SEOULTECH

SeoulTech UCS Lab
Ubiquitous Computing & Security Laboratory

Machine learning-based IDS for software define 5G network

2019.09.24
SeoulTech
Jose costa Sapalo Sicato

Table of contents

1. Introduction
2. Related work
3. Architecture
4. Intelligent intrusion detection process
5. Experiment result
6. Conclusion

Abstract

- Software-defined architecture has many advantages in providing centralized control and flexible resource management.
- As the focus of network security, intrusion detection systems (IDSs) are usually deployed separately without collaboration
- They are also unable to detect novel attacks with limited intelligent abilities, which are hard to meet the needs of software-defined 5G
- Evaluation results prove that the intelligent IDS achieves better performance with lower overhead.

Introduction

- Software-defined fifth generation (5G) architecture will be a crucial tendency in the development of future 5G networks.
- As a result, new network security architecture and systems are desperately needed to enhance the security of software-defined 5G networks
- As an essential technology in network security, intrusion detection systems (IDSs) have received more and more concerns in efficiently detecting malicious attacks
- To overcome the limitation of traditional IDS, artificial intelligence (AI) has been employed for intelligent detection.

Introduction

At present, there have been a few researches combining IDS and AI.

In this paper, was propose an intelligent IDS based on software defined 5G architecture. Benefit from the software-defined technology, it integrates relevant security function modules into a unified platform which are dynamically invoked under centralized management and control.

