

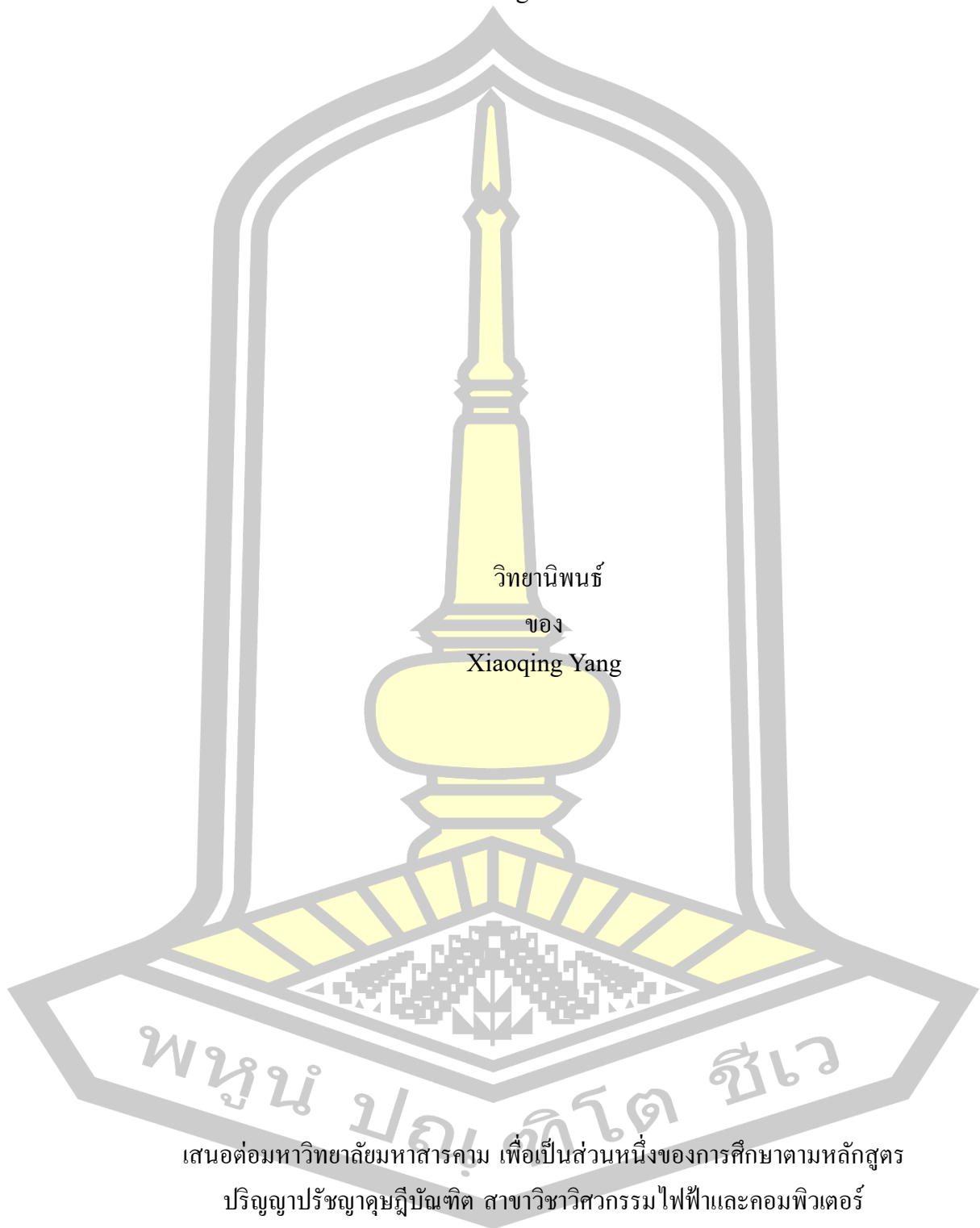
Research on Encrypted Traffic Analysis Technology based on Machine Learning
Algorithm

Xiaoqing Yang

A Thesis Submitted in Partial Fulfillment of Requirements for
degree of Doctor of Philosophy in Electrical and Computer Engineering
June 2025

Copyright of Mahasarakham University

Research on Encrypted Traffic Analysis Technology based on Machine Learning
Algorithm



วิทยานิพนธ์
ของ
Xiaoqing Yang

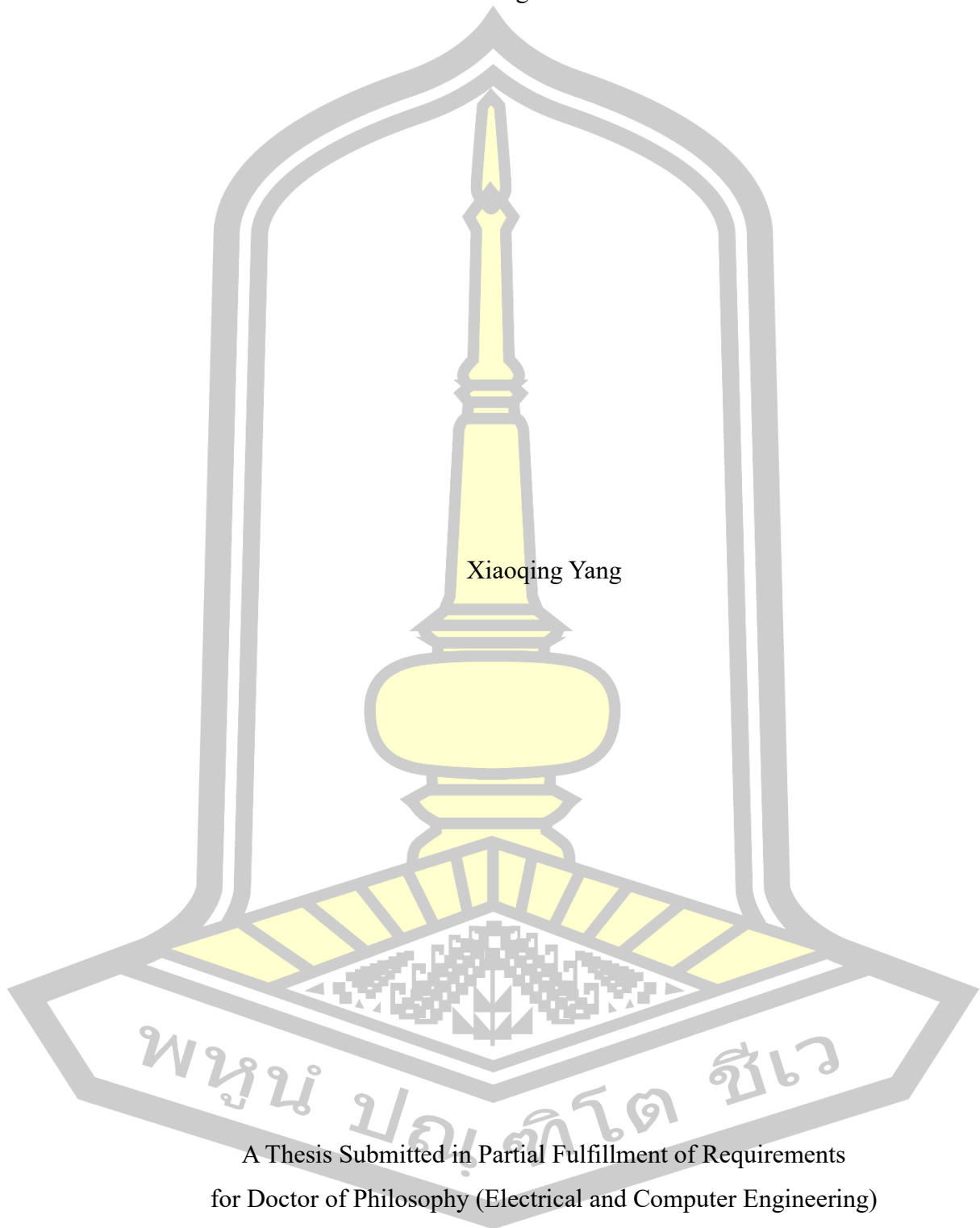
เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์

มีนาคม 2568

ลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

Research on Encrypted Traffic Analysis Technology based on Machine Learning
Algorithm

Xiaoqing Yang



A Thesis Submitted in Partial Fulfillment of Requirements
for Doctor of Philosophy (Electrical and Computer Engineering)

June 2025

Copyright of Mahasarakham University



The examining committee has unanimously approved this Thesis, submitted by Ms. Xiaoqing Yang , as a partial fulfillment of the requirements for the Doctor of Philosophy Electrical and Computer Engineering at Mahasarakham University

Examining Committee

Chairman

(Assoc. Prof. Anan Kruesubthaworn
, Ph.D.)

Advisor

(Assoc. Prof. Niwat Angkawisittpan
, Ph.D.)

Committee

(Asst. Prof. Nattawoot Suwannata ,
Ph.D.)

Committee

(Assoc. Prof. Supannika Wattana ,
Ph.D.)

Committee

(Asst. Prof. Nawarat Piladaeng ,
Ph.D.)

Mahasarakham University has granted approval to accept this Thesis as a partial fulfillment of the requirements for the Doctor of Philosophy Electrical and Computer Engineering

(Assoc. Prof. Juckamas Laohavanich ,
Ph.D.)

Dean of The Faculty of Engineering

(Asst. Prof. Pondej Chaowarat , Ph.D.)
Dean of Graduate School

TITLE	Research on Encrypted Traffic Analysis Technology based on Machine Learning Algorithm		
AUTHOR	Xiaoqing Yang		
ADVISORS	Associate Professor Niwat Angkawisittpan , Ph.D.		
DEGREE	Doctor of Philosophy	MAJOR	Electrical and Computer Engineering
UNIVERSITY	Maharakham University	YEAR	2025

ABSTRACT

The focus of this research is to apply an improved machine learning algorithm to realize the efficient and reliable identification and classification of network communication encrypted traffic, and to solve the challenges faced by traditional algorithms in analyzing encrypted traffic after adding encryption protocols. In this study, an enhanced random forest (ERF) algorithm is introduced to optimize the accuracy and efficiency of the identification and classification of encrypted network traffic. Compared with traditional methods, it aims to improve the identification ability of encrypted traffic and fill the knowledge gap in this field. Using the publicly available datasets and preprocessing the original PCAP format packets, the optimal combination of the relevant parameters of the tree was determined by grid search cross-validation, and the experimental results were evaluated in terms of performance using accuracy, precision, recall and F1 score, which showed that the average precision was more than 98%, and that compared with the traditional algorithm, the error rate of the traffic test set was reduced, and the data of each performance evaluation index were better, which It shows that the advantages of the improved algorithm are obvious. In the experiment, the enhanced random forest and traditional random forest models were trained and tested on a series of data sets and the corresponding test errors were listed as the basis for judging the model quality. The experimental results show that the enhanced algorithm has good competitiveness. These findings have implications for cybersecurity professionals, researchers, and organizations, providing a practical solution to enhance threat detection and data privacy in the face of evolving encryption technologies. This study provides valuable insights for practitioners and decision-makers in the cybersecurity field.

Keyword : Enhanced Random Forest, Encrypted Network Traffic, Encrypted Traffic Identification, Encrypted Traffic Anomaly Detection

ACKNOWLEDGEMENTS

I would like to thank Mahasarakham University for providing me with the opportunity to study for PhD, all professors and staff of Faculty of Engineering for their guidance, encouragement, and assistance, and especially my advisor, Associate Professor Niwat Angkawisitpan, for teaching me about academic research methods. I would also like to thank my classmates who studied with me in the library every day, because of their encouragement and support I was able to successfully complete my graduation requirements. Finally, I would like to thank my family for their understanding and dedication so that I can study here with all my heart and soul, and I would also like to thank myself for still having the courage to face challenges and overcome myself at this age so that I can successfully complete my doctoral studies. I wish Mahasarakham University will getting better and better, and I wish all professors and students good health and all the best.

Xiaoqing Yang

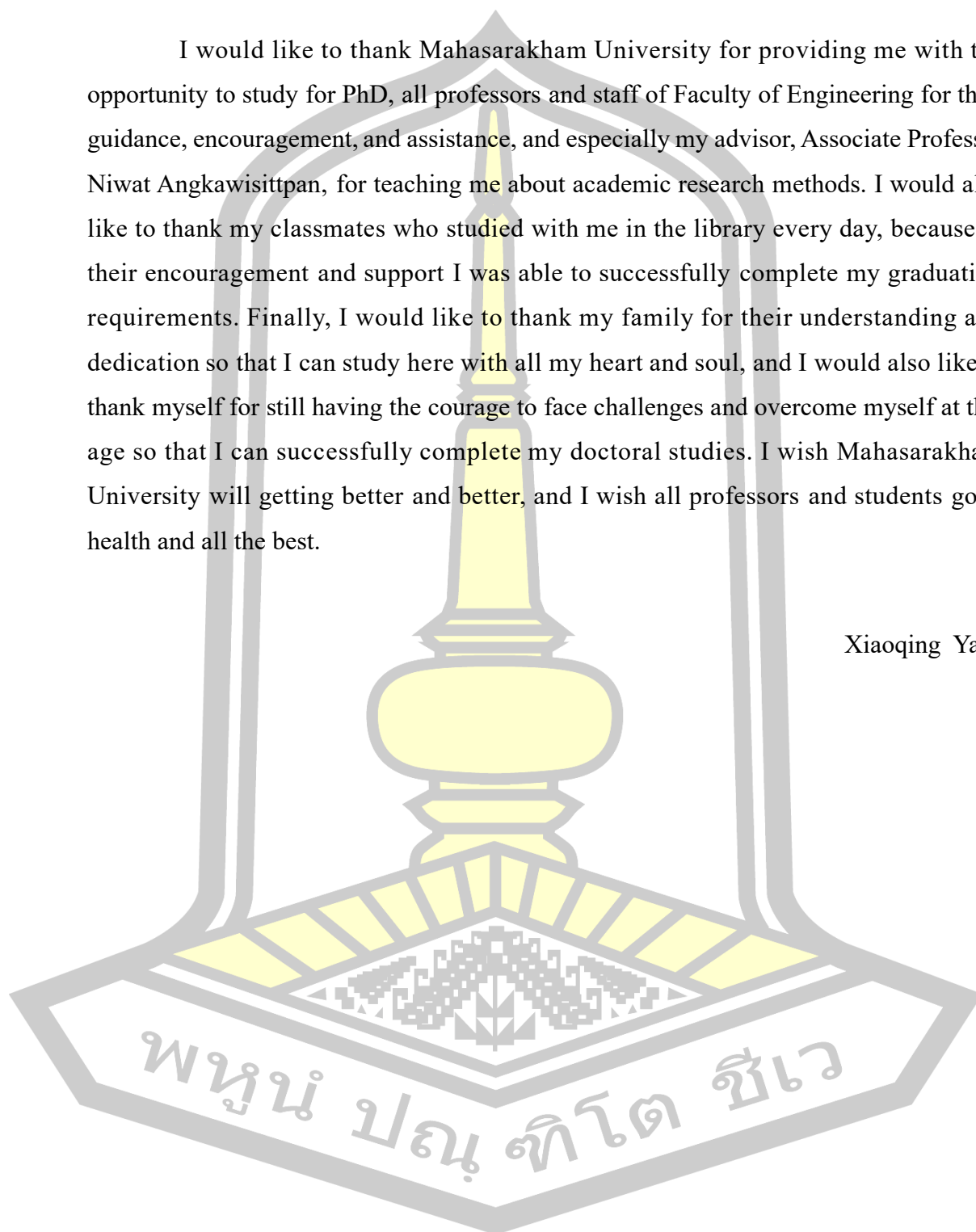


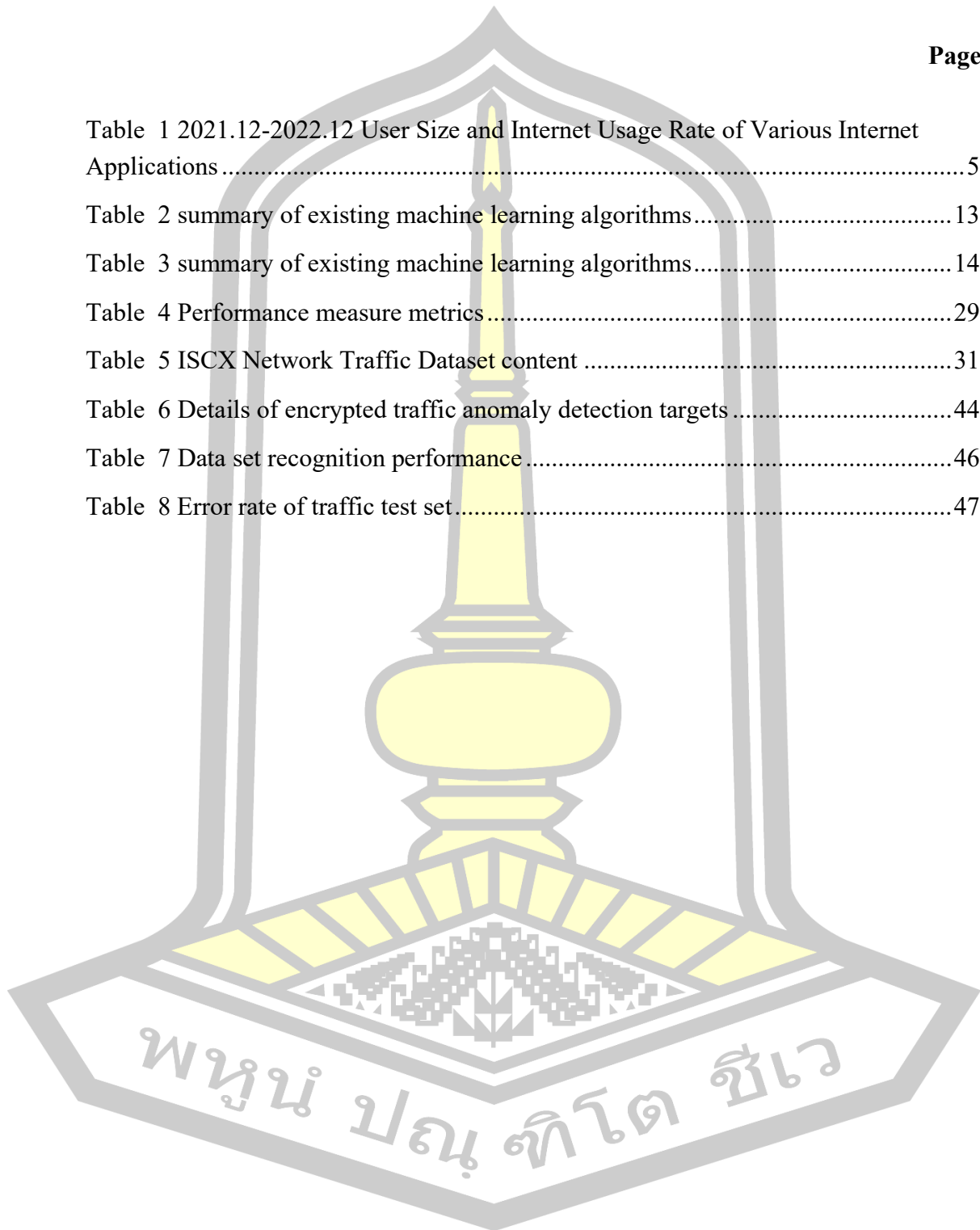
TABLE OF CONTENTS

	Page
ABSTRACT.....	D
ACKNOWLEDGEMENTS.....	E
TABLE OF CONTENTS.....	F
LIST OF TABLES.....	H
LIST OF FIGURES.....	I
Chapter 1 Introduction.....	1
I Research Background.....	1
1.1 Rapid development of the Internet.....	1
1.2 Network communication traffic encryption.....	1
1.3 Impact of traffic encryption technology.....	4
1.4 Challenges in network traffic classification.....	5
1.5 Problems to be solved by this thesis.....	6
II Research Objectives.....	7
2.1 Encrypted traffic identification.....	8
2.2 Network anomaly detection.....	9
2.3 Machine learning algorithm.....	10
III Benefits of Research.....	11
3.1 Staying Current with Research Trends.....	11
3.2 Monitoring Network Behavior.....	11
3.3 Enhancing Network Security.....	12
Chapter 2 Literature Review.....	13
I Categories of Machine Learning Algorithm.....	13
II Traditional Machine Learning Algorithm.....	14
2.1 Algorithm Introduction.....	14
2.2 Application in the encrypted traffic analysis.....	15

III Deep Learning Algorithm	18
3.1 Algorithm Introduction.....	18
3.2 Application in the Encrypted Traffic Analysis	18
Chapter 3 Research Methodology.....	21
I Relevant Concepts.....	21
1.1 Generation of encrypted traffic	21
1.2 Principles of Traffic Encryption	22
1.3 Commonly used encryption protocols.....	23
II Encrypted Traffic Analysis.....	24
2.1 Traffic Collection and Preprocessing	24
2.2 Traffic Representation	25
2.3 Traffic Analysis	27
2.4 Performance Evaluation.....	28
III Machine Learning Algorithm.....	29
3.1 Experimental Datasets.....	29
3.2 Network Protocol Analyzer WireShark.....	33
3.3 Enhanced Random Forest Algorithm	36
Chapter 4 Results and Discussion.....	41
I Encrypted Traffic Identification	41
II Raw Traffic Data Segmentation.....	41
III Encrypted Traffic Anomaly Detection.....	42
IV Results Analysis.....	44
Chapter 5 Conclusions	48
I Summary of the Research.....	48
II Contribution to Network Security.....	49
III Advantages Over Traditional Approaches	50
References.....	52
REFERENCES	61
BIOGRAPHY	63

