ICCSPA20 1570670792

Balancing Approaches towards ML for IDS: A Survey for the CSE-CIC IDS Dataset

Subiksha Srinivasa Gopalan[1], Dharshini Ravikumar[1], Dino Linekar[1], Ali Raza[1], Maheen Hasib[2]

[1]*Department of Electrical Engineering and Computing Sciences, Rochester Institute of Technology, Dubai, UAE*
[2]*Department of Liberal Arts and Mathematical Sciences, Rochester Institute of Technology, Dubai, UAE*
{ssg5920, dr3086, dxr4536, axrcada, mhcad} @rit.edu

***Abstract*—Balanced datasets play a key role in the bias observed in machine learning algorithms towards classification and prediction. The CSE-CIC IDS datasets published in 2017 and 2018 have both attracted considerable scholarly attention towards research in intrusion detection systems. Recent work published using this dataset indicates little attention paid to the imbalance of the dataset. The study presented in this paper sets out to explore the degree to which imbalance has been treated and provide a taxonomy of the machine learning approaches developed using these datasets. A survey of published works related to these datasets was done to deliver a combined qualitative and quantitative methodological approach for our analysis towards deriving a taxonomy. The research presented here confirms that the impact of bias due to the imbalance datasets is rarely addressed. This data supports further research and development of supervised machine learning techniques which reduce the impact of bias in classification or prediction due to these imbalance datasets.**

Keywords—balance, dataset, intrusion detection system, machine learning

I. Introduction

Supervised machine learning (ML) techniques require labelled datasets which are used to train a classification or prediction model. In a binary classification problem, the true or false labels need to be sufficient in quantity when each data sample has a large number of features used as an input to train the model. The performance of a ML model can be strongly affected by the dataset which is used for training. A dataset which is imbalanced can lead to a biased classification or prediction [1]. Intrusion detection systems (IDS) are able to collect samples of network traffic using fiber optical taps at various points in a network [2]. The samples can be used as inputs to a trained ML model to classify a data sample as benign or malicious in the simplest case of binary classification [3]. Predictions could also be made on potential attacks using such models [4]. In a real-world enterprise environment, the traffic collected from these taps is expected to have a high ratio of benign to malicious samples. In this paper we focus on the CSE-CIC IDS datasets which were published in 2017 and 2018. These datasets are both rich in terms of the features each data sample has and in terms of the different types of cyberattacks. The goal of this study is to analyze the contributions which use these datasets for ML. Research on the development of trained models which can accurately classify cyber intrusion events has received growing attention [5,6]. The CIC-IDS dataset [7] has been recently used in a number of studies towards ML for IDS [8,9]. In [10] and [11] the authors have reported on the accuracy of proposed ML techniques which account for the imbalance of the dataset. However, little research to date has focused on the imbalance of the dataset. The analysis presented in this paper is used to provide a taxonomy of the supervised ML techniques applied to train models for classification or prediction. The aim of this work is two-fold: (1) identify research opportunities in the development of models that account for bias and (2) identify requirements for new synthetic or semi-synthetic datasets which are balanced. The results of this analytical study suggests that the aims are well defined. This is due to a high count of published work in which a treatment of the imbalance of these datasets is missing. However, it is also encouraging to report that the imbalance issue is not completely ignored in the literature and some work has provided insights to this. Our derived taxonomy provides researchers with an opportunity to revisit algorithms developed and enhance their capabilities with techniques that account for imbalance.

This paper is organized as follows: Section II reviews published work which uses the CSE-CIC IDS dataset. Section III discussed the methodology which was used for this survey and analysis. Section IV discusses the results of the analysis and Section V concludes the study.

II. Literature Review

A large and growing body of literature has investigated ML methods applied to cybersecurity datasets. In this study we found that 75 articles were published between 2017 and 2020 on ML which use the CCSE-CIC IDS dataset.

