A Novel IoT-based Explainable Deep Learning Framework for Intrusion Detection Systems

Zakaria Abou El Houda[1], Bouziane Brik[2], and Sidi-Mohammed Senouci[2]
[1]Department of Computer Science and Operational Research, University of Montreal, Canada
[2]DRIVE EA1859, university of Bourgogne Franche-Comté, France
zakaria.abou.el.houda@umontreal.ca, {bouziane.brik,sidi-mohammed.senouci}@u-bourgogne.fr

***Abstract*—Internet of Things (IoT) is growing as a key pillar of smart city development. This growth is accompanied with serious cybersecurity risks, especially with the IoT botnets emergence. In this context, Intrusion Detection Systems (IDSs) proved their efficiency in detecting various attacks, that may target IoT networks, especially when leveraging Machine/Deep Learning (ML/DL) techniques. In fact, ML/DL-based solutions give "machine-centric" decisions about intrusion detection in IoT network, which will be then executed by humans, $i.e.$, executive cyber-security staff. However, ML/DL-based solutions do not provide any explanation of why such decisions were made, and thus their results cannot be properly understood/exploited by humans. To address this issue, Explainable Artificial Intelligence (XAI) is a promising paradigm, that helps explain the decisions of ML/DL-based IDSs to make them understandable to cyber-security experts. In this paper, we design a novel XAI-powered framework to enable not only detecting intrusions/attacks in IoT networks, but also interpret critical decisions made by ML/DL-based IDS. Therefore, we first build a ML/DL-based IDS using Deep Neural Network (DNN), to detect and predict IoT attacks in real time. Then, we develop multiple XAI models ($i.e.$, RuleFit and SHapley Additive exPlanations (SHAP)) on top of our DNN architecture, to enable more trust, transparency and explanation of the decisions made by our ML/DL-based IDS to cyber security experts. The in-depth experiments results with well-known IoT attack, show the efficiency and the explainiblity of our proposed framework.**

***Index Terms*—Internet of Things; Intrusion Detection System; Explainable Artificial Intelligence.**

I. Introduction

Internet of Things (IoT) is an emerging paradigm that has gained momentum and is now shaping our future [1] [2]. IoT aims to transform our daily live by deploying billions of smart devices, around 75 billion IoT devices by 2025 [2], to perform daily tasks. Thus, IoT is becoming a key pillar of different sectors, including Healthcare, agriculture, transportation, and factories [1] [2]. However, with the rapid deployment of IoT, numerous IoT vulnerabilities have emerged as well [3]. In fact, new sophisticated and destructive IoT attacks are increasing. For instance, Mirai IoT botnet has performed a huge attack using many compromised IoT attacks, including IoT gateways, closed circuit television cameras, and routers. This subsequently resulted in the unavailability of many Internet services such as Twitter and Amazon, for several hours [3]. In addition, such IoT attacks may cause extensive financial loss and huge damage. According to a recent report [4], it is estimated that the financial loss caused by the IoT attacks is about $20 Billion (USD) in 2021.

To deal with IoT attacks, different security measures are used, including firewalls, anti-virus, and access control, in order to filter and control incoming network traffic. However, these measures are not sufficient/efficient to protect the network [5], especially with the emergence of IoT attacks. As a second line of protection, Intrusion Detection Systems (IDSs) should be efficiently designed to secure the IoT network against various attacks ranging from Distributed denial-of-service (DDoS) to scanning attacks. In this context, IDSs have proved their efficiency in detecting various attacks, that may target IoT networks, especially when leveraging Machine/Deep Learning (ML/DL) techniques [6]. In fact, ML/DL techniques consist of learning the characteristics of each attack, so that we can quickly and efficiently identify/detect existing and new IoT attacks, without having to update traditional IDS rules. Hence, ML/DL-based IDSs systems give "machine-centric" decisions about intrusion detection, which will be then executed by humans, $i.e.$, executive staff. However, such systems do not give any explanation/interpretation about why such decisions are made and hence their results cannot be understood by humans. In other words, the main drawback of existing ML/DL-based IDSs systems, particularly the most accurate ones, are the black-box decisions, whose internal functioning is hidden and not understood.

