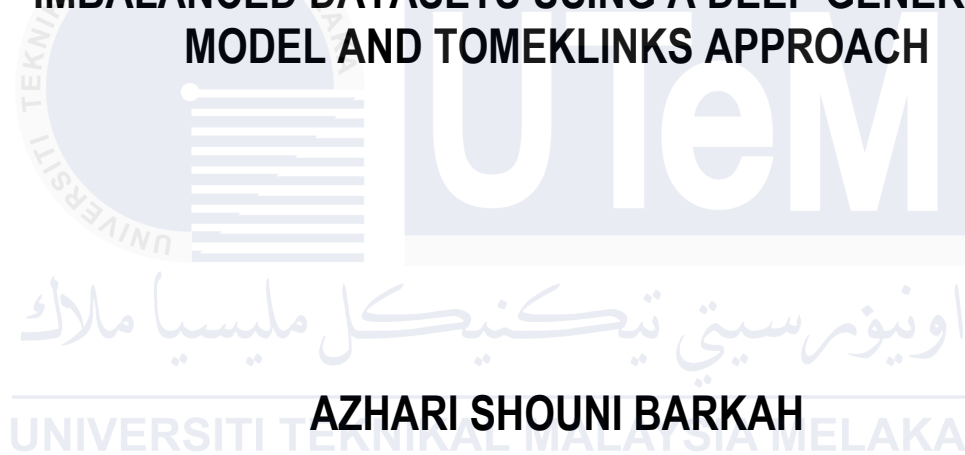




**ENHANCING ANOMALY DETECTION PERFORMANCE IN
IMBALANCED DATASETS USING A DEEP GENERATIVE
MODEL AND TOMEKLINKS APPROACH**



AZHARI SHOUNI BARKAH

DOCTOR OF PHILOSOPHY

2025



Faculty of Information and Communication Technology

**ENHANCING ANOMALY DETECTION PERFORMANCE IN
IMBALANCED DATASETS USING A DEEP GENERATIVE MODEL
AND TOMEKLINKS APPROACH**

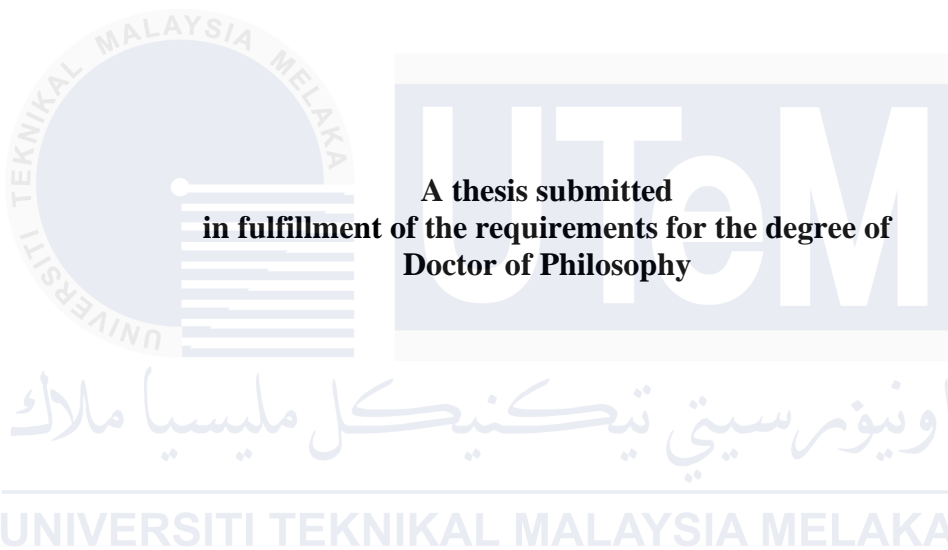
اونيورسيتي تيكنيكل مليسيا ملاك
Azhari Shouni Barkah
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Doctor of Philosophy

2025

**ENHANCING ANOMALY DETECTION PERFORMANCE IN IMBALANCED
DATASETS USING A DEEP GENERATIVE MODEL AND TOMEKLINKS
APPROACH**

AZHARI SHOUNI BARKAH



Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2025

DECLARATION

I declare that this thesis entitled “Enhancing Anomaly Detection Performance in Imbalanced Datasets Using A Deep Generative Model and TomekLinks Approach “ is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.