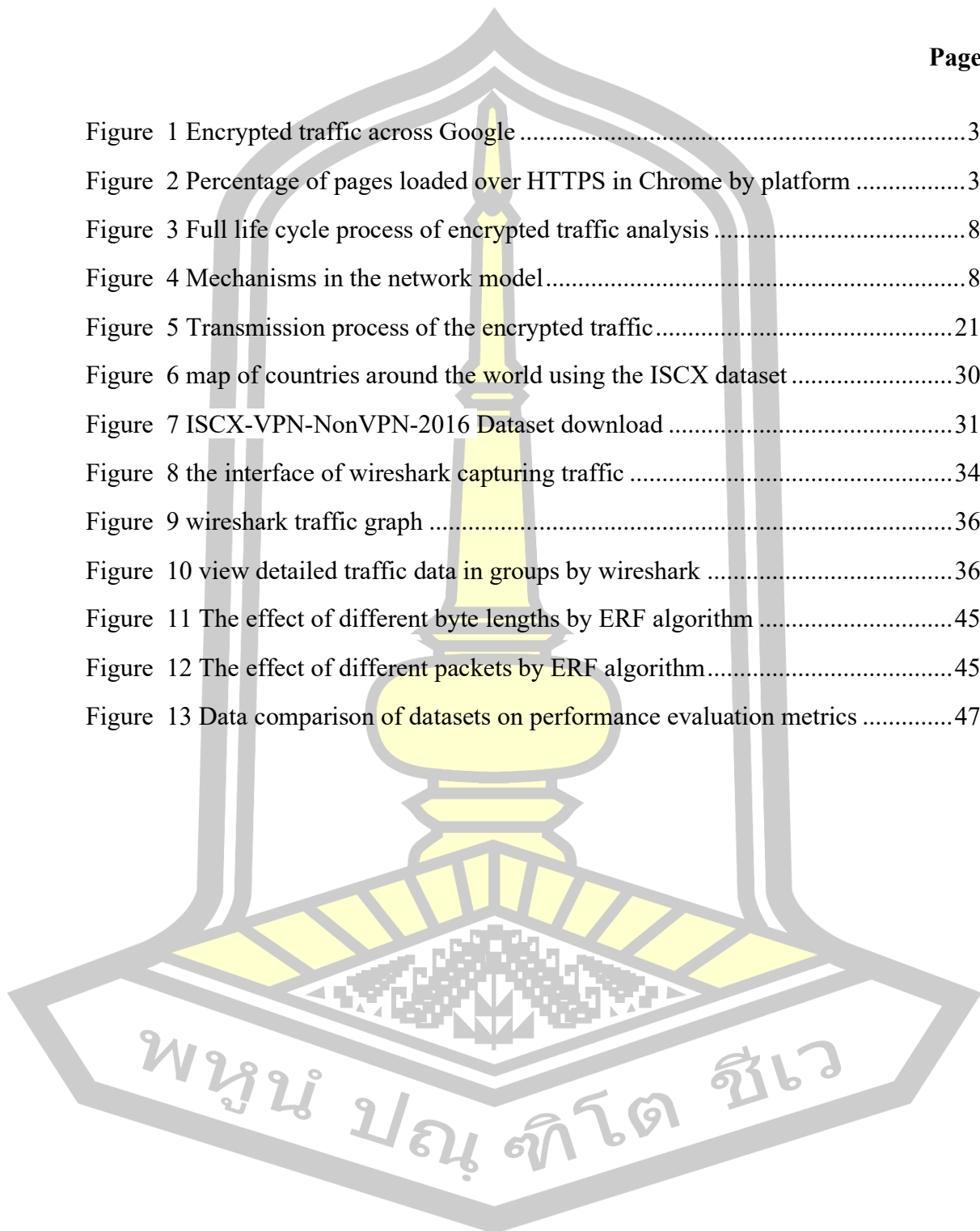
LIST OF TABLES

	Page
Table 1 2021.12-2022.12 User Size and Internet Usage Rate of Various Internet Applications	5
Table 2 summary of existing machine learning algorithms.....	13
Table 3 summary of existing machine learning algorithms.....	14
Table 4 Performance measure metrics.....	29
Table 5 ISCX Network Traffic Dataset content	31
Table 6 Details of encrypted traffic anomaly detection targets	44
Table 7 Data set recognition performance	46
Table 8 Error rate of traffic test set.....	47



LIST OF FIGURES

	Page
Figure 1 Encrypted traffic across Google	3
Figure 2 Percentage of pages loaded over HTTPS in Chrome by platform	3
Figure 3 Full life cycle process of encrypted traffic analysis	8
Figure 4 Mechanisms in the network model.....	8
Figure 5 Transmission process of the encrypted traffic.....	21
Figure 6 map of countries around the world using the ISCX dataset	30
Figure 7 ISCX-VPN-NonVPN-2016 Dataset download	31
Figure 8 the interface of wireshark capturing traffic	34
Figure 9 wireshark traffic graph	36
Figure 10 view detailed traffic data in groups by wireshark	36
Figure 11 The effect of different byte lengths by ERF algorithm	45
Figure 12 The effect of different packets by ERF algorithm.....	45
Figure 13 Data comparison of datasets on performance evaluation metrics	47



Chapter 1 Introduction

I Research Background

1.1 Rapid development of the Internet

The internet serves as a pivotal underpinning for advancing the digital economy and plays a crucial role within the sphere of network information industry development. Emerging technologies such as artificial intelligence, 5G, and virtual reality are injecting fresh impetus into the future trajectory of the internet's evolution. The progression of "5G+ Industrial Internet" is propelling the pace of upgrading conventional industrial technologies, expediting the widespread adoption of a novel production paradigm that seamlessly interconnects people, machinery, and objects. It stands as a vital cornerstone for driving the high-end, intelligent, and sustainable development of the manufacturing sector.

Internet-related fields encompass a wide range of sectors, including enterprise services, industrial internet, network equipment manufacturing, e-commerce retail, smart hardware, intelligent transportation, internet sports, entertainment, and media, metaverse, intelligent logistics, internet healthcare, internet finance, internet lifestyle services, utility software, online education, as well as internet agriculture and internet real estate, among others.

With the development of "Internet+" technologies such as big data, cloud computing and artificial intelligence, network communication exists in people's work, study and life all the time, resulting in a substantial increase in network traffic data sets.

According to the China Internet Development Report 2023[1], released by the Internet Society of China, the total number of Internet users in the world has exceeded 5 billion. The large number of Internet users raises concerns about the security of their personal information and the data security of their Internet activities.

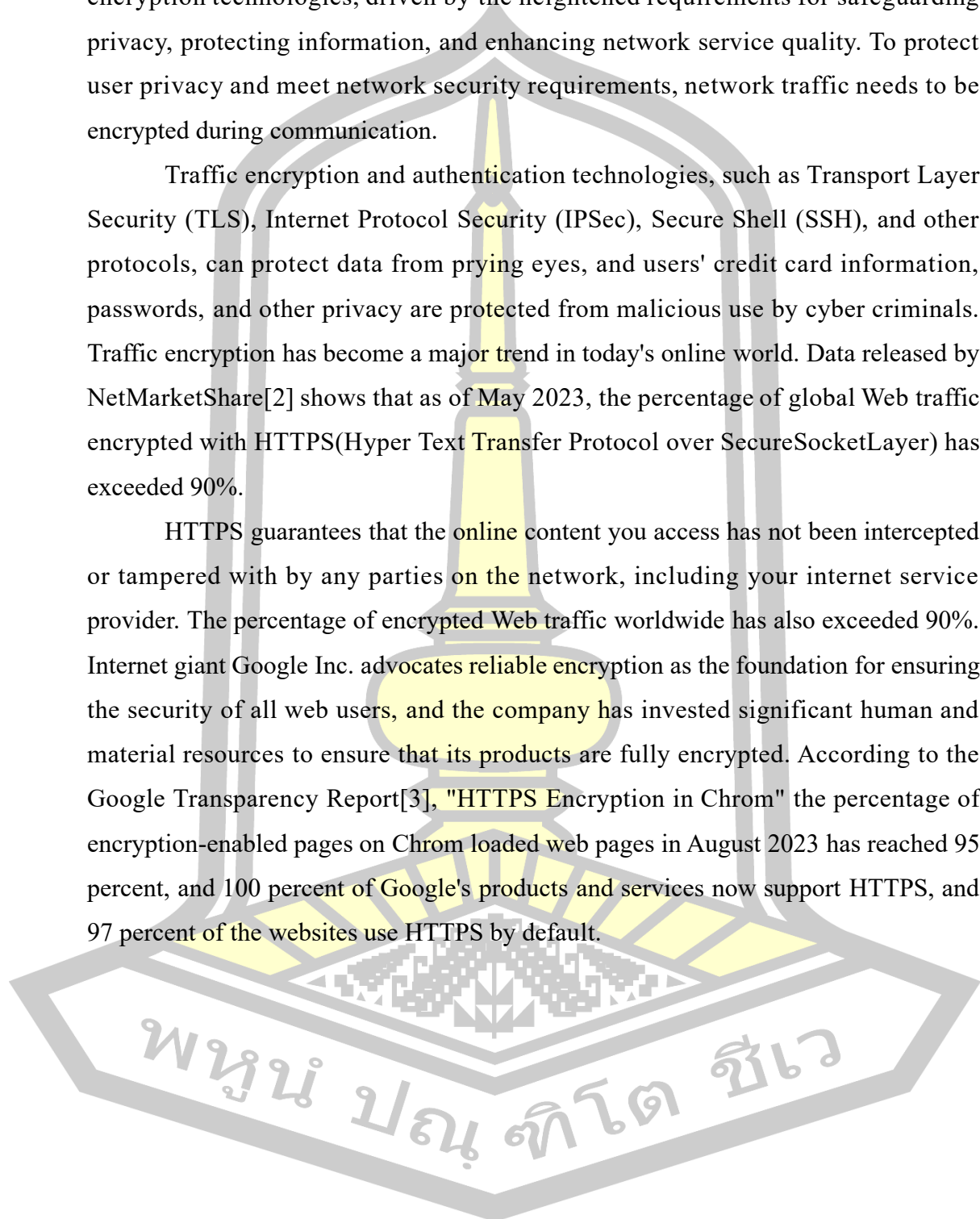
1.2 Network communication traffic encryption

The growing volume of crucial information being transmitted over the Internet presents significant challenges in terms of network monitoring and management. Concurrently, as network traffic escalates, there is an increasing demand for privacy and information security, along with elevated expectations for the quality of network

services. This has led to a robust development and widespread adoption of traffic encryption technologies, driven by the heightened requirements for safeguarding privacy, protecting information, and enhancing network service quality. To protect user privacy and meet network security requirements, network traffic needs to be encrypted during communication.

Traffic encryption and authentication technologies, such as Transport Layer Security (TLS), Internet Protocol Security (IPSec), Secure Shell (SSH), and other protocols, can protect data from prying eyes, and users' credit card information, passwords, and other privacy are protected from malicious use by cyber criminals. Traffic encryption has become a major trend in today's online world. Data released by NetMarketShare[2] shows that as of May 2023, the percentage of global Web traffic encrypted with HTTPS(Hyper Text Transfer Protocol over SecureSocketLayer) has exceeded 90%.

HTTPS guarantees that the online content you access has not been intercepted or tampered with by any parties on the network, including your internet service provider. The percentage of encrypted Web traffic worldwide has also exceeded 90%. Internet giant Google Inc. advocates reliable encryption as the foundation for ensuring the security of all web users, and the company has invested significant human and material resources to ensure that its products are fully encrypted. According to the Google Transparency Report[3], "HTTPS Encryption in Chrom" the percentage of encryption-enabled pages on Chrom loaded web pages in August 2023 has reached 95 percent, and 100 percent of Google's products and services now support HTTPS, and 97 percent of the websites use HTTPS by default.



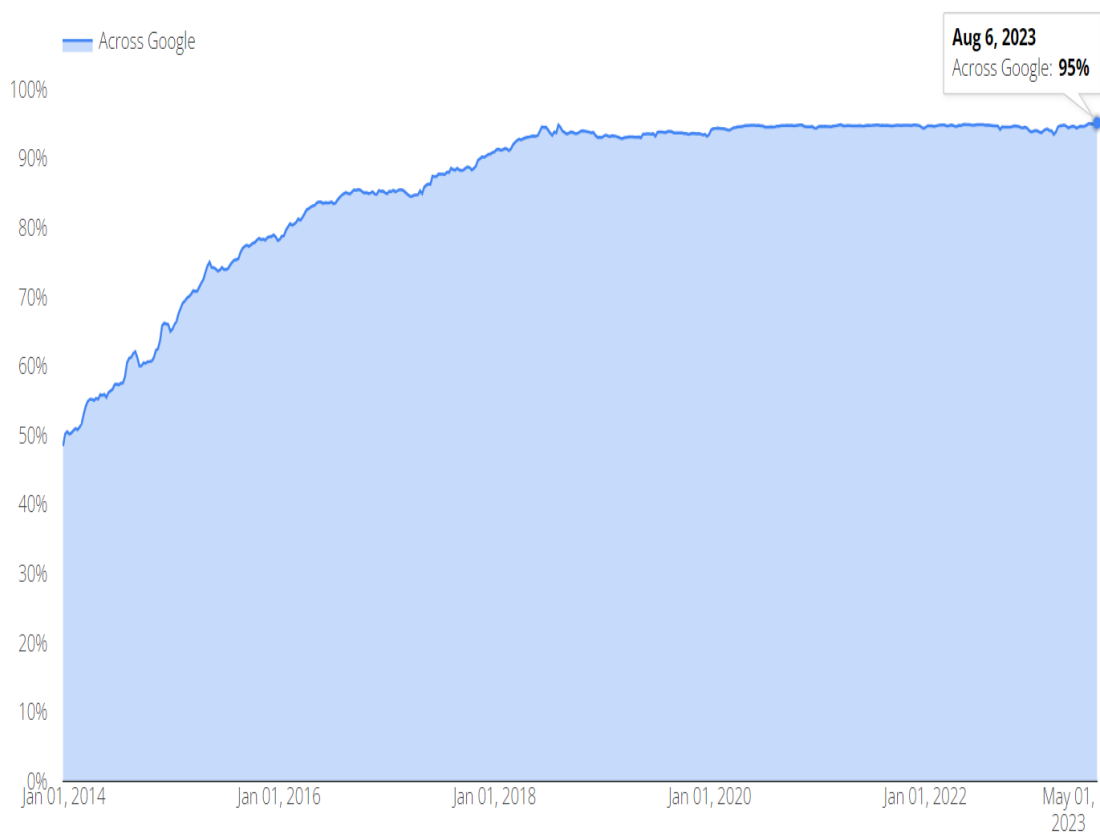


Figure 1 Encrypted traffic across Google [3]

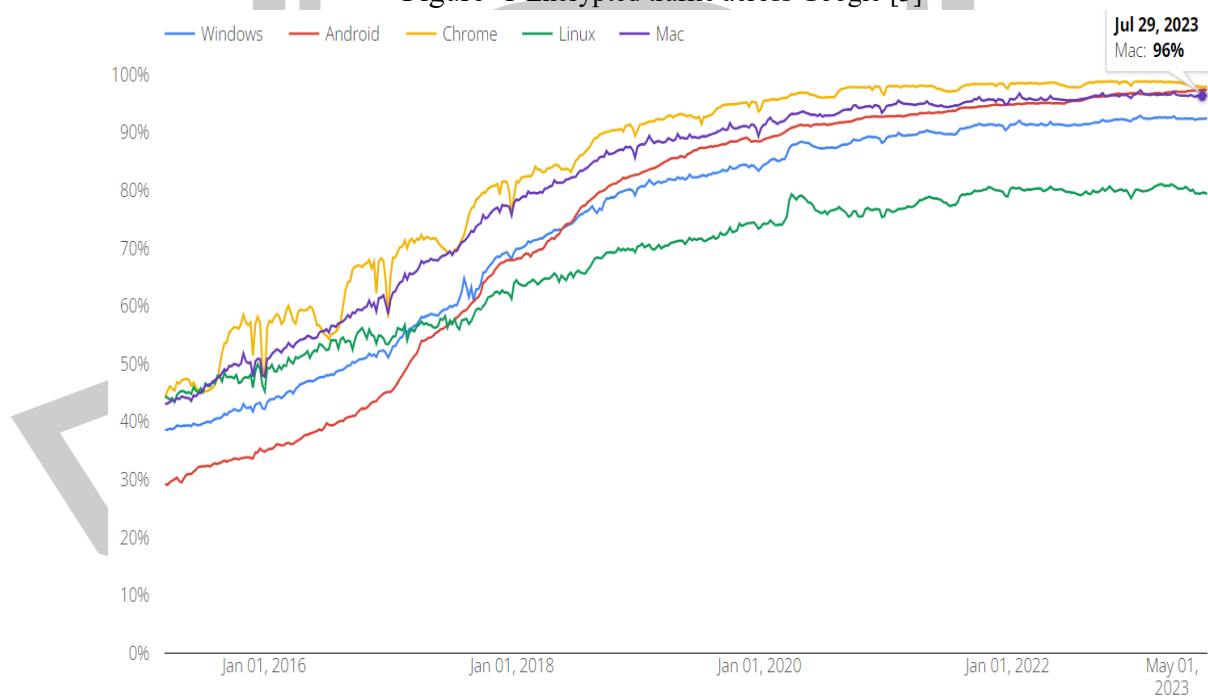


Figure 2 Percentage of pages loaded over HTTPS in Chrome by platform[3]

1.3 Impact of traffic encryption technology

While protecting data security, traffic encryption technology also brings impact and test for network service quality management and network security protection, and network and information security issues become a hotspot of increasing concern for the general public and the state.

On the one hand, in the field of quality of service management, the classification and further statistical analysis of service traffic is conducive to more effective network management by various educational organizations, corporate organizations, government organizations and Internet service providers for bandwidth abuse finding, network congestion, bandwidth planning, quality of service implementation, economic solutions, effective service pricing mechanisms, network monitoring, troubleshooting, network analysis, and lawful checking, etc. Encryption technology has made the process much harder, and how to encrypt the traffic without violating the privacy of the user's premise of the service and quality analysis of encrypted traffic has become a realistic and urgent need.

On the other hand, in the field of network security, traditionally, most malware, such as Trojans, adware, ransomware, etc., communicate with the Command and Control Server(C&C Server) to receive their commands and execute malicious behavior, and the traffic generated in this process is mostly in plaintext, which can be analyzed by Deep Packet Inspection(DPI) and other technologies to find out its IP address, domain name of the C&C Server, Uniform Resource Locator(URL) and other indicators of compromise. Deep Packet Inspection(DPI) and other technologies to resolve the IP address, C&C Server's domain name, Uniform Resource Locator (URL) and other indicators of Compromise (IoC).

Compromise (IoC) information to stop malicious behavior in time. However, nowadays, more and more malicious attackers use encryption to hide their criminal behavior, and encrypted traffic destroys the original data patterns and randomness, so traditional traffic detection techniques and devices can do nothing about the threats hidden in encrypted network attacks or malware communications.

Traditional traffic detection technologies and devices are powerless against encrypted network attacks or malware CC communications, which are hidden in encrypted traffic, and thus malicious attacks can easily evade the monitoring and

defense of firewalls and traffic management systems, and carry out more covert and damaging network attacks, such as botnets and advanced persistent threats, such as APT attacks, and so on.

1.4 Challenges in network traffic classification

Using China as an example, according to the 51st Statistical Report on Internet Development in China released by the China Internet Network Information Center (CNNIC) in March 2023, as of December 2022, the number of internet users in China reached 1.067 billion, an increase of 35.49 million compared to December 2021, with an internet penetration rate of 75.6%. In 2022, various personal internet applications in China continued to grow. Instant messaging remained the largest category with a user base of 314.1 million, growing by 97.2% compared to December 2021. Internet healthcare and online office applications saw user base increases of 64.66 million and 70.78 million, respectively, with growth rates of 21.7% and 15.1% compared to December 2021.

Table 1 2021.12-2022.12 User Size and Internet Usage Rate of Various Internet Applications

Application	2021.12 Size of users (million)	2021.12 Internet Usage Rate	2022.12 Size of users (million)	2022.12 Internet Usage Rate	Growth Rate
Instant messaging (IM)	100666	97.5%	103807	97.2%	3.1%
Web videos	97471	94.5%	103057	96.5%	5.7%
Short videos	93415	90.5%	101185	94.8%	8.3%
Online payment	90363	87.6%	91144	85.4%	0.9%
Online shopping	84210	81.6%	84529	79.2%	0.4%
Network news	77109	74.7%	78325	73.4%	1.6%
Internet music	72946	70.7%	68420	64.1%	-6.2%
Internet broadcasting	70337	68.2%	75065	70.3%	6.7%
Online game	55354	53.6%	52168	48.9%	-5.8%
Internet literature	50159	48.6%	49233	46.1%	-1.8%
Online takeaway	54416	52.7%	52116	48.8%	-4.2%
Online Office	46884	45.4%	53962	50.6%	15.1%
Online taxi	45261	43.9%	43708	40.9%	-3.4%
Online Travel Reservation	39710	38.5%	42272	39.6%	6.5%
Internet healthcare	29788	28.9%	36254	34.0%	21.7%
Online Fitness	—	—	37990	35.6%	--

In such an environment, monitoring and classifying traffic becomes increasingly crucial to enhance network service quality, strengthen network management, and ensure network security. For instance, to better safeguard the security of cyberspace, traffic can be categorized into normal and malicious traffic, enabling the detection of network traffic anomalies to ensure national security. To achieve improved network service quality, traffic can be classified based on applications, such as chat applications, file transfers, and so on. Following this, different applications can be allocated different bandwidths based on their priorities, thereby enhancing network service quality.