In [12] the authors detail the features and attacks which constitute their CSE-CIC IDS 2017 dataset. In [13], the authors apply Deep Multilayer Perceptron (DMPL) structure which uses recursive feature elimination performed using random forest (RFE-RF) and reports on the accuracy metric. In this work the authors do not indicate any treatment of the imbalance of the dataset. An approach towards addressing dataset imbalance was provided in [14] where a data level technique is applied in which the minority malicious traffic data is oversampled. A similar technique in [15] also applies a data level technique in which downsampling of the majority benign traffic is used to mitigate the class imbalance. In [16], an algorithm level approach is proposed as a solution to mitigating the imbalance. In this work the metrics reported include the accuracy, precision, F-Measure, Detection Rate

978-1-7281-6535-6/20/$31.00 ©2020 IEEE

(DR), Attack Detection Rate (ADR), False Alarm Rate (FAR), Model Building Time (MBT).

The CSE-CIC IDS 2018 dataset is thoroughly covered in [17] similar to [12] in which the authors detail the improvement over the 2017 dataset. Later in 2018, [18] is one of the many published works that apply supervised ML classifiers and reports on the well-known metrics related to the performance of the ML model. In 2019, [19] is one of the few papers found that applied the data level technique to the CSE-CIC IDS 2018 dataset. In the same year, in [20] an algorithm level technique was applied.

A combination of data level techniques and algorithm level techniques were reported in [21] and [22]. Such a combination of techniques was first reported in [22] as a hybrid technique. However, we note that the techniques in [21] and [22] are applied on the CSE-CIC IDS 2017 dataset.

III. Methods

Our approach to deriving a new taxonomy is based on the study of work published in the cybersecurity domain anchored on the CSE-CIC IDS dataset. A criteria for selecting published work for this dataset, related to AI based intrusion detection systems was developed. The criteria used for selecting papers related to IDS and dataset balancing is specified in this section as follows:

Criteria for selecting the papers related to IDS were as follows:

- Publications were only included if they were relevant to the CSE-CIC IDS datasets.
- Publications were only included if certain keywords and phrases related to dataset balancing were found. These included but not limited to: imbalance; minority class(es); majority class(es), sampling, downsampling, upsampling, balance(ing).
- Publications were only included if they were published between 2018 and 2020 to align with the first public announcement of the dataset.
- The number of citations and venue of publication was considered for each work assessed.
- Publications were only included if they approached imbalance and cited the other non-IDS related work on dataset balancing.

To come up with this methodology, previous approaches published in [24] and [25] were reviewed and analyzed. The two papers were compared, in [24] more advanced techniques were proposed due to the recent advancements provided by google scholar platform reported in this paper.

From [24], the google scholar as a platform provides access to key information about the citation of a paper. The name of the primary dataset paper is first entered in the google scholar search engine. The number of times the paper has been cited is displayed and clicking on it leads to the total list of cited papers. The papers are filtered down by clicking the checkbox **"Search within cited articles"**, the keyword search and custom range process used for the datasets are explained below:

The keyword **"imbalance"** was used for the CSE-CIC IDS dataset and a total of 75 papers were generated. A custom range option is also available in google scholar to select the key papers. The range applied for our research is *2018-2020.*

IV. Results And Discussion

Much of the literature on ML methods reviewed in this work, applied these methods to CIC-IDS dataset with a focus on reporting the accuracy of an algorithm or method. The research to date has been predominantly in optimizing the salgorithms used in ML, towards improving the metrics shown in Table 1. The papers surveyed in this study do not convincingly show that imbalance datasets are treated and to what degree.

TABLE I. Confusion Matrix

Confusion Matrix		**Classification**	
		Positive	***Negative***
Matrix	***Positive***	TP	FN
	Negative	FP	TN

Confusion Matrix is a performance measurement for machine learning classification. The confusion matrix in Table 1, comprises of four items for binary classifiers:

True Positives (TP) - when the classifier identifies the true positive label as positive.
True Negatives (TN) - when the classifier identifies the true negative label as negative.
False Positives (FP) - when the classifier identifies the true negative label as positive.
False Negatives (FN) - when the classifier identifies the true positive label as negative.

