Cyber Informedness: A New Metric using CVSS to Increase Trust in Intrusion Detection Systems

Robin Duraz, David Espes, Julien Francq, Sandrine Vaton

Cyber Informedness: A New Metric using CVSS to Increase Trust in Intrusion Detection Systems

Robin Duraz
robin.duraz@ecole-navale.fr
Chaire of Naval Cyberdefense
France

David Espes
david.espes@univ-brest.fr
UBO, Lab-STICC
France

Julien Francq
julien.francq@naval-group.com
Naval Group (Naval Cyber Laboratory, NCL)
France

Sandrine Vaton
sandrine.vaton@imt-atlantique.fr
IMT Atlantique, Lab-STICC
France

ABSTRACT

Intrusion Detection Systems (IDSs) are essential cybersecurity components. Previous cyberattack detection methods relied more on signatures and rules to detect cyberattacks, although there has been a change in paradigm in the last decade, with Machine Learning (ML) enabling more efficient and flexible statistical methods. However, ML is currently unable to integrate cybersecurity information into its inner workings. This paper introduces Cyber Informedness, a new metric taking into account cybersecurity information to give a more informed representation of performance, influenced by the severity of the attacks encountered. This metric uses a *de facto* standard in cybersecurity: the Common Vulnerability Scoring System (CVSS). Results on two public datasets show that this new metric validates results obtained with generic metrics. Furthermore, this new metric highlights ML-based IDSs that prioritize high performance on severe attacks, which is not visible with generic metrics. Consequently, this new metric nicely completes generic metrics by bridging the gap between ML and cybersecurity.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Security and privacy** → **Intrusion detection systems**.

KEYWORDS

Cybersecurity metrics, Machine Learning, Intrusion Detection Systems

ACM Reference Format:
Robin Duraz, David Espes, Julien Francq, and Sandrine Vaton. 2023. Cyber Informedness: A New Metric using CVSS to Increase Trust in Intrusion Detection Systems. In *European Interdisciplinary Cybersecurity Conference (EICC 2023), June 14–15, 2023, Stavanger, Norway*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3590777.3590786

EICC 2023, June 14–15, 2023, Stavanger, Norway

ACM ISBN 978-1-4503-9829-9/23/06…$15.00
https://doi.org/10.1145/3590777.3590786

1 INTRO

The world is increasingly digitalized, which brings a plethora of cybersecurity threats. While it is essential to make systems more secure by design [2], nothing is ever perfectly secure, so alternative solutions are needed. IDSs are used to monitor and analyze traffic and system logs to detect anomalies and potential attacks. Traditional IDSs are signature-based and rather successful in detecting known attacks, with a very low probability of giving false alarms with sufficiently well-crafted rules, but they can easily miss zero-day or polymorphic threats [9]. In the last decade, however, one of the most extensive research directions concerns ML algorithms. Many works [6, 11, 12, 15, 22] have expanded upon ML and Deep Learning (DL) performance relative to intrusion detection on popular datasets, and show that those approaches perform well.

As far as metrics are concerned, the intrusion detection problem is addressed from the ML viewpoint with generic metrics such as Accuracy, Precision, Recall or F1-Score. While these metrics are extensively used and tested, they suffer from significant flaws, particularly when performing multi-class classification. Indeed, imbalance in the data heavily influences results, and poor results on underrepresented classes are generally hidden. Furthermore, those metrics treat all data equally and are unable to cater to differences in the cost of mistakes (attacks missed or false alarms). Failing to detect port scans will obviously be less penalizing than failing to detect the exfiltration of confidential data. Therefore, it will be beneficial to create a new metric or modify existing ones to add cybersecurity knowledge to the evaluation process.

In this paper, new metrics based on CVSS will be introduced to provide a more informed evaluation of the actual performance of ML-based IDSs. These metrics can benefit both the ML expert that builds and validates an IDS based on ML algorithms and the cybersecurity expert that will use and integrate this IDS by showing results that both sides have more confidence in.

