2022 10th International Symposium on Digital Forensics and Security (ISDFS) | 978-1-6654-9796-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISDFS55398.2022.9800807

Network Intrusion Detection Packet Classification with the HIKARI-2021 Dataset: a study on ML Algorithms

Rui Fernandes
School of Technology, IPCA
Barcelos, Portugal
a17618@alunos.ipca.pt

Nuno Lopes
2AI - School of Technology, IPCA
Barcelos, Portugal
https://orcid.org/0000-0001-8897-5061

***Abstract*—The Intrusion Detection System is a critical part of a network infrastructure to detect and prevent cyberattacks. The use of Artificial Intelligence has the potential to improve the performance of IDS in achieving cybersecurity. However, one of the challenges nowadays is the lack of good datasets that can improve the results of AI algorithms. In this paper we study the recently published HIKARI-2021 dataset, built from real data in a lab to develop network traffic and classification models. A feature selection method was used to evaluate the relevant features, and different Machine Learning methods were tested with this dataset.**

The results show that the dataset is suitable for classification and that the feature size of the dataset can be reduced from 83 to 22 entries, while still maintaining an accuracy of 99%, for a faster algorithm execution. When using a balanced sample of this dataset, we obtained an accuracy above 80% on some ML algorithms.

***Index Terms*—Network Intrusion Detection System, Machine Learning, HIKARI-2021**

I. Introduction

As the world of technology grows, the IoT devices are getting smarter, with more potential to every user and with that the security of the user's private data becomes more and more a concern [1].

Cybersecurity is widely studied, and it keeps being more complex, increasing the chances of a problem occurring. Every day multiple people report cyberattacks in their own space, industries, and the services both in the private and public sectors get attacked, having frequently the users' data as the target or the shutdown of some service as the focus [2].

As the world of cybersecurity grows, the hackers become more expert and the attacks more complex and unknown, but the security experts also learn with that, which creates the need for more studies and using new technologies, such as *Artificial Intelligence*, to classify, predict and analysis of attacks [3].

Artificial Intelligence is being used in different industries, being Cybersecurity one of the main focus. With the processing capacity of machines, which is increasing exponentially nowadays, it is easier to apply heavy computation algorithms to tackle cybersecurity. However, the lack of reference datasets is of great importance. How can we measure and identify the attacks, and what about the new attacks that have never been done? There are questions hard to answer but AI is looking forward to investigating this domain. *Classification* is studied with great results in the identification of attacks, *Linear Regression* is predicting attacks based on network traffic and history of data, and *Reinforcement Learning* is already being applied in real-time network traffic to the analysis of unknown packets of data [3].

Having the power of processing and the algorithms developed the only thing missing is data. Data in Cybersecurity is not easy to acquire since that it is a sensibly information. The data of cyberattacks can result in dangerous situations since not every system is ready for every attack and it can get attention from hackers. Another problem is that most public datasets are outdated and with low diversity of network data. This is the most problematic subject in AI: having good data, up to date and in large volume in order to be possible to achieve great results.

A new dataset, *HIKARI-2021*, published in August 2021 was developed in order to contribute to the difficulty in evaluating Intrusion Detection Systems [4]. This dataset contains encrypted synthetic attacks and benign traffics. The process of the creation the dataset is explained in the article and it was produced by creating a Cybersecurity lab with a attacker machine and a victim machine using the open source software *Zeek* which saved data over two months.

With that dataset, it was possible to compare with others already published but outdated and different Machine Learning algorithms were used to test the performance of the dataset. This paper propose a deeper analysis of the dataset by checking the viability of the data, using different algorithms, less data than the whole dataset and feature selection in order to see which features impacts the performance of the classification algorithms that can be use in this dataset.

II. Related work

The need of studying the use of Artificial Intelligence in Cybersecurity world is becoming a global interest both academic and professional. There are different ways of exploring AI to

978-1-6654-9796-1/22/$31.00 ©2022 IEEE

Cybersecurity, it can be used in predicting attacks by analyzing current network in real-time [5], [6], classify the attacks when they happen, detect spam in emails and other information in internet [7] and also in attacks like password guessing [8]. The ways that AI, specifically ML, can be applied in Cybersecurity is countless and it is always growing with different challenges to the ones who want to protect data from intruders.