In the context of cybersecurity research, a well-known understanding is that a positive event is defined as a malicious event and the correct classification of such an event is deemed as a true positive outcome. A negative event is a benign event and the correct classification is deemed as true negative. Inaccurate classification can mean that a benign event is classified as a malicious event. This misclassification is deemed as a false positive. Likewise, for a malicious event to be classified as a benign event is deemed a false negative [26].

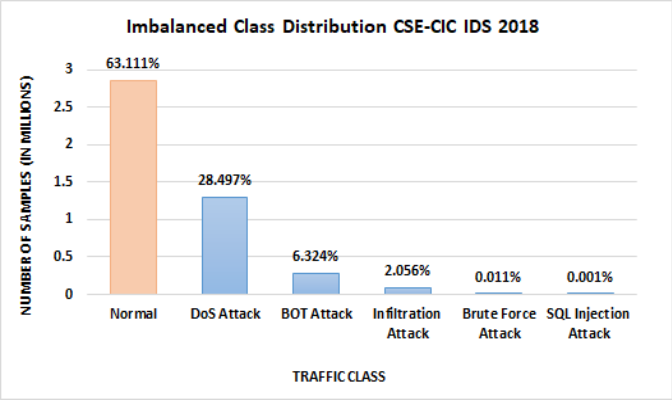

Fig. 1. Imbalanced Class Distribution CSE-CIC IDS 2018

Figure 1 highlights the extent of the imbalance observed in the CSE-CIC IDS 2018 dataset in [27]. There are different types of malicious events such as BOT, DoS, Brute Force, Infiltration, SQL injection. Heartbleed is not included in Figure 1 as the number of samples is very small. By far the most typical trend observed when evaluating the distribution of samples per class in datasets is exponential. This trend indicates that the percentage of benign samples is higher than malicious samples. A dataset best serves in machine learning algorithms when the distribution is normal or close to normal.

The issue of class imbalance distribution is dominant across all domains [28]. In this paper, two metrics - precision and recall - are studied and referenced for issues that relate to class imbalance. A dataset contains imbalanced class distribution when one class - which is often the one of interest - is a minority class or not represented adequately. A number of publications have been reviewed to determine the impact of dataset imbalance. By far the most widely accepted account can be found in [28] which states that issues arising from this are poor accuracy in classification and a general bias in the results obtained. Classifier algorithms used in machine learning as mentioned in [29] require a balanced dataset. In spite of these findings, there are recent developments in algorithms that account for the imbalance in a dataset. Such approaches are generalized as algorithm level approaches towards the mitigation of bias [30].

In this paper, we determine the CSE-CIC IDS 2018 dataset of having a benign to malicious imbalance ratio of 1.71 ***to*** 1. Table 2 represents the imbalance taking into account all the attack types included in the CSE-CIC IDS 2018 dataset in [27]. The benign traffic in this dataset is found to be 1.71 times more than the total malicious traffic.

TABLE II. CLASS DISTRIBUTION FOR CSE-CIC IDS 2018

Class Label	Count
Benign	2,856,035
BoT	286,191
Brute Force	513
DoS	1,289,544
Infiltration	93,063
SQL Injection	53
Total	*4,525,399*

TABLE III. IMBALANCE RATIO DISTRIBUTION FOR CSE-CIC IDS 2018

Normal Traffic Count	Attack Traffic Type	Count	Normal to Attack Ratio
2856035 **(63.11%)**	BoT	286191 **(6.324%)**	10 ***to*** 1
	Brute Force	513 **(0.011%)**	5567 ***to*** 1
	DoS	1289544 **(28.49%)**	2 ***to*** 1
	Infiltration	93063 **(2.056%)**	31 ***to*** 1
	SQL Injection	53 **(0.001%)**	53887 ***to*** 1

In Table 3 we provide an approximation of the ratios of benign to malicious traffic in the various minority classes provided in Table 2 from [4]. For the most populous minority class, DoS, a ratio of more than 2 *to* 1 is observed. The next minority class, BOT has a ratio of approximately ten *to* one. Infiltration has a ratio of 31 *to* 1 followed by Brute Force with a ratio of 5567 to 1. SQL Injection has a severely imbalanced ratio of 53887 to 1. Hence, a ML algorithm trained on this dataset will tend to bias towards classifying a SQL Injection event as a benign event, a false negative. The general classification of malicious will have a ratio of 1.71 *to* 1 which is an improvement but still a distance from a much required balanced dataset. Cross validation techniques may be used to mitigate the effect but will not be very accurate in the case of minority classes as discussed in [12]. Hence, from the above table it can be inferred that the attack traffic will yield a high rate of False Negatives and a low rate of False Positives.