The contribution of this research is twofold:

- Introduce Miss Cost (MC), False Alarm Cost (FAC) and Cyber Informedness (CI), three cybersecurity-related quantitative metrics that complete ML-related metrics.
- Evaluate the newly defined metrics, concurrently to other generic metrics, with various ML algorithms on the UNSW-NB15 and CIC-IDS2017 public datasets.

The rest of the paper is organized as follows: Section 2 presents related works. Section 3 describes the proposed approach, while Section 4 presents the experimental setup. Section 5 presents and analyzes the results. Finally, Section 6 discusses results and their limitations, and Section 7 concludes the paper and discusses future avenues of research.

2 RELATED WORK

For ML-based IDSs, cybersecurity datasets are required. Unfortunately, it is difficult to obtain realistic data, i.e., with at least a diversity of up-to-date attacks, a complete environment, as well as traffic representative of the real world (imbalanced, with errors, etc.). Using real world data is an obvious choice, but is often impossible to obtain because of confidentiality and/or security reasons. Another solution is to create a synthetic dataset. While it eliminates the previous problems, it is much more difficult to make it realistic. It is important to follow a thorough methodology, such as highlighted by [3, 16], to ensure quality of the data created.

2.1 Datasets

KDD'99 [1] and NSL-KDD [20] are the two most used datasets [9]. However, these datasets, and particularly the former, are heavily criticized because of their age and various other problems such as redundancy [5, 18, 21].

The UNSW-NB15 [13] and CIC-IDS2017 [17] datasets are more recent and based on quite complete environments. CIC-IDS2017 follows the methodology defined in [16] and criteria defined in [7] to ensure quality of the created dataset. Both being more recent datasets, it also ensures that the environment and simulated traffic are more representative of nowadays' real-world traffic. Furthermore, the UNSW-NB15 dataset has CVE (Common Vulnerability Exposure) information recorded for six attack categories that can be helpful in obtaining CVSS scores. While the CIC-IDS2017 dataset does not have CVE information, it is possible, given the information provided in [17], to assign CVSS scores to the respective attack classes. Although the more recent DAPT dataset [14] appears to better represent current attack methodologies, it is hard to obtain CVSS scores for this dataset.

2.2 Metrics

In order to evaluate performance of different ML-based IDSs, various metrics are generally used. The most complete representation of an IDS's performance and basis for most metrics, e.g., Accuracy, Precision, Recall, F1-Score, is the full confusion matrix. For intrusion detection problems, the metrics that are used share two major drawbacks. Firstly, they are unable to treat differently different attack classes, and this is problematic because attacks are not equally dangerous, and remediation mechanisms are different. Secondly, they are mostly not resistant to imbalance.

Imbalance in the data is a problem already highlighted in the literature. It has been described in details in [8, 10, 19], showing that many metrics might be ill-defined in case of heavily imbalanced datasets. It thus highlights the need to find better metrics, or simply account for the skewness of class distributions. However, there is no definite answer concerning the correct methodology to assess performance on imbalanced datasets.

To solve the imbalance problem, [4] has suggested the use of the Matthews Correlation Coefficient (MCC) that is probably the most complete with regard to summarizing the confusion matrix since it captures all the information contained therein, i.e., both True and False Positives and Negatives. While it is originally defined for binary classification, it can also be extended to the multi-class setting.

Although it was argued in [4] that MCC is resistant to imbalance, the authors in [23] offer a strong rebuttal to the use of MCC in case of imbalanced datasets and suggest using metrics that are more stable with regard to imbalance, such as the geometric mean of TPR (True Positive Rate) and TNR (True Negative Rate, also equal to $1 - \mathit{False\ Positive\ Rate}$) and Bookmaker Informedness (BM) equal to $TPR + TNR - 1$.

