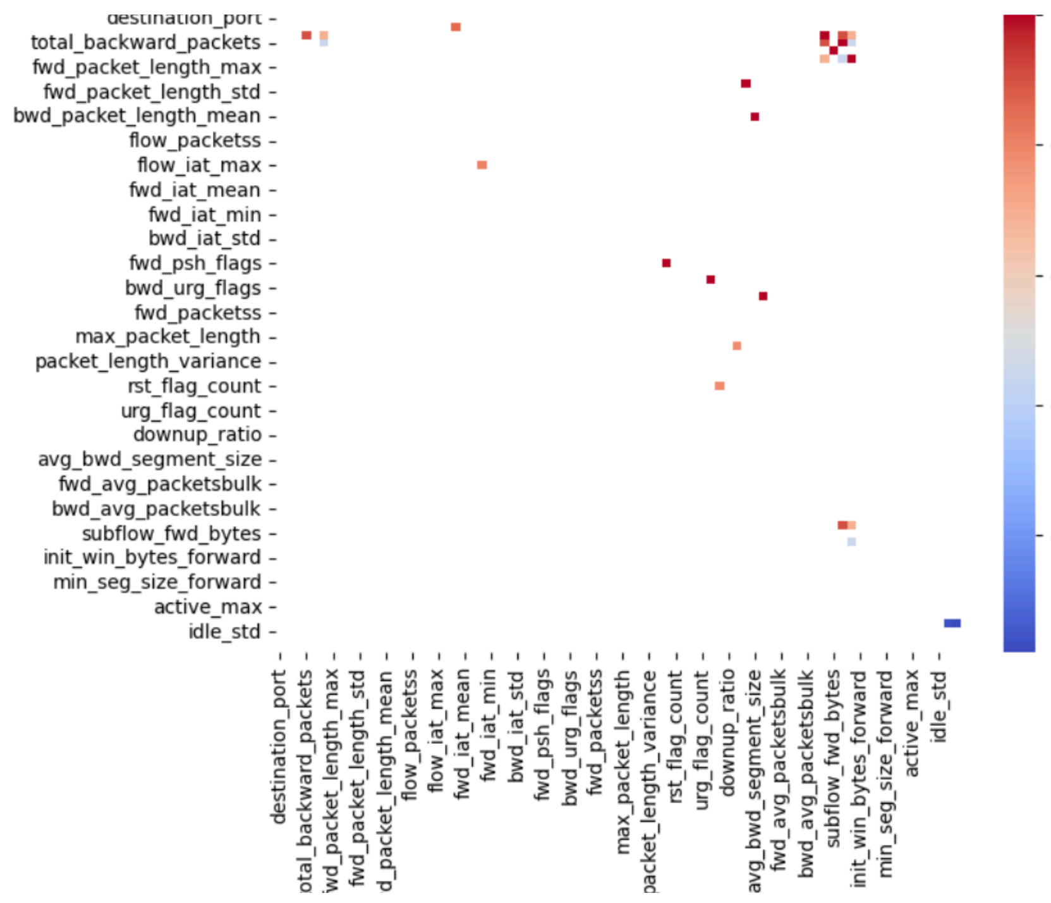


Figure 2: Heatmap showing the features with high multi-collinearity

CHALLENGES

- Heavily imbalanced class distribution (benign and attack) (figure 1)
- Curse of dimensionality
- Multi-collinearity among the features. (figure 2)