Intrusion-detection systems (IDS) is seen as a security tool that constantly monitors the host and network traffic to detect any suspicious behavior that violates the security policy and compromises its confidentiality, integrity and availability [9], [10]. It started being developed in 1980s and since that it became a fundamental research and development area of computer security, and they still growing with higher features and more safer. The history of IDS shows the evolution of the name by itself starting in Intrusion-Detection Expert Systems (IDES) [11] where those systems were based in rules and quickly researches were thinking in Next Generation Intrusion Detection Expert System (NIDES) and with the help of AI the expert systems would be a reality.

Due the subject of Cybersecurity be a sensitive area we can almost be completely sure and say that every IDS is outdated because there will be always a new exploit or attack and those systems must be adaptive and ready to learn more and with that requirement the use of Machine Learning is mandatory. Machine Learning Based Intrusion Detection System [12], [13] have been developed achieving great results in classification of network traffic. Machine Learning algorithms used to classify the attacks are in the areas of supervised and unsupervised learning as well as Reinforcement Learning [14]. Supervised learning which is the topic in this work shows great results with the KNN [15], Support Vector Machine [16], ANN [17], [18], Random Forest [19] and others algorithms but this requires having a labeled dataset which not always is possible.

Related to datasets, there are a few recent datasets [20]–[22] and the problem of being updated is a concern that makes researchers to build there own [4] with specific attacks and traffic.

III. Dataset Analysis

Creating a dataset based on a real dataset is hard since it depends on different factors to make it real and, with that requirement, the idea of *HIKARI-2021* becomes in creating a lab where network traffic was recorded with real data due to an IDS: having background traffic captured without any filter or firewall, so that's a chance of this traffic containing malicious traffic or attacks, benign profile using a profile similar to human behaviour by running a script using Selenium to run a browser like a human and the attacks data was generated in other machines with attacks categorized by Brute Force and Brute force-XML and probing.

The dataset labelling was done by *Zeek* software which produces two labels: *"traffic_category"*, which represents the name of the traffic type, and *"label"* which has a value with 0 representing Benign and 1 representing Attack. The distribution of the dataset can be seen in figure 1.

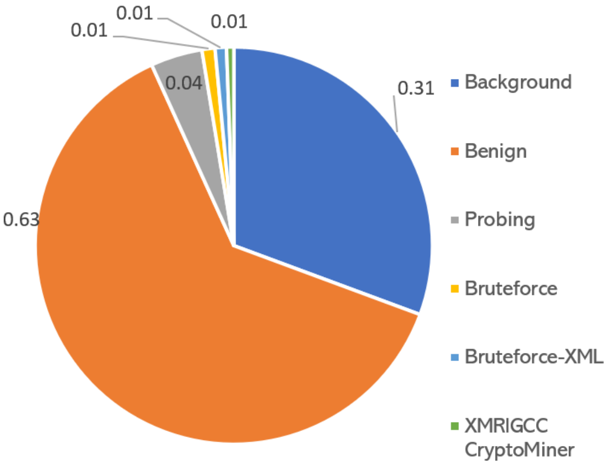

Fig. 1: Network traffic distribution on Dataset

Having a good dataset is substantial and the next step evolves to analyse the data, when dealing with a large amount of data it is necessary to clean as much as we can to provide good data for our models and test different ways of finding what counts and impacts our results so we can achieve the best results possible. The dataset is composed of *555 278 entries*and has *83 features*. Because the data source is a software of IDS, there's probably data cauterizing the traffic that will not be relevant for our classification model so a feature analysis must be done.

The first approach when analyzing a dataset with a high number of features is to see the correlation between features as it is common to encounter a high correlation and, if it happens, some of them can be excluded from the dataset because they won't have much impact on the final result.