According to [23], BM accounts for imbalance and can reflect a less biased view of performance. Its formula is simple, yet appears to offer what most other metrics cannot, i.e., it allows to capture performance on both positive and negative instances with equal importance, irrespective of the imbalance. It is something that Accuracy or MCC are not capable of doing. However, resistance to imbalance of the MCC and BM metrics is still relatively unclear in the multi-class setting, since both [4] and [23] limited their analysis to the binary setting. Furthermore, although some metrics might appear more suitable than others, it is advised in [19] to rely on multiple metrics to correctly compare two algorithms.

Finally, MCC and BM appear to offer a solution to the imbalance problem. However, there is currently no metric that offers to solve the problem of attack classes that are inherently not equally important. For example, breaches in servers holding classified data need to be prioritized much more than simple brute forcing attempts.

3 PROPOSED APPROACH

Besides being an evaluation of performance, a given metric (or set of metrics) should be able to provide objective information about the actual cost of being mistaken, particularly in critical situations.

From the ML standpoint, all the metrics mentioned in Section 2.2 are already enough to constitute a set of complementary metrics that quite exhaustively represent a ML-based IDS's performance. However, it is still lacking from the cybersecurity standpoint and difficult to adapt to different monitored systems and goals. Moreover, it does not provide information about the cost of mistakes.

Because the application domain is cybersecurity, it should be important to consider widely used and consensual cybersecurity-based metrics relative to the severity of cyberattacks. Two methods are widely used to characterize cyberattacks: the MITRE ATT&CK framework and the CVSS. While MITRE ATT&CK is a knowledge base to classify TTPs (Techniques, Tactics and Procedures), making it hard to adapt to quantitative metrics, CVSS scores are numerical values and can be easily integrated into a cybersecurity-aware metric.

3.1 False Alarm Cost and Miss Cost

Let c be a class, with $c \in \{0, 1, ..., C\}$ and 0 being the normal class. For every instance i, let G_i be the ground truth value and D_i be the decision for this instance. CVSS_i is the CVSS score corresponding to instance i.

$\mathbf{1}$ stands for the indicator function and $\overline{\bullet}$ is the averaging operator. $\overline{\text{CVSS}_c}$ thus corresponds to the mean of CVSS scores for instances belonging to class c, which is necessary when instances of the same class do not have the same CVSS score (as is the case in the UNSW-NB15 dataset).

For each attack class c ($c \neq 0$), and with N the total number of instances, we define the False Alarm Cost (FAC, Eq. 1) and the Miss Cost (MC, Eq. 2) as follows:

$$\text{FAC}_c \overset{def}{=} \frac{\sum_{i=1}^{N} \mathbf{1}_{D_i=c} \cdot \mathbf{1}_{G_i \neq D_i}}{10 \sum_{i=1}^{N} \mathbf{1}_{D_i=c}} \cdot \overline{\text{CVSS}_c} \tag{1}$$

$$\text{MC}_c \overset{def}{=} \frac{\sum_{i=1}^{N} \mathbf{1}_{D_i \neq c} \cdot \mathbf{1}_{G_i=c} \cdot \text{CVSS}_i}{10 \sum_{i=1}^{N} \mathbf{1}_{G_i=c}} \tag{2}$$

In both formulae, the number 10 in the denominator represents the maximum possible value of a CVSS score, thus acting as a normalizing constant (bounding results between 0 and 1) while also highlighting the importance of attacks having a higher score.

As such, both formulae are generalizations of Machine Learning metrics. FAC is the generalization of the False Discovery Rate, the proportion of mistakes by predicting a specific class. Intuitively, it represents the frequency of false alarms, weighted by the CVSS score of these alarms. MC is the generalization of the False Negative Rate, the proportion of class instances that are incorrectly classified. Intuitively, it represents the frequency of missed attacks, weighted by their individual CVSS scores. These newly defined metrics are equal to their ML metrics counterparts when all CVSS scores are equal to 10 for classes different from normal traffic.

3.2 Cyber Informedness

For each class $c (c \neq 0)$, the Cyber Informedness (CI) metric that contains both FAC and MC is given by (3).

$$CI_c \overset{def}{=} 1 - \text{FAC}_c - \text{MC}_c \tag{3}$$

It is thus defined analogously to BM, although the FPR is replaced by FAC, a generalization of the FDR. Therefore, although a bit different from BM, the new metric takes into account both False Positives and False Negatives, and should therefore exhibit nice properties regarding class imbalance.