Signature :

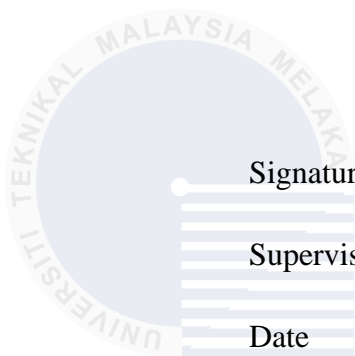
Name : Azhari Shouni Barkah

Date : 8 May 2025

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Doctor of Philosophy.



Signature

.....

Supervisor Name

: Assoc. Prof. Dr.Ts. Siti Rahayu binti Selamat

Date

: 8 May 2025

اونيورسيتي تېكنيكل مليسيا ملاك

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

DEDICATION

In the name of Allah, the most Gracious, the most beneficent, the most merciful

Al-hamdulillahi rabbil ‘alamin

All the praises and thanks be to Allah, the Lord of the ‘Alamin.

Sacrificed being made, steps taken, time passed.

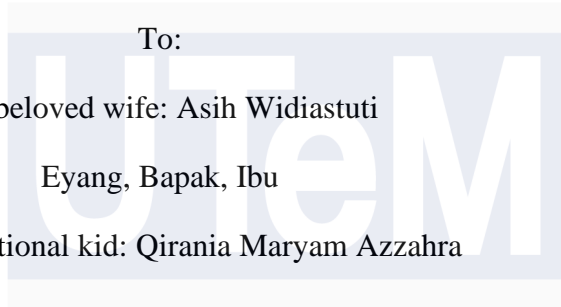
Finally, light at the end of the long tunnel.

To:

My beloved wife: Asih Widiastuti

Eyang, Bapak, Ibu

My inspirational kid: Qirania Maryam Azzahra



اونيورسيتي تيكنيكل مليسيا ملاك

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

ABSTRACT

Data imbalance is a problem in machine learning. Unbalanced classes cause a common problem in machine classification, where there is a disproportionate ratio within each class. Data imbalance results in a decrease in model quality, where the model can provide high accuracy but only applies to the majority of data and ignores minority data. Many techniques are used to deal with class imbalance problems, namely the resampling technique, which includes oversampling and undersampling. Both of these techniques aim to change the ratio between the majority and minority classes. By making the training data more balanced, resampling allows different classes to have relatively the same effect on the results of the classification model. The oversampling technique is used because of the independence of the classifier, especially with random oversampling and synthetic minority oversampling. However, this technique causes overfitting problems because random oversampling only duplicates the minority data class. Besides that, it also increases data training time. The overlapping problem caused by synthetic minority oversampling is solved by using an approach based on local information, not on the distribution of the minority class as a whole, in synthesizing new data. Besides that, it also causes data noise in the samples because the separation between the majority and minority class groups is not clear. Aiming to address the problem of dataset imbalance that improves the performance of anomaly detection in detecting new and rare attacks, this research proposes an enhanced ANIDS model called as DGT-RF using a Conditional Generative Adversarial Network (CGAN) combine with TomekLinks and Random Forest as classifier. According to test and evaluation reports, DGT-RF has proven successful in increasing the performance of anomaly detection to detect new and rare attacks on extreme imbalance minority classes. The validation results show that this model outperforms previous work by an average of 7.62% accuracy. In the future, aiming to improve the performance in detecting new and rare attacks, the use of techniques like data balancing other variants of synthetic data based on deep learning will need to be considered.

MENINGKATKAN PRESTASI PENGESANAN ANOMALI DALAM SET DATA TIDAK SEIMBANG MENGGUNAKAN MODEL GENERATIF MENDALAM DAN PENDEKATAN TOMEKLINKS