1.5 Problems to be solved by this thesis

Currently network traffic encryption analysis is facing a huge challenge, there are three main points:

1)The amount of encrypted traffic in the Internet is increasing dramatically as people have become more and more concerned about their privacy in recent years, and more applications have begun to use encryption to transmit data. Google has also been emphasizing the use of SSL encryption across the network to protect users from data theft. Because if the data is encrypted using HTTPS protocol, then even if they are intercepted, hackers won't be able to decipher them. Since 2014, Google has been using HTTPS as a reference for ranking its search results, which has led to more and more websites using encrypted connections. However, encrypted traffic poses a huge challenge to traditional traffic inspection methods, such as Deep Packet Inspection (DPI), which only applies to plaintext traffic, but not encrypted traffic. Therefore, more and more models using machine learning or deep learning to solve encrypted traffic classification have been proposed in recent years.

2)With the rapid development of short videos, webcasting and other applications, the difference in the amount of traffic between different applications has become huge, which brings about the problem of unbalanced network traffic samples. This data imbalance will make the final encrypted traffic classification model have a very serious bias, and the final result will be biased towards the traffic with a large number, which leads to no scalability of the model.

3)Machine learning has been more and more widely applied to network traffic classification problems in recent years. But as the models become more complex, the

transparency of the models is decreasing. Although the model can give a high accuracy rate, it cannot give the reason why he makes this judgment. Cybersecurity personnel usually make judgments based on the system's prompts, which makes it difficult for them to make decisions if the system only gives predictions without giving reasons.

In this paper, further research is carried out in the areas of business type identification and maliciousness detection of encrypted traffic, with a view to better automatically extracting the deep features of encrypted traffic and improving the recognition capability of the model.

This article intends to employ an enhanced random forest algorithm for the domain of encrypted traffic recognition and classification. It aims to identify and categorize encrypted traffic while preserving the privacy of the encrypted traffic data. In comparison to conventional traffic recognition and classification techniques, deep learning-based encrypted traffic recognition and classification technology can autonomously extract features from extensive datasets, diminishing the requirement for manual intervention while augmenting recognition and classification efficiency and efficacy. Additionally, it can be rapidly updated and optimized by retraining deep learning classification models, making it more adaptable to intricate network environments. This technology also holds applicability in other domains, such as the Internet of Things and blockchain. Thus, the application of deep learning to encrypted traffic recognition and classification carries substantial significance.

II Research Objectives

The research process divides the work in the field of encrypted traffic analysis into three phases: encrypted traffic detection, encrypted traffic classification and encrypted traffic identification according to the target requirements, and summarizes the process of encrypted traffic analysis throughout the whole life-cycle as shown in Figure.3 below. The research in this thesis focuses on encrypted traffic identification, network anomaly detection, and machine learning algorithms.

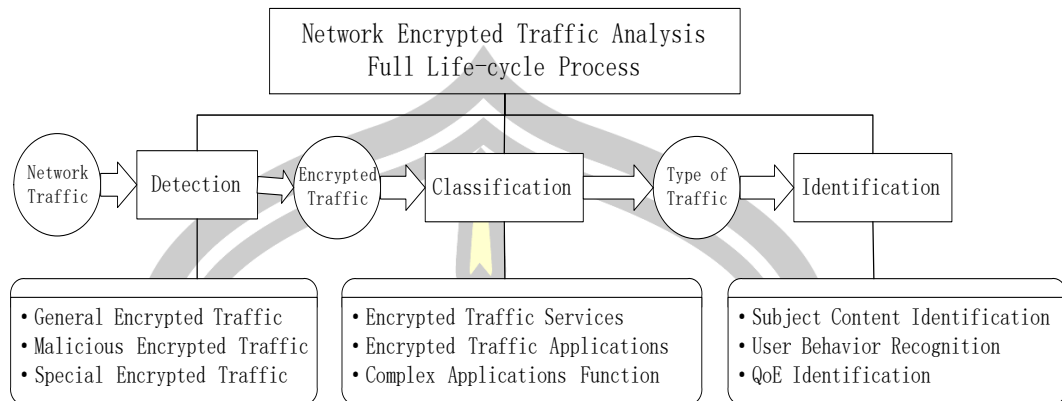


Figure 3 Full life cycle process of encrypted traffic analysis

2.1 Encrypted traffic identification

Before introducing each individual component in Figure.4, I will provide an overview of the encryption protocols and anonymity mechanisms currently employed to ensure secure communication. Without traffic encryption, sensitive information transmitted in plain text would be vulnerable to interception by malicious eavesdroppers with access to network interfaces. Furthermore, unencrypted communication can facilitate deliberate attackers in injecting false information, rendering communication unreliable. Hence, the implementation of traffic encryption mechanisms serves as a means to ensure the security of communication.

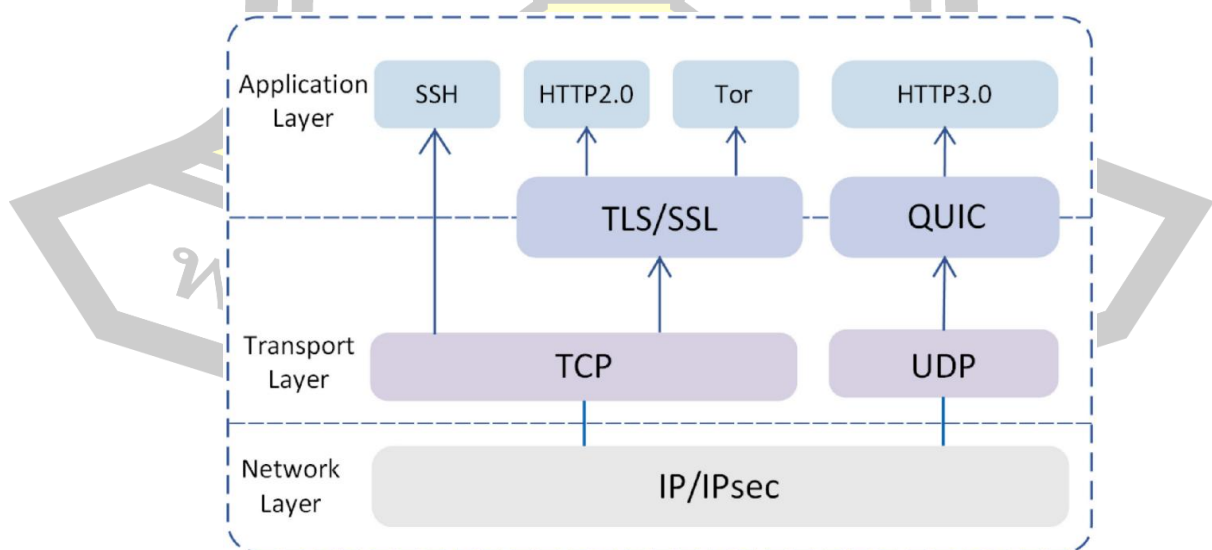


Figure 4 Mechanisms in the network model

2.2 Network anomaly detection

Network anomaly detection is an important research direction in the field of network security, aiming at identifying and analyzing anomalous behaviors and attacks in the network in order to protect the security and stability of the network. Network anomalies may include malicious attacks, abnormal traffic, device failures, etc. These abnormal behaviors may lead to serious consequences such as data leakage, service interruption, and system crash. Therefore, network anomaly detection is of great significance to protect the normal operation of the network and data security.

In the field of encrypted traffic analysis, network anomaly traffic detection refers to the identification and detection of encrypted data traffic exhibiting abnormal behavior during encrypted communications. Because encrypted communication conceals the content of data, traditional traffic analysis methods cannot directly access encrypted data, necessitating the use of other techniques to identify potential anomalies in the traffic. These anomalous traffic patterns may involve malicious activities, intrusion attempts, data breaches, and as such, anomaly traffic detection is crucial for network security.

Key areas of research focus within network anomaly traffic detection include the following:

a. Feature extraction and analysis: Although the data content in encrypted traffic is encrypted, some metadata and statistical information, such as packet size, transmission frequency, connection time, etc., can still be extracted from the traffic. These features can be used to identify anomalous traffic. For example, anomalous traffic may exhibit a different distribution of features than normal traffic.

b. Behavioral Pattern Analysis: Anomalous traffic often exhibits behavior patterns distinct from normal traffic. Researchers can monitor whether the traffic's behavior aligns with established baseline behavior models. For instance, if a user's encrypted traffic suddenly becomes unusually active, it may indicate that their account is potentially under attack.

c. Traffic Clustering: Cluster analysis of encrypted traffic to group similar traffic into clusters can help detect anomalous behavior. If the traffic in a cluster is significantly different from other clusters, it may mean that there is anomalous traffic.

d. Timing Analysis: The pattern of encrypted traffic over time can also be used to

identify anomalies. Anomalous traffic may result in unusual time intervals, frequency variations, etc.

e. Multimodal data analysis: Combining other types of data, such as system logs and user behavior, can improve the accuracy of anomalous traffic detection. This combined analysis can provide more contextual information.

f. Zero-Day Attack Detection: Zero-day attacks are new types of attacks that defenses have not yet been able to identify and stop. Anomalous traffic detection can attempt to detect unknown zero-day attacks by finding patterns of behavior that do not match known attacks.

2.3 Machine learning algorithm

Machine learning has emerged as a potent tool in encrypted traffic analysis, allowing for the extraction of valuable insights without the need to access payload data[4]. Machine learning methods can be broadly categorized into two groups: traditional machine learning and deep learning. In the context of encrypted traffic analysis, a major distinction between the two lies in their approach. Traditional machine learning requires manually crafted features as input, which involves intricate feature engineering processes. On the other hand, deep learning enables end-to-end traffic classification. The process of feature extraction in traditional machine learning often necessitates substantial prior knowledge about the task, including identifying the most significant features for distinguishing target traffic from background traffic.

Traditional machine learning techniques can be considered as shallow learning methods, and they have certain limitations in achieving high accuracy in classification tasks. Specifically, their capacity to generalize effectively is constrained when handling intricate problems. In contrast, deep learning methods have been introduced to overcome these limitations. Leveraging the feature learning capability inherent in deep learning, it allows for the utilization of raw traffic data as input.

In response to the fresh challenges arising from traffic encryption, machine learning techniques have been harnessed to extract valuable insights from encrypted traffic without requiring access to packet payloads. Machine learning-driven encrypted traffic analysis relies on statistical or behavioral features of encrypted traffic, which are less impacted by the use of encryption protocols. These methods exhibit significant capabilities in handling vast amounts of data, making them well-suited for constructing

classification models without the need for intricate programming. Furthermore, deep learning, as a subset of machine learning, eliminates the manual feature extraction process, rendering it an appealing approach for encrypted traffic analysis, particularly when dealing with dynamically shifting traffic patterns.

III Benefits of Research

3.1 Staying Current with Research Trends

The research on network traffic holds significant importance and research value in safeguarding user privacy and fortifying network security. Therefore, the classification and identification of encrypted traffic has also become a difficult and hot spot for researchers in recent years. According to the research objectives, the researchers of encrypted traffic can be divided into malicious traffic identification, identification of encrypted and non-encrypted, protocol-specific identification, and fine-grained behavior identification.

As mentioned above, fine-grained identification of encrypted traffic is of great significance for precise control of network behavior. This thesis is dedicated to the fine-grained identification of SSL/TLS encryption applications from the real needs, and selects a specific application to carry out fine-grained behavioral identification based on the application, in combination with the actual needs of the source project.

3.2 Monitoring Network Behavior

Traffic analysis involves the continuous monitoring of network activities to detect specific patterns and extract valuable insights from network traffic. It finds applications in various domains, including network asset probing and anomaly detection. However, the advent of network traffic encryption has introduced considerable challenges to traffic analysis. With the packet payload concealed, conventional traffic analysis methods reliant on extracting valuable information from plaintext payload may lose effectiveness.

Encryption serves as a critical tool for safeguarding data during transmission, rendering it inaccessible to unauthorized individuals. By implementing robust encryption protocols and methodologies, researchers and cybersecurity professionals ensure that sensitive information remains protected, thus bolstering network security.

This security measure is particularly significant in an era where cyber threats are continually evolving.

Moreover, the study of network traffic encryption encompasses the development and implementation of encryption techniques in various network protocols and communication channels. This ensures that data remains secure across diverse technological landscapes, from web traffic using HTTPS to secure remote access via VPNs. Understanding how encryption integrates with these technologies is essential for effective monitoring of network behavior.

In response, machine learning has emerged as a formidable tool for extracting informative features without needing access to payload data, making it a prevalent choice in encrypted traffic analysis.

3.3 Enhancing Network Security

Leveraging machine learning for the analysis of encrypted network traffic proves instrumental in identifying malicious traffic. Encrypted traffic analysis furnishes a valuable tool for gaining deeper insights into potential information leakage within network connections, thereby enabling the development and implementation of effective defense mechanisms.

Network security is of paramount importance in today's digital landscape, as the reliance on the internet for communication, commerce, and data exchange continues to grow. Encryption is a fundamental technique used to secure data in transit, rendering it indecipherable to unauthorized entities. This practice is particularly crucial in the context of network traffic, where vast volumes of data traverse networks daily.

By delving into the realm of network traffic encryption, researchers and cybersecurity experts aim to better understand the intricacies of securing data during transmission. This involves exploring encryption protocols, algorithms, and methodologies that can safeguard information from prying eyes and malicious actors. The ultimate goal is to ensure that sensitive data, whether it's personal information, financial transactions, or critical business communications, remains confidential and protected.

Chapter 2 Literature Review

I Categories of Machine Learning Algorithm

Techniques for analyzing encrypted traffic based on machine learning algorithms have achieved significant results in recent years. These techniques are able to efficiently identify different protocols, applications and behavioral patterns in encrypted traffic, and even detect unknown zero-day attacks through deep learning, feature extraction and pattern recognition. These research results not only expand the boundaries of network security, but also provide powerful tools for practical applications, further enhancing the monitoring and protection of encrypted traffic.

Machine learning algorithms can be categorized into several broad categories based on their learning style and functionality. From the aspect of whether labels are required, machine learning can be classified as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning and so on[5], the main categories of machine learning algorithms as shown in Table.2:

Table 2 summary of existing machine learning algorithms

Category	Description	Related Works	Examples of Algorithms
Supervised Learning	Predicts discrete class labels or continuous values.	[6-8]	Logistic Regression, Decision Trees, SVM, Neural Networks, Linear Regression, Ridge Regression, etc.
Unsupervised Learning	Identifies patterns or groups in data without labeled outputs.	[9-10]	K-Means Clustering, Hierarchical Clustering, PCA, t-SNE, DBSCAN, etc.
Semi-Supervised Learning	Combines labeled and unlabeled data for training.	[11-13]	Self-training, Label Propagation, etc.
Reinforcement Learning	Agents learn to maximize a reward signal by taking actions.	[14-15]	Q-Learning, DQN, Policy Gradient methods, etc.
Self-Supervised Learning	Generates supervised-like tasks from unlabeled data.	[16-20]	Word2Vec, Contrastive Learning, etc.
Generative Models	Generates new data instances or learns latent representations.	[21-22]	GANs (Generative Adversarial Networks), VAEs (Variational Autoencoders)

Table 3 summary of existing machine learning algorithms

Category	Description	Related Works	Examples of Algorithms
Ensemble Learning	Combines predictions from multiple models for better results.	[23-24]	Random Forests, Gradient Boosting, AdaBoost, etc.
Anomaly Detection	Identifies unusual or rare data points.	[25-27]	One-class SVM, Isolation Forest, etc.
Instance-Based Learning	Makes predictions based on similarity to stored instances.	[28-29]	k-Nearest Neighbors (k-NN), etc.
Neural Networks and Deep Learning	Deep architectures with multiple layers.	[30-32]	CNNs, RNNs, Transformers (BERT, GPT), etc.
Natural Language Processing (NLP)	Specialized for language data analysis.	[33-36]	Tokenization, Word Embeddings (Word2Vec, GloVe), LSTM, BERT, GPT
Time Series Analysis	Analyzes data that changes over time.	[37-38]	ARIMA (Autoregressive Integrated Moving Average), LSTM, etc.

II Traditional Machine Learning Algorithm

2.1 Algorithm Introduction

Common instances of traditional machine learning techniques encompass Naïve Bayes, Markov Models, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and clustering. A concise overview of each is presented below.

The Naïve Bayes method, as described in reference [39], is a classification approach grounded in Bayes' theorem and the assumption of conditional independence among features. It earns its "naïve" moniker because it presupposes that the features are mutually independent. This technique allows for probabilistic inferences in the context of classification and decision-making problems.

The Markov Model method, elaborated in reference [40], is a stochastic model that provides probabilities for transitions between random states. It comprises states, state transition probabilities, and an initial probability distribution. In the case of a first-order Markov model, it assumes that the present state is solely dependent on the preceding state and remains unrelated to earlier states.

The k-Nearest Neighbor (k-NN) method, as detailed in reference [41], is a widely employed supervised learning technique. When provided with a target sample, it identifies the closest k samples and then makes predictions based on the characteristics of these neighboring samples. Calculations of proximity can involve various methods such as Euclidean distance and Manhattan distance, among others.

Support Vector Machine (SVM), as outlined in reference [42], is a versatile supervised algorithm suitable for both classification and regression tasks. To tackle intricate classification challenges, SVM employs a kernel function that transforms features from a lower-dimensional space to a higher-dimensional one. Notable kernel functions encompass the linear kernel, polynomial kernel, Gaussian kernel, and Sigmoid kernel.

The Decision Tree (DT), detailed in reference [43], consists of directed edges and nodes, comprising the root node, internal nodes, and leaf nodes. Each path from the root node to a leaf node corresponds to a decision sequence. Common algorithms employed for creating Decision Trees include ID3, C4.5, and CART.

Random Forest (RF), as mentioned in reference [44], is essentially an ensemble of conventional Decision Trees (DTs). Addressing the susceptibility of DTs to overfitting, RF incorporates a voting mechanism based on multiple DTs to enhance performance. In practice, each tree within the RF ensemble is trained using a sub-dataset sampled from the original dataset, and the final result is determined through a collective voting process involving all the individual DTs.

The k-Means algorithm, elaborated in reference [45], is a frequently applied clustering technique. It involves assigning each object to the cluster nearest to it, determined by distances from cluster centers. Subsequently, the algorithm recalculates the cluster centers and repeats this process iteratively until convergence is achieved. Ultimately, all objects are organized into k clusters based on their proximity to the cluster centers.

2.2 Application in the encrypted traffic analysis

Akhil. et al.[46] proposed Intrusion Detection System IDS for monitoring and detecting intrusions in network traffic data or computer systems using Gaussian Plain Bayesian Classification, IDS is designed to monitor and detect suspicious operations or intrusions in computer systems or network traffic data. IDS is evaluated using the

Kyoto dataset, which shows higher accuracy, higher intrusion detection rate and lower false alarm rates compared to the existing methods.

Yi. et al.[47] proposed a new encrypted traffic classification method that introduces Markov Model encrypted traffic classification method on Gaussian mixture model constrained clustering (MET-GCC), which takes traffic classification into account in a simplified communication process, solves the limitations of traditional analysis methods, applies Gaussian mixture model (GMM) to analyze the length of the message, and establishes a model for solving the application fingerprint duplication problem is modeled. The method uses constrained clustering to classify fingerprints of the same application with similar lengths into as few clusters as possible, which improves the convergence speed and clustering effect. Experimental results show that the proposed method outperforms other encrypted traffic categorization methods with significant improvements in TPR, FPR, Precision and Recall.

Chen. et al.[48] presents a novel secure k-NN scheme with multiple keys for secure k-nn queries over encrypted cloud data using multiple keys, which solves the problem of querying users who do not fully trust or know the key of the data owner. The scheme is constructed using a Distributed Two-Live-Gate Public Key Cryptosystem (DT-PKC) and a set of secure two-party computation protocols to ensure data confidentiality, query privacy, and support offline data owners.

RAMRAJ S and Usha G[49] proposed a method that combines SVM with deep learning autoencoder for binary classification of network traffic for WhatsApp, one of the globally used applications, for an endless stream of social applications. This vector-based classifier also improves the performance in the binary classification task and also helps in evaluating the application's sender-receiver system that specifies the attributes of the network trace . By analyzing the behavior of network traces, we can scrutinize the type and nature of traffic for easy network maintenance.

Dai. et al.[50] addresses the malicious behavior that users can hide their IP and identity information through encryption algorithms during network communication and carry out various illegal activities and transactions in the network, establishes a decision tree model based on C4.5 and random forest, compares the accuracy of the two decision tree algorithms in traffic categorization, and identifies and applies them

to the special traffic that affects the network security, so as to find out the network Anomalies.

Ren. et al.[51] proposed a random forest set based on intuitionistic fuzzy theory, which is a generalized algorithm commonly used for classification under complex conditions. Although it has been widely adopted, its combination with diverse fuzzy theories is still worth exploring. In this paper, we propose the Intuitionistic Fuzzy Random Forest (IFRF), a novel random forest collection consisting of Intuitionistic Fuzzy Decision Trees (IFDTs). This forest tree uses intuitionistic fuzzy information gain to select features and takes into account indecision in information transfer. The proposed method offers the randomness of bootstrap sampling and feature selection, the flexibility of fuzzy logic and fuzzy sets, and the robustness of a multi-classifier system. Extensive experiments have demonstrated the competitive and superior performance of IFRF compared to other state-of-the-art fuzzy and ensemble algorithms. IFDT is more suitable for ensemble learning with excellent classification accuracy.

Gupta[52] proposed Naive Bayes approach enhanced with Deep Reinforcement Learning (DRL) and Fuzzy K-mean clustering to classify VPN and non-VPN data. The proposed method is validated on the publicly available UNB-CIC VPN non-VPN dataset and the results show that its accuracy is comparable to other state-of-the-art machine learning algorithms used for traffic characterization. The proposed approach for traffic characterization for traffic classes (e.g., FTP and P2P) and application identification (e.g., Netflix and Amazon Prime) analyzes the efficiency and accuracy of traffic class detection, as well as the ability to distinguish between VPN and non-VPN network traffic. Experimental results show that using "NB+DRL+fuzzy k-means clustering", k-means based class delineation converges to a local minimum based on the dissimilarity measure of the given packet features, which further improves fuzzy k-means clustering and reduces the computational cost of encrypted network traffic classification. For traffic characterization and application identification, the F1 scores of the test set reach 0.973 and 0.965, respectively.

