UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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

Azhari Shouni Barkah

Doctor of Philosophy

2025

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

AZHARI SHOUNI BARKAH

A thesis submitted
in fulfillment of the requirements for the degree of
Doctor of Philosophy

اونيۏرسيتي تيكنيكل مليسيا ملاك
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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

UTeM
اونيۏرسيتي تيكنيكل مليسيا ملاك
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

اونيۏرسيتي تيكنيكل 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

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.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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 pengelasl. 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 pengesahan 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.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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

UTeM
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF TABLES

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

اونيۏرسيتي تيكنيكل مليسيا ملاك
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF FIGURES

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

UTeM
اونيۆرسيتي تيكنيكل مليسيا ملاك
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF ABBREVIATIONS

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

UTeM
اونيۏرسيتي تيكنيكل 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.

UTeM
اونيۏرسيتي تيكنيكل مليسيا ملاك
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

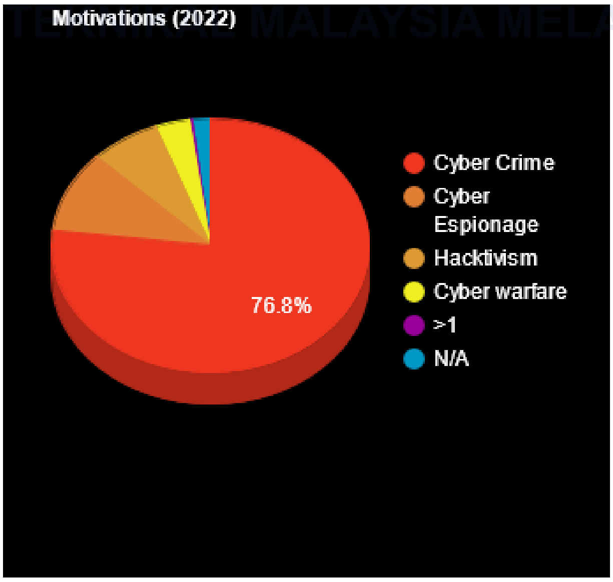

Figure 1.1 Major motivation of attacks (Source: Hackmageddon cybercrime statistics 2022)

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.

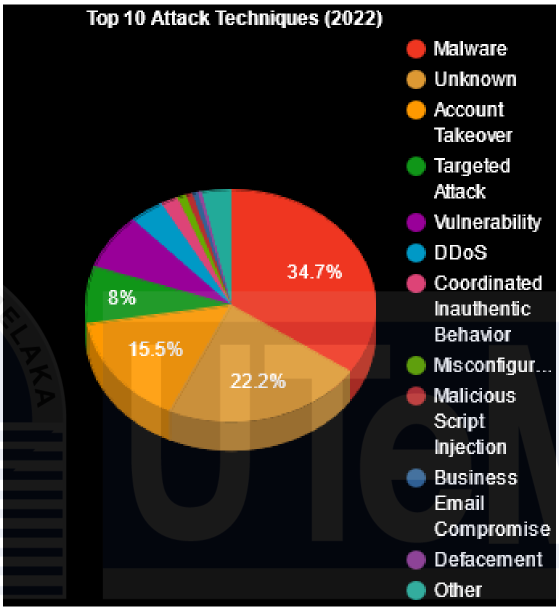

Figure 1.2 Top 10 attack techniques (source: hackmageddon cybercrime statistics 2022)

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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.

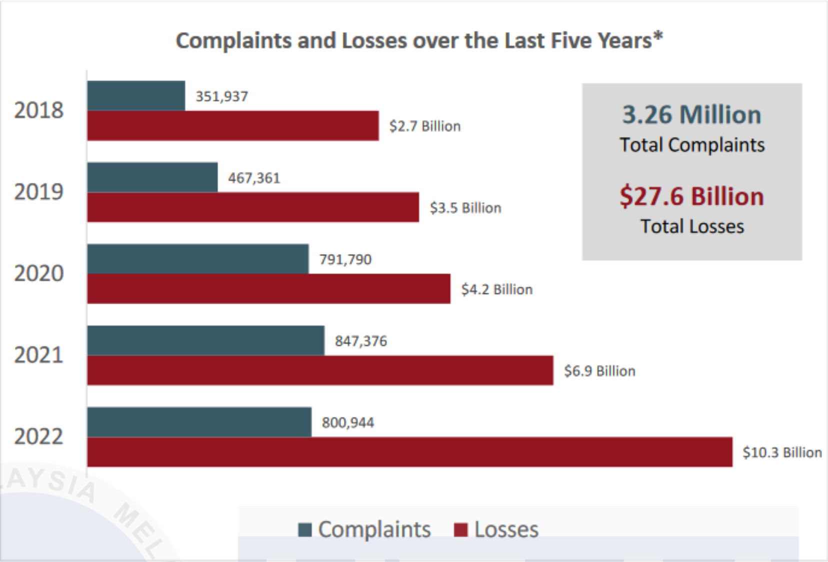

Figure 1.3 Yearly comparisons of complaints received via the IC3 website (FBI 2022)

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

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA