A ROBUST GRADIENT BOOSTING MODEL BASED ON SMOTE AND NEAR MISS METHODS FOR INTRUSION DETECTION IN IMBALANCED DATA SETS

AHMET OKAN ARIK

IŞIK UNIVERSITY
JANUARY, 2022

A ROBUST GRADIENT BOOSTING MODEL BASED ON SMOTE AND NEAR MISS METHODS FOR INTRUSION DETECTION IN IMBALANCED DATA SETS

AHMET OKAN ARIK
Işık University, School of Graduate Studies, the Degree of Master of Science in Information Technologies
2022

This thesis submitted to, Işık University, School of Graduate Studies for Master of Science Degree in Information Technologies.

IŞIK UNIVERSITY
January, 2022

IŞIK UNIVERSITY
GRADUATE SCHOOL OF SCIENCE
MASTER OF SCIENCE IN INFORMATION TECHNOLOGIES

A ROBUST GRADIENT BOOSTING MODEL BASED ON SMOTE AND NEAR MISS METHODS FOR INTRUSION DETECTION IN IMBALANCED DATA SETS

AHMET OKAN ARIK

APPROVED BY:

Asst. Prof. Gülsüm Çiğdem Çavdaroğlu Akkoç (Thesis Advisor)	Işık University / MIS
Asst. Prof. Şahin Aydın	Işık University / MIS
Asst. Prof. Zeynep Turgut Akgün	İstanbul Medeniyet University / Computer Engineering

APPROVAL DATE: 18/01/2022

A ROBUST GRADIENT BOOSTING MODEL BASED ON SMOTE AND NEAR MISS METHODS FOR INTRUSION DETECTION IN IMBALANCED DATA SETS

ABSTRACT

Novel technologies cause many security vulnerabilities and zero-day attack risks. Intrusion Detection Systems (IDS) are developed to protect computer networks from threats and attacks. Many challenging problems need to be solved in existing methods. The class imbalance problem is one of the most difficult problems of IDS, and it reduces the detection rate performance of the classifiers. The highest IDS detection rate in the literature is 96.54%. This thesis proposes a new model called ROGONG-IDS (Robust Gradient Boosting) based on Gradient Boosting. ROGONG-IDS model uses Synthetic Minority Over-Sampling Technique (SMOTE) and Near Miss methods to handle class imbalance. Three different gradient boosting-based classification algorithms (GBM, LightGBM, XGBoost) were compared. The performance of the proposed model on multiclass classification has been verified in the UNSW-NB15 dataset. It reached the highest attack detection rate and F_1 score in the literature with a 97.30% detection rate and 97.65% F_1 score. ROGONG-IDS provides a robust, efficient solution for IDS built on datasets with the imbalanced class distribution. It outperforms state-of-the-art and traditional intrusion detection methods.

Key words: Machine learning, Cyber security, Intrusion detection system, Imbalanced data, Gradient boosting.

SALDIRI TESPİT SİSTEMLERİ İÇİN DENGESİZ VERİ SETLERİNDE SMOTE VE NEAR MISS METOTLARINA DAYALI GÜÇLÜ GRADYAN ARTIRMA MODELİ

ÖZET

Yeni teknolojiler birçok güvenlik açığına ve sıfırıncı gün saldırı risklerine neden olmaktadır. Saldırı tespit sistemleri, bilgisayar ağlarını tehdit ve saldırılardan korumak için geliştirilmiştir. Mevcut yöntemlerde çözülmesi gereken birçok zorlu problem vardır. Sınıf dengesizliği problemi karşılaşılan en zorlayıcı problemlerden birisidir ve saldırı tespit sistemlerinde sınıflandırıcıların tespit oranını düşürmektedir. Literatürdeki en yüksek IDS saldırı tespit oranı 96.54%'tür. Bu tezde Gradyan Arttırma temelli ROGONG-IDS (Robust Gradient Boosting) olarak adlandırılan bir model sunulmaktadır. ROGONG-IDS modeli, sınıf dengesizliğini ele almak için Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE) ve Near Miss metotlarını kullanmaktadır. Gradyan arttırma tabanlı üç farklı sınıflandırma algoritması (GBM, LightGBM, XGBoost) karşılaştırıldı. Önerilen modelin çok sınıflı sınıflandırma üzerindeki performansı, UNSW-NB15 veri seti üzerinde test edilmiştir. ROGONG-IDS, 97.30% tespit oranı ve 97.65% F_1 skoru ile literatürdeki en yüksek saldırı tespit oranı ve F_1 skoruna erişti. ROGONG-IDS, dengesiz sınıf dağılımına sahip veri kümeleri üzerine kurulmak istenen saldırı tespit sistemleri için sağlam, verimli bir çözüm sunar. Önerilen bu modelin son teknoloji ve geleneksel yöntemler oluşturulmuş saldırı tespit sistemlerinden daha iyi performans sergilediği görülmüştür.

Anahtar Kelimeler: Makine öğrenmesi, Siber güvenlik, Saldırı tespit sistemi, Dengesiz veri, Gradyan arttırma.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Asst. Prof. Gülsüm Çiğdem Çavdaroğlu Akkoç. The door to Prof. Çavdaroğlu's was always open whenever I ran into a trouble spot or had a question about my research or writing. She consistently allowed this paper to be my work but guided me in the proper direction whenever he thought I needed it.

I would also like to thank Asst. Prof. Ali Cihan Keleş for his friendship and support, my colleague Mehmet Ali Özer for his support.

Finally, I must express my very profound gratitude to my parents and my girlfriend Beyza for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Ahmet Okan ARIK

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATONS

CHAPTER 1

1. INTRODUCTION

Cyber attack means destroying the triad of confidentiality, integrity, or availability, called the CIA triad. Many tools have been developed to combat cyberattacks, such as firewalls, anti-virus, Network Access Controllers (NAC), end-point security, Intrusion Prevention System (IPS), and Intrusion Detection System (IDS). An IDS is cyber security software that monitors the host or network to identify cyber-attacks. The continuous development of the Internet of Things (IoT), Industry 4.0, Cloud Computing, and Big Data technologies have increased the number of devices connected to the networks strikingly (Yang, Zheng, Wu, Yang, and Wang, 2020) and continues to raise its speed. Swiftly expanding accessibility and smart devices have led to an increment in cyber attacks. This increase in cyber attacks has reinforced the importance of IDS more than ever before.

IDS are divided into two different groups according to the detection method: (1) Signature-based IDS (SIDS), (2) Anomaly detection based on IDS (AIDS). While SIDS is based on marking all abnormal behavior for an entity, AIDS is a class of marking that is close to some predefined model signature of the entity (Axelsson, 2000). SIDS achieves high detection rates in known attack types because signatures are available for these attack types. However, this method is unable to identify new attack types because there are no signature patterns for these attack types (Kabiri and Ghorbani, 2005). In addition,

a large database of signatures is kept, and the incoming signatures are compared with the signatures in the database, which cannot be a resource-friendly application. (Uddin et al., 2013) Any activity other than the normal profile created in AIDS is called anomaly or abnormal behavior. The benefits of this method are that it can detect unknown attacks and new attack types and is proper for different networks and applications with a customizable normal activity profile. (Guo, Ping, Liu, and Luo, 2016) An imperfection of AIDS is that it perceives any deviation from the baseline as an attack, causing the system's unpredictable behavior to be labeled as an attack. This issue leads to a high false-positive rate.