METHODOLOGY

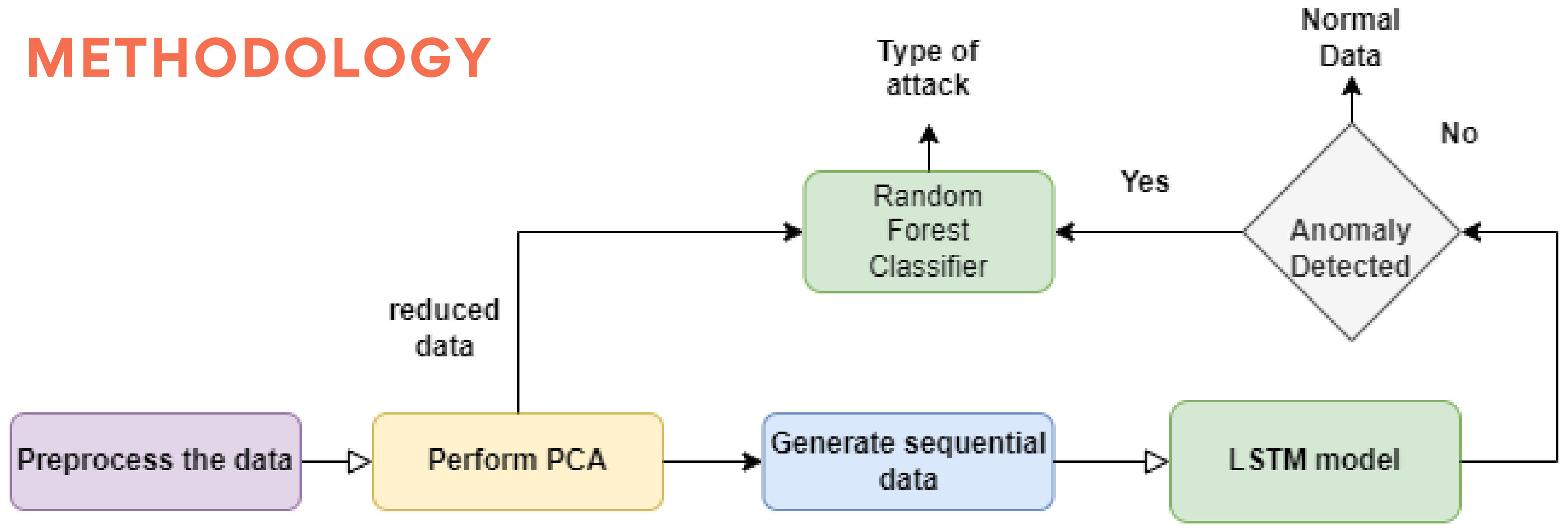


Figure 3: Methodology showing the flow of data in the system

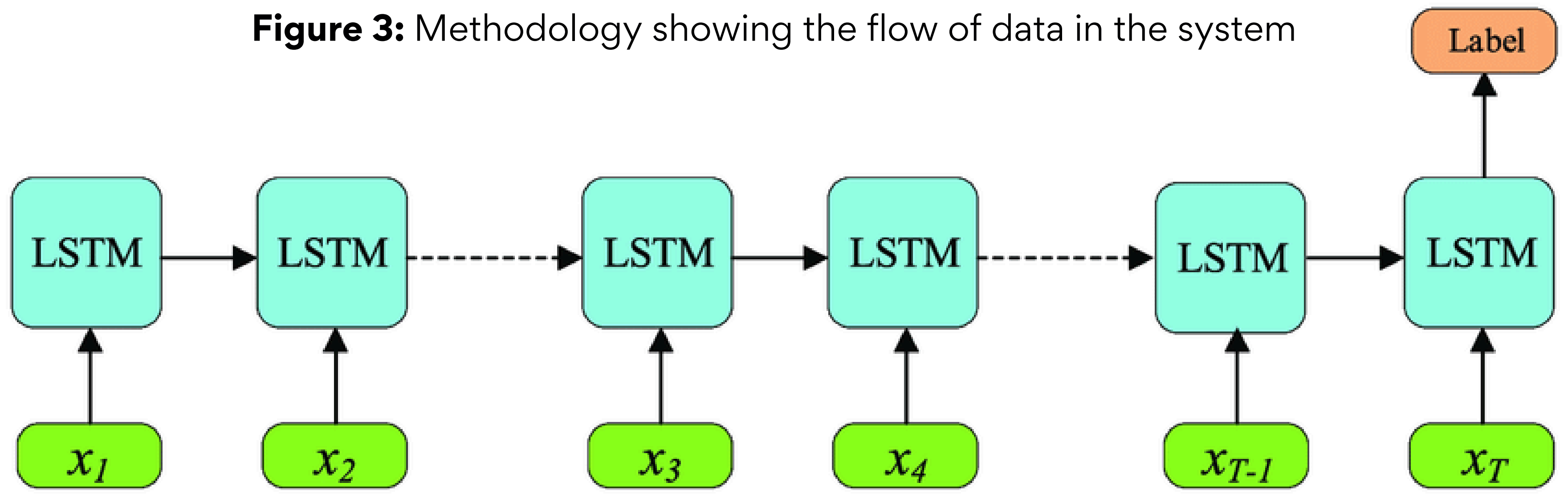


Figure 4: Many to one LSTM Architecture

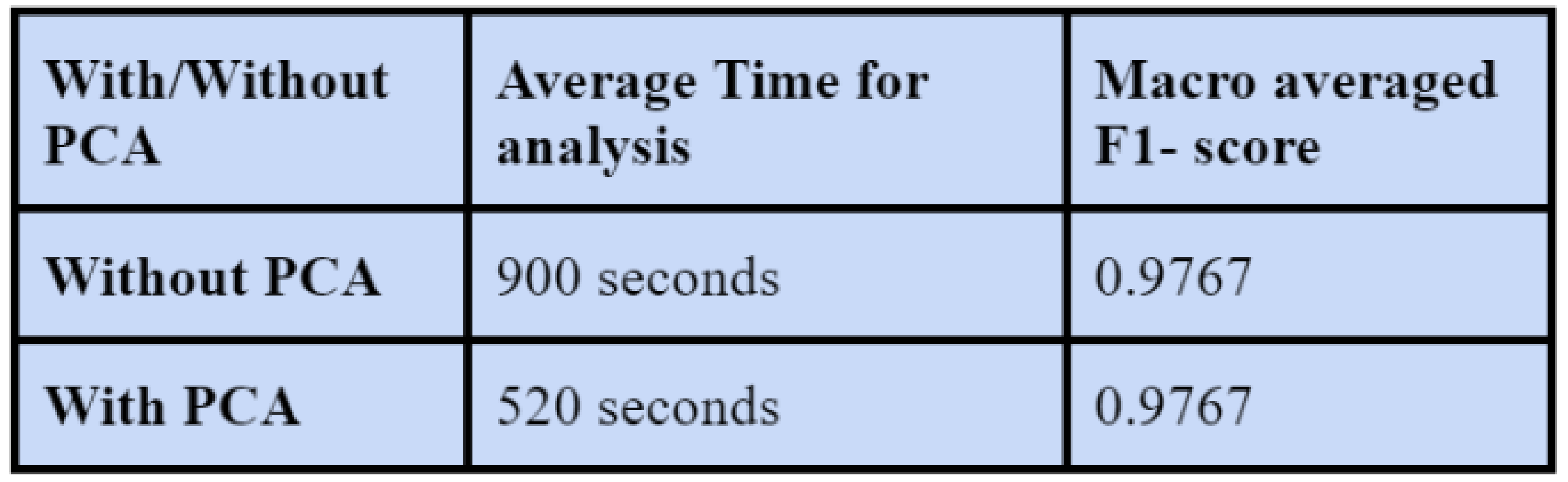
RESULTS AND KEY FINDINGS

When the **LSTM model** is trained with and without performing **PCA:**

- Time taken for analysis reduced.
- No Significant change in F1-score
- Model's performance is unchanged, but the time efficiency improved.