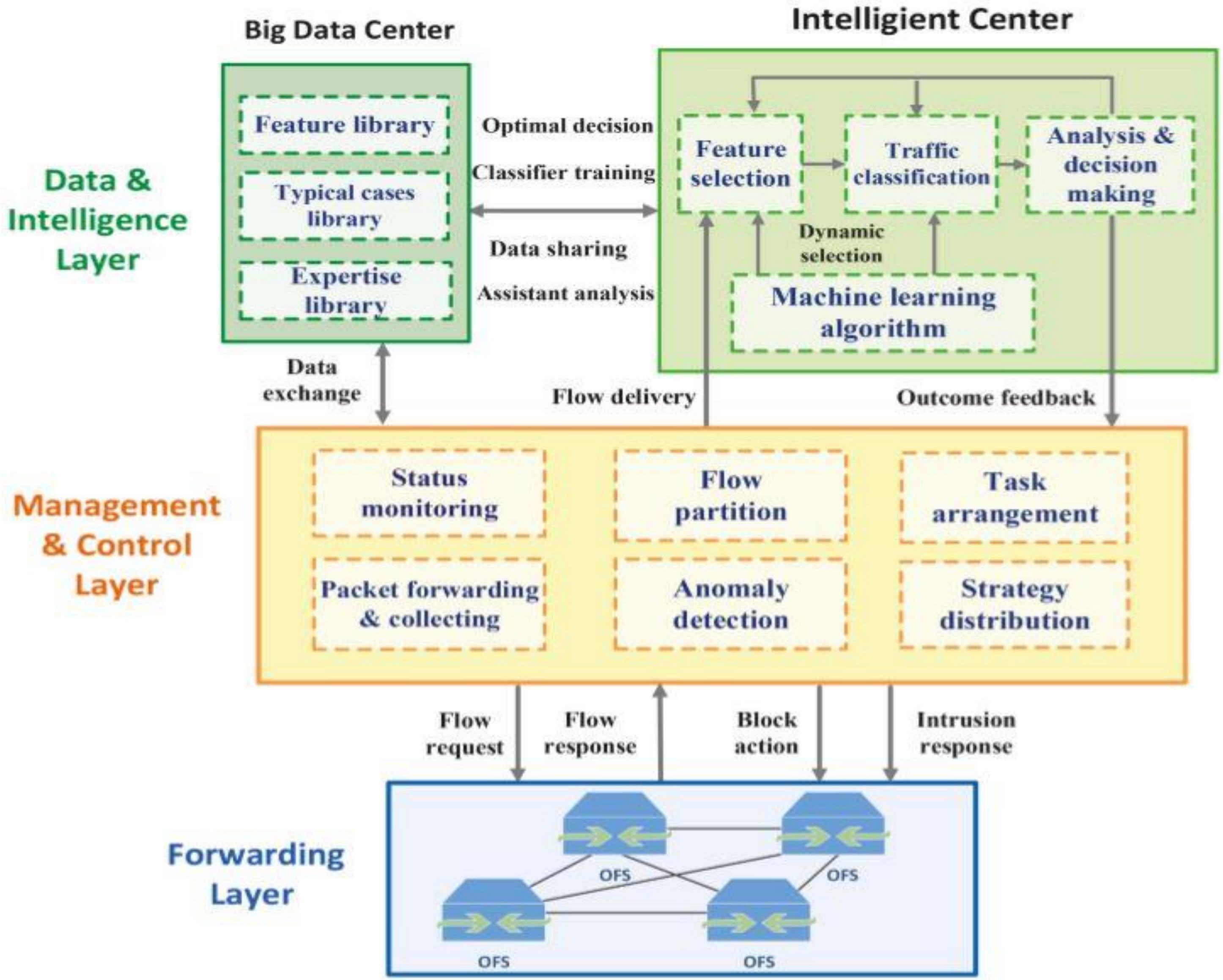
2. Related Work

- As SDN dynamically manages network configurations and controls packet processing in a centralized manner, it has well satisfied the evolution demand of cellular networks in the 5G era, which aims to provide flexible service provisioning mechanisms. Therefore, the combination of 5G and SDN has attracted a lot of research interest. A new paradigm called SoftAir toward next generation (5G) wireless networks is introduced.
- However, with the fast development of software-defined 5G networks, the emergence of unknown attacks also poses severe security challenges.