This metric aims to give a cybersecurity-informed idea about the performance of an IDS, aggregating both FAC and MC, with 1 being the best possible score. It also represents the success of an IDS to correctly identify a specific attack, with less penalties for failing to recognize less critical attacks.

3.3 Practical implications of the metrics

In order to use these newly defined metrics, the best case scenario would be to have, in the data, CVE IDs or CVSS scores related to the vulnerabilities exploited by attacks. It is, however, rarely the case. Therefore, another alternative is to directly get the CVSS scores used in this customizable metric through a publicly available calculator[1], which is doable when given enough details about the attacks.

Besides being based on a *de facto* standard in cybersecurity, the three cyber-related metrics can also provide additional benefits in practice when protecting a system. First and foremost, it inherently takes into account the severity of attacks encountered and puts more focus on attacks that are dangerous for the system. For example, failing to detect Heartbleed attacks will have much more impact than failing to detect port scans.

Secondly, it is possible to adapt the score depending on the system that needs to be protected. The CVSS score already includes this possibility with the Environmental score, where it is possible to specify Confidentiality, Integrity and Availability requirements of the system, influencing the resulting attack score.

Consequently, it gives the possibility of comparing IDSs by how adapted they are to a particular system. CVSS scores can also be modified to take into account the requirements of a system. It can thus be possible to train multiple IDSs and test their performance on different systems. This allows to pick different IDSs for different systems, depending on how adapted they are to a specific system, because the risk posed by an attack on a particular system is appropriately reflected on the new metrics.

4 EXPERIMENTAL SETUP

4.1 Choice of metrics

The finalized set of metrics chosen, both for comparison purposes and validation of the newly introduced metrics, is thus:

- Basic ML metrics: Accuracy, F1-score, TPR (Detection Rate, or Recall), PPV (Precision).
- Metrics presented as resistant to imbalance: MCC and BM. Both range between −1 and 1.
- Cyber-informed metrics: MC, FAC and CI. The former two range between 0 and 1 while the latter ranges between −1 and 1.

All metrics, except Accuracy and MCC, were computed on a per-class basis. The averaging method retained is macro-averaging, which averages irrespective of the class imbalance and thus reduces the influence of class imbalance.

4.2 Machine Learning algorithms

In order to evaluate the proposed set of metrics and understand the differences brought by the introduction of cybersecurity-based metrics, experiments were run with a wide range of algorithms, trying various hyper-parameter combinations to find the best performing IDS on the two datasets considered. The retained algorithms are:

- A dummy classifier, classifying every instance as of the most frequent class (normal traffic in both datasets) to serve as a baseline.
- Relatively simple algorithms that should give an idea about the complexity of the classification task: Gaussian Naïve Bayes (GNB), Logistic Regression (LR), Linear Support Vector Classification (LSVC), K-means, Decision Trees (DT).

[1] https://www.first.org/cvss/calculator/3.1

Table 1: CVSS scores for CIC-IDS2017

Attacks	Attack Vector	Attack Complexity	Privileges Required	User Interaction	Scope	Confidentiality	Integrity	Availability	CVSS score
DoS Slowloris DoS Slowhttptest DoS GoldenEye DoS Hulk	Network	Low	None	None	Unchanged	None	None	Low	5.3
Portscan FTP-Patator Web Attack Brute Force SSH-Patator	Network	Low	None	None	Unchanged	Low	None	None	5.3
Web Attack XSS	Network	Low	None	None	Unchanged	None	Low	None	5.3
Infiltration	Local	High	None	Required	Changed	High	None	None	5.5
Web Attack SQL Injection	Network	Low	None	None	Unchanged	Low	Low	Low	7.5
DDoS	Network	Low	None	None	Unchanged	None	None	High	7.5
Heartbleed	Network	Low	None	None	Unchanged	High	None	None	7.5
Botnet	Network	Low	None	None	Unchanged	High	High	High	9.8

Details about possible values for each category, as well as their signification, can be found at https://www.first.org/cvss/v3.1/specification-document.