ABSTRAK

Ketidakseimbangan data adalah masalah dalam pembelajaran mesin. Kelas yang tidak seimbang menyebabkan masalah biasa dalam klasifikasi mesin, di mana terdapat nisbah yang tidak seimbang dalam setiap kelas. Ketidakseimbangan data mengakibatkan penurunan kualiti model, di mana model boleh memberikan ketepatan yang tinggi tetapi hanya terpakai kepada majoriti data dan mengabaikan data minoriti. Banyak teknik digunakan untuk menangani masalah ketidakseimbangan kelas, iaitu teknik persampelan semula, yang merangkumi persampelan berlebihan dan persampelan kurang. Kedua-dua teknik ini bertujuan untuk mengubah nisbah antara kelas majoriti dan minoriti. Dengan menjadikan data latihan lebih seimbang, persampelan semula membolehkan kelas yang berbeza mempunyai kesan yang agak sama pada hasil model klasifikasi. Teknik persampelan berlebihan digunakan kerana kebebasan pengelas, terutamanya dengan persampelan berlebihan rawak dan persampelan berlebihan minoriti sintetik. Walau bagaimanapun, teknik ini menyebabkan masalah overfitting kerana persampelan berlebihan rawak hanya menduplikasi kelas data minoriti. Selain itu, ia juga meningkatkan masa latihan data. Masalah pertindihan yang disebabkan oleh persampelan berlebihan minoriti sintetik diselesaikan dengan menggunakan pendekatan berdasarkan maklumat tempatan, bukan pada taburan kelas minoriti secara keseluruhan, dalam mensintesis data baharu. Selain itu, ia juga menyebabkan gangguan data dalam sampel kerana pemisahan antara kumpulan kelas majoriti dan minoriti tidak jelas. Bertujuan untuk menangani masalah ketidakseimbangan set data yang meningkatkan prestasi pengesanan anomali dalam mengesan serangan baharu dan jarang berlaku, penyelidikan ini mencadangkan model ANIDS yang dipertingkatkan yang dipanggil sebagai DGT-RF menggunakan Conditional Generative Adversarial Network (CGAN) digabungkan dengan TomekLinks dan Random Forest sebagai pengelas. Menurut laporan ujian dan penilaian, DGT-RF telah terbukti berjaya meningkatkan prestasi pengesanan anomali untuk mengesan serangan baharu dan jarang berlaku terhadap kelas minoriti ketidakseimbangan yang melampau. Hasil pengesanan menunjukkan bahawa model ini mengatasi prestasi kerja sebelumnya dengan purata ketepatan 7.62%. Pada masa hadapan, bertujuan untuk meningkatkan prestasi dalam mengesan serangan baharu dan jarang berlaku, penggunaan teknik seperti mengimbangi data varian lain data sintetik berdasarkan pembelajaran mendalam perlu dipertimbangkan.

ACKNOWLEDGEMENT

In the Name of Allah, the Most Gracious, the Most Merciful. First and foremost, I would like to express my gratitude and sincere appreciation to my supervisor, Assoc. Prof. Dr.Ts. Siti Rahayu binti Selamat, through close interaction for strong support, encouragement, guidance and criticism continuously influenced my way of thinking about research and living life. The gratitude also goes to my co-supervisor, Dr. Zaheera Zainal Abidin, for her endless energy source for sharing knowledge, expertise and experiences.

I want to extend my special appreciation to Dr. Berlilana, M.Kom., M.Si., who provided this opportunity, resources and support to make it happen. I also thank all staff from the Faculty of Information and Communication Technology and the International Office in UTeM and Universitas AMIKOM Purwokerto for providing the excellent service and a comfortable environment during my study.

I am also grateful to all the people that I had the pleasure to meet during my stay at Melaka: Fandy Setyo Utomo, Rujianto Eko Saputro, Giat Karyono, Purwadi and all my fellow postgraduate students for the discussions, support, and friendships.

I would especially like to thank my wife (Asih Widiastuti), my daughter (Qirania Maryam Azzahra), and my parents for always supporting, motivating and inspiring me. May the Almighty Allah always guide you in every step you take, bless you with more than enough, and make things easy for you when life gets rough.

TABLE OF CONTENTS

	PAGES
DECLARATION	
APPROVAL	
DEDICATION	
ABSTRACT	i
ABSTRAK	ii
ACKNOWLEDGEMENT	iii
TABLE OF CONTENTS	iv
LIST OF TABLES	vii
LIST OF FIGURES	x
LIST OF ABBREVIATIONS	xii
LIST OF PUBLICATIONS	xiii
 CHAPTER	
 1. INTRODUCTION	 14
1.1 Background	14
1.2 Research Problem	20
1.3 Research Question	24
1.4 Research Objective	26
1.5 Scope of Research	29
1.6 Research Contribution	29
1.7 Thesis Outline	30
1.8 Summary	32
 2. LITERATURE REVIEW	 33
2.1 Introduction	33
2.2 Anomaly Network Intrusion Detection Systems (ANIDS) Model	34
2.2.1 Analysis ANIDS Model	42
2.3 Data Imbalance in Intrusion Detection	47
2.3.1 Definition of Data Imbalance in ANIDS	48
2.3.2 Data Imbalance for Detecting New and Rare Attacks	50
2.3.3 Imbalance Datasets Issues	51
2.3.4 Analysis of Data Imbalance Issues in ANIDS	55
2.4 Current Approach to Handling Data Imbalance in ANIDS	56
2.4.1 Random Oversampling	59
2.4.2 Random Undersampling	60
2.4.3 Machine learning-based Synthetics Data	62
2.4.4 Deep Learning-based Synthetics data	75
2.4.5 Analysis of Current Approach Limitations for Handling Data Imbalance in ANIDS	85
2.5 Hyper-parameter optimization	87
2.5.1 Analysis of Hyperparameter Optimization	89