In this thesis, a model named ROGONG-IDS (Robust Gradient Boosting IDS), which is an AIDS, is proposed. This model consists of (1) preprocessing module, (2) handling imbalance data module where two-stage data resampling is performed, and (3) classification decision module. The number of 47 independent variables in the UNSW-NB15 dataset was reduced to 12 after feature selection method is created Zhang et al. by using Denoising Autoencoder (DAE). Three different gradient boosting-based classifiers (LightGBM, GBM, XGBoost) were tested in the classification decision module after one-hot encoding, label encoding, and data standardization, and XGBoost, which provides the most successful result in classifying attacks successfully, was chosen as the classifier method of the ROGONG-IDS model. Section 2 includes literature review, Section 3 method, Section 4 experiments and results, and Section 5 includes conclusion and future work.

CHAPTER 2

2. LITERATURE REVIEW

2.1 Thesis Main Domain

IDS consists of two types: (1) Network Intrusion Detection System (NIDS), (2) Host Intrusion Detection System (HIDS). NIDS is positioned inside the network to keep track of all traffic on the network. It examines the incoming packages within the scope of the developed IDS model. When an attack or abnormal movement is detected, it alerts the administrator. HIDS is a type of IDS that monitors processes and applications on workstations or servers. HIDS monitors essential system configuration files, log and content files, registry files and reports any unauthorized or abnormal behavior. In this thesis, a NIDS that monitors incoming network packets is studied.

AIDS is being developed by machine learning and deep learning methods, as seen in Figure 2.1 provided. Deep learning methods have started to be used frequently in the development of IDS models today, with the ability to handle the feature engineering feature independently and produce more successful results in high-volume data. Machine learning algorithms used in this domain cannot be left behind due to the mentioned features of deep learning methods. ROGONG-IDS proposed in this thesis achieved the most successful multiclass classification results in the literature with gradient boosting methods.

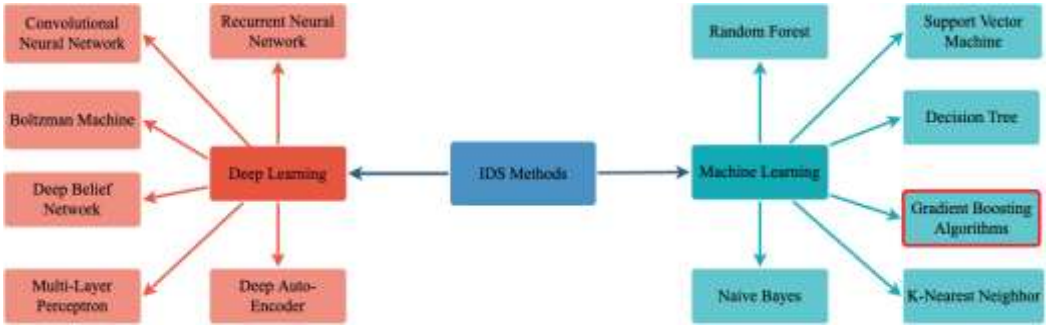

Figure 2.1 Methods used in developing IDS models.

2.2 Related Work

Ming, Zhou, and Chen (2021) designed a sequential model for the IoT system IDSs using deep learning methods. The parameters used for the created model are obtained from the packets at the network layer in the Open System Interconnection (OSI) architecture via Tcpdump. Text-Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) algorithms are preferred for the sequential-based model. It is aimed to reach a higher F_1 score which is proper measurement to evaluate model success on imbalanced data set as it could extract new parameters from the data. In the study in which the KDD99 dataset was used for testing, the generated model was compared with the Support Vector Machine (SVM), Naive Bayes, and C4.5 algorithms. The studied algorithms outperformed traditional machine learning models whose F_1 score performance was compared on multiclass classification. The model achieved an F_1 score performance of over 90% in the attack classes in the KDD99 dataset.

Thaseen, Banu, and Lavanya, Ghalib, and Abhishek (2021) used the Artificial Neural Network (ANN) to devise the IDS model. The feature selection process was applied by studying the correlation between the features, and the model was built with features that correlate 0,5. UNSW-NB15 and KDD99 datasets were used for testing. The overall accuracy was 96,44% for the UNSW-NB15 dataset and 98,45% for the KDD-99 dataset. They stated that traditional machine learning methods are insufficient to handle extensive network data and adopted the use of neural networks.

ROGONG-IDS has reduced the training and testing time by creating a model with high impact and numerically fewer features in the feature selection part. In addition, ROGONG-IDS offers a two-stage imbalance data solution for the

imbalanced data problem, one of IDS models' most significant problems and is not addressed in the mentioned study.

Mulyanto, Faisal, Prakosa, and Leu (2021) developed a model that solves the imbalanced data problem that needs to be solved in IDS models with the Focal Loss method (FL-NIDS). This method, which was studied with Deep Neural Network (DNN) and CNN architectures, was examined on three benchmark IDS datasets, NSL-KDD, UNSW-NB15, Bot-IoT, and it was observed that the detection rate increased as the number of layers increased in these two architectures. Since the accuracy score does not reflect the detection rate of the minority classes, the evaluation was made with the F_1 score. In the UNSW-NB15 dataset, it was seen that the CNN-SMOTE model reached a 36% F_1 score, while FL-NIDS reached a 39% F_1 score.

This study, which has a low F_1 score according to ROGONG-IDS, could not be said to have successfully solved the problem of imbalanced data. In addition, the study's lack of data standardization and feature selection methods resulted in low classification scores.

Vigneswaran, Vinayakumar, Soman, and Poornachandran (2018) used 3-layer DNN for their IDS model. Binary classification model using KDDCup-99 data set; Compared with Ada Boost, Decision Tree, K-Nearest Neighborhood (K-NN), Linear Regression, Naive Bayes, Random-Forest, SVM methods. One layer DNN architecture has been shown to achieve 92,9% accuracy, 95,4% F_1 score and outperform traditional machine learning methods.

In this study, some of the reasons for obtaining low binary classification scores were not to solve the imbalanced data problem and feature selection methods. ROGONG-IDS offers a robust solution for imbalanced data.

Kaja, Shaout, and Ma (2019) used a two-stage method for attack classification in the IDS model. The model performs the detection process with K-Means on the first packet coming from the network, and then the classification process is performed. This method is aimed to reduce the false-positive rate. This study using the KDD dataset reached 99,97% accuracy, which is the most successful score in the literature in this dataset.

Yin, Zhu, Fei, and He (2017) studied the RNN and Multilayer Perceptron method to generate IDS. In the NSL-KDD dataset study, a comparison was made with traditional machine learning methods. The study reached 81,29% accuracy with RNN and 78,10% accuracy with Multilayer Perceptron (MLP).