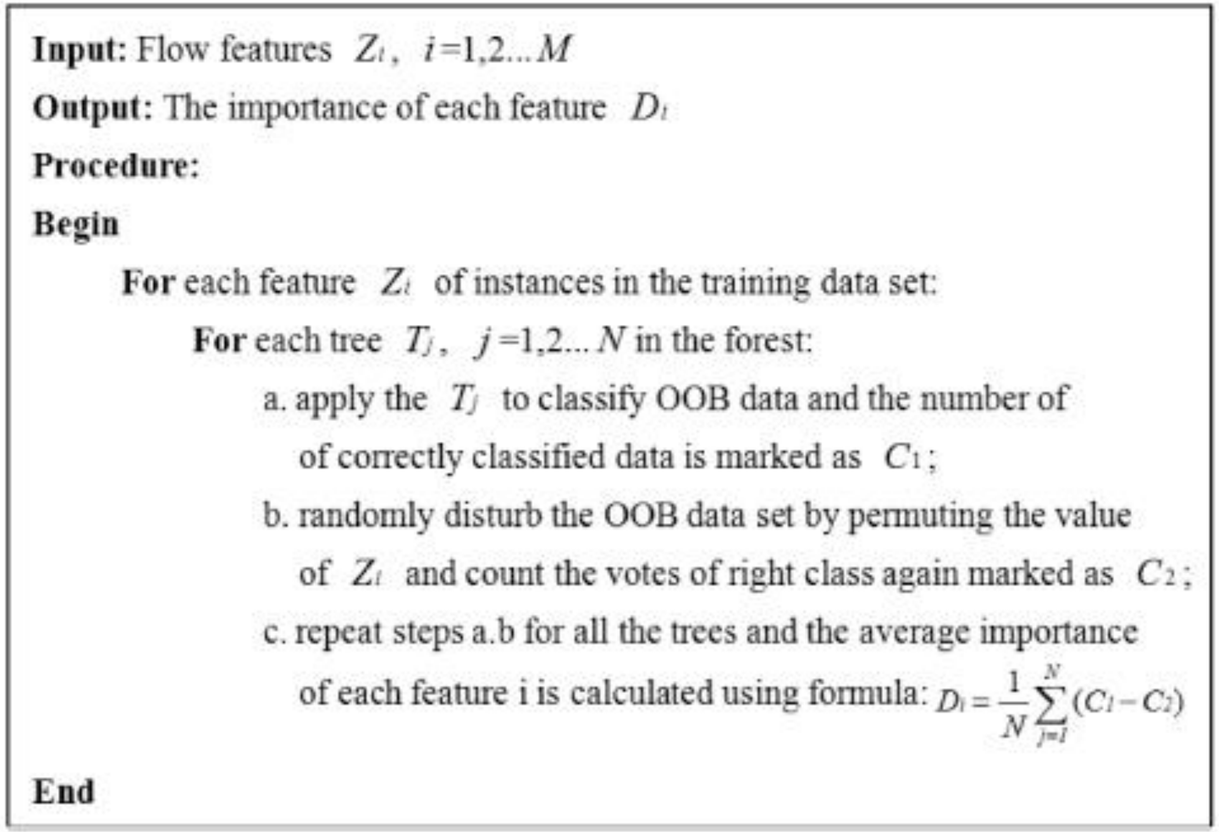
2. Related Work

- It also provides network-layer security services such as packet routing, identity authentication and automated security management in a global view which facilitates the detection and prevention of attacks
- A comprehensive survey of existing SDN-based distributed denial of service (DDoS) attack detection solutions and present an SDN-based proactive DDoS defence framework (ProDefense).

3. Architecture



4. Intelligent intrusion detection process



```
Input: Flow features $Z_i$, $i=1,2...M$
Output: The importance of each feature $D_i$
Procedure:
Begin
    For each feature $Z_i$ of instances in the training data set:
        For each tree $T_j$, $j=1,2...N$ in the forest:
            a. apply the $T_j$ to classify OOB data and the number of
               of correctly classified data is marked as $C_1$;
            b. randomly disturb the OOB data set by permuting the value
               of $Z_i$ and count the votes of right class again marked as $C_2$;
            c. repeat steps a.b for all the trees and the average importance
               of each feature i is calculated using formula: $D_i = \frac{1}{N}\sum_{j=1}^{N}(C_1 - C_2)$
End
```

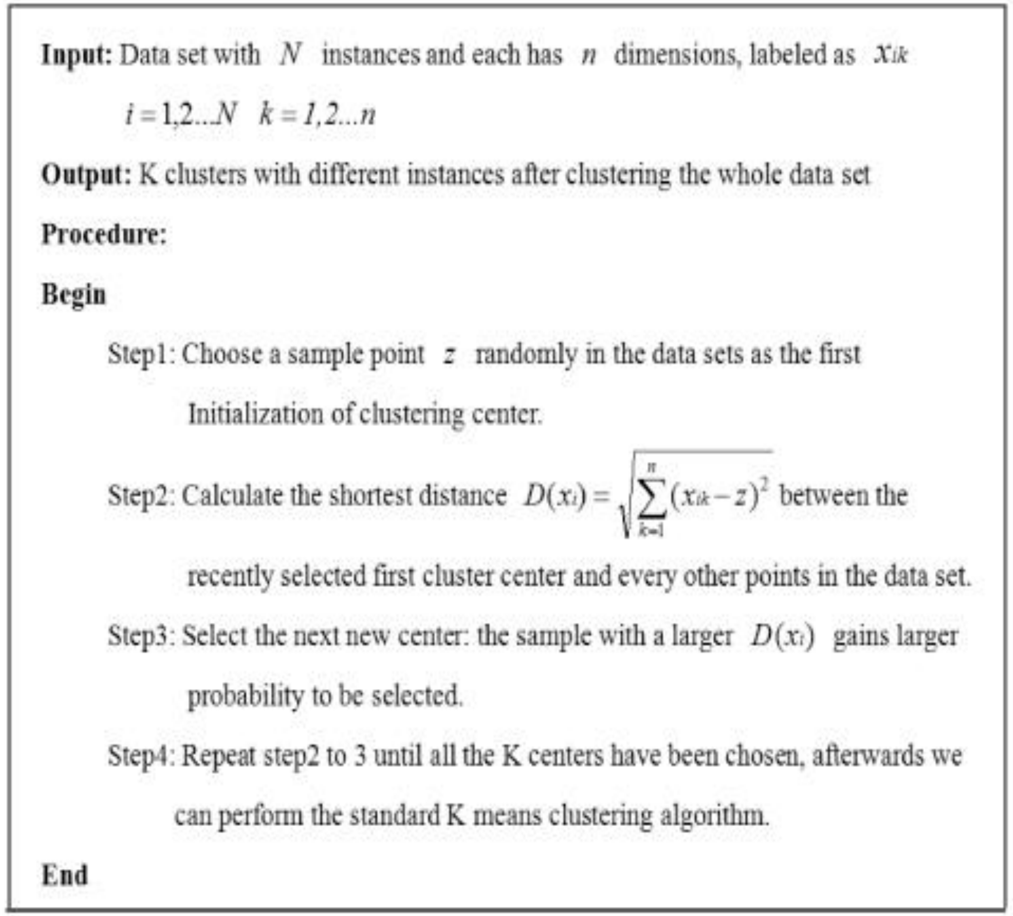
2 *Importance measurement of flow features using RF*