2.6	Feature Selection for ANIDS	90
2.7	Overall Analysis	93
2.7.1	Analysis of ANIDS Model to Handle Data Imbalance	93
2.7.2	Analysis of Current Approach to Handling Data Imbalance	97
2.8	Proposed Solution on ANIDS to handle imbalance data in Detection New and Rare Attack	99
2.9	Proposed Detection Model Evaluation and Validation Approach	101
2.9.1	Cross-Validation	102
2.9.2	Performance Measurement	102
2.9.3	Statistical Test	104
2.10	Summary	105
3.	METHODOLOGY	107
3.1	Introduction	107
3.2	Research Methodology	108
3.2.1	Literature Study	109
3.2.2	Data Collection	112
3.2.3	Data Analysis	113
3.3	Preliminary Analysis	114
3.3.1	Impact of Data Balancing and Feature Selection on Machine learning	114
3.3.2	Detecting imbalanced anomalies CICIDS2017 IDS Dataset using Data Generative Model	124
3.3.3	Preliminary Results	131
3.4	Development of the Proposed ANIDS to handle data imbalance and detect rare and new attacks	131
3.5	Research Design	133
3.5.1	Data Preprocessing	133
3.5.2	Splitting data and removing normal data	136
3.5.3	Data Imbalance Handling	136
3.5.4	Classification Processes	143
3.6	Validation and Evaluation Model	144
3.5.5	Cross-validation	144
3.5.6	Performance metrics	146
3.5.7	Statistical Test	147
3.7	Summary	149
4.	IMPLEMENTATION OF THE PROPOSED DGT-RF	151
4.1	Introduction	151
4.2	Experimental environment	151
4.2.1	Research Tool and Research Equipment	152
4.2.2	Dataset	153
4.3	Data Preprocessing	157
4.3.1	Data Cleaning	157
4.3.2	Feature Selection	158
4.3.3	Normalization and Standardization	160
4.3.4	Splitting Dataset	162

4.4	Data Imbalance Handling	166
4.5	Classification Process	170
4.6	Summary	174
5.	RESULT EVALUATION AND VALIDATION	176
5.1	Introduction	176
5.2	Dataset for Evaluation and Validation	176
5.3	Evaluation and Validation Procedures	177
5.4	Result of Evaluation and Validation	178
5.4.1	Evaluation and Comparison of the DGT-RF with Previous Model	178
5.4.2	Benchmarking DGT-RF using CICIDS2018 Dataset	184
5.4.3	Hyperparameter Optimization	185
5.5	Analysis of result validation	191
5.5.1	Comparison of the DGT-RF with previous Model	191
5.5.2	Benchmarking DGT-RF using CICIDS2018	192
5.5.3	Optimizer hyperparameter	192
5.5.4	Learning rate hyperparameter	194
5.5.5	Batch Size hyperparameter	195
5.6	Discussion of Result Validation	196
5.6.1	Validation Model	197
5.6.2	Statistical Test	198
5.7	Summary	206
6.	CONCLUSION AND RECOMMENDATIONS FOR FUTURE RESEARCH	208
6.1	Introduction	208
6.2	Summary of the Completed Work	208
6.3	Research Contributions	209
6.4	Research Constrains and Limitations	210
6.5	Future Works	212
6.6	Summary	213

LIST OF TABLES

TABLE	TITLE	PAGE
Table 1.1	Summary of the research problem	24
Table 1.2	Summary of research questions	26
Table 1.3	Summary of research problems (RP), research questions (RQ) and research objectives (RO)	28
Table 1.4	Summary of research contributions	30
Table 2.1	ANIDS Model Current Approach	38
Table 2.2	Summary comparison machine learning algorithms for IDS	43
Table 2.3	Summary of Dataset Comparison	52
Table 2.4	Extremely rare classes in datasets.	55
Table 2.5	Comparison oversampling machine learning-based synthetic data.	63
Table 2.6	Comparison deep learning-based synthetics data	76
Table 2.7	Summary of resampling methods	86
Table 2.8	Comparison of feature selection method for ANIDS	91
Table 2.9	ANIDS Model to Handle Data Imbalance	93
Table 3.1	Inclusion and Exclusion Criteria	110
Table 3.2	Characteristic of the datasets	112
Table 3.3	Result of classification on baseline	116
Table 3.4	The result of classification with balancing data SMOTE	117
Table 3.5	Result of Classification Result with Balancing Data ADASYN	117
Table 3.6	Result of Balancing Dataset Using SMOTE with Feature Selection	118
Table 3.7	Result of Balancing Dataset Using ADASYN with Feature Selection	118
Table 3.8	Result of Imbalanced Data with RFE	120

