IEEE Open Journal of the Communications Society

Received 11 June 2022; accepted 30 June 2022. Date of publication 11 July 2022; date of current version 29 July 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3188750

"Why Should I Trust Your IDS?": An Explainable Deep Learning Framework for Intrusion Detection Systems in Internet of Things Networks

ZAKARIA ABOU EL HOUDA [1] (Member, IEEE), BOUZIANE BRIK [2] (Member, IEEE), AND LYES KHOUKHI [3] (Senior Member, IEEE)

[1] L@bISEN, ISEN Yncréa Ouest, 44470 Carquefou, France

[2] DRIVE EA1859, University of Bourgogne Franche-Comté, 25000 Besançon, France

[3] GREYC CNRS, ENSICAEN, Normandie University, 44470 Caen, France

CORRESPONDING AUTHOR: B. BRIK (e-mail: bouziane.brik@u-bourgogne.fr)

ABSTRACT Internet of Things (IoT) is an emerging paradigm that is turning and revolutionizing worldwide cities into smart cities. However, this emergence is accompanied with several cybersecurity concerns due mainly to the data sharing and constant connectivity of IoT networks. To address this problem, multiple Intrusion Detection Systems (IDSs) have been designed as security mechanisms, which showed their efficiency in mitigating several IoT-related attacks, especially when using deep learning (DL) algorithms. Indeed, Deep Neural Networks (DNNs) significantly improve the detection rate of IoT-related intrusions. However, DL-based models are becoming more and more complex, and their decisions are hardly interpreted by users, especially companies' executive staff and cybersecurity experts. Hence, the corresponding users cannot neither understand and trust DL models decisions, nor optimize their decisions (users) based on DL models outputs. To overcome these limits, Explainable Artificial Intelligence (XAI) is an emerging paradigm of Artificial Intelligence (AI), that provides a set of techniques to help interpreting and understanding predictions made by DL models. Thus, XAI enables to explain the decisions of DL-based IDSs to make them interpretable by cybersecurity experts. In this paper, we design a new XAI-based framework to give explanations to any critical DL-based decisions for IoT-related IDSs. Our framework relies on a novel IDS for IoT networks, that we also develop by leveraging deep neural network, to detect IoT-related intrusions. In addition, our framework uses three main XAI techniques (*i.e.*, RuleFit, Local Interpretable Model-Agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP)), on top of our DNN-based model. Our framework can provide both local and global explanations to optimize the interpretation of DL-based decisions. The local explanations target a single/particular DL output, while global explanations focus on deducing the most important features that have conducted to each made decision (e.g., intrusion detection). Thus, our proposed framework introduces more transparency and trust between the decisions made by our DL-based IDS model and cybersecurity experts. Both NSL-KDD and UNSW-NB15 datasets are used to validate the feasibility of our XAI framework. The experimental results show the efficiency of our framework to improve the interpretability of the IoT IDS against well-known IoT attacks, and help the cybersecurity experts get a better understanding of IDS decisions.

INDEX TERMS Internet of Things, intrusion detection system, deep learning, explainable artificial intelligence, local and global explanations.

I. INTRODUCTION

INTERNET of Things (IoT) is an emerging technology that is becoming an integral part of our everyday life [1], [2]. IoT is shaping our future by revolutionizing worldwide cities into smart cities [3]. IoT consists to connect and deploy billions of devices, estimated at 75 billion IoT devices by 2025 [3], through emergent communication technologies, to realize various applications related to multiple industries, including agriculture, Healthcare, factories, and transportation [2], [3].

However, with this rapid revolution, various cybersecurity attacks are also increasing, which is mainly due to data sharing and constant connectivity, in addition to resource-limited nature of IoT networks [4], [5]. For instance, Mirai IoT botnet attack succeeded to remotely control several bots (or zombies), that were then used to perform large-scale Distributed Denial-of-Service (DDoS) attacks [6]. Such attack targeted multiple IoT devices, including IP cameras, IoT gateways, and home routers. As results, many service providers, Amazon and Twitter, were unavailable for several hours [4]. Thus, these attacks are causing significant business losses and damage, estimated at $20 Billion (USD) in 2021 [7]. To deal with the IoT attacks, research and industrial actors are investing to provide new intelligent solutions, improving security of IoT networks, like our previous intrusion detection mechanisms [8]–[10].

Thus, designing new security mechanisms becomes more than necessary to deal with various IoT attacks, ranging from DDoS attacks to scanning attacks. In this context, Intrusion Detection Systems (IDS) are promising solutions to protect IoT networks against multiple attacks. In addition, Deep Learning (DL) algorithms are recently leveraged on top of IDS to design intelligent IDS, optimizing clearly the detection rate of IoT-related intrusions. Indeed, DL-based IDSs consist to learn the signature of each IoT attack, in order to be efficiently and timely predicted/detected by the system. Once an attack is detected, precautionary measures should be taken by staff (e.g., cybersecurity experts or executive staff), to deal with such attack. However, recent DL-based IDSs are based on Deep Neural Network (DNN) models, which are becoming more and more complex, i.e., it is difficult to understand the inner working of such models, especially by not-expert users in data science. Thus, such models are provided/deployed as black-box models. In addition, decisions made by such models are provided to users, without any explanations or interpretations on how and why such decisions are made. Therefore, the corresponding users cannot neither understand and trust DL models decisions, nor optimize their decisions (users), based on DL model outputs. To overcome these limits, eXplainable Artificial Intelligence (XAI) is an emerging paradigm of Artificial Intelligence (AI), that provides a set of techniques to help interpreting and understanding predictions made by DL models [11]. Thus, XAI enables to explain the decisions of DL-based IDSs to make them interpretable by cybersecurity experts. This also enables experts to trust and adapt such models and hence perform their decisions (models) [12]–[28].

In this paper, we design a novel two stages XAI-empowered framework that uses DL-based architecture to detect IoT-based attacks and three main XAI techniques, on the top of our DNN-based model; the objective is to provide both local and global explanations to optimize the interpretation of DL-based decisions. First, we propose a novel DL-based architecture that uses Deep Neural Networks (DNNs) to protect IoT-based networks against the new emerging IoT-based attacks. Then, we develop three main XAI techniques: SHapley Additive exPlanations (SHAP) [29], RuleFit [30], and Local Interpretable Model-Agnostic Explanations (LIME) [31], on the top of our proposed DL-based architecture to provide both local and global explanations to optimize the interpretation of DL-based decisions. The local explanations target a single/particular DL output, while global explanations focus on deducing the most important features that have conducted to each made decision (e.g., intrusion detection). Hence, our framework introduces more transparency and trust between the decisions made by our DL-based IDS model and cybersecurity experts. NSL-KDD and UNSW-NB15 datasets were used to validate the feasibility of our XAI framework [32]. The experimental results show the efficiency of our framework to improve the interpretability of the IoT IDS against well-known IoT attacks, and help the cybersecurity experts get a better understanding of IDSs' decisions.