4.1 Random forest

is a collection of uncorrelated structured decision trees deemed as forest.

- If the number of input training data is *N*, we take *N* samples randomly with replacement from the original data.
- For each tree, we choose *m* (*m* < *M*, usually *m* = *M*) features out of *M*-attribute entire set randomly as input variables without replacement.

4. Intelligent intrusion detection process



```
Input: Data set with $N$ instances and each has $n$ dimensions, labeled as $x_{ik}$
    $i = 1,2...N$  $k = 1,2...n$
Output: K clusters with different instances after clustering the whole data set
Procedure:
Begin
    Step1: Choose a sample point $z$ randomly in the data sets as the first
        Initialization of clustering center.
    Step2: Calculate the shortest distance $D(x_i) = \sqrt{\sum_{k=1}^{n}(x_{ik} - z)^2}$ between the
        recently selected first cluster center and every other points in the data set.
    Step3: Select the next new center: the sample with a larger $D(x_i)$ gains larger
        probability to be selected.
    Step4: Repeat step2 to 3 until all the K centers have been chosen, afterwards we
        can perform the standard K means clustering algorithm.
End
```

g. 3 *Main steps of k-means++ algorithm*

4.2 Hybrid clustering-based AdaBoost

- For the first stage, we make a preliminary judgement by adopting *k*-means++ to divide the traffic into two clusters which most probably represent the normal and abnormal instances.
- Later, we further partition the anomaly clusters into four main classes of attacks using the ensemble algorithm AdaBoost.

5. Experiment result

- *5.1 Dataset* KDD Cup 1999 dataset- It contains ~5,000,000 network connections in the training set and nearly 2,000,000 instances in the testing set.
- *5.2 Evaluation metrics*

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$F = \frac{2}{(1/P) + (1/R)}$$

5.3 Performance analysis

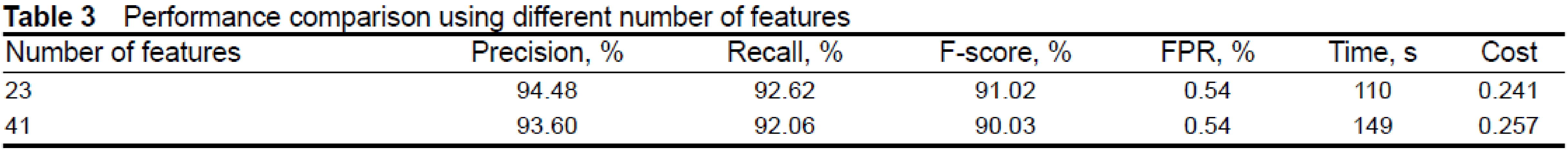


Table 3 Performance comparison using different number of features

Number of features	Precision, %	Recall, %	F-score, %	FPR, %	Time, s	Cost
23	94.48	92.62	91.02	0.54	110	0.241
41	93.60	92.06	90.03	0.54	149	0.257