A. Overview of the CSE-CIC IDS used in this study

CSE-CIC IDS 2018 is a popular dataset with a large number of published works explored in the field of IDS. CSE-CIC IDS 2018 is developed using the AWS platform (Amazon Web Services). This dataset provides numerous attack profiles that can be used in the field of intelligent security and applied to network topologies and protocols in a generic approach [18]**.** This dataset was enhanced in consideration with the standards of CSE-CIC IDS2017. CSE-CIC IDS2018 is a public dataset which is currently in use that has 2 profiles classified, and consists of 5 different attack methods. Various data scenarios were collected and the raw was edited on a daily basis. 80 statistical properties such as packet length, number of packets, number of bytes etc were calculated in forward and reverse direction separately while creating data. Finally, the dataset was published over the internet to all the researchers out there. The dataset is published in two formats

CSV and PCAP with approximately 5 million records. CSV format is mainly used in the field of AI and PCAP format is used for extracting new features [4].

B. Analysis of Published Work

A systematic review of the literature in the cohort of published works from Section III allowed us to divide the work into groups according to the level of contribution each work makes towards balancing of a dataset. This grouping has been done in a hierarchical order as shown in Figure 4 with the first level determining whether the paper has a contribution or not. The second level identifies the extent of the contribution in terms of a proposed method or the application of an existing method. In the case of non-contributing work, the second level identifies whether imbalance has been recognized and mentioned or not.

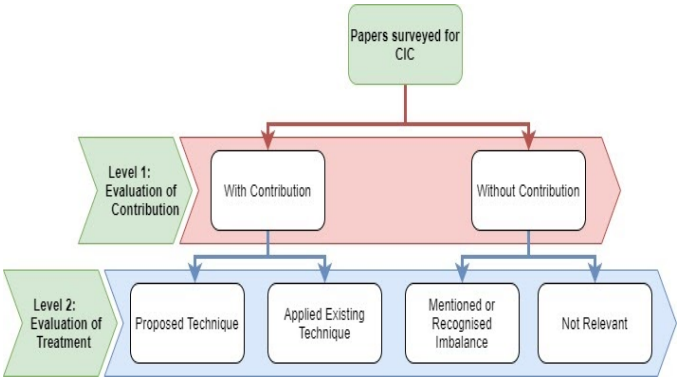

Fig. 2. Hierarchical grouping of published work

A more thorough definition of the grouping has been provided below:

- **Proposed:** The authors have proposed a technique to solve imbalance.
- **Applied Existing:** The authors have applied existing techniques in solving imbalance.
- **With Contribution:** This is a cumulative of papers in which the authors have either proposed or applied existing techniques.
- **Mentioned:** These are the papers in which the authors have mentioned an imbalance technique with respect to our analysis but have not treated it.
- **Not Relevant:** The authors have mentioned imbalance in general and not with respect to our analysis of dataset imbalance.
- **Without Contribution:** The authors of these papers have either mentioned imbalance or did not have any discussion relevant to the imbalance of the datasets.