The main contributions of this paper can be summarized as follows:

- We propose a novel XAI-empowered framework that uses advanced DL-based techniques and well-known XAI techniques to provide cybersecurity experts with the ability to systematically explain local/global DL-based IDS decisions.
- We propose a novel DL-based architecture that uses Deep Neural Networks (DNNs) to protect IoT-based networks against the new emerging IoT-based attacks.
- We develop three main XAI techniques, namely SHAP, RuleFit, and LIME, on the top of our proposed DL-based architecture, that investigates the use of both of local and global explanations to optimize the interpretation of DL-based decisions.
- We evaluate the performance/feasibility of our proposed XAI-empowered framework using NSL-KDD and UNSW-NB15 datasets. The experimental results show the efficiency of our framework to improve the interpretability of the IoT-based IDS against well-known IoT attacks, and help the cybersecurity experts get a better understanding of IDSs' decisions.

This paper is organized as follows. Section II gives an overview on existing related solutions. We describe our XAI-based framework with its main components, in Section III. Section IV gives the performance evaluation of our XAI-based framework. Finally, Section V concludes the paper.

II. RELATED WORK

In this section, we give the few existing solutions that addressed the explainability of DL-based IDS systems [33]–[37]. In [33], the authors addressed the challenge of how to explain IDS in computer networks. They first leveraged deep learning to build a DL-based IDS. Then, they designed an XAI framework to optimize the transparency of their DL-based IDS. The authors used $NSL - KDD$ dataset not only for creating the DL-based IDS, but also to validate many XAI techniques, such as, LIME, SHAP, ProtoDash, and contrastive explanations method.

Similarly, another XAI framework was designed using SHAP approach, to add more transparency and explainability to any IDS system, in [34]. The authors aimed to combine local and global explanations to improve the interpretability. They also built two classifiers (one class and multi-class), and compare the interpretations of both classifiers. The NSL-KDD dataset is used to demonstrate the feasibility of the designed framework.

In [35], another XAI framework is designed to deal with adversarial attacks on top of machine learning-based IDS. The authors first built a random forest classifier to identify intrusions in the network. Then, global explanations are associated to each classifier prediction using SHAP approach. The performance of the framework are evaluated on hop skip jump attack and CICIDS dataset. Moreover, the developed machine learning-based IDS is validated against other learning algorithms, through several metrics, including precision, recall, F1-score, and accuracy.

Besides, the authors aimed to improve user trust against deep learning-based IDS, by optimizing its transparency, in [38]. To do so, the authors first trained a deep learning-based IDS using the KDD-NSL dataset. They then implemented a layer-wise relevance propagation (LRP) method, to generate both offline and online interpretations. The offline explanations give the users the most relevant input features, in detecting each intrusion, while the online interpretations give the users the inputs features contributing more on the detection.

In [39], the authors targeted to generate the explanations of incorrect classifications made by deep learning-based IDS classifiers. In addition, an adversarial approach is designed to find the modifications of the input features, needed to correct the classification. Moreover, such approach also enables to show the most relevant features, resulting the incorrect classification. Thus, this approach enables to give more explanations about the main reasons of the misclassifications. Noting that the designed approach is validated using NSL-KDD dataset.

The authors addressed the challenge of dynamic network access in Software Defined Networking (SDN) era, in [40]. They built a Recurrent Neural Network (RNN) based IDS, to detect network anomalies and generate SDN flow rules to enable then dynamic network access control. In addition, they also train an interpretable model to explain the RNN-based model's outcome. Based on the explanation, the authors derived access control policies. In [41], the authors designed a new autoencoder-based detection framework that uses Convolutional Neural Network (CNN) and Recurrent Neural Networks (*i.e.*, LSTM), to discover attacks in Industrial IoT (IIoT) networks and explain the model. The main advantage of this framework is that it combines both LSTM and CNN to detect both traditional attacks and new (zero-day) attacks related to IIoT. Moreover, this framework leverages LIME approach to provide local explanation that matches each prediction made by the CNN-LSTM model. Although these existing works [33]–[41] considered XAI approaches, such as LIME and SHAPE, to interpret and explain DL-enabled IDS; however, they focused only on shallow machine learning algorithms, which are not complicated to interpret compared to DL algorithms. Reference [35], also they were designed for a general setup, without considering the DL algorithm implemented for IoT-based networks (*i.e.*, IoT-based attacks) [34], [38], [39]. These existing works may not be realistic for some cases, since the XAI model should take into account the main features of the DL algorithm, to be able then to explain its decisions. Moreover, most of these works did not target the IoT networks and used the NSL-KDD dataset which does not cover the IoT attacks, especially the emerging ones. To overcome these limits, in this work, we designed a novel framework that leverages RuleFit, LIME, and SHAPE as XAI approaches, to explain and interpret a deep neural network-based IDS for IoT networks, that we also develop in this work by leveraging both NSL-KDD and UNSW-NB15 datasets. We note that our framework enables not only to deduce the most relevant features conducting to each DL-based prediction, but also providing both local and global explanations related to each IDS decision.

III. XAI-BASED FRAMEWORK FOR DEEP LEARNING-BASED IDS OF IOT NETWORKS

In this section, we present our XAI-empowered framework. First, we give an overview about the architecture of our framework. Then, we describe our deep neural architecture we build to detect intrusions related to the IoT networks, and our XAI approaches we applied to interpret and explain the outputs of our deep learning model of IDS.

A. SYSTEM ARCHITECTURE

Fig. 1 shows the architecture of our AI-Empowered Framework for IoT IDS; it covers different IoT devices that may be deployed in various sectors, such as agriculture, healthcare, factories, and transportation. We first exploit sensed data by these devices, to create deep learning model that is able to identify/predict intrusions in such IoT networks. In addition, the deep learning model and sensed data from the IoT networks are also combined and leveraged by XAI approaches, in order to interpret and explain predictions made by our deep learning model. In particular, we develop three different XAI approaches: LIME, SHAPE and RuleFit, to generate local, global, and feature importance-based explanations, respectively. Thus, our framework enables to show not only how and why predictions are made, but also how the deep learning model works. Furthermore, our framework may target different users, via an explanation interface, including model users and security experts.

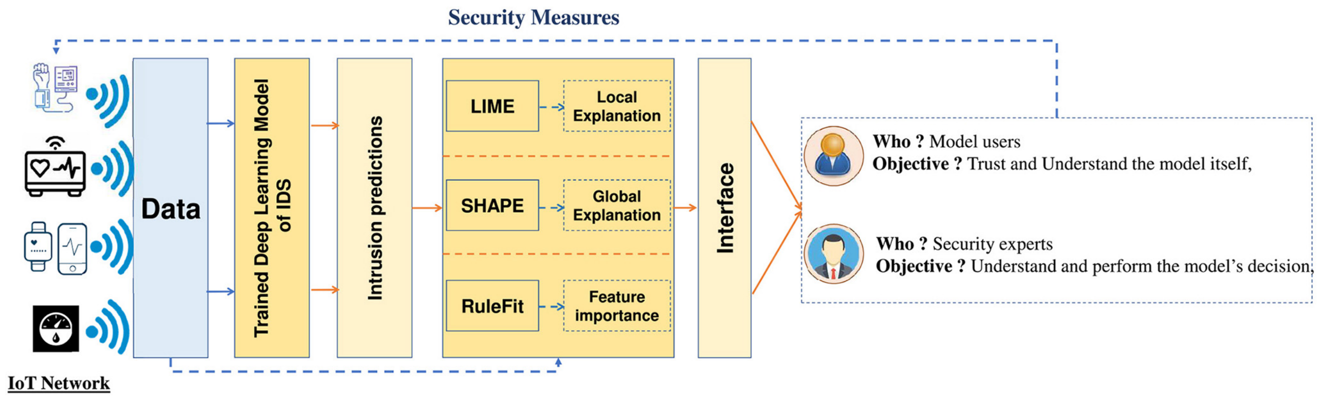


FIGURE 1. Architecture of our XAI-Empowered Framework for IoT IDS.