This study uses an old dataset and does not discuss NSL-KDD, imbalanced data, data standardization, and feature selection methods. Nevertheless, ROGONG-IDS has achieved high accuracy, F_1 scores using unique and new methods..

Naseer et al. (2018) examined IDS models' CNN, RNN, and Autoencoders architectures. They found that the deep convolutional neural network (DCNN) and Long Short-Term Memory (LSTM) were the most successful deep learning methods for the NSLKDD dataset. DCNN reached 85% accuracy while LSTM reached 89% accuracy.

Since the accuracy value does not reflect the detection rate of the minority classes, the F_1 score criterion should also be evaluated. Imbalanced data, feature selection, and data standardization are other methods that could increase the success rate of this study.

Xu, Shen, Du, and Zhang (2018) used GRU and LSTM algorithms in IDS models. This study used KDD99 and NSL-KDD datasets, and they reached 99,42% accuracy in the KDD99 dataset and 99,31% accuracy in the NSL-KDD dataset. Another significant result is that GRU outperforms the LSTM algorithm for the aforementioned datasets.

In their study, Jiang, Wang W., Wang A., and Wu (2020) stated that the imbalanced class problem causes a high false detection rate. Therefore, they propose an IDS model combined with hybrid sampling and a deep hierarchical network to avoid this. To solve the imbalanced data problem, One-Side-Selection (OSS) is used to reduce noise samples in majority classes and then SMOTE method is used to increase samples of minority classes. The proposed model was tested on NSL-KDD and UNSW-NB15 datasets, and accuracy values of 83,58% and 77,16% were achieved, respectively.

In this study, the imbalanced data problem was solved with a two-stage method. Although the created model has a complex structure, some of its low accuracies are that data standardization and feature selection were not performed.

Yang et al. (2020) propose an IDS model called SAVAER-DNN. They state that the minority classes should be increased to solve the imbalanced data problem, and the data augmentation method used shows a successful performance than other well-known methods. The number of observations of the minority classes has been increased up to the class with the most samples in the majority class. They compared the created model with other oversampling methods and classification algorithms. The

proposed model reached 93,01% accuracy, while Random Over Sampling (ROS) - DNN 81,45%, Synthetic Minority Over-sampling Technique (SMOTE) - DNN 81,94%, Adaptive Synthetic (ADASYN) - DNN 81,76% accuracy.

In this study, it could be said that increasing the number of observations of the minority class increases the training time. The long training time is unsuitable for implementing the IDS model in real-time network environments. On the other hand, ROGONG-IDS is suitable for real-time network environments due to its short execution time.

Belouch, Hadaj, and Idhammad (2018) used Apache Spark, a big data processing engine, to preprocess the IDS model they created. They reduced the number of 49 features to 42 by applying the feature selection method. SVM, Naive Bayes, Decision Tree, Random Forest algorithms have been tested. Random Forest achieved the best binary classification result with 97,49% accuracy.

For real-time network implementation, execution time needs to be reduced. This process requires a more extensive feature selection process. In the ROGONG-IDS feature selection part, the number of features has been reduced to 12. In this way, execution time has decreased significantly.

Chkirbene, Eltanbouly, Bashendy, AlNaimi, and Erbad (2020) implemented a comprehensive feature selection on the IDS model. The number of features has been reduced to 13 with the Random Forest algorithm. Classification And Regression Tree (CART) is used as the classification method. UNSW-NB15 and KDD99 datasets were used to test the success of the model. In multiclass classification; They reached 95,73% accuracy in the UNSW-NB15 dataset and 97,03% accuracy in the KDD99 dataset.

Using over-sampling and random sampling methods to solve the imbalanced class problem can achieve higher accuracy on the accuracy value.

Vinayakumar et al. (2019) propose an IDS model with DNN architecture using various layers from one to five. They tested the model created on UNSW-NB15, KDDCup99, NSL-KDD, WSN-DS, CICIDS 2017 datasets. They achieved the most successful results for the UNSW-NB15 dataset with a one-layer DNN architecture. In this dataset, 78% accuracy score and 82% F_1 score were achieved in binary classification. In multiclass classification, they reached 64% accuracy.

In the mentioned study, to increase the low accuracy in multiclass classification and reduce the execution time, it is necessary to solve the imbalanced data problem and use feature selection and data standardization methods.

Zhang, Huang, Wu, and Li (2020) propose an IDS model called SGM-CNN. This model offers a two-stage imbalanced data solution. First, the sample number of the minority classes is equal to the $I_{resample}$ value we included in our study with the SMOTE method. In the second stage, the Gaussian Mixture Model (GMM) involves equating the sample number of the majority classes to the $I_{resample}$ value. This study, which proposes the CNN architecture, reached 96,54% accuracy in multiclass classification and 98,82% accuracy in binary classification.

Lee, Amaresh, Green, and Engels (2018) examined deep learning approaches on IDSs in their studies. Vanilla DNN, Self-Taught Learning (STL), and LSTM-based RNN algorithms were compared on the KDD Cup 99 dataset. According to the results, it is known that Autoencoder reaches 98,9% accuracy and LSTM reaches 79,2% accuracy.

Al-Yaseen, Othman, and Nazri (2017) propose a multi-level hybrid IDS model to understand whether incoming network packets are standard or attack packages. K-Means was used to reduce the dataset by 10 percent. SVM classifier is used on the reduced dataset. It was stated that it achieved 95,75% accuracy on the KDD Cup 1999 dataset and outperformed the other methods compared.

Elmasry, Akbulut, and Zaim (2020) state that the most critical problems of IDS, false alarm rate (FAR), and low detection rate (DR), are datasets consisting of irrelevant features. Particle Swarm Optimization (PSO) based method has been proposed to overcome this problem. It aims to tune hyperparameters automatically and define feature subsets with this method. They evaluated this proposed method with three different deep learning methods, DNN, LSTM, and Deep Belief Networks (DBN), respectively. NSL-KDD and CICIDS2017 datasets were used for testing in the study. According to the results, the proposed method increases the detection rate by 4%-6% and reduces the FAR by 1%-5%. The highest success in multiclass classification was achieved with the DBN 86,53% accuracy score in the NSL-KDD dataset. In the CICIDS-2017 dataset, the most successful result was again achieved with DBN with an accuracy score of 82%.

Zhiqianq et al. (2019) state that traditional machine learning algorithms are not sufficiently practical on IDSs. In this direction, they propose a deep learning model tested with the UNSW-NB15 dataset. The proposed model is a 10-layer Feedforward ANN model consisting of 100 neurons. It is stated that the proposed method

outperforms algorithms such as Logistic Regression, Naive Bayes, ANN. Furthermore, the accuracy score is 99,5% in binary classification, and the FAR is 0,47.

Moustafa and Slay (2017) propose a hybrid feature selection method for feature selection, which has an important place in developing the advanced IDS model. First, it reduces processing time by selecting the most frequently used observations with Central Points (CP). After this process, best-ranked features are obtained with Association Rule Mining (ARM). Finally, irrelevant, noisy features are removed. In the evaluation made on binary classification in UNSW-NB15 and NSL-KDD dataset, it was seen that this feature selection method performed more successfully in the UNSW-NB15 dataset.