III Deep Learning Algorithm

3.1 Algorithm Introduction

Traditional machine learning technologies are essentially shallow learning, which has some limitations in terms of accuracy for classification tasks. In particular, their generalization abilities are restricted when dealing with complex problems. Hence, deep learning is introduced to address these limitations. This paper mainly introduce three typical deep learning methods that are widely employed in encrypted traffic analysis.

Convolutional Neural Network (CNN) [53] is a kind of neural network with convolution operation added, which can weaken local features and enhance generalization. The special thing about CNN is that this method has the convolutional layer, which uses the convolution operation to filter the input features. Normally, the convolutional layer and the pooling layer appear alternately. After the combination of several sets of convolutional and pooling layers, a fully connected layer will be added at the end.

Graph Neural Network (GNN) [54] is a deep learning method that processes the data represented in the graph domain. As an effective graph analysis method, GNN is widely used in social network and physics system analysis, which leverage graphs as the denotation of the relationships underlying data.

Long Short-Term Memory (LSTM) [55] is a special kind of Recurrent Neural Network (RNN) used to process sequences, which has achieved excellent performances in the field of natural language processing. LSTM has the ability to selectively remember the previous output results and superimpose them on different positions of the current cell to achieve the purpose of data analysis.

3.2 Application in the Encrypted Traffic Analysis

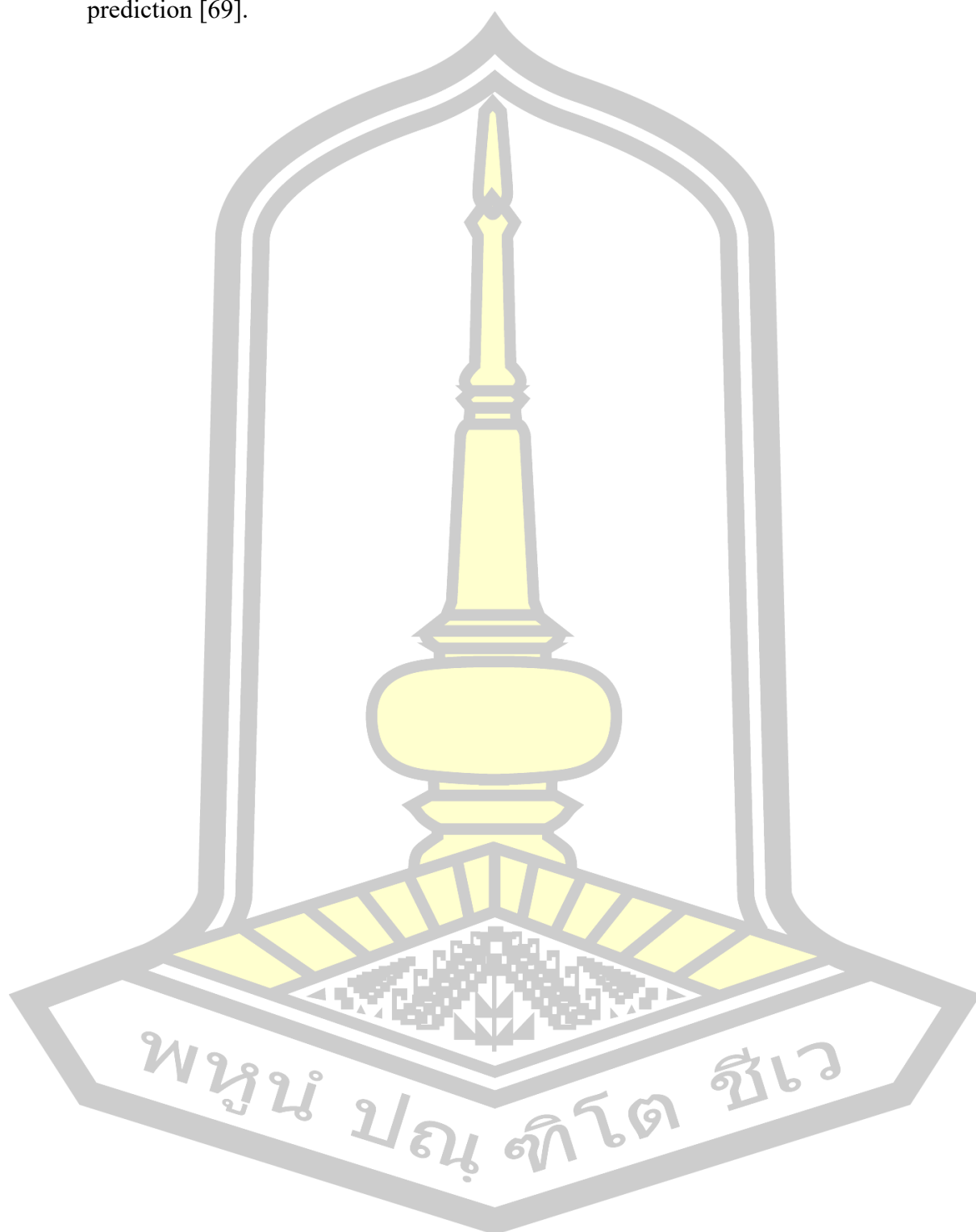
CNN (Convolutional Neural Network) have been widely used for encrypted traffic classification. These networks have shown high accuracy in identifying different types of encrypted traffic, including known and unknown types. CNNs focus on specific parts of the traffic flow, such as the initial parts of traces, rather than longer windows of continuous uploads or downloads [56]. They also exhibit transfer-learning capabilities, allowing them to identify new websites with fewer data [57]. CNNs outperform RNNs (Recurrent Neural Networks) in traffic fingerprinting due to their

resilience to random shifts in data caused by varying network conditions [58]. Additionally, CNNs can be combined with other techniques, such as SAE (Stacked Autoencoder) and LSTM (Long Short-Term Memory), to further enhance the accuracy of encrypted traffic classification [59] [60]. The proposed CLD-Net model, which integrates CNN and LSTM, has shown high accuracy in distinguishing different classes of network encrypted traffic. The ETCC method, using CNN, achieves high accuracy in classifying encrypted applications based on the encryption protocols used.

GNN(Graph Neural Network) have been used in various studies related to traffic analysis and management. One study proposed a novel system for encrypted traffic classification using the eXtreme gradient boosting algorithm [61]. Another study designed a Simplified Spatio-temporal Traffic forecasting GNN (SST-GNN) to capture spatial relationships and temporal dynamics in traffic data [62]. Additionally, a deep reinforcement learning (DRL) algorithm integrated with GNN was proposed for intelligent traffic control systems, which showed effectiveness in relieving traffic congestion [63]. Furthermore, GNNs have been explored for data-driven routing in intradomain traffic engineering, specifically in minimizing link congestion in networks using knowledge of historic traffic flows [64]. These studies demonstrate the potential of GNNs in addressing various challenges in traffic analysis and management.

LSTM-based models are used for various applications such as traffic flow prediction, traffic situation characterization, and encrypted malicious traffic detection. LSTM networks combined with Long Short-Term Memory and Support Vector Machine achieve high prediction accuracy for car following behavior and lane changing behavior in traffic flow modeling [65]. Machine learning approaches are crucial for detecting encrypted malicious traffic, especially in the post-COVID-19 environment, where traditional security solutions are ineffective. These approaches rely on datasets from different sources to train models and improve performance [66]. LSTM networks equipped with an attention mechanism outperform other models in short-term traffic flow forecasting by focusing on critical information and capturing long-term dependencies [67]. Deep reinforcement learning-based LSTM models accurately predict traffic flow data, while fuzzy comprehensive evaluation models effectively represent the road traffic situation [68]. Spatial attention mechanisms and

sequence to sequence models capture spatio-temporal dynamics for traffic flow prediction [69].



Chapter 3 Research Methodology

I Relevant Concepts

1.1 Generation of encrypted traffic

Network traffic is the data moving across the network, which is usually generated from user's network activities or active pushes from application servers and is encapsulated in packets for transmission. To characterize network traffic comprehensively, it is necessary to acknowledge what might shape the network data during the process of generation and transmission. By our experience, we summarize this process in this Figure.5, and describe some basic components as follows.

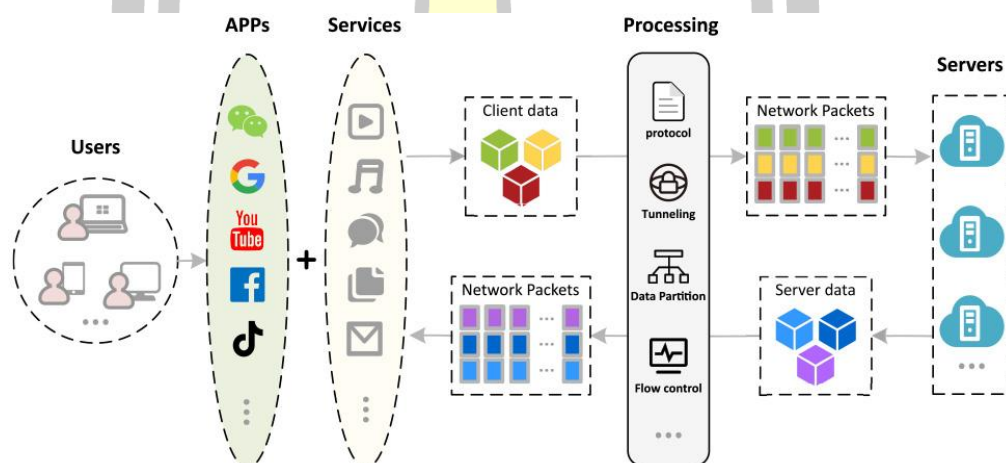


Figure 5 Transmission process of the encrypted traffic

Encrypted traffic is the process of converting plaintext data into ciphertext by using encryption algorithms to protect the privacy and security of network communication content. This includes steps such as selecting appropriate encryption algorithms and key management, establishing secure communication channels, data authentication, and ongoing key management to ensure the confidentiality, integrity, and trustworthiness of communications. Encrypted traffic is a core component of network security, preventing unauthorized access, eavesdropping, and tampering and securing communications.

In network communication, encrypted traffic is a crucial security measure used to safeguard the privacy and integrity of information. This process begins with the

selection of suitable encryption algorithms, which transform the original readable data (plaintext) into an unreadable character sequence (ciphertext) through mathematical transformations. This ensures that even if intercepted during data transmission, the content cannot be easily understood. Encrypted traffic also involves the generation and management of keys, which are essential for encrypting and decrypting data, and their security is of paramount importance.

The establishment of a secure communication channel is also an indispensable step to ensure that both communicating parties can exchange information securely. This is typically achieved through protocols such as SSL/TLS or IPsec (Internet Protocol Security), which provide comprehensive security measures including data encryption, authentication, and data integrity checks.

Furthermore, data authentication and continuous key management are critical components of encrypted traffic. The former ensures the integrity and trustworthiness of data sources, while the latter guarantees the security of long-term communication. In summary, encrypted traffic serves as the cornerstone of network security, defending against unauthorized access and eavesdropping while ensuring the privacy and security of communications.

1.2 Principles of Traffic Encryption

The principle of traffic encryption during network communication involves the use of encryption algorithms and keys to transform data so that it remains confidential and secure during transmission. The principle of traffic encryption in network communication involves the following three key aspects:

- 1) Selection and Use of Encryption Algorithms and Keys: The first principle is to select appropriate encryption algorithms and generate keys. An encryption algorithm is a collection of mathematical functions and rules that change plaintext data into ciphertext, and a key is used to perform encryption and decryption. Suitable algorithms and generation of strong ciphers are essential for data protection.

- 2) Establishment of a secure communication channel: Establishing a secure communication channel is at the heart of traffic encryption. This includes key negotiation, authentication and the use of secure protocols to ensure the confidentiality and integrity of communications. The communicating parties need to ensure that only they can access and decrypt the content of the communication and no one else can.

3) Encryption and decryption of data: Data is encrypted during transmission, the sender uses an encryption algorithm and key to convert the plaintext data into ciphertext, and the receiver uses the appropriate key and algorithm to decrypt the data and restore the original plaintext data. This ensures that even if the data is intercepted during transmission, its content cannot be understood.

In summary, traffic encryption ensures the confidentiality and integrity of network communications through encryption algorithms, key management, the establishment of secure communication channels, and the encryption and decryption of data. It is a core principle of network security and is used to secure communications against unauthorized access, eavesdropping and tampering.

1.3 Commonly used encryption protocols

Traffic encryption protocols in network communications play a key role in protecting the confidentiality of communication content, ensuring data integrity, enabling authentication of both parties, establishing secure communication channels, and protecting user privacy[70]. These protocols prevent unauthorized access and data tampering by using encryption algorithms and key management mechanisms to convert plaintext data into ciphertext. They also reduce the risk of impersonation and man-in-the-middle attacks by ensuring data integrity and source trustworthiness through digital signatures and authentication. These protocols not only play a role in sensitive exchanges over the Internet, but also provide security in a variety of scenarios such as remote access, file transfers, and virtual private networks, helping to maintain the confidentiality and security of communications.

Traffic encryption protocols commonly used for network communications include the following:

1) TLS/SSL (Transport Layer Security/Secure Sockets Layer): TLS and SSL are among the most common traffic encryption protocols used to protect communications for websites, e-mail, instant messaging and other applications. They provide confidentiality, integrity and authentication, securing data through the use of asymmetric and symmetric key encryption algorithms.

2) IPsec (Internet Protocol Security): IPsec is a set of protocols for encrypting data at the network layer, typically used for virtual private network (VPN) connections.

It encrypts the entire communication session, including data messages and network packets, to ensure end-to-end confidentiality and integrity.

3) SSH (Secure Shell Protocol): SSH is used to securely access and control servers remotely. It provides encryption and authentication of end-to-end communications, as well as security for file transfers.

4) HTTPS (Hypertext Transfer Protocol Secure): HTTPS is a secure version of HTTP that encrypts Web pages and applications via the TLS/SSL protocol. This ensures the privacy and security of sensitive information while browsing websites.

5) OpenVPN: OpenVPN is an open source VPN solution that uses the SSL/TLS protocol to encrypt communications. It is widely used to establish secure remote access connections, including telecommuting and remote server management.

6) S/MIME (Secure Multipurpose Internet Mail Extensions): S/MIME is an encryption and digital signature protocol for e-mail that ensures the confidentiality and integrity of e-mail messages and verifies the identity of the sender.

These traffic encryption protocols are widely used in different communication scenarios and provide a variety of ways to secure communication content. Their selection depends on communication needs, security requirements, and application requirements. Using a combination of these protocols can help ensure privacy and security of network communications.

II Encrypted Traffic Analysis

2.1 Traffic Collection and Preprocessing

Traffic data collection is an important step in the field of encrypted traffic analysis, as a data set is the foundation of experiments. Therefore, data collection often acts as the starting point of encrypted traffic analysis.

Encrypted traffic collection is a key network security practice designed to capture and analyze encrypted communications data transmitted over a network. The process typically involves capturing encrypted packets from network devices or sensors, then using legitimate means to obtain the keys needed for decryption and performing the necessary preprocessing steps such as filtering, protocol parsing, and session reconstruction. The ultimate goal is to store and analyze this data in order to monitor network security threats, detect anomalous behavior, support forensic efforts,

and enhance network security. Encrypted traffic collection plays an important role in network security operations, incident response, and threat intelligence analysis, helping organizations maintain the security and integrity of their networks. Traffic can be captured at various nodes in a network, such as switches, routers, and gateways. Traffic can be captured at various nodes in a network, such as switches, routers, and gateways.

Encrypted traffic preprocessing is a necessary part of network security and data analysis, designed to provide initial processing of captured encrypted communication data for subsequent analysis. This process includes steps such as data screening, protocol identification, session reconstruction, timestamp addition, and data de-duplication to prepare the data for deeper analysis. Through preprocessing, we are able to better understand the context of the communication, remove noise, extract valuable information, and more effectively detect threats, analyze anomalous behaviors, and support security decisions in subsequent phases.

2.2 Traffic Representation

Effective traffic analysis relies on the accurate representation of traffic, which may vary across different application scenarios. A high-quality traffic representation is instrumental in enhancing effectiveness while minimizing overhead. This representation serves as the input for classifiers and can encompass various features, including packet-based and statistical features, which are utilized by traditional machine learning classifiers.

Traffic Representation Level

Traffic Representation Level refers to the degree or granularity at which network traffic data is represented or analyzed in a network security or data analysis context. It encompasses how detailed or abstract the data representation is, which can impact the depth and specificity of insights gained from the analysis. The common representation levels are as follows:

a. **Packet-Level Representation:** Analyzing network traffic at the packet level involves examining individual data packets, which provides very granular information. This level of representation is useful for tasks like intrusion detection, where specific patterns or anomalies in packet contents need to be identified.

b. **Flow-Level Representation:** Flow-level analysis aggregates packets into flows or connections between devices, focusing on the source and destination,

protocols used, and data volume exchanged. It offers a more abstract view of network traffic, suitable for tasks like network performance monitoring.

c. Session-Level Representation: Going a step further, session-level analysis groups flows into sessions or higher-level interactions. This is particularly useful for understanding user behavior or application-level patterns.

d. Application-Level Representation: At the highest level of abstraction, traffic can be represented based on the specific applications or services being used, like web browsing, email, or file sharing. This is often seen in firewall rules or application-level gateway configurations.

The choice of representation level depends on the objectives of network analysis or security monitoring. More detailed levels, such as packet-level analysis, provide fine-grained insights but generate a large volume of data and can be resource-intensive. On the other hand, higher-level representations are more abstract, efficient, and can offer insights at a broader scope. The level chosen should align with the specific use case and goals of the analysis.

2) The Forms of Representations

Raw representations of traffic such as sequences or graphs may be used as the input of deep learning models. Different forms of traffic representations for encrypted traffic analysis are introduced as follows.

a. Packet-Based Features: The basic packet-based features are the information of packet header captured directly. Generally, the five-tuple of IP packets, including source and destination IP addresses, source and destination ports and protocol types, are the most commonly leveraged

packet-based features. Moreover, the Time-To-Live of the IP header, the initial window size of the TCP header can also be extracted as features. These are the simplest features obtained directly from the packet.

b. Statistical Features: Statistical features are defined as features extracted from a group of packets through abstraction or computation. The statistic values include expectations, deviations, average, minimum, maximum, medians, accumulation, etc. Combining these statistics with traffic attributes, such as packet size, inter-arrival time, packet count, results in a large number of statistical features,

e.g., average of cumulative packet sizes, quartile of inter-arrival time and burst packet count.

c. Raw Traffic Representation: In addition to selecting network traffic features manually, researchers can also use frameworks or models like deep learning algorithms to extract features automatically. Artificial feature extraction is based on experience, however, the raw representation transfers the offload of feature selection to the models. Several studies encode traffic into sequences, graphs or images for classification. For example, the traffic can be abstracted into graphs. The structure of the graph can be regarded as a set of interconnected nodes, each of which is treated as a packet. This is similar to the graph definition in data structure and can be augmented with directions, weights, etc.

2.3 Traffic Analysis

The investigated encrypted traffic analysis methods are summarized into two categories, namely traditional machine learning and deep learning. In encrypted traffic analysis, the major difference between the two categories is that the former requires manually-crafted features as input, which needs complex feature engineering processes, while the latter enables end-to-end traffic classification. The process of feature extraction requires much prior knowledge about the task, e.g., the most contributing features for distinguishing target traffic from background traffic.

Traditional machine learning technologies are essentially shallow learning, which has some limitations in the accuracy for classification tasks. In particular, their generalization abilities are restricted when dealing with complex problems. In contrast, deep learning methods are introduced to address these limitations. Benefiting from the feature learning ability of deep learning, it enables raw traffic representation as input.

Moreover, network traffic can also be represented by a graph structure, which digs out the hidden information and is combined with GNN for traffic classification. In addition, as there are correlations between the packets at the current moment and those in the previous period of time, LSTM can be used to capture timing features of packet sequences.

The rapid development of machine learning, especially deep learning, provides a large number of analytical methods for encrypted traffic analysis. Unsupervised

learning has even the ability to identify unknown targets, which provides more possibilities for abnormal traffic detection.

2.4 Performance Evaluation

In order to verify the effectiveness of encrypted traffic analysis methods, several effective validation methods and evaluation metrics are proposed. The validation method is used to confirm that the analytical method is suitable for its intended use. In machine learning, there are two commonly used datasets, training set and test set, which are utilized for model building and model validating.

The effectiveness of encrypted traffic identification will be assessed using the following metrics: TP stands for The number of samples correctly identified as normal, FP stands for The number of samples incorrectly identified as normal, FN stands for The number of samples incorrectly identified as malicious, and TN stands for The number of samples correctly identified as malicious.

Based on these four elements, the commonly-used evaluation metrics for effectiveness can be calculated. TPR (Recall) is the ratio between the number of positive samples correctly classified and the number of actual positive samples, ranging between 0 and 1. FPR is the ratio between the number of negative samples incorrectly classified and the number of actual negative samples, ranging between 0 and 1. P (Precision) is the ratio between the number of positive samples correctly classified and the number of samples classified as positive, ranging between 0 and 1.

Moreover, some metrics that consider two of the above metrics comprehensively are proposed.

F1 considers both P and TPR, which is represented as the harmonic mean of P and TPR.