B. EXPLAINABLE DEEP LEARNING-BASED IDS FOR IOT APPLICATIONS

First, to evaluate the efficiency of our XAI-powered framework, we built two deep neural network (DNN) models with an input layer of 122 dimensions and 49 dimensions, that corresponds to the dimension of the input features for the NSL-KDD and UNSW-NB15 datasets, respectively. Each DNN architecture is composed of five hidden layers with Leaky Rectified Linear Unit, and an output layer of two dimensions, that corresponds to the dimension of the class label (*i.e.*, Attack or Normal). This work aims to effectively explain the decision made by this DL-based IDS, the objective is, through these explanations, to answer this question: "Why should I trust your IDS?". The emphasis is on exploring linear and non-linear techniques, including both local and global explanations; it consists of three techniques, namely, RuleFit, Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP).

1) RULEFIT

The RuleFit algorithm was originally designed by Friedman and Popescu [30] to learn sparse linear forms (*i.e.*, models) that contain the interaction effects in a form of decision-making rules. The objective is to create a simple yet interpretable model that integrates the interactions between the features; it learns a sparse linear model, including the original features and new features (*i.e.*, decision rules). RuleFit generates the new features/rules automatically from decision trees, where each path in the decision tree represents a decision rule. The new features are designed to capture the interactions between the existing features. RuleFit includes two components: (1) the decision rules that are created based on decision trees; and (2) it fits a sparse linear model with the original model as well as the new ones (*i.e.*, decision rules).

First, we use an ensemble of decisions trees to generate a variety of meaningful decision rules, a tree ensemble can be presented as follows:

$$F(x) = \widehat{i}_0 + \sum_{i=1}^{M} \widehat{i}_i F_i(y) \tag{1}$$

where M is the number of trees, $\widehat{i}$ are the weights, y is the original feature vector, and $F_i(y)$ is the prediction function of the i^{th} decision tree.

Then, we create the decision rules as follows:

$$r_i(y) = \Pi_{j\in T_i} I\left(y_j \in \zeta_{ji}\right) \tag{2}$$

where T_i is the set of features used in the i^{th} decision tree, and ζ_{ji} is an interval in the range of values of the features, $I(y_j \in \zeta_{ji})$ is 1 when y_j is in this interval value and 0 otherwise.

Thus, the number of decision trees created is defined as follows:

$$N = \sum_{i=1}^{M} 2(t_i - 1) \tag{3}$$

where t_i is the number of terminal node of the i^{th} decision tree.

Once this first phase of decision rules generation is completed, we train a sparse linear model, using the original features and the generated rules (*i.e.*, new features).

First, we winsorize the original features as follows, to make them robust against outliers:

$$l_j(y_j) = \min\left(\gamma_j^+, \max\left(\gamma_j^-, y_j\right)\right) \tag{4}$$

where $gamma_j^+$ and $gamma_j^-$ are the quantiles of the distribution of the data of the feature y_j.

Then, we normalize this linear term as follows:

$$l_j^*(y_j) = 0.4 \frac{l_j(y_j)}{std\left(l_j(y_j)\right)} \tag{5}$$

Finally, we combine both type of features and train the sparse linear model as follows:

$$F(x) = \lambda_0 + \sum_{k=1}^{K} \phi_k r_k(y) + \sum_{j=1}^{J} \lambda_j l_j^*(y_j) \tag{6}$$

Algorithm 1: RuleFit Algorithm

Input: Sequence of M Decision Trees (DTs), $i = 1, \ldots, M$
Sequence of T_i data samples $\{(x_j, y_j)\}$, $j = 1, \ldots, T_i$
We define a DT: $F(x) = \widehat{i}_0 + \sum_{i=1}^{M} \widehat{i}_i F_i(y)$
for $j \leftarrow 1$ *to* T_i **do**
  We create the decision rules: $r_i(y) = \Pi_{j \in T_i} I(y_j \in \zeta_{ji})$
  Then, we train a sparse linear model, using the original features and the generated rules.
  Afterwards, we winsorize the original features: $l_j(y_j) = \min(\gamma_j^+, \max(\gamma_j^-, y_j))$
  Then, we normalize this linear term: $l_j^*(y_j) = 0.4 \frac{l_j(y_j)}{std(l_j(y_j))}$
  Finally, we combine both type of features and train the sparse linear model: $F(x) = \lambda_0 + \sum_{k=1}^{K} \phi_k r_k(y) + \sum_{j=1}^{J} \lambda_j l_j^*(y_j)$
  Finally, we calculate the total importance score of the j^{th} feature:$IF_j(y) = I_j(y) + \sum_{y_j \in r_k} I_k(x)/m_k$
end
Make final feature importance: $IF(Y) = \sum_{j=1}^{J} IF_j(y_j)$

where λ and ϕ is the estimated weights for the original features and the new generated rules, receptively.

Since RuleFit is based on the Lasso, the loss function has the following additional constraint:

$$\left(\{\lambda\}_1^J, \{\phi\}_1^K\right) = \underset{\{\lambda\}_1^J, \{\phi\}_1^K}{arg\,min} \sum_{m=1}^{n} \mathcal{L}(y_m, f(x_m)) + \mu \left(\sum_{k=1}^{K} |\phi_k| + \sum_{j=1}^{J} |\lambda_j| \right) \tag{7}$$

For the original input features, the features importance score is calculated as follows:

$$I_j = |\lambda_j|.std\left(l_j^*(y_j)\right) \tag{8}$$

For the new generated features *i.e.*, decision rules the features importance score is calculated as follows:

$$I_k = |\phi_k|.\sqrt{\zeta_k(1 - \zeta_k)} \tag{9}$$

Finally, the total importance score of the j^{th} feature is calculated as follows:

$$IF_j(y) = I_j(y) + \sum_{y_j \in r_k} I_k(x)/m_k \tag{10}$$

where m_k is the number of input features constituting the decisions rule r_k.

And the global feature importance score is calculated as follows:

$$IF(Y) = \sum_{j=1}^{J} IF_j(y_j) \tag{11}$$

Algorithm 1 shows the algorithmic representation of the RuleFit algorithm.

2) LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS (LIME)

LIME stands for Local Interpretable Model-agnostic Explanations. The main goal of LIME is to find an interpretable model over the interpretable representation (*i.e.*, understandable by humans) which is locally faithful/truthful to the classifier. Let $x \in R^d$ be the original representation of an instance, and ler $g \in G$ be an explanation model, where G is a class of interpretable models that can be visually presented to a user (e.g., linear model). LIME's explanation can be obtained by the following:

$$\varphi(x) = arg\,min_{g \in G} \{\mathcal{L}(f, g, \omega_x) + \Omega(g)\} \tag{12}$$

where f is the model used for classification, ω_x is a proximity measure/weight between the original and the new instance; the higher the value of ω_x, the more the new instances are similar to original instances, $\mathcal{L}$ is the loss function that measures the proximity between the predictions of the explanation model and the original model, and $\Omega(g)$ is a measure of complexity of the model g.