Recently, eXplainable Artificial Intelligence (XAI)

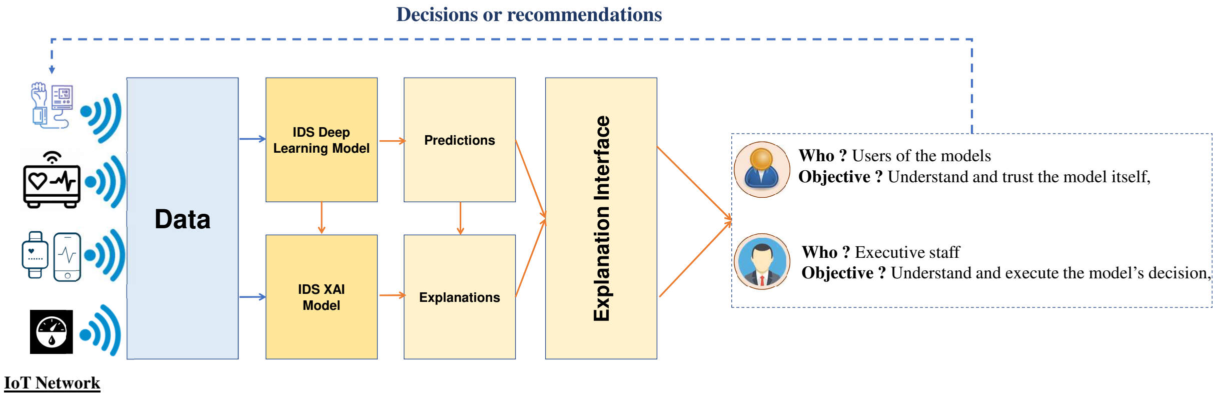

Decisions or recommendations
Data
IDS Deep Learning Model
Predictions
IDS XAI Model
Explanations
Explanation Interface
Who ? Users of the models
Objective ? Understand and trust the model itself,
Who ? Executive staff
Objective ? Understand and execute the model's decision,
IoT Network

Fig. 1. General Architecture of our XAI-based Framework for IoT IDS.

has emerged as a promising paradigm to develop new approaches explaining how ML/DL models work. XAI aims to make ML/DL models understandable for experts in the domain [7]. This also enables experts to trust and adapt such models and hence release their decisions (models).

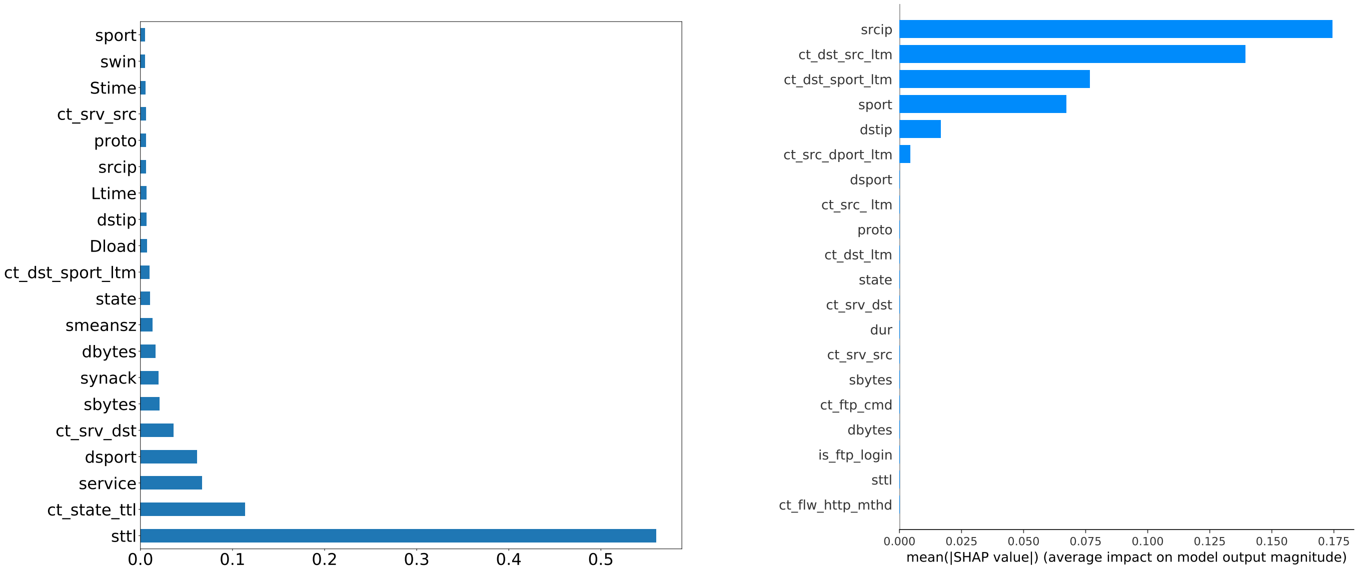
In this paper, we design a new XAI-powered framework that comprises two main modules: (1) DL-based IDS system: we first build a DL-based prediction model of intrusions for IoT applications. To do so, we leverage UNSW-NB15 dataset [8] and design a neural network to create our prediction model; and (2) XAI-enabled IDS system: we develop several XAI models on top of our DL-based IDS to add more transparency and interpretability to our DL-based IDS's decisions. Specifically, we implement RuleFit and SHAP (SHapley Additive exPlanations) as white-box models related to our DL-based black-box model. Therefore, our framework enables not only timely detecting intrusions in IoT networks, but also interpret decisions made by our DL-based model. This introduces more trust and transparency among our DL-based IDS system and experts, that will execute its decisions. This paper is organized as follows. Section II gives a review of related work. Section III describes the design and specification of our proposed XAI-powered framework. Section IV presents the performance evaluation of our proposed XAI-powered framework. Finally, section V concludes the paper.