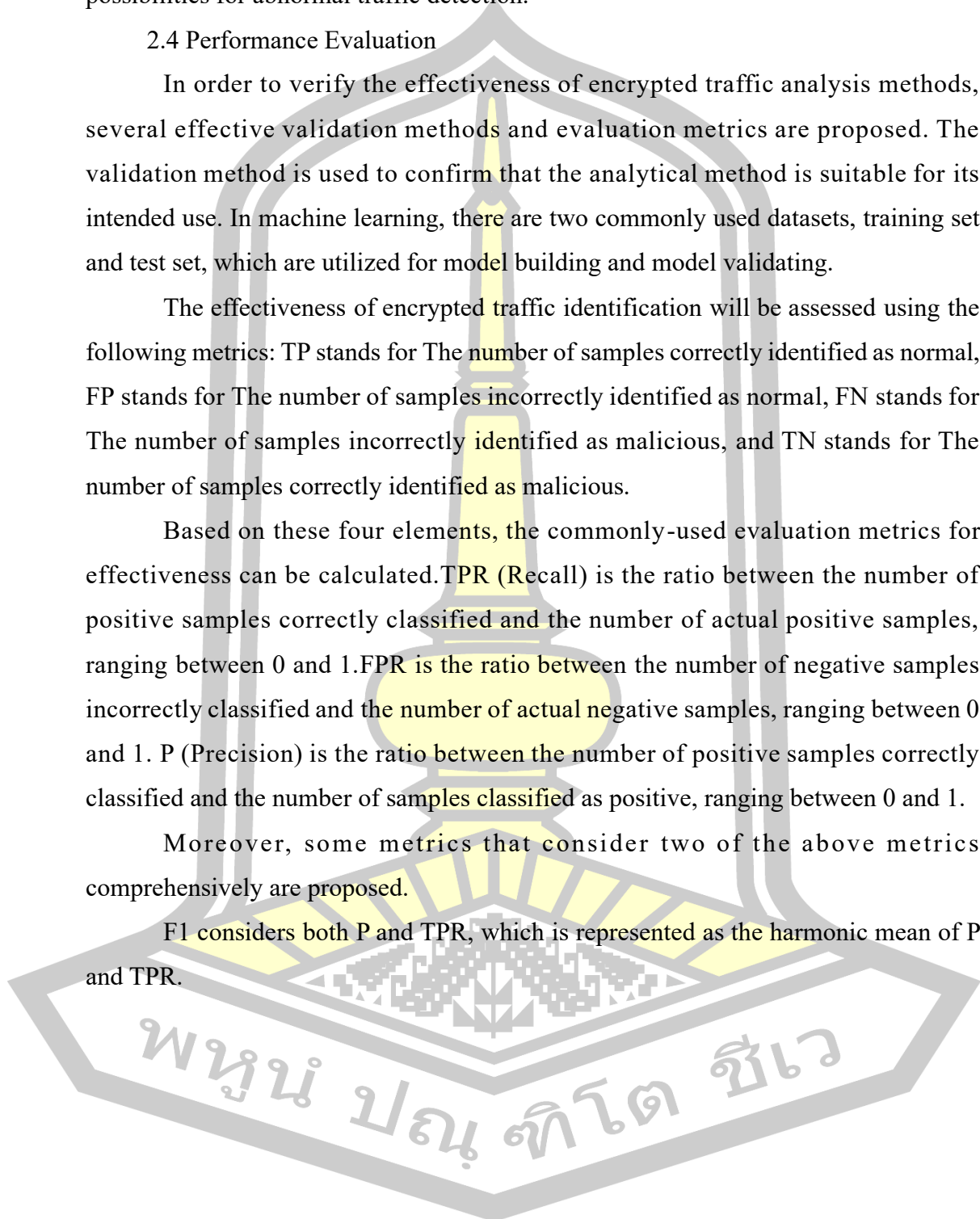


Table 4 Performance measure metrics

Parameter	Calculation formula
ACY(Accuracy)	$ACY = \frac{TP + TN}{TP + TN + FP + FN}$
P(Precision)	$P = \frac{TP}{TP + FP}$
TPR(TP Rate)	$TPR = \frac{TP}{TP + FN}$
FPR(FP Rate)	$FPR = \frac{FP}{FP + TN}$
FNR(FN Rate)	$FNR = \frac{FN}{TP + FN}$
F1	$F_1 = \frac{2 * P * TPR}{P + TPR}$

III Machine Learning Algorithm

3.1 Experimental Datasets

The ISCX-VPN-NonVPN-2016 dataset[71] is an open data resource for the study of network encrypted traffic and is widely used by research institutes and universities around the world because of its diversity and richness. This openness and multi-country usage has made the ISCX dataset popular among the international community of researchers on network encrypted traffic and has facilitated collaboration and knowledge sharing on a global scale. Researchers from different countries have been able to use this dataset to carry out a variety of research work on cybersecurity and traffic analysis, contributing to global cybersecurity, which reflects the international and global nature of cybersecurity issues. The ISCX dataset is used by countries around the world, as shown in Figure.6 below:

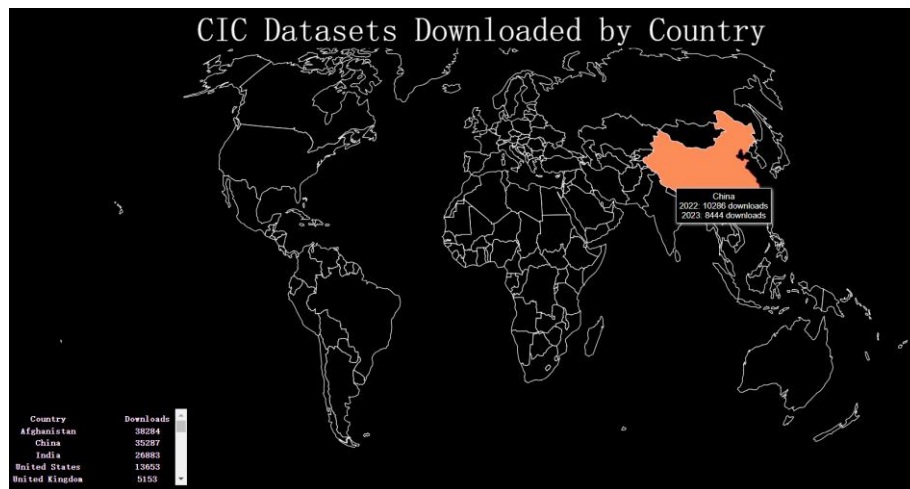


Figure 6 map of countries around the world using the ISCX dataset

The ISCX-VPN-NonVPN-2016 is a network traffic dataset designed to facilitate research and analysis of the characteristics and behaviors of Virtual Private Network(VPN) and non-VPN traffic. This dataset comprises real-world network traffic data from both VPN and non-VPN connections, with explicit labeling, allowing researchers to perform supervised learning and comparative analysis. It serves as a valuable resource for gaining insights into the distinctions between different network connection types, advancing research in network security, intrusion detection, privacy analysis, and VPN performance evaluation.

The ISCX dataset covers encrypted traffic data for a wide range of network communication scenarios and protocols, which is not only large in scale, but also provides rich metadata and annotations, which helps to study network activities in depth. Due to its versatility, the ISCX dataset has been widely used in a variety of research areas such as network security, traffic analysis, machine learning, etc., which promotes worldwide collaboration and knowledge sharing, and provides strong support for network encrypted traffic research.

This dataset is publicly available and download at <http://205.174.165.80/CICDataset/ISCX-VPN-NonVPN-2016/Dataset/>. The downloaded VPN and nonVPN datasets are shown in Figure.7.

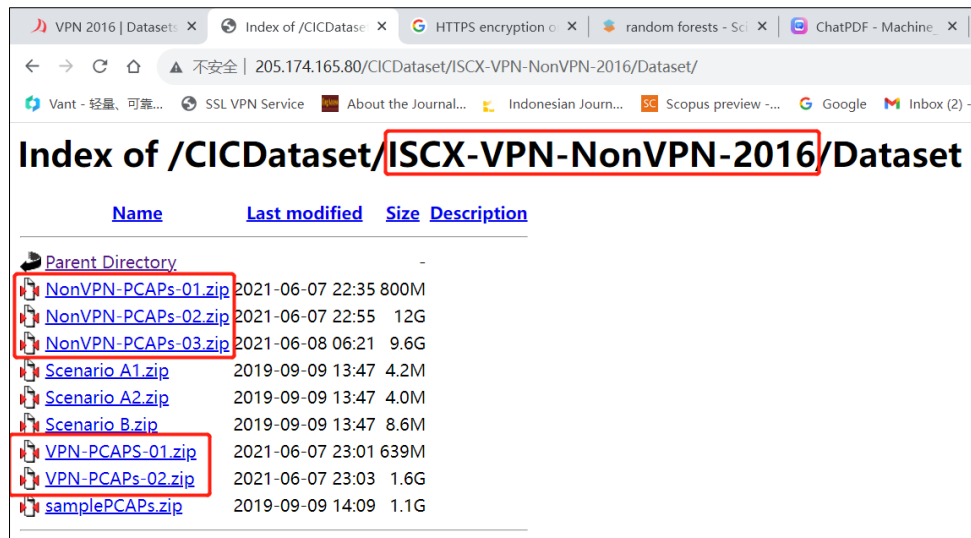


Figure 7 ISCX-VPN-NonVPN-2016 Dataset download

The types of network traffic applications included in the network traffic dataset and what they contain are shown in Table.4 below.

Table 5 ISCX Network Traffic Dataset content

Traffic Type	Content
Web Browsing	Firefox and Chrome
Email	SMTPS, POP3S and IMAPS
Chat	ICQ, AIM, Skype, Facebook and Hangouts
Streaming	Vimeo and Youtube
File Transfer	Skype, FTPS and SFTP using Filezilla and an external service
VoIP	Facebook, Skype and Hangouts voice calls (1h duration)
P2P	uTorrent and Transmission (Bittorrent)

During the research, we have a total of 14 traffic categories captured by Wireshark, the different traffic including browsing, email, chat, streaming, file

transfer, VoIP, and P2P. External VPN and file transfer service providers were used to generate VPN, SFTP, and FTPS traffic. To ensure accurate labeling, unnecessary services and applications were closed during the traffic capture process. The detailed description of the different types of traffic generated as following:

Browsing: Under this label we have HTTPS traffic generated by users while browsing or performing any task that includes the use of a browser. For instance, when we captured voice-calls using hangouts, even though browsing is not the main activity, we captured several browsing flows.

Email: The traffic samples generated using a Thunderbird client, and Alice and Bob Gmail accounts. The clients were configured to deliver mail through SMTP/S, and receive it using POP3/SSL in one client and IMAP/SSL in the other.

Chat: The chat label identifies instant-messaging applications. Under this label we have Facebook and Hangouts via web browsers, Skype, and IAM and ICQ using an application called pidgin.

Streaming: The streaming label identifies multimedia applications that require a continuous and steady stream of data. We captured traffic from Youtube (HTML5 and flash versions) and Vimeo services using Chrome and Firefox.

File Transfer: This label identifies traffic applications whose main purpose is to send or receive files and documents. For our dataset we captured Skype file transfers, FTP over SSH (SFTP) and FTP over SSL (FTPS) traffic sessions.

VoIP: The Voice over IP label groups all traffic generated by voice applications. Within this label we captured voice calls using Facebook, Hangouts and Skype.

TraP2P: This label is used to identify file-sharing protocols like Bittorrent. To generate this traffic we downloaded different .torrent files from a public repository and captured traffic sessions using the uTorrent and Transmission applications.

To facilitate the labeling process, when capturing the traffic all unnecessary services and applications were closed. (The only application executed was the objective of the capture, e.g., Skype voice-call, SFTP file transfer, etc.) We used a

filter to capture only the packets with source or destination IP, the address of the local client.

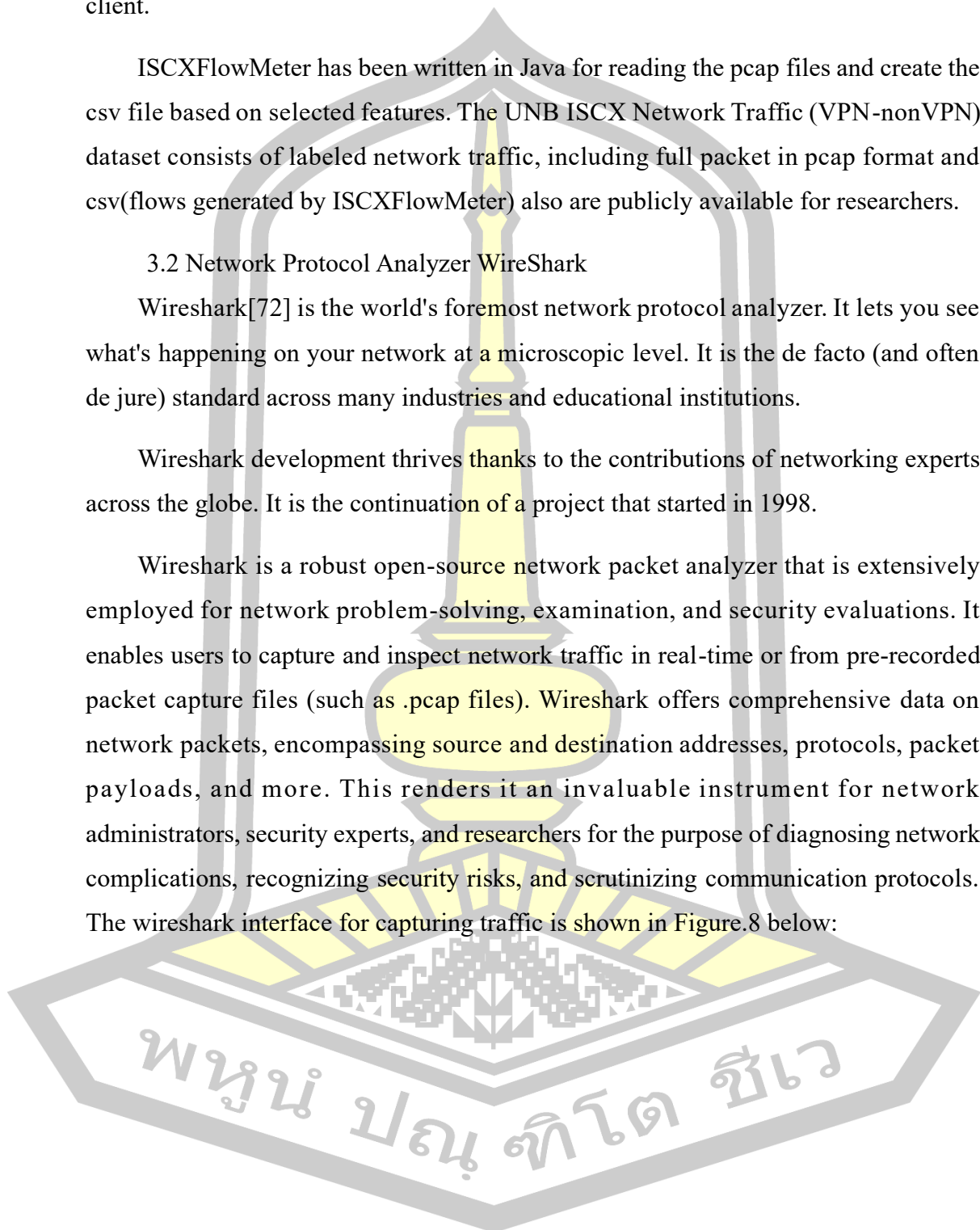
ISCXFlowMeter has been written in Java for reading the pcap files and create the csv file based on selected features. The UNB ISCX Network Traffic (VPN-nonVPN) dataset consists of labeled network traffic, including full packet in pcap format and csv(flows generated by ISCXFlowMeter) also are publicly available for researchers.

3.2 Network Protocol Analyzer WireShark

Wireshark[72] is the world's foremost network protocol analyzer. It lets you see what's happening on your network at a microscopic level. It is the de facto (and often de jure) standard across many industries and educational institutions.

Wireshark development thrives thanks to the contributions of networking experts across the globe. It is the continuation of a project that started in 1998.

Wireshark is a robust open-source network packet analyzer that is extensively employed for network problem-solving, examination, and security evaluations. It enables users to capture and inspect network traffic in real-time or from pre-recorded packet capture files (such as .pcap files). Wireshark offers comprehensive data on network packets, encompassing source and destination addresses, protocols, packet payloads, and more. This renders it an invaluable instrument for network administrators, security experts, and researchers for the purpose of diagnosing network complications, recognizing security risks, and scrutinizing communication protocols. The wireshark interface for capturing traffic is shown in Figure.8 below:



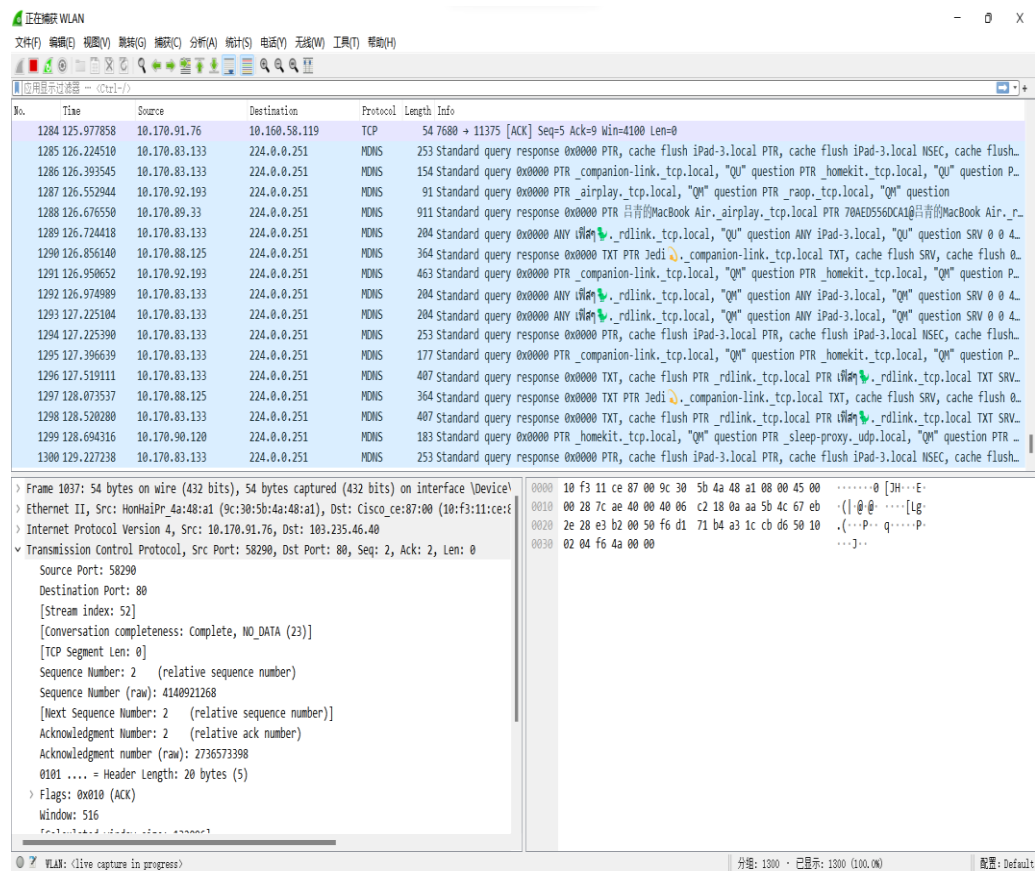


Figure 8 the interface of wireshark capturing traffic

The menu of the Wireshark Capture Traffic interface consists of the following main sections:

a. Main Window: The main interface of Wireshark typically includes a menu bar, toolbar, and a display area. The display area is where captured network packets are primarily shown.

b. Capture Options: By clicking the "Capture Options" button in the Wireshark main window, you can select the network interface to capture from (such as Ethernet, Wi-Fi, etc.). You can also configure capture filters to limit the range of captured packets.

c. Start Capture: Once capture options are configured, clicking the "Start" button initiates the network traffic capture. Captured packets are displayed in the main window.

d. **Packet List:** The packet list in the main window displays summarized information about the captured packets, including timestamps, source and destination addresses, protocols, and more. You can click on a packet to view its detailed information.

e. **Packet Details:** Selecting a specific packet results in Wireshark displaying detailed information about that packet at the bottom of the window. This detailed information includes protocol-layer analysis, data field contents, and more. This information is typically referred to as the "packet decode."

f. **Filter Bar:** Wireshark allows you to quickly filter and search for specific types of packets using the filter bar. You can filter based on criteria like protocol, source or destination address, port, etc.

g. **Statistics:** Wireshark provides various network statistics, including the number of captured packets, protocol distribution, traffic graphs, and more. These statistics help you analyze network performance and issues.

h. **Menu Bar and Toolbar:** The menu bar contains various function options, such as file operations, capture options, and view settings. The toolbar includes commonly used capture and analysis tools, like start capture, stop capture, and clear data.

i. **Color Coding:** Wireshark uses color coding to represent different types of packets, making it easy to identify protocols and issues.

j. **Navigation Panel:** Wireshark's navigation panel allows you to easily browse through captured packets, quickly jump to specific packets, or sessions.

These components of the Wireshark traffic capture interface enable users to effectively monitor and analyze network traffic for diagnosing issues, analyzing performance, and ensuring network security.

problem of decision trees being prone to overfitting. The main idea of Random Forest is to increase the diversity among trees by introducing randomness, thus improving the performance and robustness of the model. Random Forest is an integrated learning method that consists of multiple decision trees. Each decision tree is trained on a randomly sampled subsample and is partitioned at each node considering only randomly selected features. These two kinds of randomness help to reduce the variance (through sample randomness) and bias (through feature randomness) of the model, which improves the generalization ability of the model. The final prediction is usually obtained by voting or averaging all the decision trees.