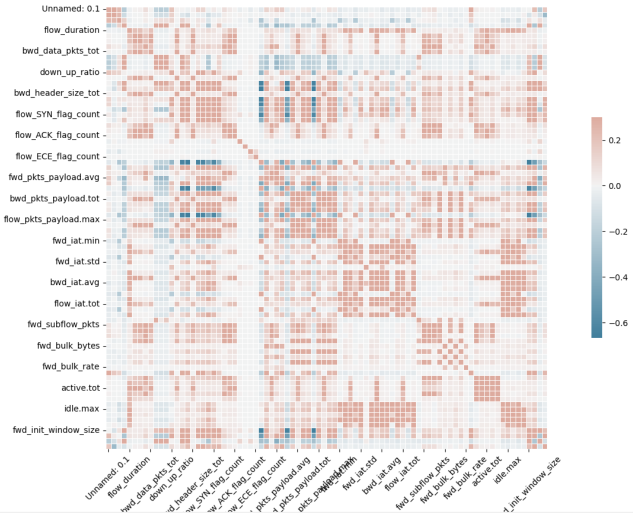

Fig. 2: Correlation graphic of the Dataset

A matrix graphic was obtained to visualize the correlation (figure 2). There is some difficulty to identify higher correlations because none is relevant (high values of correlation coefficients). With that, we can conclude that the features are independent of each other and not repeated. Discarding the first approach we decided to go deeper using statistical analysis, using the *Chi-squared* test, which measures dependence between stochastic variables to get the k features that are relevant to test the model. For that, the whole dataset was used with Random Forest and KNN models with different numbers of features selecting always the best k number of features to the chi-squared test and we realised (figure 3) that for both 22 features algorithms give approximately the same accuracy (99%) as using 83 features so it will shorten the processing time required.

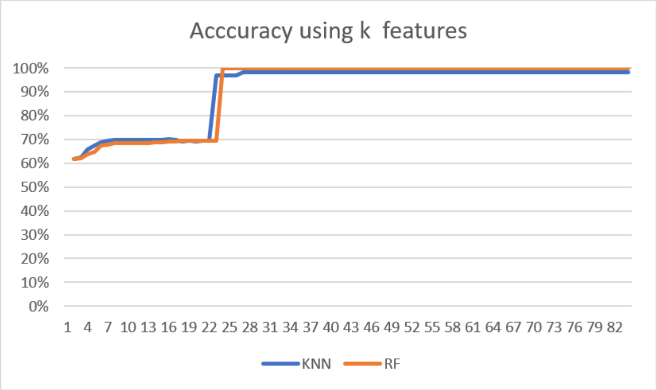

Fig. 3: Accuracy according k number of features

IV. Machine Learning Algorithms

In order to evaluate the dataset [4] the same algorithms used in the publication were tested and the metrics *(accuracy, balanced accuracy, precision, recall, f1 score and time)* were collected. Those test were made to both approaches using all features and the 22 features selected using the chi-squared tests.

As explained before, the dataset is not balanced so two tests were made, one considering the full dataset not balanced (equally done in the paper mentioned) and one using the same entries per label.

Those tests were made in a ASUS GL503-GE with a i7-8750H, 16Gb RAM and a NVIDIA GEForce GTX1050 TI 4Gb with no powerfully background apps running simultaneously. The library scikit-learn 1.1.0 was used with Python 3.10 to use different ML algorithms and functions to split the dataset, train the model and evaluate results. The dataset was shuffled and divided into 80% for training and 20% for test.

V. Results and Discussion

The Table 1 shows that the results are similar to the original paper of the dataset [4] although the Support Vector Machine and the Neural Network Classifier algorithms show worse results in the balanced accuracy when compared to the original paper's results. In the same table it is possible to compare the results between the use of all features and the use of only the 22 features that were previously selected. The feature comparison shows that in the KNN and RF algorithms the results are both similar between them and with the tests made in the original paper. The SVM and MLP algorithms show that to obtain better results, the 22 features aren't enough.

The Table 2 shows the results of testing the four algorithms using the entire dataset, corresponding to 555278 entries, or half of the dataset (with entries chosen randomly) and with those two cases a comparison was made using all of the features and just the 22 previously selected features. The results of using the full dataset or the half dataset shows that there aren't significant differences between them when using the KNN, MLP or RF algorithms. With the SVM algorithm, the results can decrease almost 10% when using half dataset instead of the full. When comparing the results with the two feature groups (with sizes 83 and 22 respectively), we can see that for the KNN and RF algorithms the results can vary up to 2%, although they are above 96%. For the MLP and SVM algorithms, the drop of accuracy can be up to 30%. Relative to the time needed for training, we assume that there is a strong dependency between the time, the data quantity and the number of features used when usigng the KNN, SVM and RF algorithms.

The Table 3 shows the results obtained with a balanced dataset, given that the category ”XMRIGCC CryptoMiner” only has less than 3000 entries. Two different tests were made with 1500 and 750 entries per class on the dataset. The dataset has highly relevant features that even with fewer data entries, they can achieve results with an accuracy and precision above 80% using the KNN or RF algorithms and above 77% using the MLP algorithm. The SVM algorithm shows worse results with 34% of accuracy and 38% of precision which means that the SVM would need more data and more features to achieve better results.