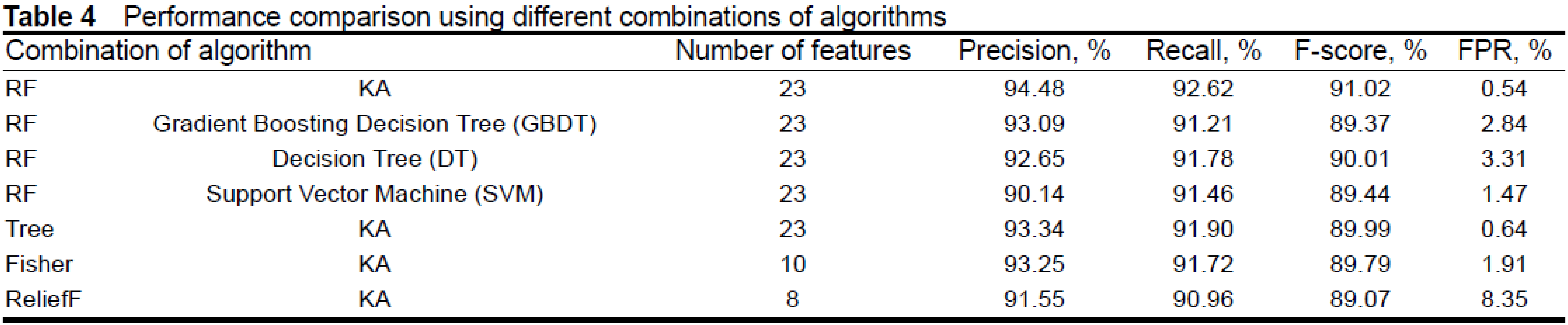


Table 4 Performance comparison using different combinations of algorithms

Combination of algorithm		Number of features	Precision, %	Recall, %	F-score, %	FPR, %
RF	KA	23	94.48	92.62	91.02	0.54
RF	Gradient Boosting Decision Tree (GBDT)	23	93.09	91.21	89.37	2.84
RF	Decision Tree (DT)	23	92.65	91.78	90.01	3.31
RF	Support Vector Machine (SVM)	23	90.14	91.46	89.44	1.47
Tree	KA	23	93.34	91.90	89.99	0.64
Fisher	KA	10	93.25	91.72	89.79	1.91
ReliefF	KA	8	91.55	90.96	89.07	8.35