Random forest is a remarkable machine learning algorithm widely used in the field of encrypted traffic identification for network communication. One of its main roles is to help identify the encrypted and non-encrypted parts of network traffic, and specific applications include threat detection, traffic classification, and encryption detection. First, Random Forest can be used for early threat detection, which helps to find potential network threats such as malware, intrusion, or data leakage. By analyzing the characteristics of network traffic, it is able to identify patterns that differ from normal communication patterns, thus providing early warning before threatening behavior occurs. Second, random forests play an important role in traffic classification by categorizing network traffic into different applications, protocols, or services.

This is critical for network management, bandwidth optimization, and monitoring traffic compliance, helping to improve network efficiency and resource utilization. Finally, Random Forests are also able to detect encrypted communications, which is critical for identifying encryption in malicious campaigns as well as ensuring network security. Overall, Random Forest has a wide range of applications in analyzing encrypted traffic for network communications, improving identification accuracy, reducing false positives, enhancing network security, and providing network security professionals with powerful tools to protect and maintain the security and privacy of communications.

(2) Enhanced Random Forest Algorithm

Random forest is a model established using a large number of decision trees. Among them, W represents a random vector that follows independent and identically

distributed distribution, K is the number of decision trees, and each tree is the known variable x for optimal voting. N represents the total number of samples, the objects in X have M -dimensional feature vectors, and Y includes F different categories of information.

The classifier excessively emphasizes the classification of training samples, resulting in poor prediction of test samples, which is called overfitting. In the process of model construction, various error analyses will be introduced, and for random forests, generalization error is a key point in expressing overfitting problems. The generalization error is described by the edge function, as follows.

$$mg(X, Y) = \alpha v_k I(h_k(X) = Y) - \max_{j \neq Y} \alpha v_k I(h_k(X) = j). \quad (1)$$

Among them, I represents the indicator function, and αv_k represents the average value. This function expresses the difference between the average number of correct votes obtained by correctly classifying the random vector X into Y and the average number of votes obtained by other categories. The larger the function value, the higher the accuracy of the classification. The definition of generalization error is as follows.

$$PE = P_{X,Y}(mg(X, Y) < 0). \quad (2)$$

Among them, the subscripts X and Y represent the generalization error obtained in the random variables X and Y .

By solving the expected value of the edge function and inferring the Chebyshev inequality, an upper bound for the generalization error can be obtained.

$$PE \leq \frac{\bar{\rho}(1-s^2)}{s^2}. \quad (3)$$

Among them, $\bar{\rho}$ represents the mean correlation factor ρ between trees, and s represents the performance strength of the classifier. In order to better analyze the final

performance of the classifier, the correlation factors and performance strength are described by the c/s^2 ratio, and the smaller the ratio, the better the performance of the classifier. It is defined as

$$c/s^2 = \frac{\bar{\rho}}{s^2}. \quad (4)$$

The random forest cleverly integrates the bagging framework and CART process. Specifically, each tree randomly selects some variables as candidates for the split variables during each split to produce child nodes, in order to reduce the correlation between individuals.

Build more trees than traditional models, and then select some more accurate trees for aggregation. There is a variable interval between the correlation coefficient and the strength, within which the correlation coefficient increases and the strength will increase as compensation, and vice versa. Therefore, from the perspective of balancing the degree of correlation and strength, the above modifications attempt to find individuals with higher strength, while ensuring that the correlation coefficient between these individuals does not increase significantly. In addition, building more trees will inevitably increase the training time of the model, but as discussed earlier, due to the good parallelism of bagging, the time cost can easily be solved by increasing computational resources.

Due to the fact that the enhanced random forest aggregates the most accurate subset of individuals in the early stages, it can achieve more accurate prediction results than traditional algorithms. As the number of aggregated trees increased, both models eventually aggregated all the constructed trees. The algorithm steps for enhancing random forests are as follows.

Step1. Generate a bootstrap sample Z with a sample size of N from the training set.

Step2. Generate a tree in a random forest based on each bootstrap sample during the process of tree growth, recursively repeat the following steps for each node until the node reaches the predetermined size limit.

Step3. Randomly select r variables from all p variables, and select the best splitting variable and splitting point from r variables as candidates.

Step4. Split the current node into two sub nodes based on the selected split points.

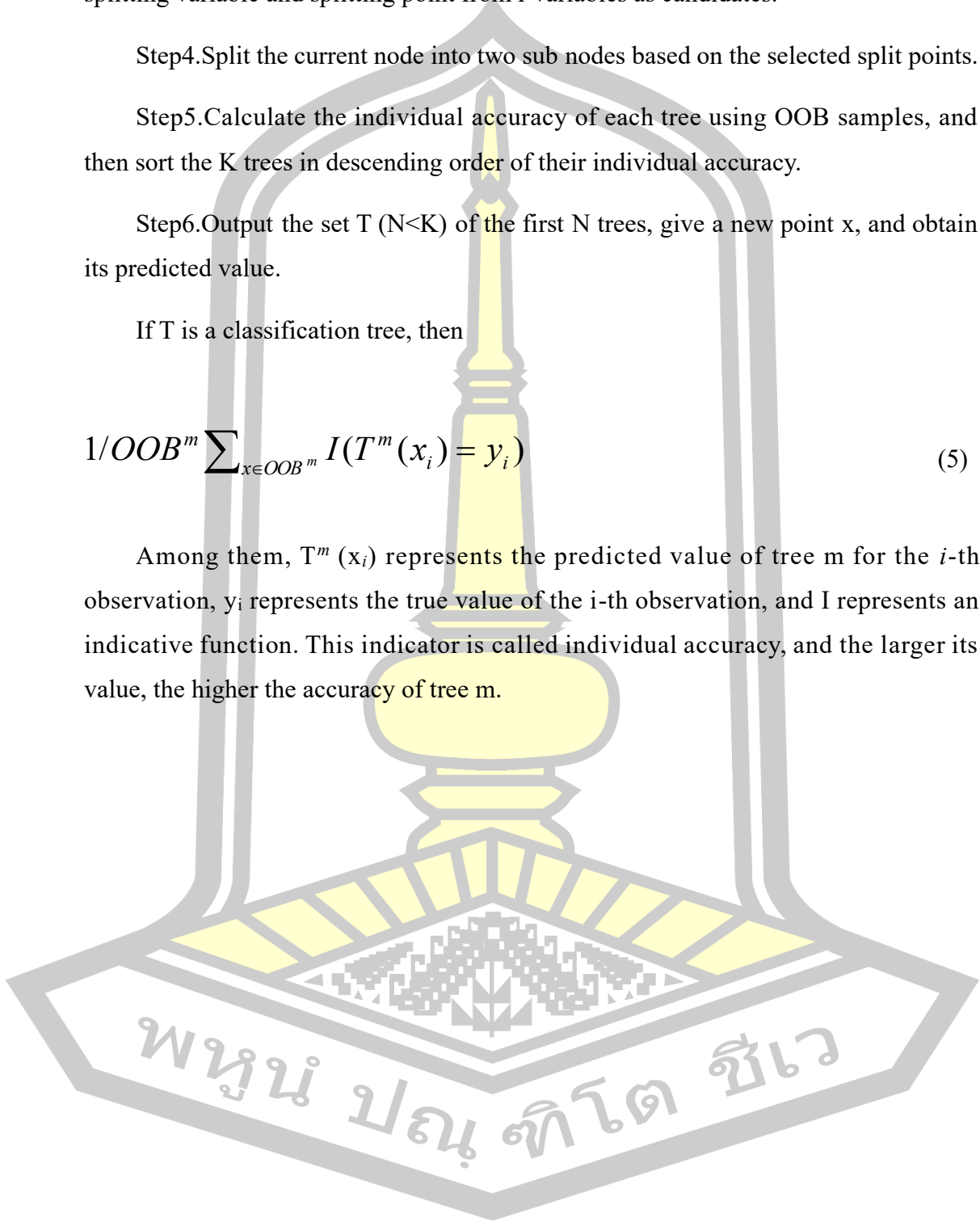
Step5. Calculate the individual accuracy of each tree using OOB samples, and then sort the K trees in descending order of their individual accuracy.

Step6. Output the set T ($N < K$) of the first N trees, give a new point x , and obtain its predicted value.

If T is a classification tree, then

$$1/OOB^m \sum_{x \in OOB^m} I(T^m(x_i) = y_i) \quad (5)$$

Among them, $T^m(x_i)$ represents the predicted value of tree m for the i -th observation, y_i represents the true value of the i -th observation, and I represents an indicative function. This indicator is called individual accuracy, and the larger its value, the higher the accuracy of tree m .



Chapter 4 Results and Discussion

I Encrypted Traffic Identification

Encrypted traffic identification, also known as encrypted traffic analysis, is a technique that analyzes the content of network packets to determine whether they have been encrypted and the type of encryption. In the field of network security, encrypted traffic identification is of great significance because it can help us better understand network attacks, track malicious behavior, and identify potential security threats [73].

The principle of encrypted traffic recognition is mainly based on the header information of network packets. The header information contains key information such as protocol type, source address, destination address, etc. By analyzing this information, we can determine whether the data packet has been encrypted and the type of encryption (such as SSL, TLS, UDP encryption, etc.). In addition, some advanced encryption traffic recognition technologies can also be combined with the content of data packets to further improve the accuracy of recognition.

There are many methods for identifying encrypted traffic, mainly including but not limited to the following:

- 1) Protocol analysis: By analyzing the packet structure of common encryption protocols such as SSL and TLS, it is possible to determine whether the data packet has been encrypted [74].
- 2) Fingerprint recognition: Using known encryption algorithm features, by comparing the information in the data packet, determine its encryption type.
- 3) Deep learning: Using machine learning algorithms to train a large number of encrypted data packets and establish classification models to improve recognition accuracy.

II Raw Traffic Data Segmentation

Whether it is the captured datasets or the publicly available datasets downloaded from the internet, the most primitive data format is the PCAP package format. It is obvious that the data format of the PCAP package cannot be used directly. Before use, it must be segmented and necessary cleaning work must be carried out. It is necessary

for us to understand the PCAP package format before splitting [75]. Each PCAP file consists of three parts: GlobalHeader, PacketHeader, and PacketData. The initial position of the file is GlobalHeader, followed by paired PacketHeader and PacketData.

When dividing traffic, the first step is to determine whether to divide it according to bidirectional flow or unidirectional flow. Traffic data has a direction, which refers to whether a specific message is sent from the source IP to the destination or from the destination IP to the source IP. In the actual conversation process, these two directions of flow exist simultaneously. In the process of capturing traffic, real-time data flowing through the host network card is read out, and these two different directions of flow are not distinguished. However, in practical use, sometimes it may be necessary to distinguish between traffic data from two different directions. Therefore, dividing the direction of flow here can provide more targeted data analysis. In other scenarios such as application recognition, whether it is traffic data from the client to the server or from the server to the client, they all contain specific application characteristics. In this scenario, there is no need to distinguish the direction of diversion.

In order to extract as many features as possible from the original data, a huge feature library is established, which contains a large number of original features, of course, its dimensionality is also very high. However, in this step, we will reduce the dimensionality of samples in high-dimensional space by removing redundant and irrelevant features through feature selection. The availability of data tends to focus on its features which have capability for its considerable representation. Therefore, in the step of feature selection, the main task is to filter out features that have good discriminative properties for the datasets from the feature library. And there are some basic principles that should be followed when performing feature selection work here, such as obtaining the smallest possible subset of features, not significantly reducing classification accuracy, not affecting class distribution, and the feature subset should have stable and strong adaptability.

III Encrypted Traffic Anomaly Detection

Encrypted traffic identification objects mainly include stream level, host level, packet level and other modes [76]. Among them, stream-level and packet-level identification objects are the most widely used [77]. Table 5 shows the details of

encrypted traffic anomaly detection targets. My research mainly takes stream-level encrypted traffic as the detection object.

There are two main reasons. On the one hand, the traffic characteristics are obvious. Stream-level encrypted traffic usually exists in the form of continuous data packets in network communication, and these data packets have clear correlation and time sequence. This feature enables the stream-level encrypted traffic to reflect the overall state and behavior pattern of network traffic more accurately, which is conducive to detecting abnormal traffic in the network. On the other hand, it is rich in information. Stream-level encrypted traffic contains rich information, such as source address, destination address, port number and transmission protocol, which is of great significance for identifying abnormal behavior in the network.

Through the analysis of stream-level encrypted traffic, we can get the overall characteristics of network traffic and the characteristics of abnormal behavior, which provides strong support for anomaly detection. However, choosing stream-level encrypted traffic as the main detection target also has potential limitations: the detection of stream-level encrypted traffic depends on the data of network traffic, and if the data of network traffic is incomplete or has errors, the accuracy of detection results will be affected. However, the data of network traffic in this paper is complete, and the influence of limitations can be ignored.

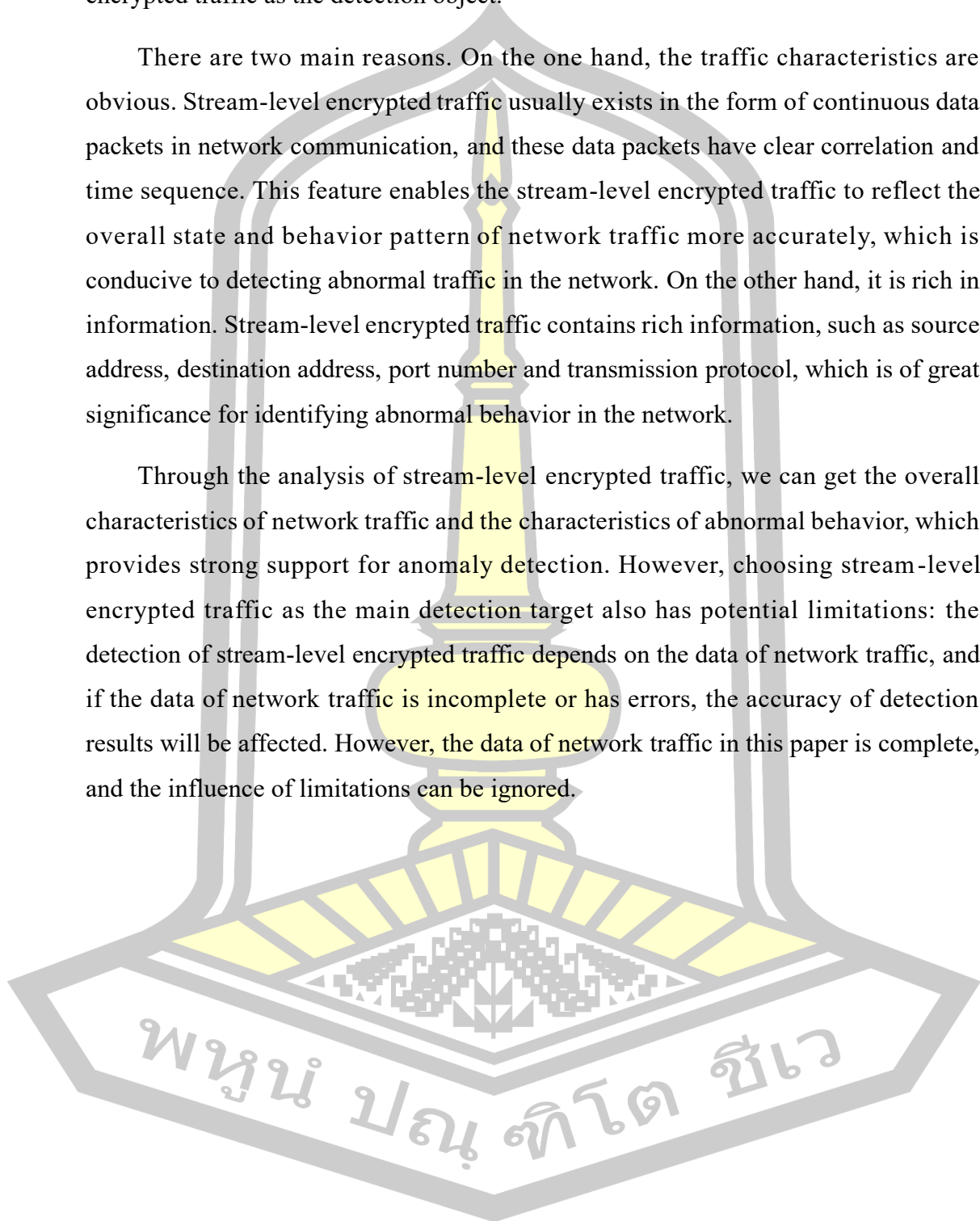


Table 6 Details of encrypted traffic anomaly detection targets

Target type	Detection content	Core features
Flow level	Characteristics of flow and transmission process	Duration, number of bytes, etc
Packet-level	The characteristics and transmission process of traffic grouping	Group size, transmission time, time interval
Host level	Correlation between hosts	Number of connections and ports
Session level	The characteristics and transmission process of conversations	Bytes, session time

IV Results Analysis

Based on public datasets, three main types of datasets have been released: malicious traffic data, normal traffic data, and mixed traffic data. In this paper, combined with the research direction of this paper, encrypted malicious traffic data is mainly used. Whether encrypted traffic is collected is determined by determining whether the PCAP packet in the collected traffic uses an encryption protocol.

The experimental hardware environment is based on Windows 10 64-bit operating system with Intel Core i7 CPU and 8GB of RAM, message queue kafka, and the Python machine learning library auto-sklearn is installed. All models including the comparison methods are trained using the enhanced random forest algorithm proposed in this paper.

In order to ensure the rapid convergence of the Enhanced Random Forest (ERF) algorithm and eliminate redundant information as much as possible, it is necessary to unify the length of the input encrypted traffic. When unifying the length of the encrypted traffic, it is necessary to determine the number of intercepted data packets and the byte length in the data packet. Different data packet numbers and byte lengths will have an impact on the classification performance of the model. Therefore, the evaluation index value is used as a performance evaluation indicator for data packet

numbers, respectively, Extract bytes of different lengths for comparative experiments, and display the average values of experiments. The experimental results are shown in Figure 11.

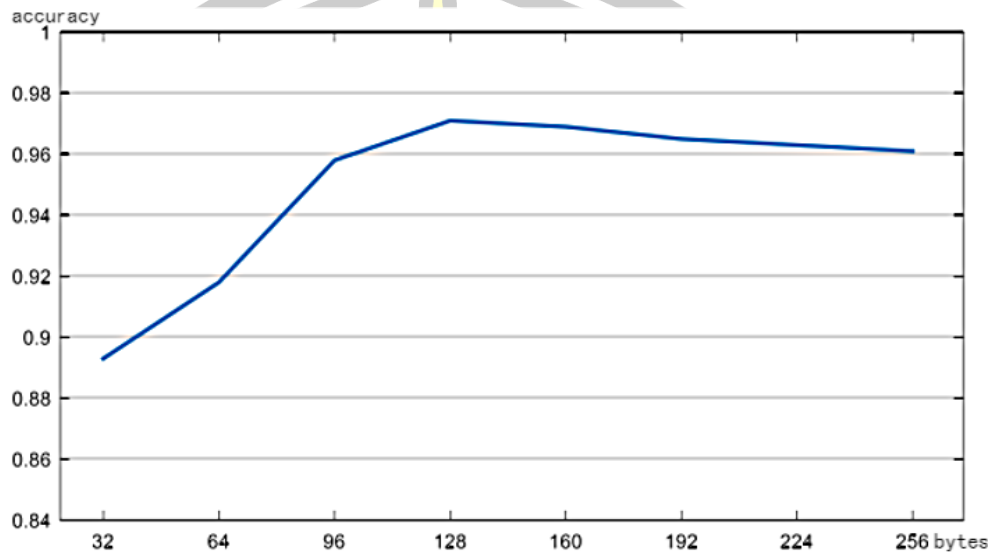


Figure 11 The effect of different byte lengths by ERF algorithm

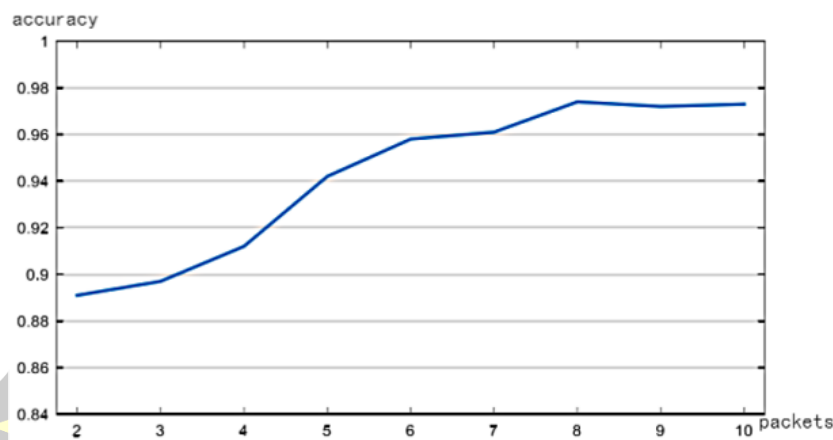


Figure 12 The effect of different packets by ERF algorithm

The horizontal axis represents the number of data packets, and the vertical axis use accuracy as the evaluation index. From the comparative experiment, it can be seen that the model performs best in classification when intercepting the first 8 data packets of the session.

Evaluation metrics for the effectiveness of encrypted traffic identification involves measuring TP (true positives, correctly identified as normal), FP (false

positives, incorrectly identified as normal), FN (false negatives, incorrectly identified as malicious), and TN (true negatives, correctly identified as malicious). These metrics gauge the accuracy in distinguishing between normal and malicious traffic types.

Recall is the ratio between the number of positive samples correctly classified and the number of actual positive samples, ranging between 0 and 1. FPR is the ratio between the number of negative samples incorrectly classified and the number of actual negative samples, ranging between 0 and 1. P (Precision) is the ratio between the number of positive samples correctly classified and the number of samples classified as positive, ranging between 0 and 1. F1 considers both P and Recall, which is represented as the harmonic mean of P and Recall.