II. Related Work

In this section, we briefly present the main works that addresses the explainability of ML/DL-based IDS systems, along with their limitations. In [9], deep neural network is first used for network IDS and then XAI-based framework is designed to improve the transparency deep learning model. The authors leveraged NSL KDD dataset to implement and validate several XAI approaches including, SHAP, contrastive explanations method, LIME, and ProtoDash. Similarly, another framework used SHAP approach, to improve the transparency of IDSs of any ML/DL-based IDS system, in [10]. The authors used also the NSL-KDD dataset to test the performance of the framework.

In [11], XAI is integrated with ML-based IDS to deal with adversarial attacks. First, a random forest classifier is built to detect network intrusions. Then, SHAP approach is applied to explain and interpret the outputs of the random forest-based model. The performance of this scheme is evaluated using CICIDS dataset. Besides, a layer-wise relevance propagation (LRP) method is used in [12] to determine input feature relevance and send offline and online feedback to end-users, to help them deduce which features have more impact on the predictions made by IDS. In [13], an explanation approach is proposed to deal with incorrect classifications made by ML/DL-based IDSs. This approach helps to determine the suitable modifications needed to correctly classify a given dataset sample. These modifications are also exploited to deduce the most important features, that justify the reason for the incorrect classification. the designed approach is evaluated and tested using NSL-KDD dataset. A local explanation method is used in [14], to explain/interpret each prediction separately of a ML-based IDS.

Even the above works leveraged XAI to explain and interpret ML/DL-based IDS, however, some of these works are limited to traditional machine learning algorithms, which are less complex and easy to interpret, as compared to deep learning [11]. In addition, most of them designed a general XAI framework whatever the targeted ML/DL-based IDS [10] [12] [13]. This may not be realistic, since each ML/DL-based IDS model has its specific input features and performance, and the XAI framework should consider such characteristics as



(a)

(b)

Fig. 2. Feature Importance Scores on $UNSW-NB15$ dataset for: a) RuleFit; and b) SHAP.

input, to be able then in explaining ML/DL-based IDS's decisions.

III. Our Framework to explain DL-based IDS of IoT Applications

This section describes the design of our proposed XAI-empowered framework. First, we present our system architecture. Then, we present our DNN architecture to predict IoT intrusions/attacks in real-time, and our XAI models to explain/interpret our DL-based IDS.