The results in all tables show that the dataset is adequately built for this purpose, being used as a classification dataset.

TABLE I: Comparison of the results obtained with the original paper's results [4].

Algorithm	Features	Accuracy	Accuracy [4]	Balanced Accuracy	Precision	Recall	F1
KNN	83	0.98	0.98	0.97	0.98	0.98	0.98
	22	0.97	-	0.95	0.97	0.97	0.97
MLP	83	0.90	0.99	0.61	0.90	0.90	0.89
	22	0.63	-	0.17	0.39	0.63	0.48
SVM	83	0.92	0.99	0.44	0.9	0.92	0.91
	22	0.57	-	0.53	0.48	0.57	0.49
RF	83	1	0.99	1	1	1	1
	22	1	-	1	1	1	1

TABLE II: Comparison results using 22 features and full or half dataset

Algorithm	Entries	Features	Accuracy	Balanced Accuracy	Precision	Recall	F1	Time
KNN	Full DS	83	0.98	0.97	0.98	0.98	0.98	3.39
		22	0.97	0.95	0.97	0.97	0.97	1.47
	Half DS	83	0.98	0.95	0.98	0.98	0.98	1.41
		22	0.96	0.93	0.96	0.96	0.96	0.54
MLP	Full DS	83	0.9	0.61	0.9	0.9	0.89	99.99
		22	0.63	0.17	0.39	0.63	0.48	531.13
	Half DS	83	0.91	0.62	0.9	0.91	0.9	95.82
		22	0.68	0.22	0.62	0.68	0.64	322.77
SVM	Full DS	83	0.92	0.44	0.9	0.92	0.91	806.69
		22	0.57	0.53	0.48	0.57	0.49	633.67
	Half DS	83	0.83	0.63	0.87	0.83	0.83	297.55
		22	0.62	0.2	0.57	0.62	0.59	261.97
RF	Full DS	83	1	1	1	1	1	133.68
		22	1	1	1	1	1	74.71
	Half DS	83	1	1	1	1	1	48.47
		22	1	1	1	1	1	36.71

TABLE III: Comparison results using 22 features with different amounts of data

Algorithm	Entries per category	Accuracy	Balanced Accuracy	Precision	Recall	F1	Time
KNN	1500	0.86	0.86	0.85	0.86	0.85	0.01
	750	0.82	0.82	0.82	0.82	0.82	0
MLP	1500	0.77	0.77	0.76	0.77	0.72	2.59
	750	0.58	0.56	0.51	0.58	0.54	1.04
SVM	1500	0.34	0.34	0.38	0.34	0.29	2
	750	0.16	0.16	0.05	0.16	0.07	0.9
RF	1500	0.86	0.86	0.85	0.86	0.85	1.38
	750	0.86	0.86	0.85	0.86	0.85	0.64

The choice of using fewer data to have a balanced dataset resulted in using only 1.2% of the overall dataset which is not ideal but is the most accurate way of having the real dataset. Another possibility is to apply a technique of ”oversampling” or ”data augmentation”, which consists in generating artificial data to the classes that have the lowest registers but it can result in over-fitting, so we decided not to use it.

With those ML algorithms tested and with those results, other ML algorithms should be tested, such as Neural Networks and the results must be compared using other datasets. As explained in the paper this dataset was built without having a firewall or other security devices or applications so unknown traffic that was not recognized by Zeek software can be hidden in the dataset. This can be a research topic using Reinforcement Learning and giving the dataset and the model learn by himself to auto-detect new malign traffic.

VI. Conclusions

This paper presented a thorough analysis of the HIKARI-2021 dataset, the most recent public dataset released to the community. The study has revealed HIKARI-2021 to be a well-built dataset with a large number of entries and features, labeled with five distinct attack types and up to date.

The analysis showed different results were obtained in the MLP and SVM algorithms when compared to the original paper of the dataset. The remaining algorithms reveal identical results as expected. When studying the feature reduction size, we achieved similar results when using only 22 features instead of all the features, with the MLP and SVM models.

When reducing the amount of training data to half of the original size, we achieved similar results in the KNN, MLP and RF algorithms. In particular, the MLP presents better results with half of the data instead of the entire dataset. When comparing the time taken to train each model, we can notice that time decreases with fewer data and fewer features but in case of MLP the time increases when training with fewer features than using all of them.