Thus, the objective of LIME is to train a local yet interpretable model by minimizing the function $\mathcal{L}(f, g, \omega_x) + \Omega(g)$. Then, explain the prediction of an instance using he locally computed interpretation model $\varphi(x)$.

3) SHAPLEY ADDITIVE EXPLANATIONS (SHAP)

SHAP stands for SHapley Additive exPlanations, is defined as a well-known unified framework for the interpretation of models. SHAP explains the predictions of an instance by calculating the contribution of each feature to the final decision/prediction. The contribution can be either negative or positive. The major strength of SHAP is that it can be applied to any model/classifier, instead of linear models/classifiers. Rather than examining only local decisions/interpretations, SHAP examines global interpretations by summing the input values of features and averaging all columns/features individually. SHAP's explanation for an instance can be obtained by the following:

$$g(s) = \upsilon_0 + \sum_{i=1}^{N} \upsilon_i s_i \tag{13}$$

where s is the simplified feature, it represents the new features that are similar to the original ones, N is the maximum size, and υ_j is the Shapley value; the higher the value of υ_j of feature j, the more this feature has a large contribution on the final prediction of the model.

Finally, we select the most important features as follows:

$$IF_j = \sum_{i=1}^{n} \left\| \upsilon_j(x_i) \right\| \tag{14}$$

where n the total number of data samples, IF_j is to the average Shapley value of the i^{th} input feature.

TABLE 1. Performance metrics of Our XAI-empowered framework and state-of-the-art ML/DL-based models using $NSL-KDDTest^{+}$.

Methods	Accuracy	Precision	Recall	F1
J48 [42]	0.81	N/A	N/A	N/A
NB [42]	0.76	N/A	N/A	N/A
RF [42]	0.80	N/A	N/A	N/A
MLP [42]	0.77	N/A	N/A	N/A
SVM [42]	0.70	N/A	N/A	N/A
CNN [43]	0.85	0.91	0.81	0.86
ResNet architecture [44]	0.79	0.91	0.69	0.79
GoogleNet architecture [44]	0.77	0.91	0.65	0.76
DNN architecture [45]	0.75	0.83	0.75	0.74
RNN [46]	0.83	N/A	0.83	N/A
SVM-IDS [47]	0.78	N/A	0.78	N/A
Our XAI-empowered framework	**0.88**	**0.96**	**0.88**	**0.88**

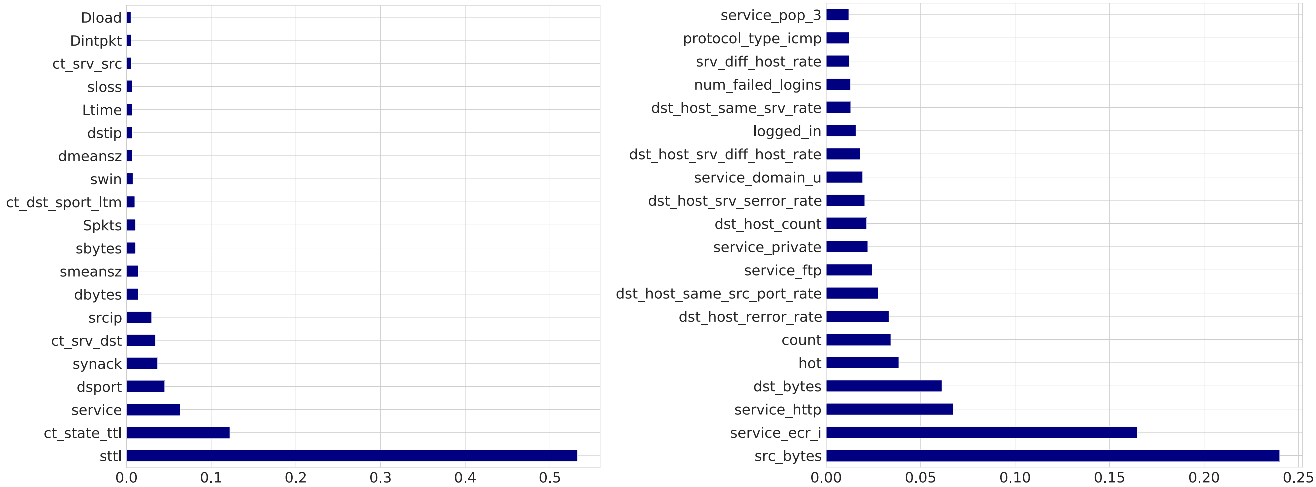
TABLE 2. Performance metrics of Our XAI-empowered framework and state-of-the-art ML/DL-based models using UNSW-NB15.

Methods	Accuracy	TPR
Fuzziness semi-supervised Architecture [48]	0.86	0.85
Random Forest Architecture [49]	0.93	0.92
Generalized Outlier Gaussian Mixture [50]	0.95	0.94
Mixture-Hidden Markov Model [51]	0.96	0.95
Our XAI-empowered framework	**0.99**	**0.99**