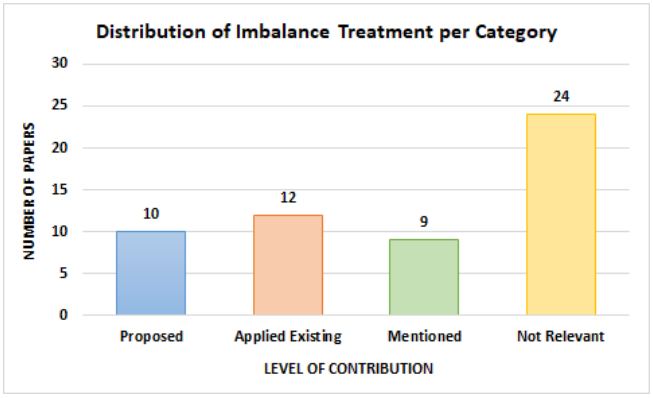

Fig. 3. Comparision of the imbalance treatment catogories in CSE-CIC IDS 2017 and 2018

The categories: *with contribution* and *without contribution* have been added for the ease of presenting the analysis. Papers that have proposed a new technique or applied an existing technique to deal with the imbalance are labelled as ***With Contribution***. Furthermore, papers that have no relevance to dataset imbalance or have reservedly mentioned the **word** imbalance are collectively labelled as ***Without Contribution***.

Figure 5 shows the distribution of published work across the four groups defined for CSE-CIC IDS dataset. Figure 6 presents a cumulative percentage distribution across the four groups irrespective of the datasets used. It has been observed that only 40% of the papers have either proposed or applied techniques to treat dataset imbalance and the remaining 60% of the papers have not contributed to this study. These results further support the idea that there is a lack of attention to imbalance because 44% of the papers are ***not relevant*** which takes precedence over other groups.

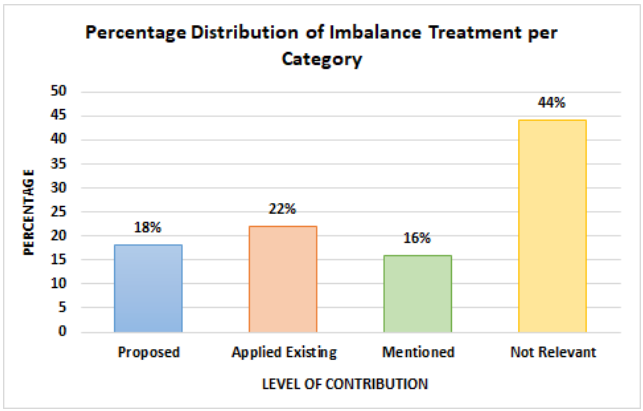

Fig. 4. Categorization of Imbalance Treatment – A comparison of all papers surveyed, across all categories of imbalance treatment.

The "***without contribution***" category shown in Figure 5 takes precedence with the highest count. In the "***with contribution***" category the ***applied existing*** takes precedence as shown in Figure 7

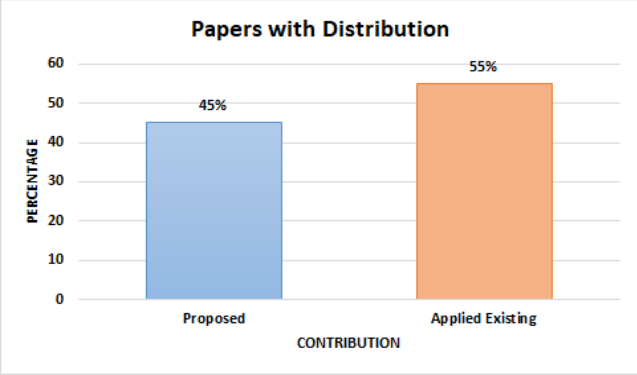

Fig. 5. Comparision of papers in which proposed techniques are compared with the exisiting techniques that have been applied.

C. Level Distribution

A study of the proposed and applied existing papers was undertaken to determine the technique which has been used. In contrast to the findings in the literature review, the outcome of this particular study demonstrates that there are three distinct classifications of techniques for balancing of datasets as shown in Figure 8. The grouping or classification of these techniques are defined as levels to be consistent with published literature presented in Section II.