- More complex algorithms that should reflect the expected performance of IDSs relying on ML: Random Forests (RF), Multi-Layer Perceptron (MLP), Deep Neural Networks (DNN).

All algorithms are from the *scikit-learn*[2] library except DNNs that were programmed using the *PyTorch*[3] and *PyTorch Lightning*[4] libraries. In order to evaluate K-means, which is an unsupervised algorithm and does not predict a label, labels were attributed to individual clusters by a majority vote, i.e., the class that is the most represented inside a cluster is the class assigned to it. For both datasets, the K-means algorithm was parameterized with the number of classes as the number of clusters. This is due to the fact that for both datasets, the higher the number of clusters, the better the performance. This is an extreme behavior that is unwanted here because it means the algorithm is completely unable to group instances of the same classes together while excluding other classes.

4.3 Dataset Pre-processing

Both the UNSW-NB15 and CIC-IDS2017 datasets were split using a stratified scheme into 70% train (60% and 10% validation for DNN) and 30% test sets.

For the UNSW-NB15 dataset, features such as IP addresses, timestamps, *attack_cat* were removed, while categorical features or features having a small number of unique values, were one-hot encoded. The resulting dataset has 229 features.

For the CIC-IDS2017 dataset, two features and 5792 instances were removed because of problematic or missing values. A further eight features were removed because they only had one value. The resulting dataset has 70 features.

[2]https://scikit-learn.org/stable/index.html
[3]https://pytorch.org/
[4]https://www.pytorchlightning.ai/

4.4 CVSS Scores for Cyber-related Metrics

Since it is possible to obtain CVSS scores for the UNSW-NB15 dataset with ground truth information, these scores were collected and assigned to the corresponding instances for the computation of the new metrics. However, the CVSS scores assigned to attacks of this dataset used the CVSS 2.1 standard and thus can be a bit different from the more recent CVSS 3.1 standard used for the other dataset. CVE IDs were thus used to get CVSS scores following the 3.1 standard when possible.

For the CIC-IDS2017 dataset, however, there is no such information. Although this information is missing, the attacks performed were described in sufficient detail in the original paper [17], allowing to manually craft CVSS scores since the described attacks seem to exploit the same vulnerabilities for a given attack category. It is possible because the given classes individually contain very similar attacks and thus do not suffer much from heterogeneity in a given class. For this dataset, the scores obtained through the CVSS calculator[5], as well as the vector used for computation, are visible in Table 1. Some choices can appear misleading because the name chosen for a given attack class in the original dataset misrepresents its actual execution mechanisms.

5 RESULTS

Results according to the relevant metrics are presented in Table 2. For each category of ML algorithm, a coarse grid-search scheme was used to pick hyper-parameter values and the IDS obtaining the best results were kept. For those IDSs, results are shown for the retained metrics. Considering only some of the metrics, particularly Accuracy and MCC, it is difficult to see which IDS perform better than others. The most significant differences on both datasets can be seen with PPV and the newly defined metrics.