Table 3.9	Result of Classification using RFE+SMOTE	120
Table 3.10	Result of Classification Using RFE+ADASYN	121
Table 3.11	The result of Classification using CICIDS2017 Dataset	127
Table 3.12	Comparison results of weighted F1-score in the CICIDS2017 dataset.	128
Table 3.13	Accuracy result of existing methods	149
Table 4.1	Hardware Description	152
Table 4.2	Software Description	152
Table 4.3	Distribution of CICIDS2017 dataset	154
Table 4.4	Distribution of CSI-CICIDS2018 dataset	156
Table 4.5	Distribution of CICIDS2017 training and testing dataset	162
Table 4.6	Distribution of training and testing CICIDS2018 Dataset	163
Table 4.7	Distribution training and testing set CICIDS2017 after removing normal data.	164
Table 4.8	Distribution training and testing set CICIDS2018 after removing normal data	165
Table 4.9	Distribution of upsampling strategy CICIDS2017 training dataset	167
Table 4.10	Distribution of upsampling strategy for CICIDS2018 training dataset	168
Table 4.11	Tuning Hyperparameter of CGAN Model	170
Table 5.1	Comparison result ACURACY for CICIDS2017 dataset with 10x iteration using cross-validation value of k=10	179
Table 5.2	Comparison result accuracy dan weighted f1-score	179
Table 5.3	Recall test comparison results using cross-validation, with a value of k = 10.	181
Table 5.4	Comparison results with previous research	183
Table 5.5	Comparison results of accuracy and weighted F1-score the CICIDS2018.	184

Table 5.6	Comparison result optimizer hyperparameter SGD, Adam and RMSProp	186
Table 5.7	Comparison Result Learning Rate 0.1, 0.01, 0.001, 0.0001, 0.00001 with SGD Optimizer.	188
Table 5.8	Experiment result with comparing batch size 32, 64, 128 with SGD optimizer.	190
Table 5.9	Comparison results 10-fold cross-validation with SGD, Adam, RMSProp optimizer	193
Table 5.10	Comparison Result 10-fold cross-validation for F1-Score with Difference Learning Rate Value.	194
Table 5.11	Average of accuracy DGT-RF compared to other popular methods	199
Table 5.12	ANOVA Test	199
Table 5.13	Comparison of DGT-RF model Post-Hoc, Analysis against SMOTE-RF, SMOTEENN-RF and Borderline-SMOTE-RF	200
Table 5.14	Average Recall on rare class	201
Table 5.15	ANOVA Test for heartbleed attack class Recall	202
Table 5.16	ANOVA Test for Web_Attack_SQL_Injection class Recall	202
Table 5.17	ANOVA Test of Infiltration class Recall	202
Table 5.18	ANOVA Test of Web_Attack_XSS class Recall	203
Table 5.19	ANOVA Test of Web_Attack_Brute_Force class Recall	203
Table 5.20	ANOVA Test of Bot class Recall	203
Table 5.21	Multiple comparison between data balancing methods with 95% Confidence interval	204
Table 5.22	Comparison summary of DGT-RF with previous Model	205

LIST OF FIGURES

FIGURE	TITLE	PAGE
Figure 1.1	Major motivation of attacks (Source: Hackmageddon cybercrime statistics 2022)	14
Figure 1.2	Top 10 attack techniques (source: hackmageddon cybercrime statistics 2022)	15
Figure 1.3	Yearly comparisons of complaints received via the IC3 website (FBI 2022)	16
Figure 2.1	Literature Review Phase	34
Figure 2.2	ANIDS based on Machine Learning and Deep Learning Methodology (Ahmad et al. 2021)	36
Figure 2.3	ANIDS Handling Data Imbalance Methodology	43
Figure 2.4	Taxonomy of balancing data approaches	56
Figure 3.1	Research framework	108
Figure 3.2	Research framework balancing before RFE	114
Figure 3.3	Research framework RFE before imbalance handling	115
Figure 3.4	Accuracy result on the initial experimental scenario	116
Figure 3.5	Accuracy classification of second scenario	119
Figure 3.6	Research Framework Data Generative Model	125
Figure 3.7	Design of Proposed Model	132
Figure 3.8	Conditional Generative Adversarial Network (CGAN) Architecture	141
Figure 3.9	K-Fold cross-validation procedure	145
Figure 3.10	Confusion matrix	146
Figure 4.1	Data cleaning algorithm	158
Figure 4.2	Feature selection algorithm	159
Figure 4.3	Normalization and standardization algorithm	161