5.3.2 Evaluations of the proposed solutions:

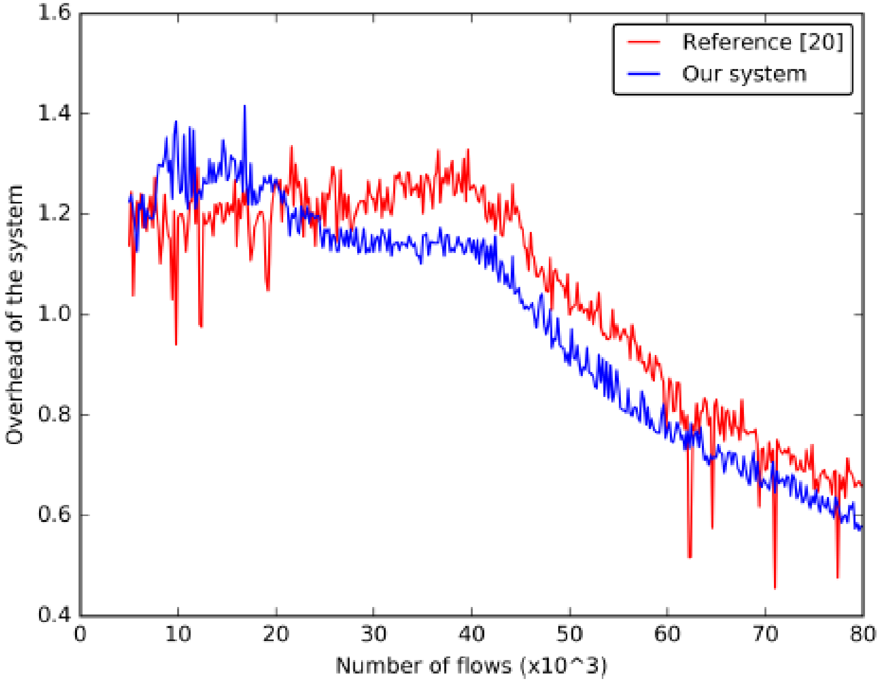


Fig. 4 *Overhead produced by different systems with different numbers of flows*

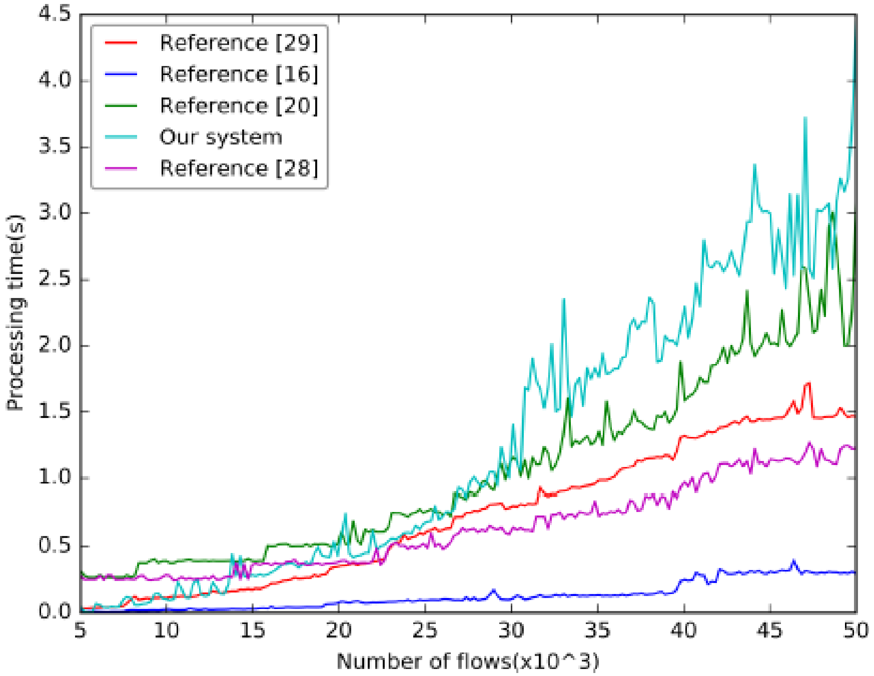


Fig. 5 *Processing times of different systems*

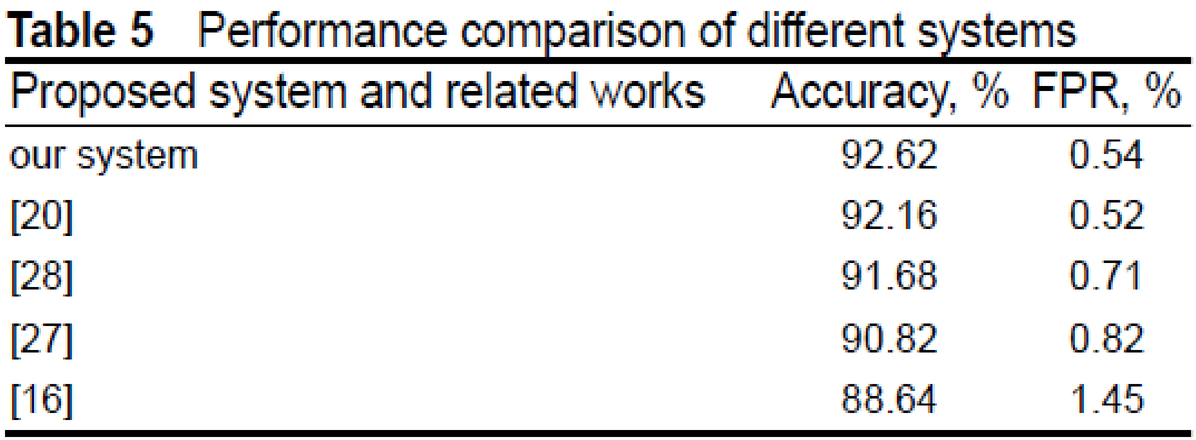