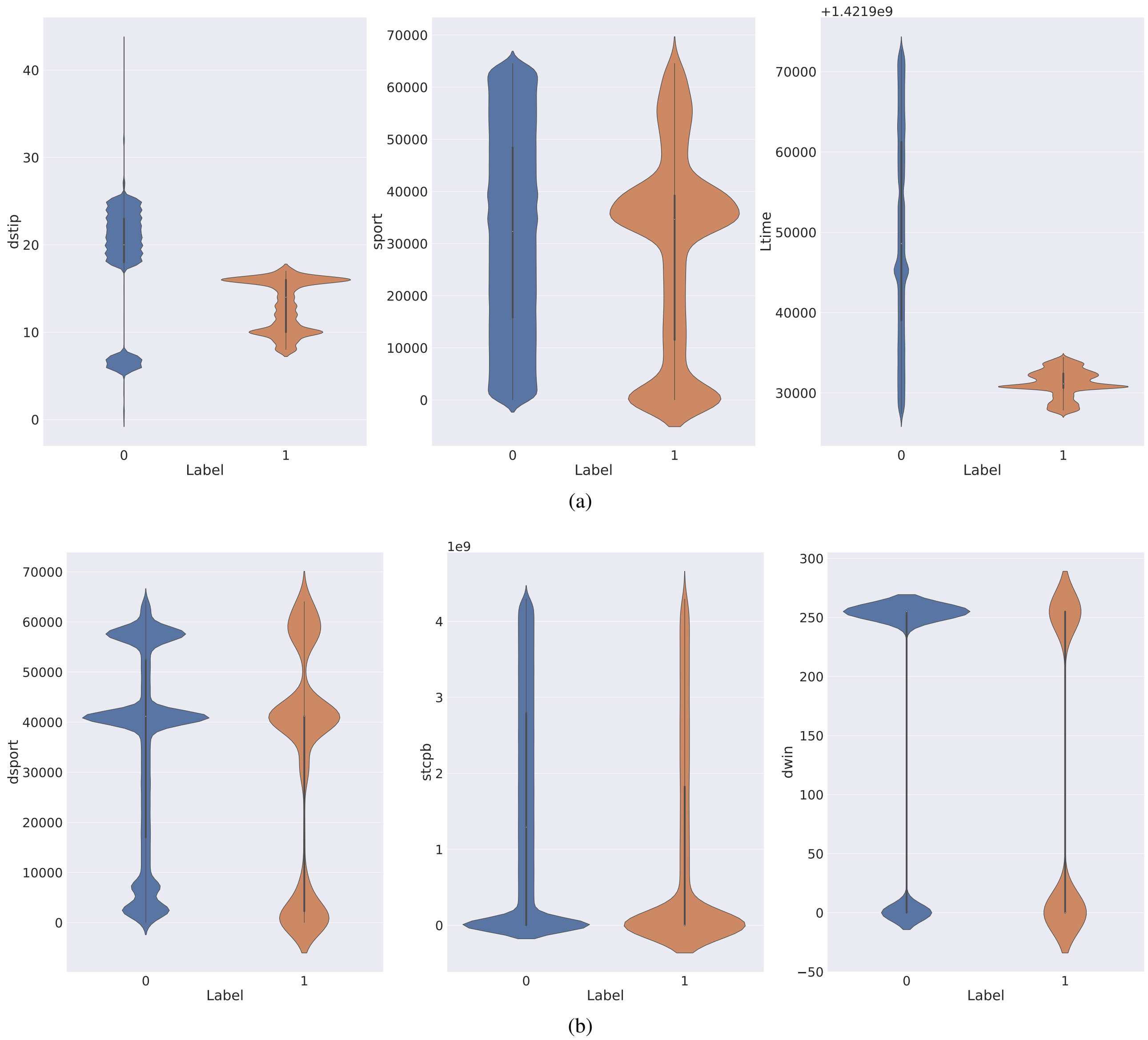
A. *System Architecture*

Fig. 1 shows the general system architecture of our designed framework. The data collected from the IoT network will be exploited, on one hand, to build a deep learning-based model to predict/detect intrusions in the IoT network. On the other hand, an XAI-model is created which leveraged both the sensed data and DL-based model's predictions, in order to explain/interpret such predictions. This enables not only to explain how the DL-based model works, but also why its predictions and hence decisions are made. Noting that performed predictions with their explanations are showed to different audiences through an explanation interface. Moreover, our framework targets both users of the DL-based model and executive staff. The users of the model should understand and trust the model predictions, before transferring model's decisions to the executive staff, that should also understand the received decisions and execute them. In the following, we present our Explainable Deep Learning-based IDS suitable for IoT applications

B. *Explainable Deep Learning-based IDS of IoT Applications*

In this work, we leverage the UNSW-NB15 dataset for the attack traffic. UNSW-NB15 is a synthetic network security dataset that contains 100 GB of network data samples, including several IoT attacks ($e.g.$, backdoors, DoS, and worms). For the pre-processing phase, we have encoded the categorical/non-numeric input features ($i.e.$, 'service', 'proto', and 'state') into numeric values using one hot encoding techniques. Some of the the UNSW-NB15 features ($e.g.$, 'Destination TCP sequence number (dtcpb) [0;$4*10^9$]' and 'Source TCP sequence number (stcpb) [0;$4*10^9$]') have higher values than other features ($e.g.$, 'Source IP address (srcip) [0;39]' and ' Destination to source time to live (dttl) [0;254]'); which may impact the final model decisions. This latter may miss out important features that have minimum values $i.e.$, source time to live (dttl). This, we have applied the standardization technique to overcome this issue. Finally, we encoded the Labels/output features ($e.g.$, backdoors, Shellcode, and Fuzzers) into numerical values.