Figure 4.4	Splitting data and removing normal data after normalization and standardization data	162
Figure 4.5	Handling imbalance data algorithm	169
Figure 4.6	Training Model Algorithm	172
Figure 4.7	Testing Model Algorithm	173
Figure 5.1	Evaluation and Validation Procedures	177



LIST OF ABBREVIATIONS

<i>UTeM</i>	-	Universiti Teknikal Malaysia Melaka
<i>CGAN</i>	-	Conditional Generative Adversarial Network
<i>ANIDS</i>	-	Anomaly Network Intrusion Detection System
<i>SMOTE</i>	-	Synthetic Minority Oversampling Technique
<i>DGT-RF</i>	-	Deep Generative Model TomekLinks Random Forest
<i>ADASYN</i>	-	Adaptive Synthetic Sampling
<i>ROS</i>	-	Random Oversampling

اونيورسيتي تېكنيكل مليسيا ملاك

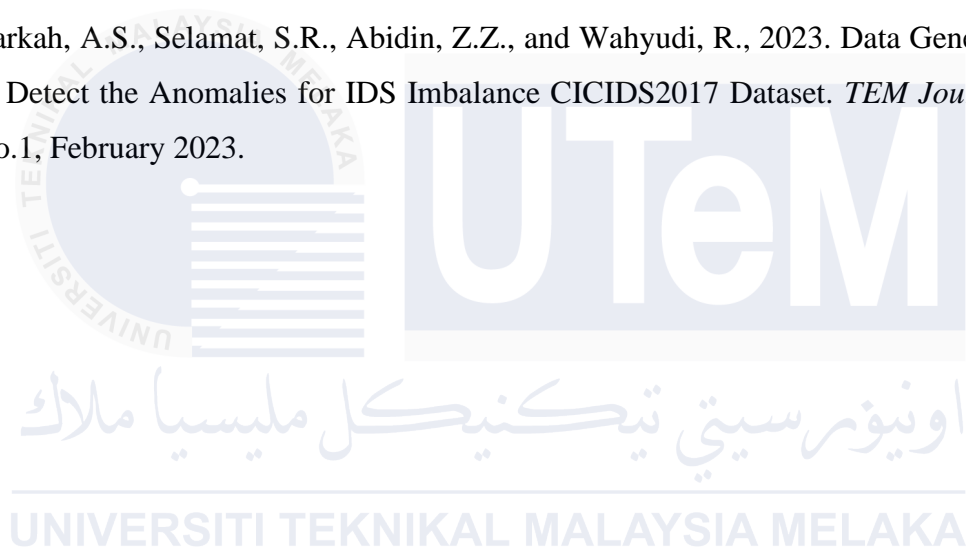
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF PUBLICATIONS

The followings are the list of publications related to the work on this thesis:

Barkah, A.S., Selamat, S.R., Abidin, Z.Z., and Wahyudi, R., 2023. Impact of Data Balancing and Feature Selection on Machine Learning-based Network Intrusion Detection. *International Journal on Informatics Visualization*. Vol. 7, No. 1, March 2023.

Barkah, A.S., Selamat, S.R., Abidin, Z.Z., and Wahyudi, R., 2023. Data Generative Model to Detect the Anomalies for IDS Imbalance CICIDS2017 Dataset. *TEM Journal*. Vol. 12, No.1, February 2023.



CHAPTER 1

INTRODUCTION

1.1 Background

After land, sea, and air, cyberspace is considered an area that needs to be explored and understood (Karie et al., 2019). The primary cause of this phenomenon is the intermittent surge in cybercrime and the proliferation of cybercriminals. The increase in cybercrime has been driven by the expansion of technology and the internet. According to cyber-attack statistics from Hackmageddon in 2022, cybercrime continues to be the main motivation for cyber-attacks, which contributed 76.8%, although it tends to decrease from 2021, which contributed 84.1%, while cyber espionage is the second contributor at 10.4%, and hacktivism has increased from 1.3% in 2021 to 7% in 2022, as depicted in Figure 1.1.