[5]https://www.first.org/cvss/calculator/3.1

Table 2: Performances on the UNSW-NB15 and CIC-IDS2017 datasets

Dataset	Algorithm	Acc.	F1	TPR	PPV	MCC	BM	MC	FAC	CI
UNSW-NB15	**Dummy**	0.8735	0.0932	0.1	0.0873	NaN	−0.1020	0.3959	0.6239	−0.0198
	GNB	0.4902	0.1308	0.2965	0.2648	−0.0395	−0.2539	0.3296	0.4947	0.1756
	LR	0.9745	0.4478	0.4443	0.5861	0.8880	0.4059	0.2212	0.2862	0.4924
	LSVC	0.9728	0.4455	0.4362	0.4806	0.8796	0.3947	0.2643	0.3443	0.3913
	K-means	0.8747	0.1594	0.1879	0.1881	0.5216	0.0430	0.3332	0.5582	0.1085
	DT	0.9794	**0.5864**	**0.5657**	0.6665	0.9102	**0.5351**	**0.1869**	0.2289	0.5840
	RF	**0.9816**	0.5713	0.5498	0.7383	**0.9197**	0.5218	0.2055	0.1801	0.6143
	MLP	0.9800	0.5205	0.5189	**0.7721**	0.9128	0.4889	0.2206	**0.1538**	**0.6254**
	DNN	0.9789	0.5056	0.5118	0.6277	0.9084	0.4807	0.2068	0.2575	0.5356
CIC-IDS2017	**Dummy**	0.8030	0.0593	0.0666	0.0535	NaN	−0.1725	0.5633	0.5633	−0.1266
	GNB	0.7232	0.4997	**0.8480**	0.4692	0.5729	0.5792	**0.0692**	0.3211	0.6096
	LR	0.9908	0.6394	0.6335	0.7880	0.9733	0.6236	0.2337	0.1474	0.6187
	LSVC	0.9865	0.5466	0.5897	0.6021	0.9606	0.5746	0.2564	0.2530	0.4905
	K-means	0.8687	0.1460	0.1383	0.1588	0.5499	−0.0196	0.5181	0.5003	−0.0185
	DT	0.9984	0.8399	0.8439	0.8367	0.9954	**0.8421**	0.1013	0.1041	0.7944
	RF	**0.9986**	**0.8502**	0.8362	0.8707	**0.9959**	0.8346	0.1064	0.0850	**0.8085**
	MLP	0.9965	0.7254	0.7219	0.8356	0.9898	0.7182	0.1771	0.1035	0.7192
	DNN	0.9970	0.7576	0.7393	**0.8968**	0.9915	0.7361	0.1686	**0.0670**	0.7643

Values were truncated to the fourth decimal. Best results for a given metric and dataset are in **bold**.
MCC is undefined (NaN) for dummy because all predictions are the same, causing the denominator to become 0.

5.1 Zoom comparison of two IDS' performances

Contrarily to generic metrics, a significant difference can be seen with the newly defined metrics. The following example compares results presented in Table 2 for the LSVC and MLP on the UNSW-NB15 dataset.

In the UNSW-NB15 dataset, for attacks that do have CVE IDs and thus an assigned CVSS score, Exploits is the class with the highest average CVSS score because most instances have a high CVSS score (9.3 or 10). DoS attacks, on the contrary, generally have CVSS scores between 5 and 8.

For this particular case, both IDSs have relatively similar performance (under a 5% difference) on all classes, except Exploits and DoS. For those two specific classes, LSVC outperforms MLP by correctly classifying 69% of DoS instances versus 37%. On the other hand, MLP significantly outperforms LSVC by correctly classifying 74% of Exploits instances versus 46%.

When looking at results, the Accuracy of both IDSs is extremely close, which is understandable since results on most classes (including the overrepresented normal class) are similar. However, when looking at the FAC and CI metrics, the results are very different. Exploits attacks are generally more dangerous, i.e., have a higher CVSS score, which is directly translated into those two metrics. Indeed, the MLP that performs better on Exploits has results that are more than two times better for FAC and close to 60% better on the CI metric. This difference is the most significant on these two metrics, showing they manage to capture much needed information: the better performance on more dangerous attacks.

Operationally, it means the MLP-based IDS will more often detect attacks that are critical and might endanger the system. When using such an IDS, automated mitigation strategies can be used with more confidence, and human operators will be able to divert their energy in investigating other more relevant alarms.

6 DISCUSSION

In the multi-class setting, most metrics, including the MCC and BM metrics chosen to resist imbalance, still appear to suffer from it. It seems that macro-averaging class results also managed to reduce the influence of imbalance. Results with metrics such as Accuracy and MCC, that were not macro-averaged appear very similar for most IDSs. Furthermore, while the new metrics also seem to suffer from imbalance, their formulation and the use of CVSS scores managed to reduce this influence further.