From Table 6, it can be seen that the average checking Accuracy reaches 98.95% in the experimental results for different datasets, and the Precision, Recall rate, and the F1 data, which comprehensively reflect the overall performance, all show better results.

Table 7 Data set recognition performance

Data Set	P (%)	Recall (%)	F1 (%)	Accuracy (%)
N1	98.5	97.8	98.1	98.2
N2	99.4	98.2	98.4	99.7
N3	98.2	97.6	97.7	97.5
N4	97.3	99.5	99.8	98.4
N5	99.6	98.6	97.9	97.9
N6	98.5	99.3	99.1	98.9
N7	98.5	98.8	98.1	98.2
N8	99.4	98.2	98.4	99.7
N9	98.2	97.6	97.7	98.5

Comparison of the data results of the different datasets for the four performance evaluation metrics are shown in Figure 13.

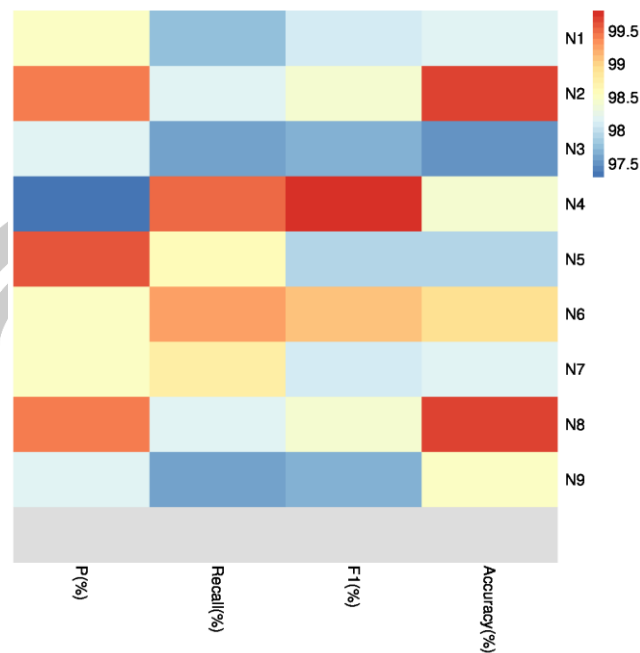


Figure 13 Data comparison of datasets on performance evaluation metrics

From Table 7, it can be seen that the ERF algorithm has a good recognition effect on encrypted traffic. Whether in conventional encrypted traffic or VPN encrypted traffic, the algorithm has a recognition accuracy of over 90%, with an average recognition accuracy of 92% in several conventional encrypted traffic and 94% in several VPN encrypted traffic. This indicates that the model extracts more features from VPN encrypted traffic, Therefore, the classification effect is better.

Table 8 Error rate of traffic test set

Data Set	SVM	GBM	RF	ERF1	ERF2	ERF3	ERF4
N1	3.36	3.96	3.16	3.47	3.37	3.57	3.37
N2	23.52	25.51	25.25	25.25	24.91	25.15	25.11
N3	23.89	23.32	23.54	23.78	23.28	23.53	23.27
N4	37.68	3.26	2.11	2.12	2.03	2.01	2.02
N5	4.60	3.72	3.98	4.00	4.02	4.07	4.02
N6	29.82	25.80	26.29	27.74	27.2	28.06	27.43
N7	6.85	5.82	4.9	4.86	4.82	4.87	4.79
N8	5.72	20.53	3.86	4.03	3.79	4.03	3.76
N9	10.38	11.74	8.69	8.83	8.64	8.83	8.65

Chapter 5 Conclusions

I Summary of the Research

In this research, we explored the effectiveness of an Enhanced Random Forest (ERF) algorithm for network traffic classification, particularly in identifying malicious traffic and improving defense mechanisms against network attacks. The study utilized publicly available datasets and employed a comprehensive data preprocessing strategy, including data segmentation and data cleaning, to refine the original PCAP package format. The primary goal was to extract and filter features with high discriminative capabilities from the feature library, ensuring that only the most relevant attributes were considered for classification.

Building upon the traditional Random Forest model, this research introduced several enhancements to optimize its performance. One of the primary improvements involved constructing a larger number of decision trees than the conventional model and subsequently selecting the most accurate ones for aggregation. Additionally, to fine-tune the model's predictive capability, we employed grid search cross-validation, systematically adjusting key parameters such as the number of trees, tree depth, and feature selection criteria. The iterative evaluation of different parameter combinations allowed us to identify the optimal settings, leading to a more robust classification model.

The performance of the enhanced Random Forest algorithm was rigorously evaluated using well-established metrics, including accuracy, precision, recall, and F1-score. The experimental results demonstrated that the improved model significantly outperforms the traditional RF approach. Specifically, the average precision reached an impressive 98.95%, and other performance evaluation metrics also indicated strong classification capabilities. These results highlight the substantial benefits of our enhanced RF model, confirming its suitability for real-world network security applications.

The high accuracy of our model underscores its ability to effectively differentiate between normal and malicious traffic, even in complex network environments. Furthermore, the improved recall suggests that the model successfully captures a high proportion of malicious traffic without missing significant instances,

thereby reducing false negatives. The high F1-score indicates a strong balance between precision and recall, ensuring both reliability and robustness in classification.

Future research in identifying encrypted traffic will continue to search for algorithm choices and improvements with optimal efficiency, and explore refined classification and anomaly detection techniques for encrypted traffic. Further research will be conducted on how to efficiently deploy algorithms in real-time environments to reduce latency and resource consumption, as well as to improve the robustness and generalization ability of the model under different network environments and traffic patterns.

II Contribution to Network Security

One of the most significant contributions of this study is its applicability to network security, particularly in the detection and classification of encrypted traffic and emerging attack patterns. Traditional intrusion detection systems (IDS) often struggle with evolving threats due to the increasing complexity of network protocols and encryption techniques. Our enhanced model demonstrates adaptive capabilities, allowing it to classify encrypted traffic and new types of network traffic characteristics effectively.

This adaptability is critical in modern cybersecurity frameworks, as cyber threats continue to evolve rapidly. Attackers frequently modify their methodologies to evade detection, utilizing advanced obfuscation techniques and encrypted communication channels. The ability of our enhanced RF model to recognize and classify such traffic ensures a proactive defense mechanism, providing security analysts with valuable insights to mitigate threats before they escalate.

Furthermore, by optimizing the feature selection process and employing grid search cross-validation, we have demonstrated that the efficiency and effectiveness of network traffic classification can be significantly improved. The reduction in false positives and false negatives enhances the reliability of security monitoring systems, ensuring that legitimate traffic is not erroneously flagged while still maintaining high detection rates for malicious activities.

III Advantages Over Traditional Approaches

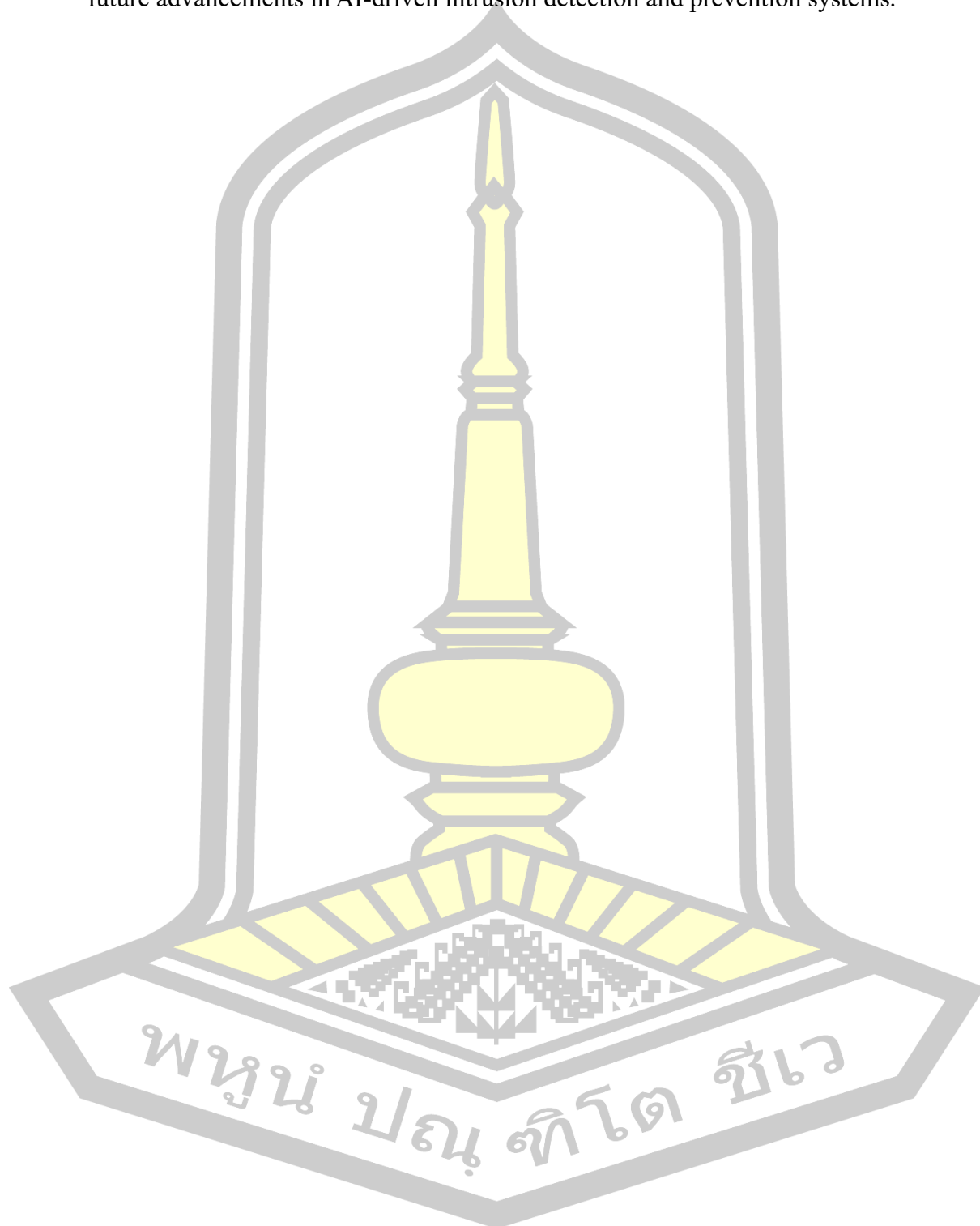
Compared to traditional Random Forest models and other machine learning-based intrusion detection methods, our ERF algorithm offers several distinct advantages. By increasing the number of trees and selecting only the most accurate ones for aggregation, our model achieves significantly higher precision and recall rates. The enhanced model ensures a more effective classification process, leading to improved identification of malicious activities while maintaining computational efficiency. Additionally, the grid search cross-validation technique allows for fine-tuning of key parameters, reducing manual intervention and optimizing the model's overall performance.

Another notable advantage is the model's ability to adapt to encrypted traffic, a common challenge in modern network security. Many contemporary cyberattacks utilize encryption to evade detection, making traditional classification methods less effective. Our ERF model effectively identifies malicious activities within encrypted communication channels, demonstrating its robustness in handling emerging security threats. This adaptability makes the model a valuable tool for modern cybersecurity infrastructures, particularly in environments where encrypted and obfuscated traffic is prevalent.

Lastly, the model demonstrates strong scalability and generalization capabilities, making it suitable for diverse real-world applications. The random forest model inherently supports parallelization, and our enhancements ensure that the training process remains computationally efficient even with large-scale datasets. Additionally, the model performs consistently well across different datasets, proving its ability to generalize effectively. These advantages collectively highlight the superiority of the ERF algorithm in addressing contemporary network security challenges.

In conclusion, this research has successfully demonstrated the potential of an enhanced Random Forest algorithm in improving network security through accurate malicious traffic detection and classification. As cybersecurity challenges continue to evolve, machine learning-based solutions will play an increasingly vital role in proactive threat detection and network defense mechanisms. The findings of this study

contribute to the growing field of intelligent network security, paving the way for future advancements in AI-driven intrusion detection and prevention systems.



References

- [1] China Internet Network Information Center (CNNIC)[EB/OL].<https://cnnic.cn/n4/2023/0302/c199-10755.html>. 2023-08-08
- [2] Netmarketshare.HTTPS[EB/OL].<https://netmarketshare.com/>.2023-08-10
- [3] Google. HTTPS encryption on the web Google Transparency Report.[EB/OL].<https://transparencyreport.google.com/https/overview>.2023-08-13
- [4] Y. Xin et al., "Machine learning and deep learning methods for cybersecurity," IEEE Access, vol. 6, pp. 35365–35381, 2018
- [5] M. Shen et al., "Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 791-824, Firstquarter 2023, doi: 10.1109/COMST.2022.3208196.
- [6] Ma, S., & Yang, J. J. (2023). Image-based vehicle classification by synergizing features from supervised and self-supervised learning paradigms. Eng, 4(1), 444. doi:<https://doi.org/10.3390/eng4010027>
- [7] Zhuang, B., Zhang, C., & Hu, Z. (2023). Self-supervised skill learning for semi-supervised long-horizon instruction following. Electronics, 12(7), 1587. doi:<https://doi.org/10.3390/electronics12071587>
- [8] Gong, Y., Shushrruth, S. S., Zhang, R., Kessenbrock, K., & Zhang, J. (2022). scEpiLock: A weakly supervised learning framework for cis-regulatory element localization and variant impact quantification for single-cell epigenetic data. Biomolecules, 12(7), 874. doi:<https://doi.org/10.3390/biom12070874>
- [9] Yolwas, N., & Meng, W. (2023). JSUM: A multitask learning speech recognition model for jointly supervised and unsupervised learning. Applied Sciences, 13(9), 5239. doi:<https://doi.org/10.3390/app13095239>
- [10] Alhadlaq, A. (2023). Computer-based simulated learning activities: Exploring saudi students' attitude and experience of using simulations to facilitate unsupervised learning of science concepts. Applied Sciences, 13(7), 4583. doi:<https://doi.org/10.3390/app13074583>
- [11] Park, S., Ye, J. C., Eun, S. L., Cho, G., Jin, W. Y., Choi, J. H., . . . Yoon, J. L. (2023). Deep learning-enabled detection of pneumoperitoneum in supine and erect abdominal radiography: Modeling using transfer learning and semi-supervised

learning. *Korean Journal of Radiology*, 24(6), 541-552.
doi:<https://doi.org/10.3348/kjr.2022.1032>

[12] Kurita, Y., Meguro, S., Tsuyama, N., Kosugi, I., Enomoto, Y., Kawasaki, H., . . . Iwashita, T. (2023). Accurate deep learning model using semi-supervised learning and noisy student for cervical cancer screening in low magnification images. *PLoS One*, 18(5) doi:<https://doi.org/10.1371/journal.pone.0285996>

[13] Buschi, D., Curti, N., Cola, V., Carlini, G., Sala, C., Daniele Dall'Olio, . . . Giampieri, E. (2023). Automated wound image segmentation: Transfer learning from human to pet via active semi-supervised learning. *Animals*, 13(6), 956. doi:<https://doi.org/10.3390/ani13060956>

[14] Sahu, S. K., Mokhade, A., & Bokde, N. D. (2023). An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: Recent progress and challenges. *Applied Sciences*, 13(3), 1956. doi:<https://doi.org/10.3390/app13031956>

[15] Mustapha, R., Soukaina, G., Mohammed, Q., & Es-Sâadia, A. (2023). Towards an adaptive e-learning system based on deep learner profile, machine learning approach, and reinforcement learning. *International Journal of Advanced Computer Science and Applications*, 14(5) doi:<https://doi.org/10.14569/IJACSA.2023.0140528>

[16] S.Roy and A. Etemad, "Self-supervised contrastive learning of multi-view facial expressions", *Proc. Int. Conf. Multimodal Interact.*, pp. 253-257, Oct. 2021.

[17] Y.Li, Y. Gao, B. Chen, Z. Zhang, G. Lu and D. Zhang, "Self-supervised exclusive-inclusive interactive learning for multi-label facial expression recognition in the wild", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3190-3202, May 2022.

[18] S. Jawed, J. Grabocka and L. Schmidt-Thieme, "Self-supervised learning for semi-supervised time series classification", *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, pp. 499-511, 2020.

[19] X. Zhai, A. Oliver, A. Kolesnikov and L. Beyer, "S4L: Self-supervised semi-supervised learning", *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 1476-1485, Oct. 2019.

- [20] T. Chen, S. Kornblith, K. Swersky, M. Norouzi and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners", *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 22243-22255, 2020.
- [21] Bhandari, A., Tripathy, B., Adate, A., Saxena, R., & Thippa, R. G. (2023). From beginning to BEGANing: Role of adversarial learning in reshaping generative models. *Electronics*, 12(1), 155. doi:<https://doi.org/10.3390/electronics12010155>
- [22] Templin, T., Memarzadeh, M., Vinci, W., Lott, P. A., Ata, A. A., Anthony, A. A., & Rieffel, E. (2023). Anomaly detection in aeronautics data with quantum-compatible discrete deep generative model. *Machine Learning : Science and Technology*, 4(3), 035018. doi:<https://doi.org/10.1088/2632-2153/ace756>
- [23] Cao, G., & Zhang, H. (2023). Bayesian network structure learning by ensemble learning and frequent item mining. *Mathematical Problems in Engineering*, 2023 doi:<https://doi.org/10.1155/2023/3119316>
- [24] Khattak, A., Pak-Wai Chan, Chen, F., Peng, H., & Matara, C. M. (2023). Missed approach, a safety-critical go-around procedure in aviation: Prediction based on machine learning-ensemble imbalance learning. *Advances in Meteorology*, 2023 doi:<https://doi.org/10.1155/2023/9119521>
- [25] Long, G., & Zhang, Z. (2023). Deep encrypted traffic detection: An anomaly detection framework for encryption traffic based on parallel automatic feature extraction. *Computational Intelligence and Neuroscience : CIN*, 2023 doi:<https://doi.org/10.1155/2023/3316642>
- [26] Wang, H., Zhang, Y., Liu, Y., Liu, F., Zhang, H., Xing, B., . . . Chen, L. (2022). ASAD: Adaptive seasonality anomaly detection algorithm under intricate KPI profiles. *Applied Sciences*, 12(12), 5855. doi:<https://doi.org/10.3390/app12125855>
- [27] Jiang, S., Dong, R., Wang, J., & Xia, M. (2023). Credit card fraud detection based on unsupervised attentional anomaly detection network. *Systems*, 11(6), 305. doi:<https://doi.org/10.3390/systems11060305>
- [28] Zhang, C., Wang, J., Yan, T., Lu, X., Lu, G., Tang, X., & Huang, B. (2023). An instance-based deep transfer learning method for quality identification of longjing tea from multiple geographical origins. *Complex & Intelligent Systems*, 9(3), 3409-3428. doi:<https://doi.org/10.1007/s40747-023-01024-4>

[29] Instance-based learning following physician reasoning for assistance during medical consultation. (2021). *Applied Sciences*, 11(13), 5886. doi:<https://doi.org/10.3390/app11135886>

[30] Wan, Z., Chang, Z., Xu, Y., Huang, Y., & Šavija, B. (2023). Inverse design of digital materials using corrected generative deep neural network and generative deep convolutional neural network. *Advanced Intelligent Systems*, 5(1) doi:<https://doi.org/10.1002/aisy.202200333>

[31] Getahun, W. G., & Ding, J. (2023). Elephant flows detection using deep neural network, convolutional neural network, long short-term memory, and autoencoder. *Journal of Computer Networks and Communications*, 2023 doi:<https://doi.org/10.1155/2023/1495642>

[32] Boronina, A., Maksimenko, V., & Hramov, A. E. (2023). Convolutional neural network outperforms graph neural network on the spatially variant graph data. *Mathematics*, 11(11), 2515. doi:<https://doi.org/10.3390/math11112515>

[33] Yang, C., & Huang, C. (2023). Natural language processing (NLP) in aviation safety: Systematic review of research and outlook into the future. *Aerospace*, 10(7), 600. doi:<https://doi.org/10.3390/aerospace10070600>

[34] Dash, G., Sharma, C., & Sharma, S. (2023). Sustainable marketing and the role of social media: An experimental study using natural language processing (NLP). *Sustainability*, 15(6), 5443. doi:<https://doi.org/10.3390/su15065443>

[35] Luander Cipriano de Jesus Falcão, Lopes, B., & Renato, R. S. (2022). Absorção das tarefas de processamento de linguagem natural (NLP) pela ciência da informação (CI): Uma revisão da literatura para tangibilização do uso de NLP pela CI. [Absorption of Natural Language Processing (NLP) Tasks by Information Science (IS): a literature review for tangibilizing the use of NLP by IS] *Em Questão*, 28(1), 13-34. doi:<https://doi.org/10.19132/1808-5245281.13-34>

[36] Arowosegbe, A., & Oyelade, T. (2023). Application of natural language processing (NLP) in detecting and preventing suicide ideation: A systematic review. *International Journal of Environmental Research and Public Health*, 20(2), 1514. doi:<https://doi.org/10.3390/ijerph20021514>

[37] López-Oriona, Á., & Vilar, J. A. (2023). Ordinal time series analysis with the R package *otsfeatures*. *Mathematics*, 11(11), 2565. doi:<https://doi.org/10.3390/math11112565>

[38] Bao, H., Zhu, H., Yu, P., Luo, G., Zhang, R., Yue, Q., & Fang, J. (2022). Time-series transcriptomic analysis reveals the molecular profiles of diapause termination induced by long photoperiods and high temperature in *chilo suppressalis*. *International Journal of Molecular Sciences*, 23(20), 12322. doi:<https://doi.org/10.3390/ijms232012322>

[39] V. Vijay and P. Verma, "Variants of Naïve Bayes Algorithm for Hate Speech Detection in Text Documents," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 18-21, doi: 10.1109/AISC56616.2023.10085511.