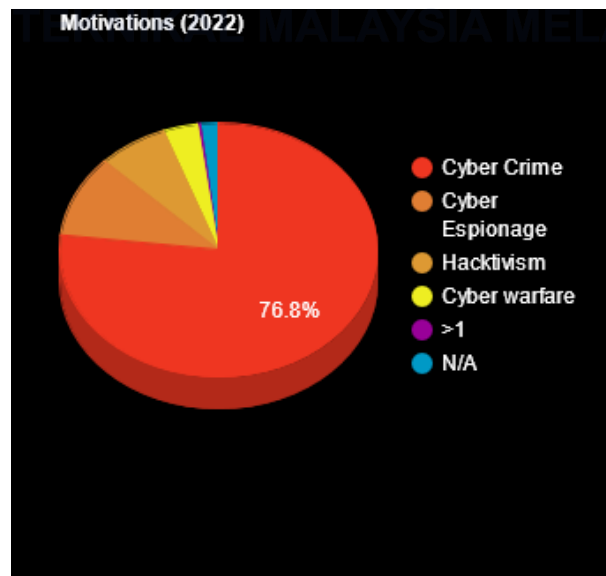


Figure 1.1 Major motivation of attacks (Source: Hackmageddon cybercrime statistics 2022)

Malware attacks continue to be the most frequent cyberattacks, accounting for 34.7% of all attacks, with unknown attacks coming in at 22.2% and phishing attacks at 15.5%, as shown in Figure 1.2.

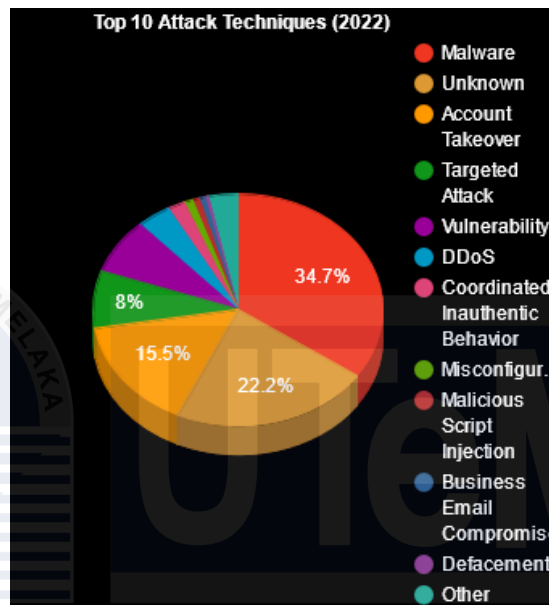


Figure 1.2 Top 10 attack techniques (source: hackmageddon cybercrime statistics 2022)

According to Steve Morgan (2022), founder of Cybersecurity Ventures, losses due to global cybercrime will increase by 15% per year over the next five years. Losses in 2023 are expected to reach 8 trillion dollars. In 2025, it is estimated that it will be 10.5 trillion dollars, up significantly from 3 trillion dollars in 2015. In the United States, the number of complaints received and processed by the Federal Bureau of Investigation (FBI) (2022) in collaboration with the Internet Crime Complaint Centre (IC3) is increasing, as depicted in Figure 1.3. In 2022, IC3 received a total of 800,944 complaints, with losses reaching 10.3 billion dollars. This shows a significant increase compared to 2018, which received 351,937 complaints, with losses reaching 2.7 billion dollars.

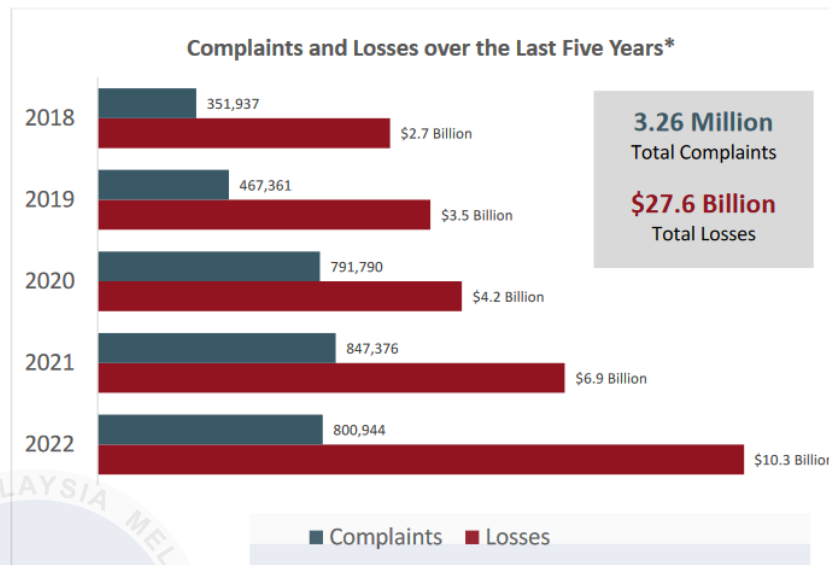


Figure 1.3 Yearly comparisons of complaints received via the IC3 website (FBI 2022)