Shende and Thorat (2020) state that more effective deep learning methods should be used instead of traditional machine learning methods in their study. Using the NSL-KDD dataset, IDS models were created with MLP, LW-MLP, ANN, CNN methods. They achieved accuracy values of 97,79% with CNN and 97,14% with MLP.

Gupta, Jindal, and Bedi (2021) propose a model called LIO-IDS. This model uses LSTM classifier and Improved One-vs-One (I-OVO) techniques. Consisting of two layers, the LIO-IDS anomaly is a one-based IDS. In the first layer, packets are detected as an attack or normal with LSTM, and in the second layer, the ensemble method is used to classify attacks. I-OVO, used for multiclass classification in the second layer, differs from the traditional OVO method by using only three classifiers for each observation, thus reducing the test time. Over-sampling methods SVM-SMOTE, Borderline-SMOTE, and Random Over-Sampling (ROS) were used to improve detection in the second layer. NSL-KDD, CIDDS-001, and CICIDS2017 datasets were used for evaluation. Looking at the results, it has been determined that the proposed LIO-IDS model makes a significant difference from other IDS models. It has been stated that it is suitable for real-world deployment with its high DR and short computational time. LIO-IDS achieved 87% accuracy in the NSL-KDD dataset, 96% accuracy in the CIDDS-01 dataset, and 86% accuracy in the CICIDS2017 dataset.

Kasongo and Sun (2019) state that IDSs get weaker and heavier as the feature space grows. They propose a Feedforward Deep Neural Network (FFDNN) to solve this problem for wireless IDSs and, with it, a filter-based feature selection method. Compared with SVM, Decision Tree, K-NN, and Naive Bayes methods using NSL-KDD dataset. The proposed model outperformed these methods and reached 86.19% accuracy on multiclass classification.

Kang M-J. and Kang J-W. (2016) propose a new IDS model for vehicular networks using DNN. The proposed IDS model monitors the packets broadcast in the Controller Area Network and tests an attack. Probability-based feature vectors were trained using DBN. The experiments carried out can provide real-time response to network traffic with an accuracy score of 98%.

Liu, Gao, and Hu (2021) propose an IDS model that addresses the imbalanced data problem. An ensemble model is proposed to solve the Imbalanced data problem. This model uses ADASYN for oversampling and LightGBM for classification. After performing the normalization process on the data, experiments were performed on KDD, UNSW-NB15, CICIDS2017 datasets. The model, which offers a more prosperous and shorter training time than other IDS models, reached 85.89% accuracy in the UNSW-NB15 dataset.

Khan, Gumaei, Derhab, and Hussain (2019) suggest a two-stage IDS model. First, this model detects that network packets are not normal or abnormal based on the probability score generated by the Stacked Auto-Encoder. In the second stage, attacks are classified using the Softmax classifier. Thus, the proposed system can classify unlabeled data. The model, evaluated with different algorithms, reached 89,13% accuracy, 0,74 FAR in the UNSW-NB15 dataset.

2.3 Contributions

This thesis proposes an IDS model called ROGONG-IDS, which has the highest accuracy and F_1 score in the literature. Furthermore, as seen in the literature review, it may help to include gradient boosting methods and resampling studies, which are rarely used in IDS models, frequently in IDS development. Thus, it may be possible to examine the development of AIDS in a wide range with this thesis.

CHAPTER 3

3. METHOD

ROGONG-IDS architecture consists of three modules as shown in Figure 3.1: (1) Data preprocessing module, (2) imbalanced data handling module, and (3) classification decision module. The data preprocessing module aims to make the data proper for modeling. Operations performed in the data preprocessing module are: (a) one-hot encoding and label encoding transactions to process categorical data, (b) feature selection process to reduce model training time and increase accuracy by eliminating redundant features, and (c) data standardization process to examine many different types of measurable features in a common standard. The imbalanced data handling module involves resampling operations to ensure class balance. The two-stage resampling method uses the Near Miss method for undersampling and SMOTE for oversampling. This method is the most critical factor in increasing the model accuracy. The SMOTE method was used to increase the number of sample minority classes, whereas the Near-Miss method was used to reduce bias by undersampling majority classes. Finally, the XGBoost algorithm based on Gradient boosting, whose hyperparameters were optimized using Bayesian optimization, was used in the classification decision stage. Then it was experimented with in the UNSW-NB15 dataset to consider the model on the networking environment.

Figure 3.2 Architecture of ROGONG-IDS.

3.1 Description of UNSW-NB15 Dataset

IDS studies suffer from the unavailability of data sets without structured network information, which does not cover current network traffic scenarios. Datasets such as KDD98, KDDCUP99, NSLKDD, which are still used to evaluate the created IDS models, do not reflect today's modern network traffic scenarios. The Australian Center for CyberSecurity (ACCS) research group developed the UNSW-NB15 dataset (Moustafa and Slay, 2015). The research group combined the current standard network data and synthetic attack data to generate this dataset. It is a deep, comprehensive dataset that contains 2.54 million rows of traffic data and nine attack types and can reflect today's modern network traffic scenarios. In the dataset, which has 49 features, two features are class labels. There is a high level of class imbalance in the data set. Regular traffic is 87,35%, and attack traffic is only 12,65%. The whole dataset was used for the modeling, and it was divided into training and testing at a ratio of 7:3. It could be examined the attack class distributions in the dataset in detail in Table 3.1 below.

Table 3.1 Attack class distributions.

Class	Description	Training set - size	Test set - size	Total
Analysis	Port-based attack for web applications.	1.874	803	2.677
Backdoor	Penetration remote attack to obtain unauthorized access to a system.	1.630	699	2.329
DoS	An attack that aims to disrupt the services of the system temporarily or indefinitely.	11.447	4.906	16.353
Exploits	A penetration attack that aims to exploit a bug or vulnerability through code.	31.167	13.358	44.525
Fuzzers	A type of attack that scans the target system for information using software testing technique.	16.972	7.274	24.246

Table 3.2 Attack class distributions. (cont.)

Generic	A "generic attack" toward a cryptographical primitive is one that can be run separately from the circumstances of how that cryptographical primitive is implemented.	150.837	64.644	215.481
Normal	Real transction data.	1.553.134	665.630	2.218.764
Reconnaissance	Reconnaissance attacks are information-gathering attacks.	9.791	4.196	13.987
Shellcode	Shellcode is a collection of directions that performs a command in software to gain control of or exploit a compromised machine.	1.058	453	1.511

Table 3.3 Attack class distributions. (cont.)

Worms	The worm is software carrying malicious code that attacks host machines and lies via a network.	122	52	174
Total	10 classes.	1.778.032	762.015	2.540.047

3.2 Data Preprocessing

Feature selection, one-hot encoding, label encoding, and data standardization processes were carried out in this phase.

DAE model (Zhang et al., 2018) was used in the feature selection stage. DAE aims to reduce feature dimensionality by specifying a limited number of critical features. The feature size was limited, with 12 features determined as a result of DAE. Table 3.2 shows these 12 selected features. Feature selection was performed at the beginning of the process to accelerate the preprocessing phase.