IV. PERFORMANCE EVALUATION

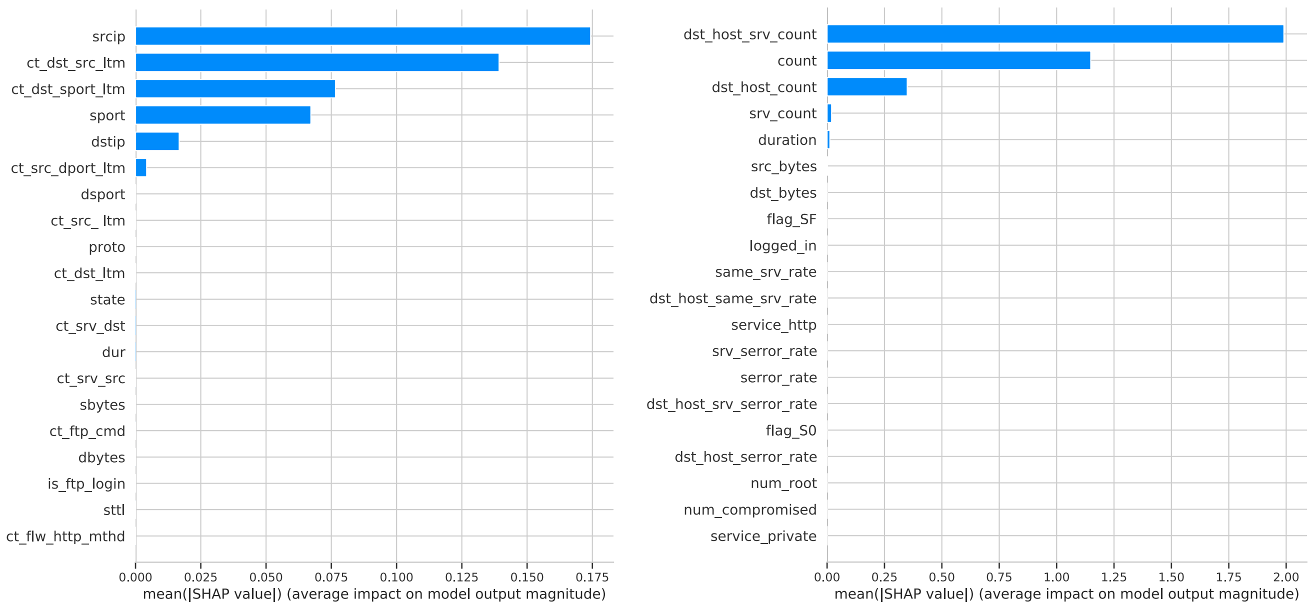
In this work, we consider two well-known public network security datasets, namely NSL-KDD and UNSW-NB15. The NSL-KDD dataset contains real-world network security attacks; it is an improved version of the KDD'99 dataset where all redundant features have been removed. NSL-KDD dataset includes the following attacks: Distributed Denial of Service (DDoS), User to Root (U2R), Probe (Probing) and Root to Local (R2L). The UNSW-NB15 dataset is a synthetic network security dataset that includes more than 100 GB of network data; it includes the following attacks: analysis, fuzzers, DoS, backdoors, reconnaissance, generic, exploits, shellcode, and worms. We have implemented the proposed XAI framework using Pytorch and the XAI libraries, including SHAP [52]. First, we have encoded the categorical data (e.g., 'proto') into numeric ones using one hot encoding techniques. Some features of both NSL-KDD and UNSW-NB15 datasets (e.g., Source jitter (mSec) (sjit) [0;11*10^5]' and 'Destination jitter (mSec) (djit) [0;7.8*10^9]') have higher values than others; which may have an impact on the final decisions of the model, where the model may miss out important features, *i.e.*, ct_flw_http_mthd (number of flows that have the Get and Post methods in the http service). Thus, we used standardization technique to address this problem. Finally, we encoded the labels of both NSL-KDD and UNSW-NB15 datasets (e.g., DDoS, Probe, backdoors, and Fuzzers) into numerical values. At the first stage, We tested the performance of our DNN architecture in terms of accuracy and F1 score. we also compared the results obtained with the state-of-the-art schemes, using both datasets, NSL-KDD and UNSW-NB15. Tables 1 and 2 show the results of our proposed XAI-empowered framework and the most relevant works in the state-of-the-art; we observe that our proposed XAI-empowered framework achieves the highest accuracy and detection rate on both datasets. The experimental results confirm that our proposed XAI framework outperforms the state-of-the-art works in terms of accuracy and F1-score on both datasets. The feature importance scores includes the importance of the original features and the decision rules where the features appears; it shows the most relevant features/rules that have important/significant impact on the model predictions. The proposed framework studies the use of linear and non-linear techniques, including both local and global explanations, to identify the most informative features and investigate their impact on the final model's predictions; it consists of three techniques, namely, RuleFit, Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP).

Fig. 2 shows the important features that have the highest scores using RuleFit method on UNSW-NB15 and NSL-KDD datasets, respectively. For the UNSW-NB15 dataset, the highest scoring features corresponds to the following features: (1) sttl: which is the Source to destination time to live; (2) ct_state_ttl: which is the Number of each state according to a range of values for source/destination time to live (ttl); (3) service: which is the protocol used, e.g., http, dns, ssh; and (4) dsport: which is the destination port number. For the NSL-KDD dataset, the highest scoring features corresponds to the following features: (1) src_bytes: which is the number of data bytes from source to destination; (2) service: which is the network service of the destination host/machine, e.g., http; (3) dst_bytes: which is the number of data bytes from destination to source; and (4) hot feature: which is the number of "hot" indicators (e.g., directory accesses). Fig. 3 shows the important features that have the highest scores using SHAP method on UNSW-NB15 and NSL-KDD datasets, respectively. For the UNSW-NB15 dataset, the highest scoring features corresponds to the following features: (1) srcip: corresponds to the Source IP address of the source machine; (2) ct_dst_src_ltm: corresponds to the number of connections that contain the same service and destination address in the last hundred connections; and (3) ct_dst_sport_ltm: corresponds to the number of connections of the same destination address and the source port in the last hundred connections. For the NSL-KDD dataset, the highest scoring features corresponds to the following features: (1) dst_host_srv_count: corresponds to the feature Srv-count for destination host;



(a) (b)

FIGURE 2. **Feature Importance Scores using RuleFit on: a) UNSW-NB15; and b) NSL-KDD.**



(a) (b)

FIGURE 3. **Feature importance scores using SHAP on: a) UNSW-NB15; and b) NSL-KDD.**

(2) count: corresponds to the number of connections to the same host as the current connection in the past two seconds; and (3) dst_host_count: corresponds to the feature fount for the destination host. Fig. 4 shows the 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, while Fig. 5 shows the data samples distribution of features of NSL-KDD dataset in terms for: (a) the highest scoring features using RuleFit and SHAP; and (b) the other non-irrelevant features. In both figures, we observe that the most relevant features, computed based on RuleFit and SHAP methods, respectively, 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.

Figs. 6 and 7 show the interpretation of our DL-based IDS on UNSW-NB15 and NSL-KDD datasets using SHAP method, receptively. In our experiments we have examined two observations for each dataset. 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