In addition, based on the Cisco Internet Annual Report 2018–2023 (Cisco and Internet, 2020), by 2023, it is projected that almost two-thirds of the world's population will have access to the Internet. The number of internet users is projected to reach 5.3 billion by 2023, increasing from 3.9 billion in 2018. By 2023, the global population will be outnumbered threefold by the number of devices linked to IP networks. The predicted number of network devices is expected to reach 29.3 billion in 2023, an increase from 18.4 billion in 2018. IoT devices, smart devices, and home applications are growing at a rapid pace of 48% in 2023. Plus, by 2023, more than 70% of the global population will have cellular connectivity, with the global cellular subscriber population growing from 5.1 billion in 2018 to 5.7 billion in 2023. Therefore, with the increasing number of users and devices connected to the internet, computer networks have become an infrastructure that is inseparable from human life. Its role has evolved beyond being solely a digital information exchange platform and now offers several crucial services to its users. Cybercriminals are

attracted to individuals and organizations that heavily rely on computer networks due to the potential for financial gain. Cybercriminals attempt to undermine the secrecy, accuracy, and accessibility of data and online services by executing diverse network intrusions. It is crucial to identify the origin of intrusions in order to ensure network security and minimize the frequency of attacks.

As a result, an intrusion detection system (IDS) was constructed to identify the intrusion. IDS is an important component of network security (Li et al., 2015; Kwon et al., 2019). The main purpose of IDS is to detect anomalous activities and attempts caused by attackers in computer networks and computer systems; in other words, IDS monitors and analyses network traffic to separate normal and malicious data (Denning, 1987; Kwon et al., 2019; Liu and Lang, 2019; Vinayakumar et al., 2019; Ahmad et al., 2021; Kocher and Kumar, 2021). To achieve these goals, designing and implementing an IDS is a major challenge (Salama et al., 2011; Li et al., 2015; Kwon et al., 2019). Some intrusion detection techniques used to implement IDS include statistically based anomalies, pattern matching, data mining, machine learning, and deep learning (Kwon et al., 2019; Liu and Lang, 2019; Kocher and Kumar, 2021).

IDS application is divided into two (Khraisat et al., 2019; Kwon et al., 2019; Singh and Khare, 2022). First is an application on hardware connected to a network called a host-based detection system (HIDS). Second, an application on a network is used to detect intrusions, and this application is known as a network-based intrusion detection system (NIDS). This technology gathers and analyzes network traffic to identify and detect malicious assaults. Implementing NIDS provides a significant advantage in its ability to monitor data traffic from various devices on the network. In detecting malicious network

traffic, NIDS applies two models, namely signature-based (SNIDS) and anomaly-based (ANIDS) (Ahmed and Garcia, 2005; Salama et al., 2011; Li et al., 2015; Niyaz et al., 2015; Khraisat et al., 2019). SNIDS stores records of established attack patterns referred to as signatures. The system detects network traffic by comparing it with pre-existing signatures and triggers an alarm whenever there is a match. While this system may provide a small number of incorrect alerts, it is unable to detect novel forms of network intrusions that do not have predefined patterns stored in the attack database. ANIDS first establishes a standard network traffic profile and subsequently compares the network traffic to this established profile. The system is able to detect both known and novel threats by flagging any change from the standard profile as an intrusion. Therefore, ANIDS is more effective in detecting known and unknown attacks (Ning and Jajodia, 2004; Deka et al., 2015; Khraisat et al., 2019; Kwon et al., 2019).

An ANIDS model has to be trained using data samples that follow real-world network traffic characteristics before it is implemented in the actual world. Except from rare instances that produce harmful data, most of this network traffic is benign—that is, regular. Developed from a real-world network with a mix of regular and malicious network traffic, the intrusion detection dataset. The intrusion detection data set has significant differences in the number of samples in different classes, causing the data set to be unbalanced. In an imbalanced data set, the class that has the majority of data from the sample is called the majority class. Intrusion detection data usually includes normal network traffic data and frequently occurring attacks. Meanwhile, the minority class consists of attack data samples that rarely occur.