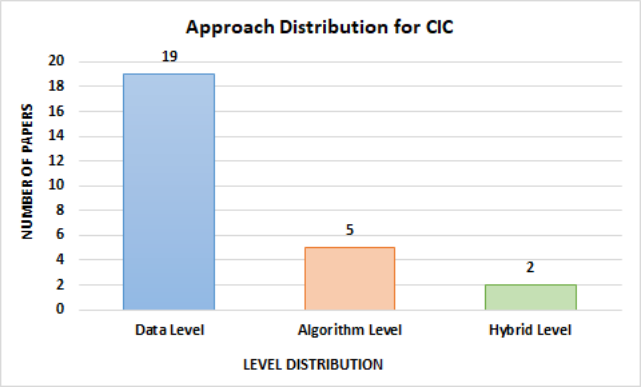

Fig. 6. Comparing the number of papers that propose, apply or mention approaches in CSE-CIC IDS dataset 2017 and 2018

The statistical representation in Figure 8 spans the 26 papers published in the CSE-CIC IDS. This dataset shows that the majority of papers use data level techniques to solve the issue of imbalance.

D. Ranking of undertaken approaches

The percentage distribution for ***proposed*** and ***applied existing*** from the total amount of papers ***with contribution*** is depicted in Figure 7. Papers with contribution had 45% of new methods proposed and 55% of existing methods applied. This result may be explained by the fact that a majority of the papers have not focused on proposing a new technique to overcome bias. This discrepancy can be a dominant focus area for future research directions.

CONCLUSION

The findings in this work clearly suggests an opportunity for research work which proposes novel techniques to handle the dataset imbalance in the CSE-CIC IDS datasets. The formal definition of the hybrid level technique in the proposed taxonomy can be used by researchers to explore new combinations that have not been discovered in literature related to these datasets. Most interestingly, we found an opportunity to apply hybrid level techniques to the CSE-CIC IDS 2018 dataset as these have only been found in the CSE-CIC IDS 2017 dataset. Furthermore, researchers may consider developing novel hybrid techniques to improve the ML model performance metrics for classification and prediction.

REFERENCES

[1] R. Malhotra and S. Kamal, "An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data," *Neurocomputing*, vol. 343, pp. 120–140, 2019, doi: 10.1016/j.neucom.2018.04.090.

[2] "Introduction to Network-Based Intrusion Detection Systems | Network-Based Intrusion Detection | InformIT." [Online]. Available: https://www.informit.com/articles/article.aspx?p=782118. [Accessed: 08-Aug-2020].

[3] C. Yin, Y. Zhu, J. Fei, and X. He, "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks," *IEEE Access*, vol. 5, pp. 21954–21961, 2017, doi: 10.1109/ACCESS.2017.2762418.

[4] V. Jaganathan, P. Cherurveettil, and P. Muthu Sivashanmugam, "Using a prediction model to manage cyber security threats," *Sci. World J.*, vol. 2015, 2015, doi: 10.1155/2015/703713.

[5] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, 2019, doi: 10.1186/s42400-019-0038-7.

[6] R. Mitchell and I. R. Chen, "A survey of intrusion detection techniques for cyber-physical systems," *ACM Comput. Surv.*, vol. 46, no. 4, 2014, doi: 10.1145/2542049.

[7] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *ICISSP 2018 - Proc. 4th Int. Conf. Inf. Syst. Secur. Priv.*, vol. 2018-Janua, no. Cic, pp. 108–116, 2018, doi: 10.5220/0006639801080116.

[8] C. F. Tsai, Y. F. Hsu, C. Y. Lin, and W. Y. Lin, "Intrusion detection by machine learning: A review," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 11994–12000, 2009, doi: 10.1016/j.eswa.2009.05.029.

[9] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," *Proc. - IEEE Symp. Secur. Priv.*, pp. 305–316, 2010, doi: 10.1109/SP.2010.25.

[10] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.

[11] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018, doi: 10.1016/j.neunet.2018.07.011.

[12] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *ICISSP 2018 - Proc. 4th Int. Conf. Inf. Syst. Secur. Priv.*, vol. 2018-Janua, no. Cic, pp. 108–116, 2018, doi: 10.5220/0006639801080116.

[13] S. Ustebay, Z. Turgut, and M. A. Aydin, "Intrusion Detection System with Recursive Feature Elimination by Using Random Forest and Deep Learning Classifier," *Int. Congr. Big Data, Deep Learn. Fight. Cyber Terror. IBIGDELFT 2018 - Proc.*, pp. 71–76,

2019, doi: 10.1109/IBIGDELFT.2018.8625318.