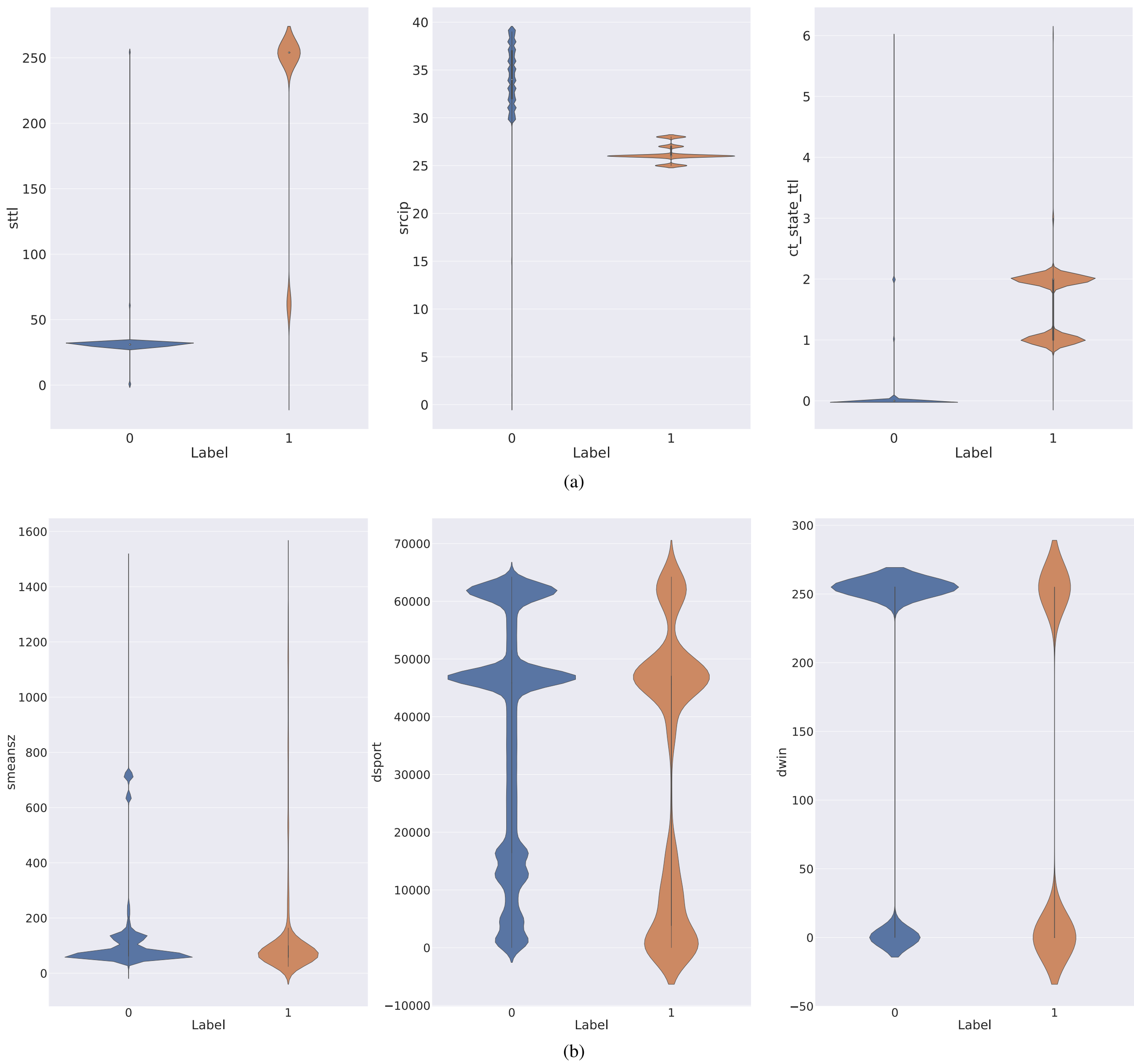
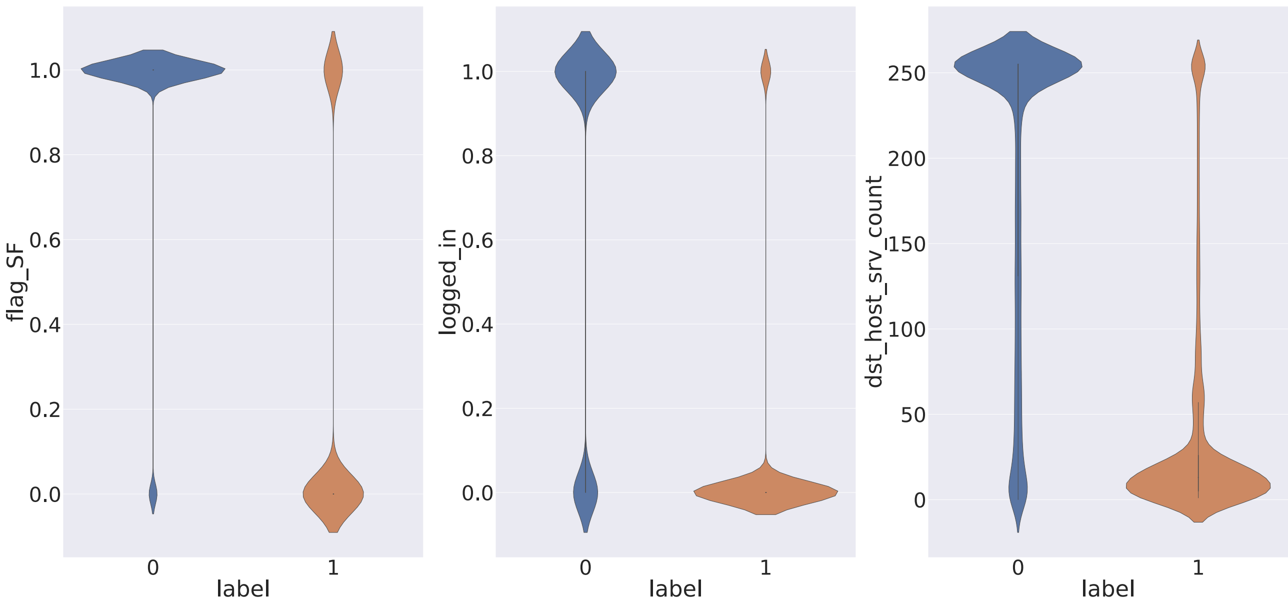
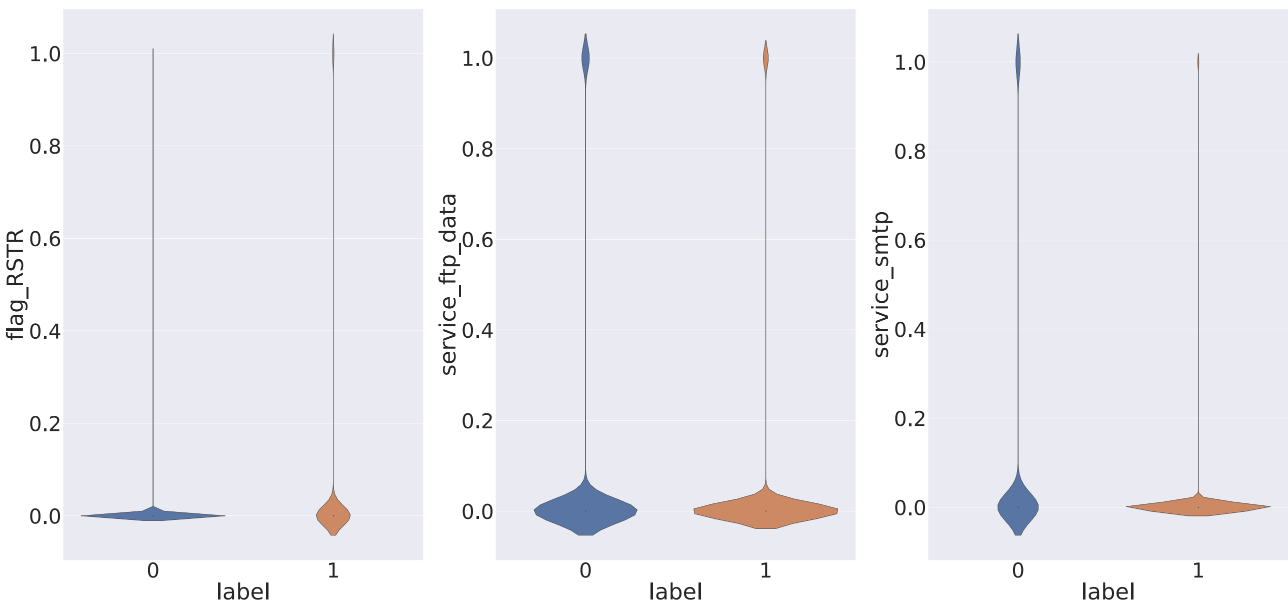