Table 3.4 Selected features of the UNSW-NB15 according to DAE.

Dtcpb	Stcpb	Service__-	Dload	Dmeansz	Service_dns
Smeansz	Sload	Trans_depth	Sttl	Service_ftp-data	Ct_ftp

UNSW-NB15 dataset has three nominal data types. These properties are: "proto", "state", "service". Respectively, each attribute has 135, 16, 14 different values. One-hot encoding technique is used to process these features by machine learning algorithms while preserving the irregular relationship. With this process, the number of features in the dataset increased from 47 to 208. Label encoding, another type of

encoding, was implemented on the target feature attack class. Data standardization, the last method implemented in the data preprocessing stage, is used to bring data into a common format. The common format allows to increase the accuracy value of the model and to perform analytical studies on multidimensional data sets. According to the Gaussian distribution, all features were standardized using the formula in Eq. (1), with a mean of 0 and a standard deviation of 1. The data standardization formula is provided below in Eq. (1):

$$Z = \frac{\chi - \mu}{\sigma}$$

3.3 Handling Imbalance Data

As seen in Table 3.1, there is a terrific imbalance between target variable classes. As there are two classes with less than 2000 instances in 2.54 million rows oversampling for minority classes or undersampling for majority classes alone will not be sufficient. There may be situations such as removing useful information and generating new observations that will increase the cost. A process is proposed that recommends the use of both methods in the ROGONG-IDS model. As you see in Table 3.3, 10 different methods were tried to find the most successful undersampling method for the UNSW-NB15 dataset with XGBoost algorithm and SMOTE for oversampling. The near Miss (v1) method gave the most successful results for undersampling, and SMOTE was used in this stage.

Table 5.3 Studied undersampling methods.

Undersampling Method	**Accuracy**	**F_1 score**	**Detection Rate**	**Algorithm**	**Oversampling Method**
AIIKNN	96,05%	96,77%	96,05%	XGBoost	SMOTE
Edited Nearest Neighbours	96,26%	96,89%	96,26%	XGBoost	SMOTE
Repeated Edited Nearest Neighbours	96,08%	96,78%	96,08%	XGBoost	SMOTE

Table 6.3 Studied undersampling methods. (cont.)

Instance Hardness Threshold	94,20%	95%	94,20%	XGBoost	SMOTE
Near Miss Undersampling (v1)	96,49%	97,10%	96,49%	XGBoost	SMOTE
Near Miss Undersampling (v3)	95,45%	96,41%	95,45%	XGBoost	SMOTE
Neighbourhood Cleaning Rule	96,03%	96,64%	96,03%	XGBoost	SMOTE
Random Undersampling	96,22%	96,97%	96,22%	XGBoost	SMOTE
Tomek Link	96,24%	96,89%	96,24%	XGBoost	SMOTE
One Sided Selection	95,44%	96,41%	95,44%	XGBoost	SMOTE

ROGONG-IDS uses a technique combining the SMOTE and Near Miss methods that resamples all classes with an equal sample count called $I_{resample}$ (Abdulmuhammed, Musafer, Alessa, Faezipour and Abuzneid, 2019) to handle the imbalanced class distribution. Explanation of $I_{resample}$ is provided in Eq. (2):

$$I_{resample} = \text{int}\left(\frac{\text{Number of samples}}{\text{Number of classes}}\right)$$

ROGONG-IDS uses SMOTE (Chawla, Bowyer, Hall, and Kegelmeyer, 2002) to oversample classes with less than $I_{resample}$ .One of the most commonly used oversampling methods, SMOTE, increases minority class instances by synthesizing. The reason for its success is that it uses the synthesis method. Instead of copying samples from the data set, this method generates samples not in the data set. Thus, the overfitting problem caused by the random oversampling method is avoided. Instead,

focusing on the feature space, SMOTE draws a line between the existing minority class instances and places the synthetic data generated with the help of interpolation on this line.

ROGONG-IDS uses the Near Miss undersampling method (Zhang and Mani, 2003) for classes with more samples than the $I_{resample}$ value. Near Miss works by selecting samples based on their distance from the majority class to the minority class. It uses the Euclidian distance or similar as the distance measurement. ROGONG-IDS uses version 1 of Near Miss, which has three different versions. This version balances classes by keeping the average closest majority class samples to the three closest minority class samples. It significantly increases the detection rate of minority classes. The method ROGONG-IDS uses to handle imbalanced data differs from the undersampling method used in the two-stage SGM (Zhang et al., 2020) method. Algorithm 1 provide the pseudocode of ROGONG-IDS handling imbalanced data method.

Algorithm 1. Method of handling imbalanced data on ROGONG-IDS.

Input:
Training set D = { $D_i , i = 1,2, \dots, C$ };
C = the total number of classes;
|D| = N; #the total number of samples

Output:
A balanced training set D' ;

$I_{resample}$ = int ($\frac{N}{C}$)
for i ← 1 to C do
  if $|D_i| < I_{resample}$ then
    D_i' = SMOTE (D_i , $I_{resample}$) #Generating new samples to minority class
  end if
  if $|D_i| > I_{resample}$ then
    D_i' = NearMiss (D_i , $I_{resample}$) #Removing values from majority class
  end if
end for
return D'

3.4 Extreme Gradient Boosting: XGBoost

The XGBoost algorithm (Chen and Guestrin, 2016) is an enhanced version of the Gradient boosting method for decision trees. Chen and Guestrin (2016) aimed to

scalability on tree boosting systems, use computational resources effectively and improve model performance in classification and regression problems. Boosting is an ensemble technique in which new models are added to fix the errors of the existing model. New models are added iteratively until no new improvement is seen. Gradient boosting is an algorithm used to estimate the residuals of previous models and make the final estimation. It uses the gradient descent algorithm to minimize the loss of the new model. Used heavily to provide state-of-the-art results for classification and regression work, XGBoost was seen winning 17 of 29 machine learning tasks published on Kaggle by 2015 (Ogunleye and Wang, 2020) .

CHAPTER 4

4. EXPERIMENTAL ANALYSIS

ROGONG-IDS model implementation was tested on the UNSW-NB15 dataset to measure the accuracy of detections. It was used Macbook Pro with a macOS Monterey operating system during the implementation. The test environment is provided in Table 4.1.

Table 7.1 Test environment.

Project	**Environment / Version**
Operating System	macOS Monterey
CPU	1,4 GHz Quad-Core Intel Core i5
GPU	Intel Iris Plus Graphics 645
Memory	16 GB

4.1 Evaluation Metrics

Accuracy (ACC), DR, FAR, F_1 score, Precision, Recall, indicators, which are frequently used in imbalance class classification evaluation, were used during the experiments. For each attack class, the samples considered as attacks were accepted positive and the others negative. ACC represents the percentage of correctly classified samples among all samples. DR is the rate of correctly predicted positive samples.