Using of CVSS scores to prioritize the correct detection of more severe attacks also seemed to work well, as was shown in Section 5.1. When models show very similar performance on the newly defined metrics, it generally means that their performance is similar on all attacks or that it is different for attacks having similar CVSS scores. The high performance on critical attacks is thus adequately highlighted. Another advantage of these newly defined metrics is that it might be more intuitive to understand what they represent. The FAC directly shows the potential cost incurred when raising an alarm, while the MC directly shows the cost of not detecting or wrongly classifying an actual attack relative to its severity.

Although these new metrics are based on a *de facto* cybersecurity standard score, research is still ongoing concerning the impact of using CVSS score to find the optimal way to integrate it into ML metrics. For example, it would be possible to integrate an α parameter to power the CVSS scores, thus increasing their influence on the results. It might prove to be better at discriminating models' performance, although choosing optimal value would be difficult.

Finally, some IDSs, although exhibiting relatively poor performance in general, can have an unexpectedly good performance in some aspects, e.g., the GNB-based IDS for CIC-IDS2017 which is the best attack detector at the cost of more false alarms. Thus, using those IDSs could be interesting when implementing ensemble methods for intrusion detection.

7 CONCLUSION AND FUTURE WORK

Common ML metrics are insufficient for cybersecurity and can bring about a false sense of security. They are generic and do not take into account the specificity, variety and severity of attacks, nor are they influenced by the operational context.

Therefore, the MC, FAC and CI metrics that use a standard cybersecurity score are proposed. Results obtained with the new metrics tend to validate those obtained on two public datasets with generic ML metrics, although some significant differences do appear.

Research is ongoing about the possibility of using such metrics to cater to specific systems. Such a possibility is given by the CVSS standard that allows the definition of Confidentiality, Availability, and Integrity requirements, as well as other Temporal and Environmental parameters, to refine further the obtained CVSS scores. This would make the newly introduced metrics adaptable to different systems, with their requirements directly impacting results.

Finally, it could be interesting to see if these new metrics could be integrated into the loss formulation of DNNs, and investigate how these metrics behave with ensemble methods, most commercial IDSs based on ML using such methods.

ACKNOWLEDGMENTS

This work is supported by the Chair of Naval Cyber Defence and its partners Ecole Navale, ENSTA-Bretagne, IMT-Atlantique, Naval Group and Thales.

REFERENCES

[1] 1999. KDD Cup 99 Data. [Online]. Available:http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[2] Yosef Ashibani and Qusay H. Mahmoud. 2017. Cyber Physical Systems Security: Analysis, Challenges and Solutions. *Computers and Security* 68 (jul 2017), 81–97. https://doi.org/10.1016/j.cose.2017.04.005

[3] Monowar H. Bhuyan, Dhruba Kumar Bhattacharyya, and Jugal Kumar Kalita. 2015. Towards Generating Real-Life Datasets for Network Intrusion Detection. *International Journal of Network Security* 17 (2015), 683–701.

[4] Davide Chicco. 2017. Ten Quick Tips for Machine Learning in Computational Biology. *BioData Mining* 10, 1 (2017), 35. https://doi.org/10.1186/s13040-017-0155-3

[5] Gideon Creech and Jiankun Hu. 2013. Generation of a new IDS test dataset: Time to retire the KDD collection. In *2013 IEEE Wireless Communications and Networking Conference (WCNC)*. 4487–4492. https://doi.org/10.1109/wcnc.2013.6555301

[6] Mohamed Amine Ferrag, Leandros Maglaras, Sotiris Moschoyiannis, and Helge Janicke. 2020. Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study. *Journal of Information Security and Applications* 50 (2020), 102419. https://doi.org/10.1016/j.jisa.2019.102419

[7] Amirhossein Gharib, Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. 2016. An Evaluation Framework for Intrusion Detection Dataset. In *2016 International Conference on Information Science and Security (ICISS)*. IEEE, 1–6. https://doi.org/10.1109/icissec.2016.7885840