FIGURE 4. **Data samples distribution of features of *UNSW − NB*15 dataset in terms of: a) the highest scoring features using RuleFit and SHAP; and b) the other non-irrelevant features.**

all the columns/features individually. In the following observations, the blue features push the prediction of an instance to be Normal, while the red features reduce the probability for a data sample to be Normal. Fig. 6(a) shows the first observation using UNSW-NB15 dataset, where the data sample is an attack and our DL-based IDS correctly predicted/detected as an attack. In this observation, the values of the input features are as follows: ct_dst_sport_ltm is equal to 1.0, ct_dst_src_ltm is equal 1.0, and srcip is equal to 38.0. In this observation, the most contributing features are: ct_dst_sport_ltm and ct_dst_src_ltm; these features drive the probability for a data sample to be an attack. Fig. 6(b) shows the second observation using UNSW-NB15 in which the data sample is Normal and our DL-based IDS correctly predicted this data sample as a Normal one. In this observation, the values of the input features are as follows: srcip is equal to 36.0, sport is equal 55806.0, and dstip is equal to 23.0. Fig. 7(a) shows the first observation using NSL-KDD dataset in which the data sample is Normal and our DL-based IDS correctly predicted/detected as a Normal data sample. In this observation, the values of the input features are as follows: dst_host_count is equal to 180.0, count is equal 1.0, and dst_host_srv_count is equal to 167.0. Fig. 7(b) shows the second observation using NSL-KDD in which the data sample is Normal and our DL-based IDS correctly predicted/detected as a Normal data sample. In this observation, the values of the input features are as follows: dst_host_count is equal to 255.0, count is equal 30.0, and dst_host_srv_count



(a)



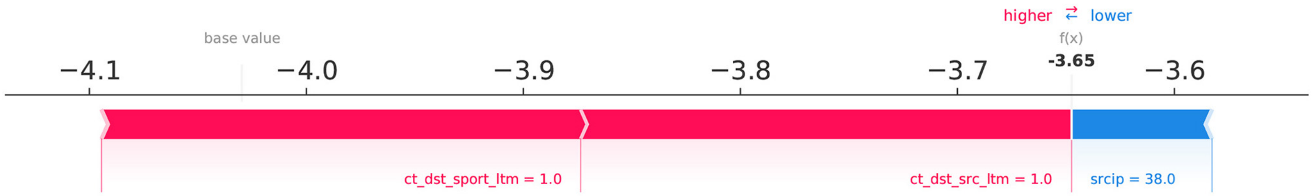
(b)

FIGURE 5. Data samples distribution of features of *NSL – KDD* dataset in terms for: a) the highest scoring features using RuleFit and SHAP; and b) the other non-irrelevant features.

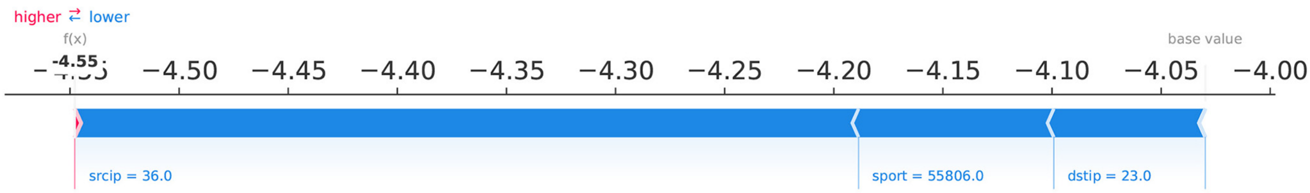
is equal to 255.0. The red feature (*i.e.*, dst_host_count) reduces the probability for a data sample to be Normal. Therefore, such solid knowledge makes cybersecurity experts more convinced of the decisions regarding ML/DL-based IDS. Fig. 8 shows the best important features using SHAP on UNSW-NB15 and NSL-KDD datasets, receptively. For a particular instance/observation, each input feature has either a positive or a negative contribution to the final decision. Fig. 9 shows the local explanation of our DL-based IDS using LIME on UNSW-NB15 dataset for (a) positive scenario; and (b) negative scenario, while Fig. 10 shows the local explanation of our DL-based IDS using LIME on NSL dataset for (a) positive scenario; and (b) negative scenario.

V. CONCLUSION

In this paper, we designed a new XAI-based Framework for intrusion detection in IoT networks. Our framework integrated first a deep neural network model to detect intrusions in real-time. Once this model makes decisions, our framework leverages three different approaches of XAI (*i.e.*, LIME, SHAPE, and RuleFit), to add more explainability,

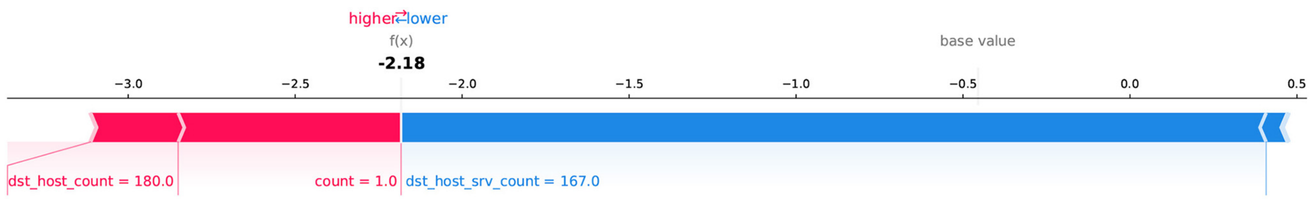


(a)

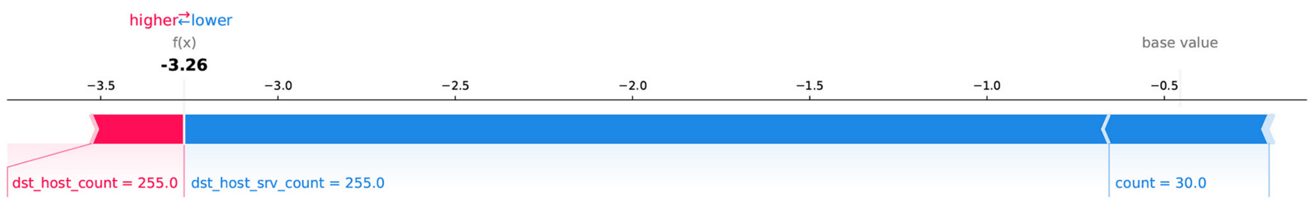


(b)

FIGURE 6. **Interpretation of our DL-based IDS on UNSW-NB15 dataset with: a) ct_dst_sport_ltm of 1.0, ct_dst_src_ltm of 1.0, and srcip of 38.0; and b) srcip of 36, sport of 55806, and dstip of 23.**



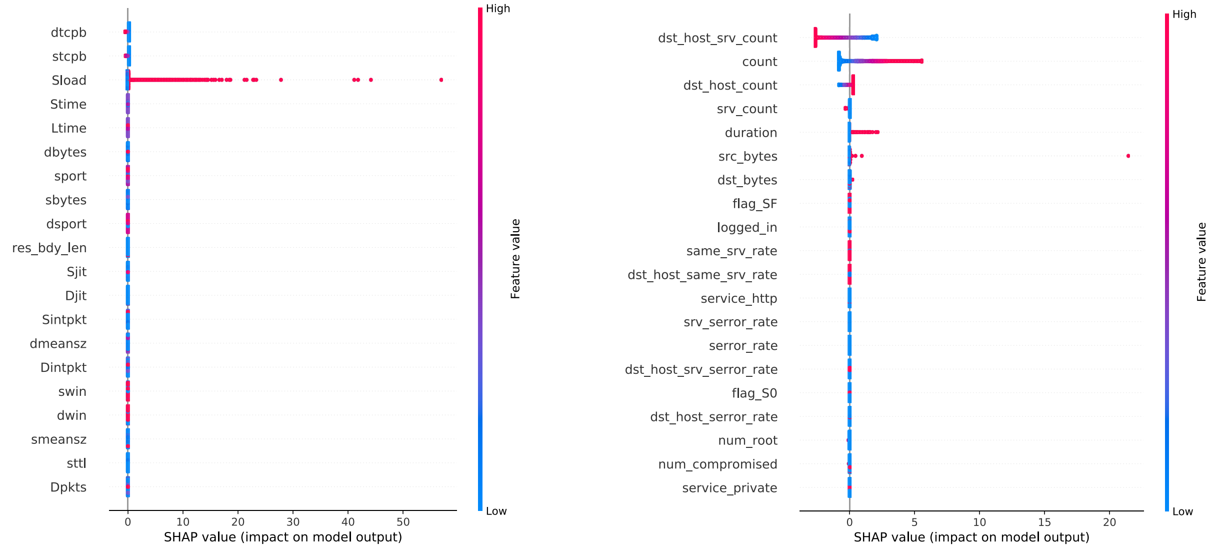
(a)