This ratio shows the success of ROGONG-IDS to detect various attack types. FAR is defined as the proportion of negative samples falsely evaluated as positive. Recall, which is DR, relates the ratio of correctly predicted positive class samples to the total number of positive class samples. Precision means how many of the samples predicted to be positive are positive samples. F_1 score is the harmonic average of Precision and Recall. In multiclass classification, each class is calculated using a weighted average method based on the number of samples in the category to understand the detection performance of the model on imbalanced data. Eqs. (3-7) shows these quality measures.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$DR = \frac{TP}{TP+FN} \tag{2}$$

$$FAR = \frac{FP}{FP+TN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$F_1\ Score = \frac{2.Precision.Recall}{Precision+Recall} \tag{5}$$

TP/FP and TN/FN are the numbers of samples correctly and incorrectly predicted to be positive and negative, respectively.

4.2 Hyper-parameter Optimization

Hyperparameters are high-impact parameters that control the learning process of the model. Since they are tuneable, they play a role in reaching the maximum performance of the model in a reasonable time. However, as the size of the processed data increases, the cost of the hyperparameter optimization process increases. Therefore, Grid Search and Random Search methods are cost-inefficient and exhaustive when used with big data. ROGONG-IDS uses Distributed Asynchronous Hyper-parameter Optimization (Hyperopt) (Bergstra, Yamins, and Cox, 2013) for

hyperparameter tuning. Hyperopt was developed to automate hyperparameter optimization based on Bayesian optimization. Hyperopt uses Bayesian optimization to define and narrow the search space and maximize the probability function. ROGONG-IDS model accuracy increased from 96,49% to 97,30% after using Hyperopt within a fair amount of time. Table 4.2 shows the XGBoost hyperparameters used after Hyperopt.

Table 8.2 XGBoost hyperparameters.

Parameters	**Value**
Learning Rate	0,5
Number of Estimators	5000
Max Depth	36
Colsample Bytree	0,61
Min Child Weight	4
Subsample	0,9

4.3 Multiclass Classification

ROGONG-IDS's performance could be analyzed in Table 4.3, which shows the DR results for each class. ROGONG-IDS essentially uses the XGBoost algorithm. However, other gradient boosting-based algorithms have experimented with the base method handling with the imbalanced data. Table 4.3 presents the performance results of the LightGBM and Gradient Boosting Machine (GBM) algorithms. ROGONG-IDS with XGBoost achieves the overall best performance in terms of DR, accuracy, and F_1 score of 97,30%, 98,16%, 97,65%, respectively.

Table 9.3 Multiclass classification perfomance comparison between LightGBM, GBM and XGBoost.

Class	ROGONG - LightGBM	ROGONG - GBM	ROGONG - XGBoost
Analysis	0.84	0.67	0.31
Backdoor	0.23	0.11	0.26
DoS	0.06	0.13	0.47
Exploits	0.46	0.48	0.54
Fuzzers	0.66	0.73	0.70
Generic	0.97	0.97	0.98
Normal	0.99	0.99	0.99
Reconnaissance	0.81	0.81	0.77
Shellcode	0.88	0.55	0.53
Worms	0.83	0.00	0.83
DR	96.55	96.26	97.30
Accuracy	96.55	96.26	97.30
Precision	98.30	97.91	98.16
F_1 Score	97.18	96.91	97.65
Train-Time (s)	15.08	4336.71	205.27
Test-Time (s)	2.2	0.73	0.81

The test results of the gradient boosting methods tried for the ROGONG-IDS model are provided in Figure 4.1 summarily.

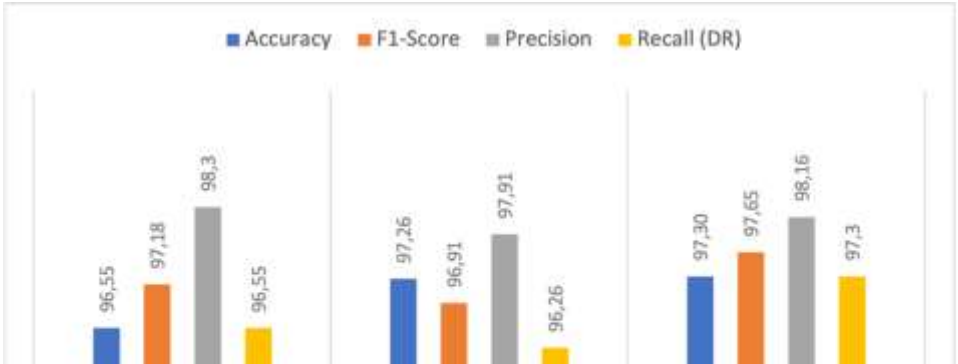

Figure 3.1 Comparison of gradient boosting methods on ROGONG-IDS.

Table 4.4 provides a performance comparison of advanced IDS models and ROGONG-IDS. Although DR development was provided for many classes, “Analysis”, “Backdoor”, “DoS” attack types remained below 50% DR. While the SGM had a test time of 8 seconds, ROGONG-IDS made a significant improvement in this regard, reducing the test time to 0,81 seconds.

Table 10.4 Comparison multiclass classification results with advanced methods on the UNSW-NB15 dataset.

Class	**SGM-CNN (Zhang et al., 2020)**	**Two stage – DL (Khan et al., 2019)**	**Hybrid Machine Learning (Chkirbene et al., 2020)**	**ICVAE-DNN (Yang, Zheng, Wu, and Yang, 2019)**	**ADASYN + LightGBM (Liu et al., 2021)**	**ROGONG - IDS**
Analysis	0,27	0,01	0,00	0,15	-	0,31
Backdoor	0,51	0,00	0,6	0,21	-	0,26
DoS	0,39	0,00	0,8	0,8	-	0,47

Table 11.4 Comparison multiclass classification results with advanced methods on the UNSW-NB15 dataset. (cont.)

Exploits	0,45	0,57	0,86	0,71	-	0,54
Fuzzers	0,67	0,40	0,53	0,35	-	0,70
Generic	0,97	0,61	0,97	0,96	-	0,98
Normal	0,98	0,82	0,80	0,81	-	0,99
Reconnaissance	0,82	0,24	0,79	0,80	-	0,77
Shellcode	0,88	0,00	0,51	0,92	-	0,53
Worms	0,83	0,00	0,59	0,79	-	0,83
DR (%)	96,54	63,27	78,65	95,68	-	97,30
Accuracy(%)	96,54	89,13	78,65	89,08	85,89	97,30
F_1 Score (%)	97,26	90,85	78,65	90,61	-	97,65
Precision (%)	98,30	89,13	78,65	86,05	-	98,16
FAR	-	-	0,11	-	0,6	0,51
Train-Time (s)	47,22	-	-	-	-	205,27
Test-Time (s)	8,26	-	-	-	-	0,81

A summary comparison of ROGONG-IDS with advanced IDS methods in the literature is provided in Figure 4.2.