[8] Qiong Gu, Li Zhu, and Zhihua Cai. 2009. *Evaluation Measures of the Classification Performance of Imbalanced Data Sets.* Springer Berlin Heidelberg, 461–471. https://doi.org/10.1007/978-3-642-04962-0_53

[9] Hanan Hindy, David Brosset, Ethan Bayne, Amar Kumar Seeam, Christos Tachtatzis, Robert Atkinson, and Xavier Bellekens. 2020. A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems. *IEEE Access* 8 (2020), 104650–104675. https://doi.org/10.1109/access.2020.3000179

[10] Laszlo A. Jeni, Jeffrey F. Cohn, and Fernando De La Torre. 2013. Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 245–251. https://doi.org/10.1109/acii.2013.47

[11] Ansam Khraisat, Iqbal Gondal, Peter Vamplew, and Joarder Kamruzzaman. 2019. Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges. *Cybersecurity* 2, 1 (2019), 20. https://doi.org/10.1186/s42400-019-0038-7

[12] Deepthi Hassan Lakshminarayana, James Philips, and Nasseh Tabrizi. 2019. A Survey of Intrusion Detection Techniques. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. https://doi.org/10.1109/icmla.2019.00187

[13] Nour Moustafa and Jill Slay. 2015. 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)*. 1–6. https://doi.org/10.1109/milcis.2015.7348942

[14] Sowmya Myneni, Ankur Chowdhary, Abdulhakim Sabur, Sailik Sengupta, Garima Agrawal, Dijiang Huang, and Myong Kang. 2020. *DAPT 2020 - Constructing a Benchmark Dataset for Advanced Persistent Threats.* Springer International Publishing, Chapter DAPT 2020 - Constructing a Benchmark Dataset for Advanced Persistent Threats, 138–163. https://doi.org/10.1007/978-3-030-59621-7_8

[15] Iqbal H. Sarker, A. S. M. Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters, and Alex Ng. 2020. Cybersecurity Data Science: an Overview From Machine Learning Perspective. *Journal of Big Data* 7, 1 (jul 2020). https://doi.org/10.1186/s40537-020-00318-5

[16] Iman Sharafaldin, Amirhossein Gharib, Arash Habibi Lashkari, and Ali A. Ghorbani. 2017. Towards a Reliable Intrusion Detection Benchmark Dataset. *Software Networking* 2017, 1 (2017), 177–200. https://doi.org/10.13052/jsn2445-9739.2017.009

[17] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. 2018. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy*. 108–116. https://doi.org/10.5220/0006639801080116

[18] Kamran Siddique, Zahid Akhtar, Farrukh Aslam Khan, and Yangwoo Kim. 2019. Kdd Cup 99 Data Sets: a Perspective on the Role of Data Sets in Network Intrusion Detection Research. *Computer* 52, 2 (2019), 41–51. https://doi.org/10.1109/mc.2018.2888764

[19] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. *Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation.* Springer Berlin Heidelberg, 1015–1021. https://doi.org/10.1007/11941439_114

[20] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. 2009. A Detailed Analysis of the KDD CUP 99 Data Set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*. https://doi.org/10.1109/cisda.2009.5356528

[21] Amjad M. Al Tobi and Ishbel Duncan. 2018. Kdd 1999 Generation Faults: a Review and Analysis. *Journal of Cyber Security Technology* 2, 3-4 (2018), 164–200. https://doi.org/10.1080/23742917.2018.1518061

[22] Yang Xin, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmiao Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, and Chunhua Wang. 2018. Machine Learning and Deep Learning Methods for Cybersecurity. *IEEE Access* 6 (2018), 35365–35381. https://doi.org/10.1109/access.2018.2836950

[23] Qiuming Zhu. 2020. On the Performance of Matthews Correlation Coefficient (MCC) for Imbalanced Dataset. *Pattern Recognition Letters* 136 (2020), 71–80. https://doi.org/10.1016/j.patrec.2020.03.030