Finally, the same algorithms were tested with the feature selection on a balanced dataset, i.e. having the same number of samples per category, and the results showed us accuracy and precision percentages above 80% for the KNN and RF algorithms.

In future work, we plan to study other recently published datasets and conduct an analysis with the same methods used in this article for further comparison between different datasets. Deep Learning algorithms will also be tested in this dataset to evaluate if the metrics here obtained can increase.

VII. Acknowledgements

This work was partially supported by the Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), within project ”Cybers SeC IP” (NORTE-01-0145-FEDER-000044).

References

[1] D. K. Alferidah and N. Jhanjhi, “Cybersecurity impact over bigdata and iot growth,” in *2020 International Conference on Computational Intelligence (ICCI)*, 2020, pp. 103–108.

[2] A. Yeboah-Ofori, U. M. Ismail, T. Swidurski, and F. Opoku-Boateng, “Cyberattack ontology: A knowledge representation for cyber supply chain security,” in *2021 International Conference on Computing, Computational Modelling and Applications (ICCMA)*, 2021, pp. 65–70.

[3] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35 365–35 381, 2018.

[4] A. Ferriyan, A. H. Thamrin, K. Takeda, and J. Murai, "Generating network intrusion detection dataset based on real and encrypted synthetic attack traffic," *Applied Sciences*, vol. 11, no. 17, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/17/7868

[5] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *2010 IEEE Symposium on Security and Privacy*, 2010, pp. 305–316.

[6] K. Wang and S. J. Stolfo, "Anomalous payload-based network intrusion detection," in *Recent Advances in Intrusion Detection*, E. Jonsson, A. Valdes, and M. Almgren, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 203–222.

[7] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. Sutton, J. D. Tygar, and K. Xia, "Exploiting machine learning to subvert your spam filter," in *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, ser. LEET'08. USA: USENIX Association, 2008.

[8] W. Melicher, B. Ur, S. M. Segreti, S. Komanduri, L. Bauer, N. Christin, and L. F. Cranor, "Fast, lean, and accurate: Modeling password guessability using neural networks," in *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, Aug. 2016, pp. 175–191. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/melicher

[9] E. Vasilomanolakis, S. Karuppayah, M. Mühlhäuser, and M. Fischer, "Taxonomy and survey of collaborative intrusion detection," *ACM Computing Surveys*, vol. 47, no. 4, may 2015. [Online]. Available: https://doi.org/10.1145/2716260

[10] D. Denning, "An intrusion-detection model," *IEEE Transactions on Software Engineering*, vol. SE-13, no. 2, pp. 222–232, 1987.

[11] J. R. Yost, "The march of ides: Early history of intrusion-detection expert systems," *IEEE Annals of the History of Computing*, vol. 38, no. 4, pp. 42–54, 2016.

[12] A. Halimaa A. and K. Sundarakantham, "Machine learning based intrusion detection system," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019, pp. 916–920.

[13] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," *IEEE Access*, vol. 6, pp. 33 789–33 795, 2018.

[14] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber security," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–17, 2021.

[15] S. S. Swarna Sugi and S. R. Ratna, "Investigation of machine learning techniques in intrusion detection system for iot network," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020, pp. 1164–1167.

[16] K. A. Taher, B. Mohammed Yasin Jisan, and M. M. Rahman, "Network intrusion detection using supervised machine learning technique with feature selection," in *2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)*, 2019, pp. 643–646.

[17] Y. Sani, A. Mohamedou, K. Ali, A. Farjamfar, M. Azman, and S. Shamsuddin, "An overview of neural networks use in anomaly intrusion detection systems," in *2009 IEEE Student Conference on Research and Development (SCOReD)*, 2009, pp. 89–92.

[18] L. Chen, X. Kuang, A. Xu, S. Suo, and Y. Yang, "A novel network intrusion detection system based on cnn," in *2020 Eighth International Conference on Advanced Cloud and Big Data (CBD)*, 2020, pp. 243–247.

[19] N. Farnaaz and M. Jabbar, "Random forest modeling for network intrusion detection system," *Procedia Computer Science*, vol. 89, pp. 213–217, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050916311127

[20] O. Yavanoglu and M. Aydos, "A review on cyber security datasets for machine learning algorithms," in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 2186–2193.

[21] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *ICISSP*, 2018.

[22] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1–6.