Isolation Forest performance was not efficient compared to LSTM

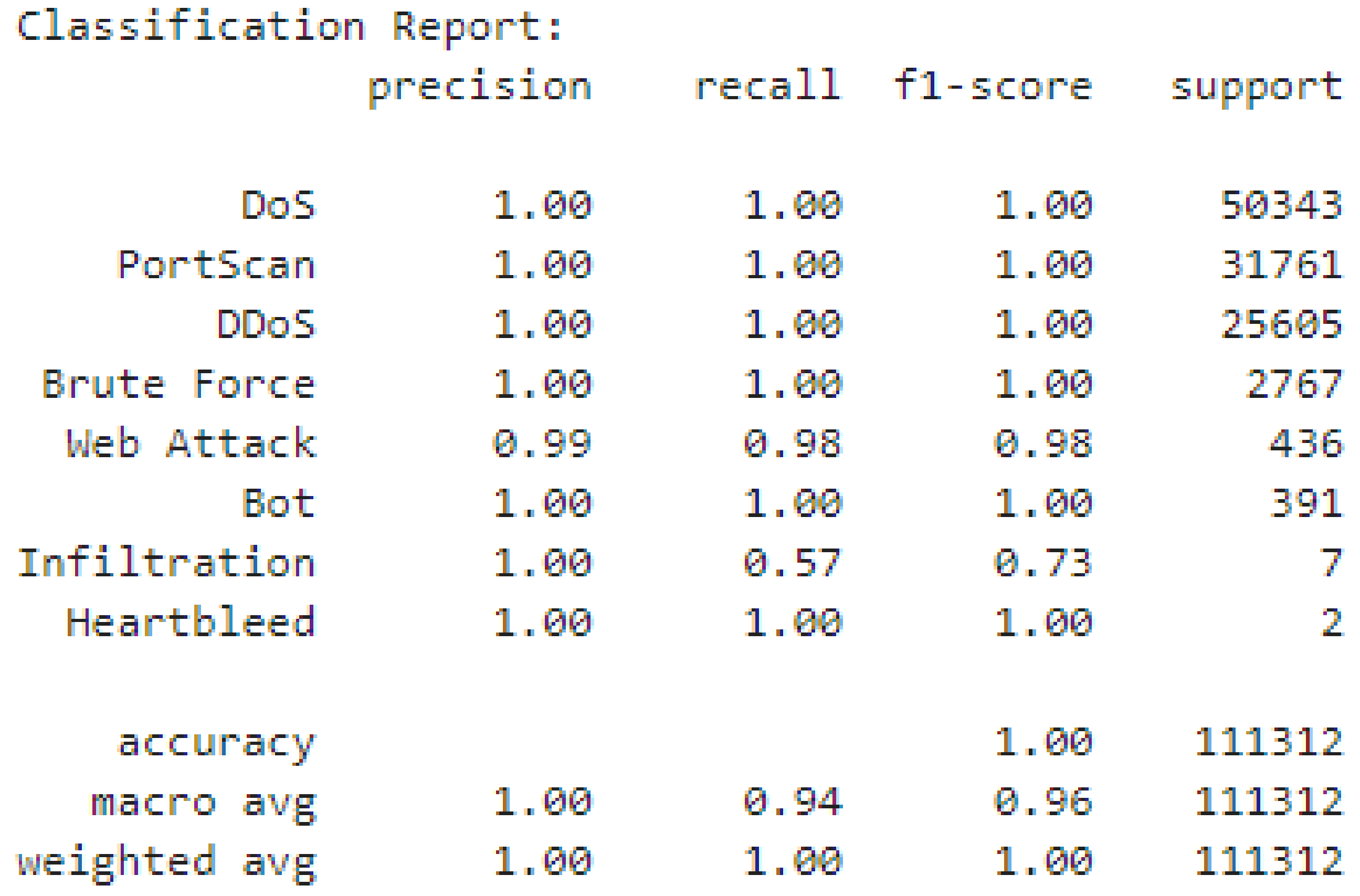
F1-Score - 0.4567, Recall - 0.5052



With/Without PCA	Average Time for analysis	Macro averaged F1- score
Without PCA	900 seconds	0.9767
With PCA	520 seconds	0.9767

Table 1: Results with and without PCA

Random Forest Classifier Results:



```
Classification Report:
               precision    recall  f1-score   support

         DoS       1.00      1.00      1.00     50343
    PortScan       1.00      1.00      1.00     31761
        DDoS       1.00      1.00      1.00     25605
 Brute Force       1.00      1.00      1.00      2767
  Web Attack       0.99      0.98      0.98       436
         Bot       1.00      1.00      1.00       391
Infiltration       1.00      0.57      0.73         7
  Heartbleed       1.00      1.00      1.00         2

    accuracy                           1.00    111312
   macro avg       1.00      0.94      0.96    111312
weighted avg       1.00      1.00      1.00    111312
```

Figure 5: Multi-class Classification Report of Random Forest

CONCLUSION

It is important to consider temporal dependencies in case of network attacks. Many-to-one LSTM captured the temporal patterns producing a good F1-score for the test data.

For a real time system like intrusion detection, time efficiency is important and as well as is it important to ensure no anomalies go undetected. So, to ensure this balance, PCA is used where time reduced without compromise in efficiency.

DISCUSSION

The network data is evolving. There can be new kinds of attacks other than ones in the dataset. So, it is important to consider unsupervised learning based approach. How can the efficiency of such approach can be improved ? Can LSTM based auto-encoder perform well.

How can real-time intrusion detection models be integrated into existing network security infrastructure to provide seamless protection against attacks without causing significant overhead.

REFERENCES

1.Smolen, T., & Benova, L. (2023). Comparing autoencoder and isolation forest in network anomaly detection. Paper presented at the 33(1), 276-282. https://doi.org/10.23919/FRUCT58615.2023.10143005

2.Iqbal, A., Amin, R., Alsubaei, F. S., & Alzahrani, A. (2024). Anomaly detection in multivariate time series data using deep ensemble models. PloS One, 19(6), e0303890. https://doi.org/10.1371/journal.pone.0303890

3. Alatawi, M. N., Alsubaie, N., Ullah Khan, H., Sadad, T., Alwageed, H. S., Ali, S., & Zada, I. (2023). Cyber security against intrusion detection using ensemble-based approaches. Security and Communication Networks, 2023, 1-7. https://doi.org/10.1155/2023/8048311