Besides, to test the effectiveness of our proposed XAI-powered framework, we constructed a deep neural network (DNN) model with input layer of 49 neurons, that corresponds to the dimension of the input sample of UNSW-NB15 dataset, five hidden layers with Leaky Rectified Linear Unit, and an output layer of one neuron. We implemented our proposed XAI-powered framework using Pytorch and SHAP Library [15], an open source library that includes various functions to explain the



(a)

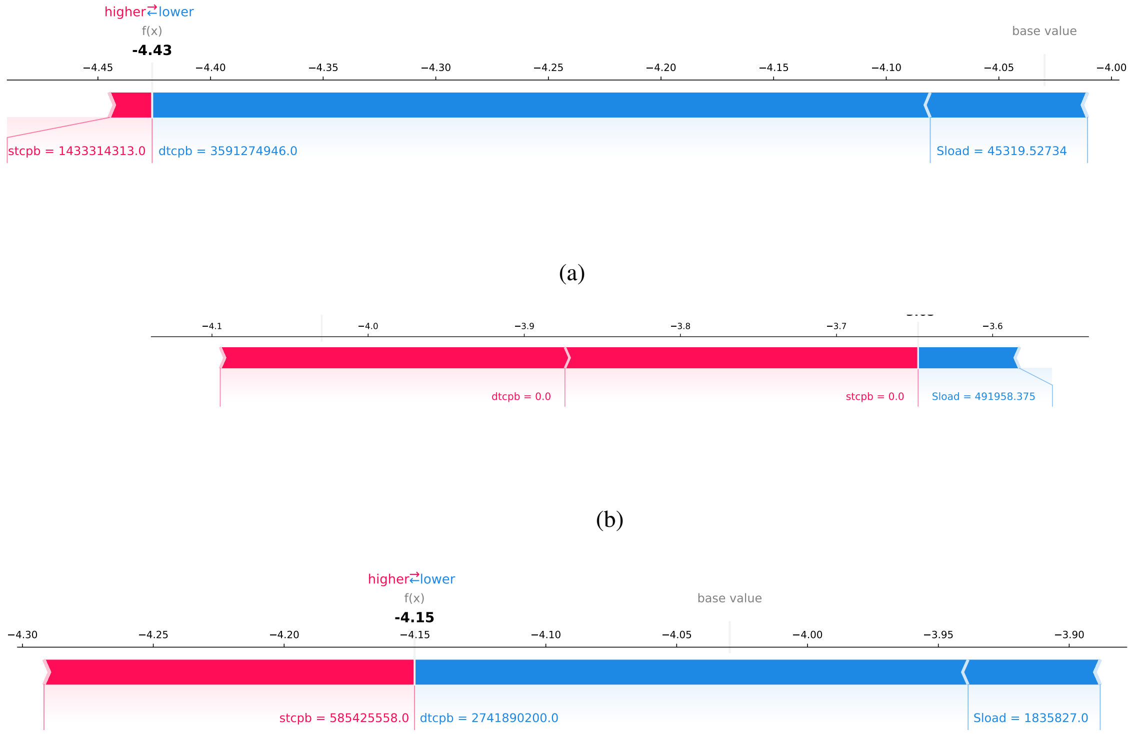
(b)

Fig. 3. Data samples distribution of features of $UNSW-NB15$ dataset in terms for: a) the highest scoring features using RuleFit and SHAP; and b) the other non-irrelevant features.

output of ML/DL-based models. In this work, we have considered two techniques, namely RuleFit and SHapley Additive exPlanations (SHAP) methods to effectively interpret a DL-based IDS model decisions/classifications. The objective is to explore linear and non-linear methods, including local and global explanations. In RuleFit, we learn sparse linear models/forms that include the effects of interaction in a decision-making rule-based form; it crates new features in the form of decision-making rules and constructs a transparent model with these features. RuleFit includes two steps: (1) it trains a tree-based model and use it to create the decision rules; and (2) it trains a sparse linear model (*e.g.*, LASSO) to select the most informative/significant features. SHAP is a well known unified framework for model interpretation; it explains the predictions of an input data sample by calculating the contribution of each feature to the final decision/prediction. This contribution can be either positive or negative. The main advantage of SHAP is that it can be applied to any model, rather than simple/linear models. Also, instead of looking only at local decisions/interpretations, SHAP looks at the overall/global interpretations by summing the input values of the features and averaging all columns/features individually.