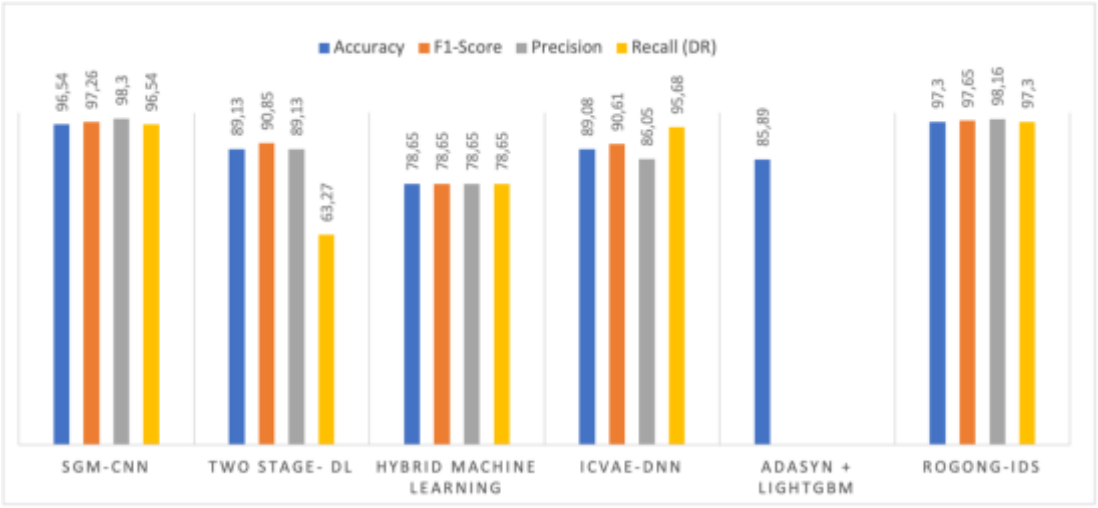

Figure 4.2 Comparison multiclass classification results with advanced methods using UNSW-NB15 dataset in the literature.

The experimental results show that ROGONG-IDS using the two-method handling imbalance data module used with XGBoost significantly improves DR. In

Table 4.3, ROGONG-IDS is compared with other gradient boosting classifiers. It has been determined that the XGBoost algorithm is more successful than other methods (GBM, LightGBM). XGBoost provided higher DR than the other two classifiers in attack types “Backdoor”, “DoS”, “Exploits”, “Generic”, “Normal”, “Worms”. When examined in general, it provided more successful results than the other two algorithms within the scope of DR, ACC, F_1 score, and test time.

As seen in Table 4.4, which includes the comparison of ROGONG-IDS with other advanced IDS models in the literature, ROGONG-IDS is seen to be the most successful IDS model in the literature in terms of DR, ACC, F_1 score, and test time. ROGONG-IDS; It is more successful than others in detecting attack types such as “Analysis”, “DoS”, “Fuzzers”, “Generic”, “Normal”, “Worms”, but in attack types such as “Backdoor”, “Exploits”, “Reconnaissance”, “Shellcode” was found to be less successful than other IDS models.

CHAPTER 5

5. CONCLUSION AND FUTURE WORK

Intrusion detection systems involve many challenges. The existence of anomalies may be specific to each field, but new anomalies and threats are created in complex ways by harmful actors in this domain. UNSW-NB15 data, which includes modern attack types and offers many different network parameters, has been used for this study to be suitable in a current network environment. However, more up-to-date data sets should be provided to develop more robust IDS models in this domain. Another difficulty is that attack packets in networks are less frequent than regular packets. This causes a considerable imbalance data problem and increases the size of the data to be used for modeling. This leads to an increase in the computing power and time required to process the data.

In this thesis, the UNSW-NB15 dataset, which includes the most up-to-date and modern scenarios and attack types in the literature, has been tested with classifiers based on Gradient boosting. This evaluation has determined that the XGBoost algorithm is more successful than other methods (GBM, LightGBM). The ROGONG-IDS model was compared with five advanced IDS models in the literature using the UNSW-NB15 dataset for testing. ROGONG-IDS with DR, ACC, and F_1 score reached 97,30%, 97,30%, 97,65%, respectively. These results prove that the ROGONG-IDS model is the most successful IDS model in the literature.

IDS studies have a different structure from the general anomaly detection and classification problems due to the size and volume of the data they encounter. It tries to handle streaming data. The proposed ROGONG-IDS model both solves the imbalanced data problem and has a fast implementation time (205s training, 0,81s test). Offering high success quickly, ROGONG-IDS is an efficient solution for real-time intrusion detection applications.

The generated ROGONG-IDS model could be used in areas that have huge data imbalance on streaming data. Accordingly, the two-stage imbalanced data handling module successful results could be achieved in diverse areas such as; smart production lines, autonomous drive, social network analysis, fraud detection, real-time stock trading. As future work, it is planned to study the optimization of attack classes, which ROGONG-IDS has difficulty in detecting, with the use of new reinforcement methods and use Apache Spark, which is used to process large-scale data, to reduce the implementation time.

REFERENCES

Abdulhammed, R., Musafer, H., Alessa, A., Faezipour, M. and Abuzneid, A. (2019) Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection. Electronics 2019, 8, 322.

Al-Yaseen, W. L., Othman, Z. A., and Nazri, M. Z. A. (2017) Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. Expert Syst. Appl. 67, C (January 2017), 296–303. DOI:https://doi.org/10.1016/j.eswa.2016.09.041

Andresini G., Appice, A., Mauro, N. D., Loglisci, C. and Malerba, D. (2020) Multi-Channel Deep Feature Learning for Intrusion Detection, in IEEE Access, vol. 8, pp. 53346-53359, 2020, doi: 10.1109/ACCESS.2020.2980937.

Axelsson, S. (2000) Intrusion Detection Systems: A Survey and Taxonomy. Technical Report 99-15. Department of Computer Engineering, Chalmers University.

Belouch, M., El hadaj, S. and Idhammad, M. (2018) Performance evaluation of intrusion detection based on machine learning using Apache Spark. Procedia Computer Science. 127. 1-6. 10.1016/j.procs.2018.01.091.

Bergstra J., Yamins, D. and Cox, D.D. (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. To appear in Proc. of the 30th International Conference on Machine Learning.

Chawla, N. V., Bowyer K. W., Hall, L. O. and Kegelmeyer, W. P. (2002) SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res., 16 (1) , pp. 321-357

Chen, T. and Guestrin C. (2016) XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 785–794, 2016, https://arxiv.org/abs/1603.02754.

Chkirbene, Z., Eltanbouly, S., Bashendy, M., AlNaimi, N. and Erbad, A. (2020) Hybrid Machine Learning for Network Anomaly Intrusion Detection, 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), 2020, pp. 163-170, doi: 10.1109/ICIoT48696.2020.9089575.

Elmasry, W., Akbulut, A., Zaim, A. H. (2020) Evolving deep learning architectures for network intrusion detection using a double pso metaheuristic. Computer Networks, 168, 107042. doi:10.1016/j.comnet.2019.107042

Guo, C., Ping, Y., Liu, N. and Luo, S. (2016) A two-level hybrid approach for intrusion detection. Neurocomputing. 214. 10.1016/j.neucom.2016.06.021.