[40] M. Schaffeld, R. Bernemann, T. Weis, B. König and V. Matkovic, "Lifecycle-Based View on Cyber-Physical System Models Using Extended Hidden Markov Models," 2022 20th ACM-IEEE International Conference on Formal Methods and Models for System Design (MEMOCODE), Shanghai, China, 2022, pp. 1-5, doi: 10.1109/MEMOCODE57689.2022.9954592.

[41] T. Mladenova and I. Valova, "Comparative analysis between the traditional K-Nearest Neighbor and Modifications with Weight-Calculation," 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISM SIT), Ankara, Turkey, 2022, pp. 961-965, doi: 10.1109/ISMST56059.2022.9932693.

[42] F. Borges et al., "An Unsupervised Method based on Support Vector Machines and Higher-Order Statistics for Mechanical Faults Detection," in *IEEE Latin America Transactions*, vol. 18, no. 06, pp. 1093-1101, Jun 2020, doi: 10.1109/TLA.2020.9099687.

[43] S. Mitrofanov and E. Semekin, "An Approach to Training Decision Trees with the Relearning of Nodes," 2021 International Conference on Information Technologies (InfoTech), Varna, Bulgaria, 2021, pp. 1-5, doi: 10.1109/InfoTech52438.2021.9548520.

[44] D. Sotnikov, M. Lyly and T. Salmi, "Prediction of 2G HTS Tape Quench Behavior by Random Forest Model Trained on 2-D FEM Simulations," in *IEEE*

Transactions on Applied Superconductivity, vol. 33, no. 5, pp. 1-5, Aug. 2023, Art no. 6602005, doi: 10.1109/TASC.2023.3262212.

[45] K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[46] Akhil, Jabbar, Meerja., A., Ashu., Aluvalu, Rajani, Kanth. (2019). Gaussian Naïve Bayes Based Intrusion Detection System. doi: 10.1007/978-3-030-49345-5_16

[47] Junkai, Yi., Guanglin, Gong., Zeyu, Liu., Yacong, Zhang. (2021). Classification of Markov Encrypted Traffic on Gaussian Mixture Model Constrained Clustering. Wireless Communications and Mobile Computing, 2021:1-11. doi: 10.1155/2021/4935108.

[48] Ke, Cheng., Liangmin, Wang., Yulong, Shen., Hua, Wang., Yongzhi, Wang., Xiaohong, Jiang., Hong, Zhong. (2021). Secure \$k\$-NN Query on Encrypted Cloud Data with Multiple Keys. IEEE Transactions on Big Data, doi: 10.1109/TBDATA.2017.2707552

[49] RAMRAJ S and Usha G, "Unsupervised Feature Learning Methodology for Tree based Classifier and SVM to Classify Encrypted Traffic" International Journal of Advanced Computer Science and Applications(IJACSA), 14(2), 2023. <http://dx.doi.org/10.14569/IJACSA.2023.01402102>

[50] Jiahao, Dai., Yongqun, Chen., Yucheng, Chen., Aodong, Meng. (2021). An analysis of Network Traffic Identification based on Decision Tree. doi: 10.1109/AIEA53260.2021.00072

[51] Y. Ren, X. Zhu, K. Bai and R. Zhang, "A New Random Forest Ensemble of Intuitionistic Fuzzy Decision Trees," in IEEE Transactions on Fuzzy Systems, vol. 31, no. 5, pp. 1729-1741, May 2023, doi: 10.1109/TFUZZ.2022.3215725.

[52] A. Gupta, "VPN-nonVPN Traffic Classification Using Deep Reinforced Naive Bayes and Fuzzy K-means Clustering," 2021 IEEE 41st International Conference on Distributed Computing Systems Workshops (ICDCSW), Washington, DC, USA, 2021, pp. 1-6, doi: 10.1109/ICDCSW53096.2021.00008.

[53] Y. Lou, R. Wu, J. Li, L. Wang, X. Li and G. Chen, "A Learning Convolutional Neural Network Approach for Network Robustness Prediction," in

IEEE Transactions on Cybernetics, vol. 53, no. 7, pp. 4531-4544, July 2023, doi: 10.1109/TCYB.2022.3207878.

[54] T. -L. Huoh, Y. Luo, P. Li and T. Zhang, "Flow-Based Encrypted Network Traffic Classification With Graph Neural Networks," in IEEE Transactions on Network and Service Management, vol. 20, no. 2, pp. 1224-1237, June 2023, doi: 10.1109/TNSM.2022.3227500.

[55] M. Saad, N. Munir and C. -H. Kim, "Pattern Recognition Based Auto-Reclosing Scheme Using Bi-Directional Long Short-Term Memory Network," in IEEE Access, vol. 10, pp. 119734-119744, 2022, doi: 10.1109/ACCESS.2022.3221818.

[56] Thilini, Dahanayaka., Guillaume, Jourjon., Suranga, Seneviratne. (2020). Understanding Traffic Fingerprinting CNNs. doi: 10.1109/LCN48667.2020.9314785

[57] Lu, Yifei., Li, Yan. (2020). Encrypted traffic identification method based on CNN and SAE.

[58] Tan, Xiaobin., Tong, Xinxin., Chen, Xiang., Yang, Jian., Zhang, Yongdong. (2020). Online encrypted traffic classification method based on CNN and LSTM.

[59] Xinyi, Hu., Chunxiang, Gu., Fushan, Wei. (2021). CLD-Net: A Network Combining CNN and LSTM for Internet Encrypted Traffic Classification. Security and Communication Networks, doi: 10.1155/2021/5518460

[60] Yan, Li., Yifei, Lu. (2021). ETCC: Encrypted Two-Label Classification Using CNN. Security and Communication Networks, doi: 10.1155/2021/6633250

[61] Neha, Gupta., Vinita, Jindal., Punam, Bedi. (2022). Encrypted Traffic Classification Using eXtreme Gradient Boosting Algorithm. doi: 10.1007/978-981-16-3071-2_20

[62] Amit, Roy., Kashob, Kumar, Roy., Amin, Ahsan, Ali., M., Ashraful, Amin., A, K, M, Mahbubur, Rahman. (2021). SST-GNN: Simplified Spatio-temporal Traffic forecasting model using Graph Neural Network. doi: 10.1007/978-3-030-75768-7_8

[63] Xiaorong, Hu., Chenguang, Zhao., Gang, Wang. (2020). A Traffic Light Dynamic Control Algorithm with Deep Reinforcement Learning Based on GNN Prediction.. arXiv: Learning,

[64] Oliver, Hope., Eiko, Yoneki. (2021). GDDR: GNN-based Data-Driven Routing. doi: 10.1109/ICDCS51616.2021.00056

[65] Xue-Cheng, Shang., Xingang, Li., Dong-Fan, Xie., Bin, Jia., Rui, Jiang., Feng, Liu. (2022). A data-driven two-lane traffic flow model based on cellular automata. *Physica A-statistical Mechanics and Its Applications*, doi: 10.1016/J.PHYSA.2021.126531

[66] Zihao, Wang., Kar, Wai, Fok., Vrizlynn, L., L., Thing. (2022). Machine Learning for Encrypted Malicious Traffic Detection: Approaches, Datasets and Comparative Study. *Computers & Security*, doi: 10.1016/J.COSE.2021.102542

[67] Weiwei, Fang., Wenhao, Zhuo., Jingwen, Yan., Youyi, Song., Dazhi, Jiang., Teng, Zhou. (2022). Attention meets long short-term memory: A deep learning network for traffic flow forecasting. *Physica A-statistical Mechanics and Its Applications*, doi: 10.1016/J.PHYSA.2021.126485

[68] Zekuan, Chen., Xiong, Luo., Ting, Wang., Weiping, Wang., Wenbing, Zhao. (2022). Deep Reinforcement Learning-Based LSTM Model for Traffic Flow Forecasting in Internet of Vehicles. doi: 10.1007/978-981-16-6372-7_57

[69] Wanpeng, Wu., Wuneng, Zhou., You, Wu., Xuan, Gu., Guang, Yang. (2022). A Novel Deep Learning Model Based on the Improved Loss Function for Traffic Flow Forecasting. doi: 10.1007/978-981-16-6324-6_27

[70] M. Shen et al., "Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey," in *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 791-824, Firstquarter 2023, doi: 10.1109/COMST.2022.3208196.

[71] Gerard Drapper Gil, Arash Habibi Lashkari, Mohammad Mamun, Ali A. Ghorbani, "Characterization of Encrypted and VPN Traffic Using Time-Related Features", In *Proceedings of the 2nd International Conference on Information Systems Security and Privacy(ICISSP 2016)* , pages 407-414, Rome, Italy.

[72] WireShark.AboutWireshark.[EB/OL].<https://www.wireshark.org/about.html>.2023-8-15.

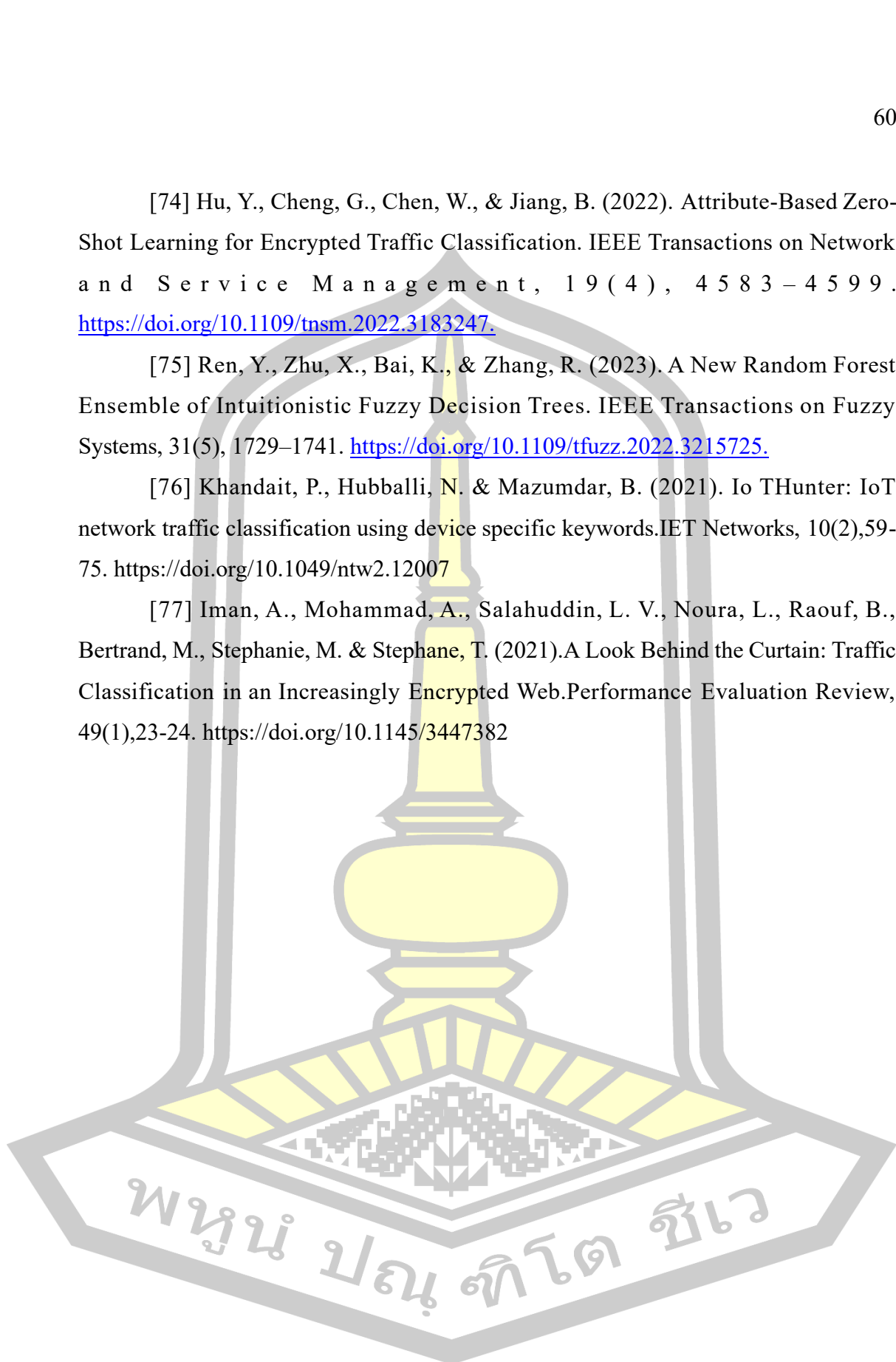
[73] Shen, M., Ye, K., Liu, X., Zhu, L., Kang, J., Yu, S., Li, Q., & Xu, K. (2023). Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials*, 25(1), 791–824. <https://doi.org/10.1109/comst.2022.3208196>.

[74] Hu, Y., Cheng, G., Chen, W., & Jiang, B. (2022). Attribute-Based Zero-Shot Learning for Encrypted Traffic Classification. *IEEE Transactions on Network and Service Management*, 19(4), 4583–4599. <https://doi.org/10.1109/tnsm.2022.3183247>.

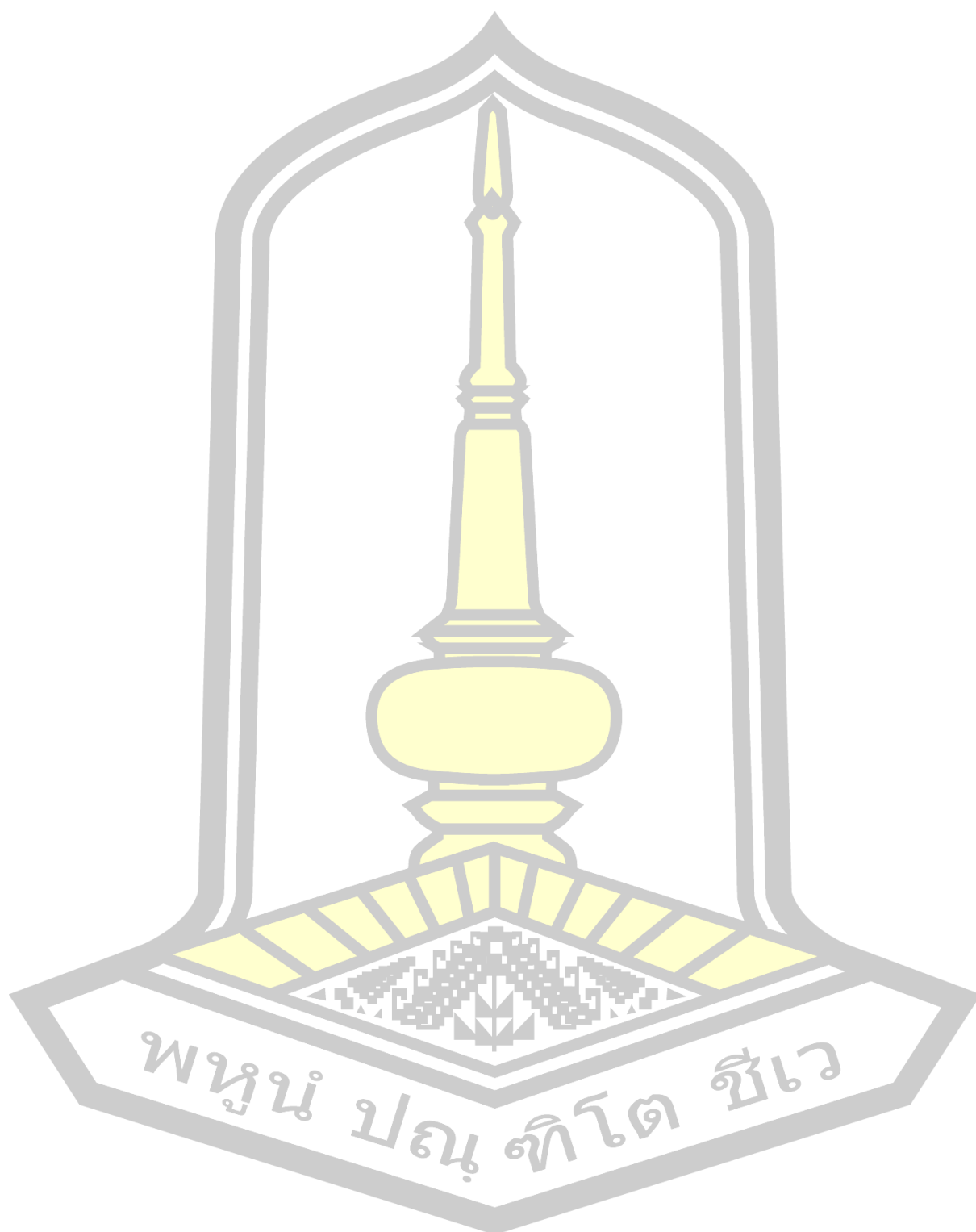
[75] Ren, Y., Zhu, X., Bai, K., & Zhang, R. (2023). A New Random Forest Ensemble of Intuitionistic Fuzzy Decision Trees. *IEEE Transactions on Fuzzy Systems*, 31(5), 1729–1741. <https://doi.org/10.1109/tfuzz.2022.3215725>.

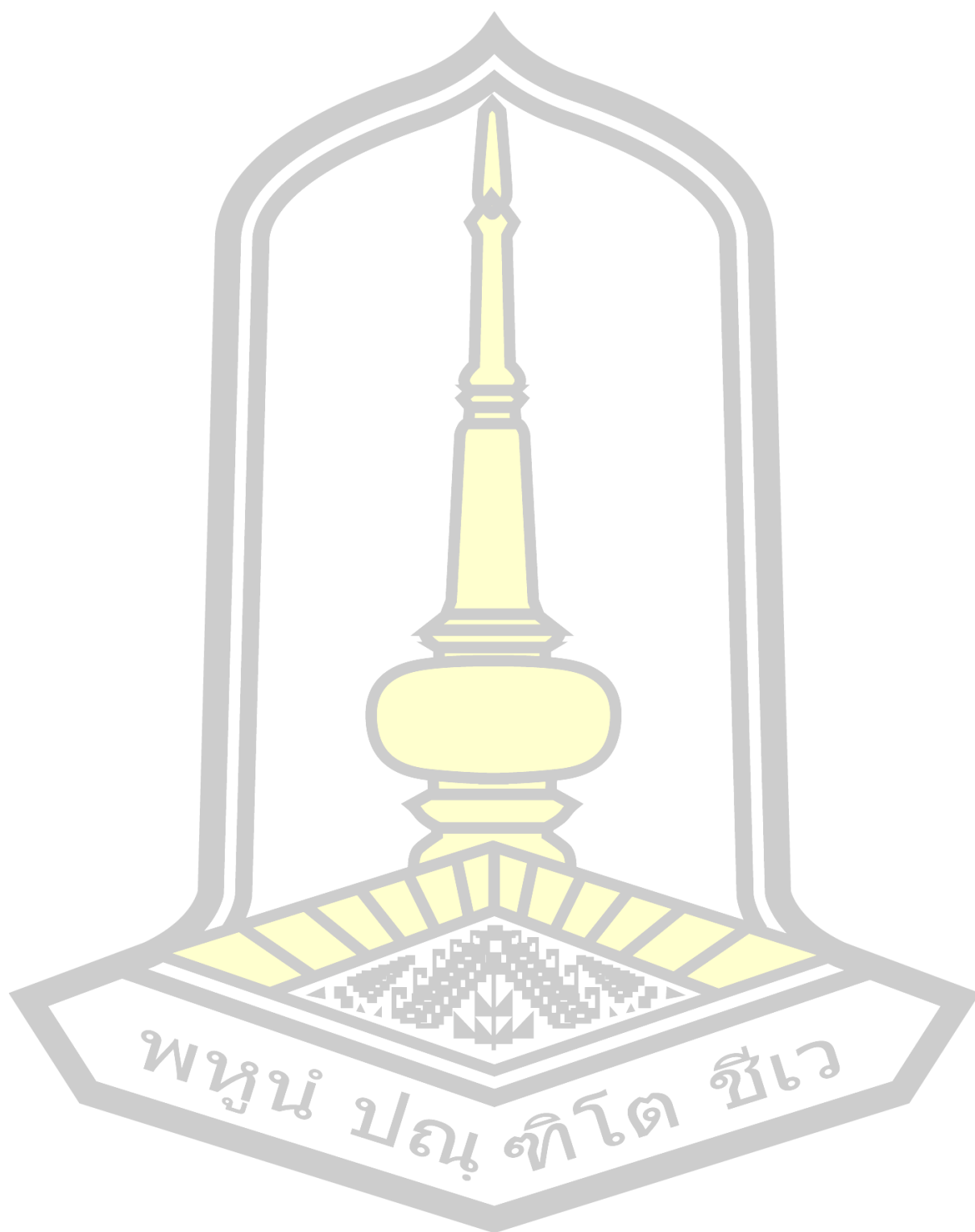
[76] Khandait, P., Hubballi, N. & Mazumdar, B. (2021). Io THunter: IoT network traffic classification using device specific keywords. *IET Networks*, 10(2), 59-75. <https://doi.org/10.1049/ntw2.12007>

[77] Iman, A., Mohammad, A., Salahuddin, L. V., Noura, L., Raouf, B., Bertrand, M., Stephanie, M. & Stephane, T. (2021). A Look Behind the Curtain: Traffic Classification in an Increasingly Encrypted Web. *Performance Evaluation Review*, 49(1), 23-24. <https://doi.org/10.1145/3447382>



REFERENCES





BIOGRAPHY

NAME	Xiaoqing Yang
DATE OF BIRTH	17/10/1990
PLACE OF BIRTH	China
ADDRESS	ShanXi Province
POSITION	TaiYuan
PLACE OF WORK	TaiYuan
EDUCATION	2006.09-2009.06 China, Lilin High School, High School 2009.09-2013.07 China, North University of China, Bachelor's Degree 2013.09-2016.07 China, North University of China, Master's Degree