IV. Performance Evaluation

The feature importance scores shows the most important/relevant features among all features of dataset; these features have significant impact on the model pre-



(a)

(b)

(c)

Fig. 4. Interpretation of our DNN model on $UNSW-NB15$ dataset with: a) Sload of $4.5*10^4$, a stcpb of $1.43*10^9$, and a tcpb of $3.5*10^9$; b) Sload of $4.9*10^5$, a stcpb of 0, and a dtcpb of 0; and c) Sload of $1.8*10^9$, a stcpb of $5.8*10^9$, and a dtcpb of 2.7*10^9.

dictions that other features. Our proposed XAI-powered framework investigates the use of both RuleFit and SHAP methods, to select the most informative/significant features and explores their effect on final model predictions. Fig. 2 shows the feature importance scores on $UNSW-NB15$ dataset using RuleFit and SHAP, respectively; it shows the highest scoring features in a descending order.

For RuleFit method, the highest scoring features on $UNSW-NB15$ dataset corresponds to the following features: (1) sttl: corresponds to the Source to destination time to live; (2) ct_state_ttl: corresponds to the Number of each state ($e.g.$, ACC, CLO) according to a range of values for source/destination time to live (ttl); (3) service: corresponds to the used protocol $e.g.$, http, dns, ssh; (4) dsport: corresponds to the destination port number; (5) ct_srv_dst: corresponds to the number of connections that contain the same service and destination address in the last 100 connections; and (6) sbytes: corresponds to Source to destination bytes. For the SHAP method the highest scoring features on $UNSW-NB15$ dataset corresponds to the following features: (1) srcip: corresponds to the Source IP address; (2) ct_dst_src_ltm: corresponds to the number of connections that contain the same service and destination address in the last 100 connections; (3) ct_dst_sport_ltm: corresponds to the number of connections pf the same destination address and the source port in the last 100 connections; (4) sport: corresponds to Source port number; and (5) dstip: corresponds to the Destination IP address. Fig. 3 shows the data distribution of UNSWNB15 dataset features. Fig. 3(a) shows some of highest scoring features on $UNSW-NB15$ dataset, based on RuleFit and SHAP; while Fig. 3(b) shows the other non-irrelevant features. We observe that the most relevant features, computed based on RuleFit and SHAP, can effectively distinguish the two classes ($i.e.$ Normal and Attack), because the data distribution of the two classes is completely different, while the data distribution of the two classes is similar for the other non-relevant features, which makes classification difficult for the IDS. Fig. 4 shows the interpretation of our DNN model on $UNSW-NB15$ dataset using SHAP method. Instead of examining decisions of our DNN model locally, we examine the overall/global feature importance of $UNSW-NB15$ dataset using SHAP, we sum up shapley the input values and we average all the columns/features individually. For a particular observation, each input feature value ($e.g.$, Sload (source bits per second), (stcpb) Source TCP sequence, and dtcpb) Destination TCP sequence) has either a positive or a negative contribution to the final decision $i.e.$, base value. In our analysis, we have examined three observations. Fig. 4(a) shows the first observation in which the data sample is Normal ($i.e.$, non-attack) and the DNN model correctly predicted/detected as a Normal data sample. In this observation, the values of the input features are as follows: Sload is equal to $4.5 * 10^4$, stcpb is equal to $1.43 * 10^9$, and dtcpb is equal $3.5 * 10^9$. Fig. 4(b) shows the second observation in which the data sample is an IoT attacks and the DNN model correctly predicted/detected as an IoT attack. In this observation, the values of the input features are as follows: Sload is equal to $4.9 * 10^5$, stcpb is equal to 0, and dtcpb is equal 0. Fig. 4(c) shows the last studied observation in which the data sample is an IoT attacks and the DNN model predicted as a Normal data sample ($i.e.$, False Negative (FN)). In this observation, the values of the input features are as follows: Sload is equal to $1.8{*}10^9$, stcpb is equal to $5.8{*}10^9$, and dtcpb is equal 2.7*10^9. In all these observations, the blue features pushes the prediction of the data sample to be Normal $i.e.$, class 0. The larger the shaft, the more effect this input feature of the $UNSW-NB15$ dataset has on the final detection/prediction. In the first scenario (see Fig. 4(a)), we observe that the most contributing/significant features are as follows: Sload and dtcpb. The red feature ($i.e.$, stcpb) reduces the probability for a data sample to be Normal. In the second scenario (see Fig. 4(b)), we observe that the most contributing/significant features are as follows: stcpb and dtcpb. The red features ($i.e.$, stcpb and dtcpb) drives the probability for a data sample to be an attack. In the last scenario (see Fig. 4(c)), we observe that the most contributing/significant features are as follows: Sload and dtcpb. The red feature ($i.e.$, stcpb) reduces the probability for a data sample to be Normal. Thus, such solid knowledge makes cybersecurity experts more convinced of the decisions regarding ML/DL-based IDS.