Table 5 Performance comparison of different systems

Proposed system and related works	Accuracy, %	FPR, %
our system	92.62	0.54
[20]	92.16	0.52
[28]	91.68	0.71
[27]	90.82	0.82
[16]	88.64	1.45

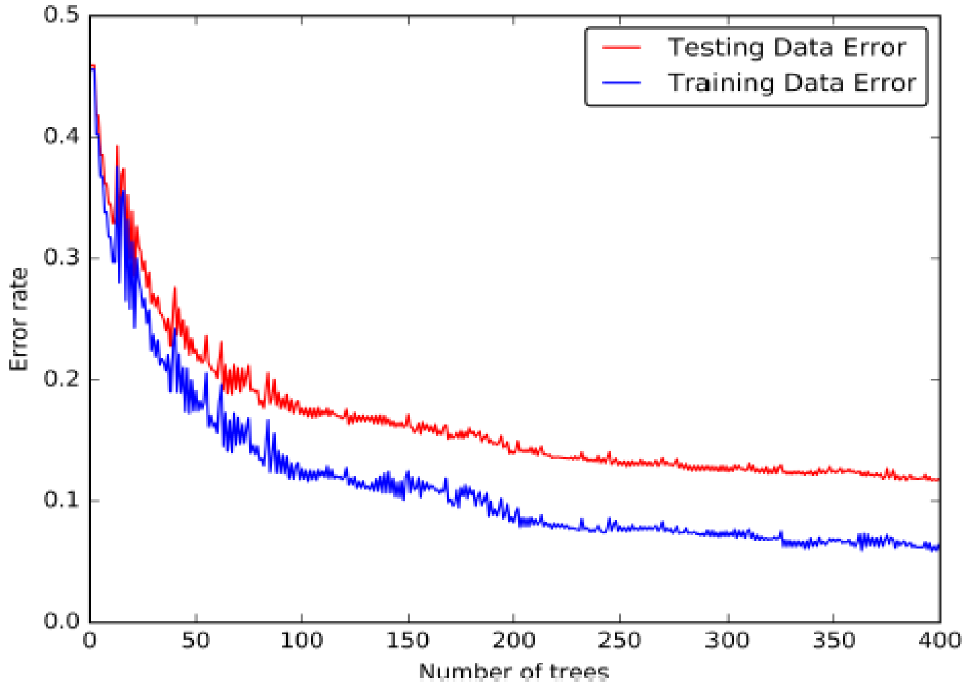
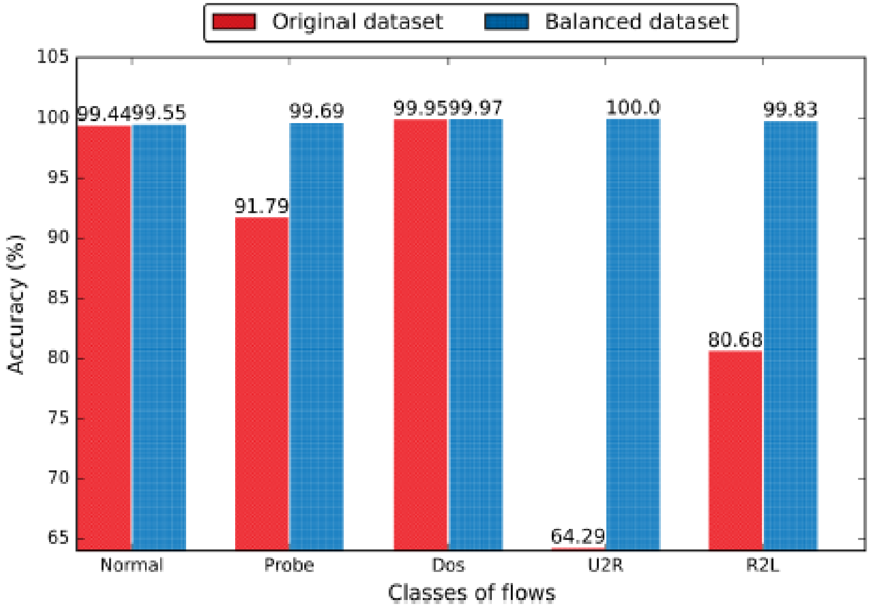
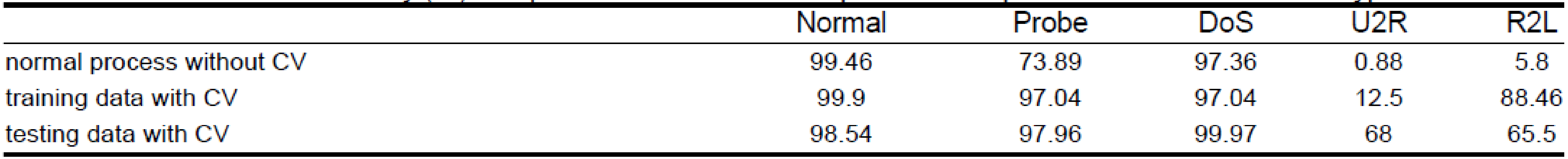