Gupta, N., Jindal, V. and Bedi P. (2021) LIO-IDS: Handling class imbalance using LSTM and improved one-vs-one technique in intrusion detection system,Computer Networks, Volume 192, 2021, 108076, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2021.108076.(https://www.sciencedirect.com/science/article/pii/S1389128621001675)

Jiang, K., Wang, W., Wang, A. and Wu, H. (2020) Network Intrusion Detection Combined Hybrid Sampling With Deep Hierarchical Network, in IEEE Access, vol. 8, pp. 32464-32476, 2020, doi: 10.1109/ACCESS.2020.2973730.

Kabiri, P. and Ghorbani, A. (2005) Research on Intrusion Detection and Response: A Survey. International Journal of Network Security. 1. 84-102.

Khan, F. A., Gumaei, A., Derhab, A. and Hussain, A. (2019) A Novel Two-Stage Deep Learning Model for Efficient Network Intrusion Detection, in IEEE Access, vol. 7, pp. 30373-30385, 2019, doi: 10.1109/ACCESS.2019.2899721.

Kasongo, S. M. and Sun, Y. (2019) A Deep Learning Method With Filter Based Feature Engineering for Wireless Intrusion Detection System, in IEEE Access, vol. 7, pp. 38597-38607, 2019, doi: 10.1109/

Kaja, N., Shaout, A. and Ma, D. (2019) An intelligent intrusion detection system. Applied Intelligence. 49. 3235-3247. 10.1007/s10489-019-01436-1.

Kang, M-J., Kang, J-W. (2016) Intrusion Detection System Using Deep Neural Network for In-Vehicle Network Security. PLOS ONE 11(6): e0155781. https://doi.org/10.1371/journal.pone.0155781

Lee, B., Amaresh, S., Green, C. and Engels, D. (2018) Comparative Study of Deep Learning Models for Network Intrusion Detection, SMU Data Science Review: Vol. 1 : No. 1 , Article 8. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss1/8

Liu, J., Gao, Y. and Hu, F. (2021) A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM, Computers & Security, Volume 106, 2021, 102289, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2021.102289.

Moustafa, N. and Slay, J. (2015) UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set), 2015 Military Communications and Information Systems Conference (MilCIS), 2015, pp. 1-6, doi: 10.1109/MilCIS.2015.7348942.

Moustafa, N. and Slay, J., (2017) A hybrid feature selection for network intrusion detection systems: Central points. [online] arXiv.org. Available at: <https://arxiv.org/abs/1707.05505> [Accessed 12 May 2021].

Mulyanto, M., Faisal, M., Prakosa S. W. and Leu, J-S. (2021) Effectiveness of Focal Loss for Minority Classification in Network Intrusion Detection Systems. Symmetry. 13(1):4. https://doi.org/10.3390/sym13010004

Naseer S., et al. (2018) Enhanced Network Anomaly Detection Based on Deep Neural Networks, in IEEE Access, vol. 6, pp. 48231-48246, 2018, doi: 10.1109/ACCESS.2018.2863036.

Ogunleye, A. and Wang, Q-G. (2020) XGBoost Model for Chronic Kidney Disease Diagnosis, in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 6, pp. 2131-2140, 1 Nov.-Dec. 2020, doi: 10.1109/TCBB.2019.2911071.

Shende, S. and Thorat, S. (2020) A Review on Deep Learning Method for Intrusion Detection in Network Security, 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020, pp. 173-177, doi: 10.1109/ICIMIA48430.2020.9074975.

Thaseen, S. I., Banu, J. S., Lavanya, K., Ghalib, R. M. and Abhishek, K. (2020) An integrated intrusion detection system using correlation-based attribute selection and artificial neural network. Trans Emerging Tel Tech. 2021; 32:e4014. https://doi.org/10.1002/ett.4014

Network-data Packet Analyzer [Tcpdump]. (2021) Retrieved from https://www.tcpdump.org/

Uddin, M., Abdul Rahman, A., Uddin, N., Memon, J. and Kazi, S. (2013). Signature-based Multi-Layer Distributed Intrusion Detection System using Mobile Agents. International Journal of Network Security. 15. 79-87.

Vigneswaran, R. K., Vinayakumar, R., Soman, K. P. and Poornachandran, P. (2018) Evaluating Shallow and Deep Neural Networks for Network Intrusion Detection Systems in Cyber Security, 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1-6, doi: 10.1109/ICCCNT.2018.8494096.

Vinayakumar, R., Alazab, M., Soman K. P., Poornachandran, P., Al-Nemrat, A. and Venkatraman, S., Deep Learning Approach for Intelligent Intrusion Detection System (2019), in IEEE Access, vol. 7, pp. 41525-41550, 2019, doi: 10.1109/ACCESS.2019.2895334.

ACCESS.2019.2905633.

Xu, C., Shen, J., Du, X., and Zhang, F. (2018) An Intrusion Detection System Using a Deep Neural Network With Gated Recurrent Units, in IEEE Access, vol. 6, pp. 48697-48707, 2018, doi: 10.1109/ACCESS.2018.2867564.

Yang, Y., Zheng, K., Wu, C. and Yang, Y. (2019) Improving the Classification Effectiveness of Intrusion Detection by Using Improved Conditional Variational AutoEncoder and Deep Neural Network. Sensors. 19. 2528. 10.3390/s19112528.

Yang, Y., Zheng, K., Wu, B., Yang, Y. and Wang, X. (2020) Network Intrusion Detection Based on Supervised Adversarial Variational Auto-Encoder With Regularization, in IEEE Access, vol. 8, pp. 42169-42184, 2020, doi: 10.1109/ACCESS.2020.2977007.

Yin, C., Zhu, Y., Fei, J. and He, X. (2017) A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks, in IEEE Access, vol. 5, pp. 21954-21961, 2017, doi: 10.1109/ACCESS.2017.2762418.

Zhang, H., Wu, C. Q., Gao, S., Wang, Z., Xu, Y. and Liu, Y. (2018) An Effective Deep Learning Based Scheme for Network Intrusion Detection, 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 682-687, doi: 10.1109/ICPR.2018.8546162.

Zhang, J. and Mani, I. (2003) KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. Proceeding of International Conference on Machine Learning (ICML 2003), Workshop on Learning from Imbalanced Data Sets, Washington DC, 21 August 2003.

Zhang, H., Huang, L., Wu, C. Q. and Li, Z. (2020) An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset, Computer Networks, Volume 177, 2020, 107315, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2020.107315.

Zhiqiang, L., Mohi-Ud-Din G., Bing, L., Jianchao, L., Ye, Z. and Zhijun, L. (2019) Modeling Network Intrusion Detection System Using Feed-Forward Neural Network Using UNSW-NB15 Dataset, 2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), 2019, pp. 299-303, doi: 10.1109/SEGE.2019.8859773.

Zhong, M., Yajin, Z. and Chen, G. (2021) Sequential Model Based Intrusion Detection System for IoT Servers Using Deep Learning Methods. Sensors. 2021; 21(4):1113. https://doi.org/10.3390/s21041113

APPENDIX

APPENDIX A) SOURCE CODE

The source codes of 10 different sampling methods and 3 different classifiers written in Python are provided in https://github.com/aokanarik/ROGONG-IDS.

RESUME