V. Conclusion

In this paper, we proposed a novel XAI-powered framework that enabled not only the detection of IoT attacks, but also the interpretation of critical decisions made by ML/DL-based IDSs. First, we built a DNN model to detect and predict IoT attacks in real time. Then, we have developed multiple XAI models ($i.e.$, RuleFit and SHapley Additive exPlanations (SHAP)) on top of our DNN architecture, to enable more trust, transparency, and explainability of the decisions taken by our ML/DL-based IDS to the cyber-security experts. The in-depth experiments results on well-known IoT attack, showed the efficiency and the explainiblity of our proposed framework. This makes it a promising cyber-security framework for accurate IoT attack detection and explainable Deep Learning Framework for IDSs.

References

[1] M. A. Jamshed, K. Ali, Q. H. Abbasi, M. A. Imran, and M. Ur-Rehman, "Challenges, applications and future of wireless sensors in internet of things: A review," *IEEE Sensors Journal*, pp. 1–1, 2022.

[2] L. Horwitz, "The future of iot miniguide: The burgeoning iot market continues." [Online]. Available: https://www.cisco.com/c/en/us/solutions/internet-of-things/future-of-iot.html

[3] M. A. M.Sadeeq, S. R. M. Zeebaree, R. Qashi, S. H. Ahmed, and K. Jacksi, "Internet of things security: A survey," in *2018 International Conference on Advanced Science and Engineering (ICOASE)*, 2018, pp. 162–166.

[4] S. Morgan, "Global ransomware damage costs predicted to reach $20 billion (usd) by 2021." [Online]. Available: https://cybersecurityventures.com/

[5] Z. Abou El Houda, L. Khoukhi, and A. Senhaji Hafid, "Bringing intelligence to software defined networks: Mitigating ddos attacks," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2523–2535, 2020.

[6] Z. A. El Houda, A. S. Hafid, and L. Khoukhi, "A novel machine learning framework for advanced attack detection using sdn," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[7] S. Wang, M. A. Qureshi, L. Miralles-Pechuán, T. Huynh-The, T. R. Gadekallu, and M. Liyanage, "Explainable ai for b5g/6g: Technical aspects, use cases, and research challenges," 2021.

[8] 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.

[9] S. Mane and D. Rao, "Explaining network intrusion detection system using explainable ai framework," 2021.

[10] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for intrusion detection systems," *IEEE Access*, vol. 8, pp. 73 127–73 141, 2020.

[11] s. wali and I. Khan, "Explainable ai and random forest based reliable intrusion detection system detection system," 12 2021.

[12] K. Amarasinghe and M. Manic, "Improving user trust on deep neural networks based intrusion detection systems," in *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, 2018, pp. 3262–3268.

[13] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable ai in intrusion detection systems," 2018.

[14] H. Li, F. Wei, and H. Hu, "Enabling dynamic network access control with anomaly-based ids and sdn," in *Proceedings of the ACM International Workshop on Security in Software Defined Networks amp; Network Function Virtualization*. Association for Computing Machinery, 2019, p. 13–16.

[15] "Shap (shapley additive explanations) library." [Online]. Available: https://shap.readthedocs.io/en/latest/index.html