Fig. 6 *Error rate with the different NTs*



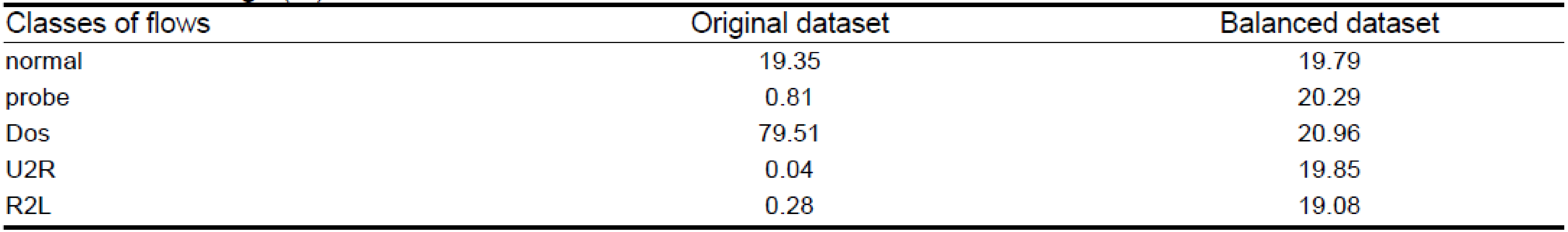
- NTs and the error rate of classification using the training and testing dataset, respectively, are plotted. It is shown that the error rate decreases as NTs become larger.

Table 7 Classification accuracy (%) comparison between normal process and processes with CV in each type of attack



	Normal	Probe	DoS	U2R	R2L
normal process without CV	99.46	73.89	97.36	0.88	5.8
training data with CV	99.9	97.04	97.04	12.5	88.46
testing data with CV	98.54	97.96	99.97	68	65.5

Table 8 Percentage (%) of flows of each class in different datasets



Classes of flows	Original dataset	Balanced dataset
normal	19.35	19.79
probe	0.81	20.29
Dos	79.51	20.96
U2R	0.04	19.85
R2L	0.28	19.08

6. Conclusion

- An intelligent IDS based on software-defined 5G architecture using machine learning algorithms.
- It integrates and coordinates security function modules under centralized management and applies machine learning algorithms to detect intrusions intelligently.

Opinion

- 5G network introduces a slew of cybersecurity concerns and problems.
- Anomaly-based **intrusion detection techniques**, that utilize **algorithms** of **machine learning**, have the capability to recognize unpredicted malicious.
- Machine Learning is the field of study that gives the capability to learn and improve from experience without being programmed explicitly automatically.