[14] J. H. Lee and K. H. Park, "AE-CGAN model based high performance network intrusion detection system," *Appl. Sci.*, vol. 9, no. 20, 2019, doi: 10.3390/app9204221.

[15] Y. Zhang, X. Chen, L. Jin, X. Wang, and D. Guo, "Network Intrusion Detection: Based on Deep Hierarchical Network and Original Flow Data," *IEEE Access*, vol. 7, pp. 37004–37016, 2019, doi: 10.1109/ACCESS.2019.2905041.

[16] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, "Building an efficient intrusion detection system based on feature selection and ensemble classifier," *Comput. Networks*, vol. 174, 2020, doi: 10.1016/j.comnet.2020.107247.

[17] G. Karatas, O. Demir, and O. K. Sahingoz, "Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset," *IEEE Access*, vol. 8, pp. 32150–32162, 2020, doi: 10.1109/ACCESS.2020.2973219.

[18] L. D'Hooge, T. Wauters, B. Volckaert, and F. De Turck, "Classification hardness for supervised learners on 20 years of intrusion detection data," *IEEE Access*, vol. 7, pp. 167455–167469, 2019, doi: 10.1109/ACCESS.2019.2953451.

[19] G. Karatas, O. Demir, and O. K. Sahingoz, "Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset," *IEEE Access*, vol. 8, pp. 32150–32162, 2020, doi: 10.1109/ACCESS.2020.2973219.

[20] V. A. Chastikova and V. V. Sotnikov, "Method of analyzing computer traffic based on recurrent neural networks," *J. Phys. Conf. Ser.*, vol. 1353, no. 1, 2019, doi: 10.1088/1742-6596/1353/1/012133.

[21] I. A. Jimoh, I. Ismaila, and M. Olalere, "Enhanced Decision Tree-J48 with SMOTE Machine Learning Algorithm for Effective Botnet Detection in Imbalance Dataset," *2019 15th Int. Conf. Electron. Comput. Comput. ICECCO 2019*, no. Icecco, 2019, doi: 10.1109/ICECCO48375.2019.9043233.

[22] R. Panigrahi and S. Borah, "Dual-stage intrusion detection for class imbalance scenarios," *Comput. Fraud Secur.*, vol. 2019, no. 12, pp. 12–19, 2019, doi: 10.1016/S1361-3723(19)30128-9.

[23] L. McNabb and R. S. Laramee, "How to Write a Visualization Survey Paper : A Starting Point," 2019, doi: 10.2312/eged.20191026.

[24] J. Beel, B. Gipp, and E. Wilde, "Academic search engine optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co," *J. Sch. Publ.*, vol. 41, no. 2, pp. 176–190, 2010, doi: 10.3138/jsp.41.2.176.

[25] M. H. Abdulraheem and N. B. Ibraheem, "A detailed analysis of new intrusion detection dataset," *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 17, pp. 4519–4537, 2019.

[26] "Battling Data Imbalance in IBM HR Attrition Challenge | by Александра Пухова | Medium." [Online]. Available: https://medium.com/@alepukhova526/battling-data-imbalance-in-ibm-hr-attrition-challenge-3a26337a4943. [Accessed: 08-Aug-2020].

[27] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A review," *Int. J. Adv. Soft Comput. its Appl.*, vol. 7, no. 3, pp. 176–204, 2015.

[28] F. Taghi M. Khoshgoftaar, Chris Seiffert, Jason Van Hulse, Amri Napolitano, "Learning with Limited Minority Class Data," *Proc. - 6th Int. Conf. Mach. Learn. Appl. ICMLA 2007*, pp. 13–18, 2007, doi: 10.1109/ICMLA.2007.76

[29] "IDS 2018 | Datasets | Research | Canadian Institute for Cybersecurity | UNB." [Online]. Available: https://www.unb.ca/cic/datasets/ids-2018.html. [Accessed: 08-Aug-2020].