(b)

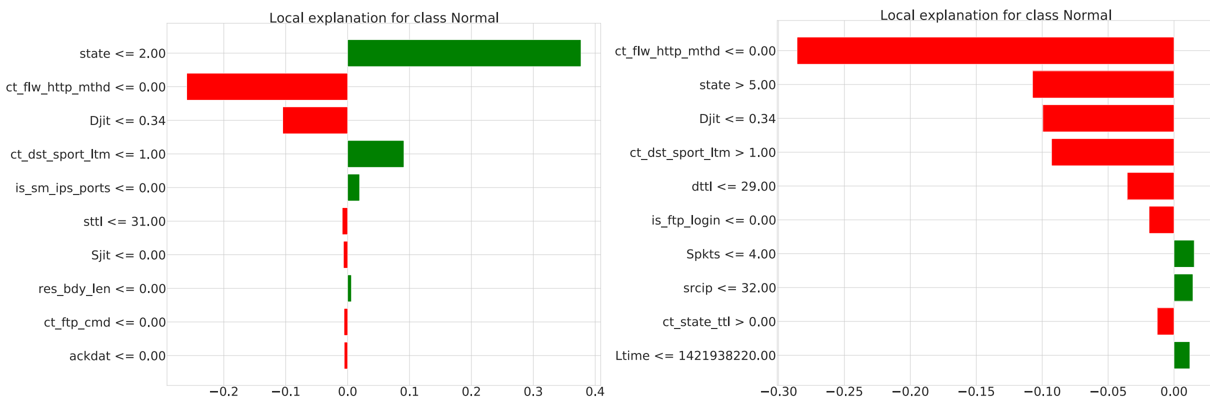
FIGURE 7. **Interpretation of our DL-based IDS on NSL-KDD dataset with: a) dst_host_count of 180.0, count of 1.0, and dst_host_srv_count of 167.0; and b) dst_host_count of 255.0, count of 30.0, and dst_host_srv_count of 255.0.**

transparency, and trust to the model's decisions. Moreover, our framework with its explainability targets two different users: users of the deep learning model that aim to understand and trust model's outputs, in order to be able to optimize their decisions, and cybersecurity experts that also aim to understand the model's outputs, in order to make the suitable recommendations, especially when an intrusion is detected. We have used both NSL-KDD and UNSW-NB15 datasets to demonstrate the feasibility/performances of our framework; the experimental results show the efficiency of our proposed XAI-based Framework in not only detecting IoT-based attacks, but also integrating more details and



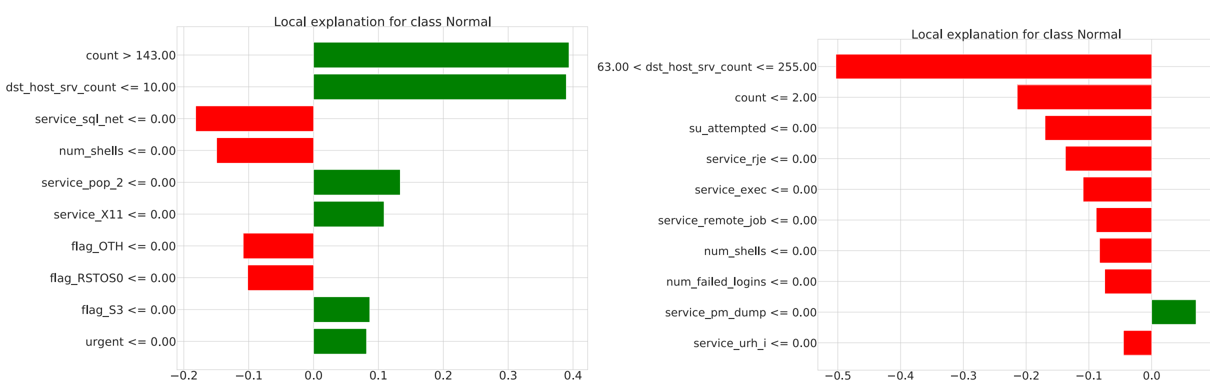
(a) (b)

FIGURE 8. **Top 20 important features using SHAP on : a) UNSW-NB15; and b) NSL-KDD.**



(a) (b)

FIGURE 9. **Local explanation of our DL-based IDS using LIME on UNSW-NB15 dataset for: a) positive scenario; and b) negative scenario.**



(a) (b)

FIGURE 10. **Local explanation of our DL-based IDS using LIME on NSL-KDD dataset for: a) positive scenario; and b) negative scenario.**

interpretation about how and why such detection decisions are made by our deep neural network model. As future work, we plan to secure our framework against adversarial attacks that may target the explainability module of our framework.

REFERENCES

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, 4th Quart., 2015.

[2] 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 J.*, vol. 22, no. 6, pp. 5482–5494, Mar. 2022.

[3] L. Horwitz. "The Future of IoT Miniguide: The Burgeoning IoT Market Continues." 2021. [Online]. Available: https://www.cisco.com/c/en/us/solutions/internet-of-things/future-of-iot.html

[4] M. A. M. Sadeeq, S. R. M. Zeebaree, R. Qashi, S. H. Ahmed, and K. Jacksi, "Internet of Things security: A survey," in *Proc. Int. Conf. Adv. Sci. Eng. (ICOASE)*, 2018, pp. 162–166.

[5] A. Mall, P. Singh, A. Thute, S. P. Khapre, and A. Shankar, "Security issues of edge computing in IoT," in *Proc. Int. Conf. Mach. Intell. Data Sci. Appl.*, 2021, pp. 567–579.

[6] A. Bhardwaj, M. Kumar, T. Stephan, A. Shankar, M. R. Ghalib, and S. Abujar, "IAF: IoT attack framework and unique taxonomy," *J. Circuits Syst. Comput.*, vol. 31, no. 2, 2022, Art. no. 2250029.

[7] S. Morgan. "Global Ransomware Damage Costs Predicted to Reach $20 Billion (USD) by 2021." 2021. [Online]. Available: https://cybersecurityventures.com/

[8] Z. A. El Houda, L. Khoukhi, and A. S. Hafid, "Bringing intelligence to software defined networks: Mitigating DDoS attacks," *IEEE Trans. Netw. Service Manag.*, vol. 17, no. 4, pp. 2523–2535, Dec. 2020.

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

[10] Z. A. El Houda, A. S. Hafid, and L. Khoukhi, "Cochain-SC: An intra- and inter-domain DDoS mitigation scheme based on blockchain using SDN and smart contract," *IEEE Access*, vol. 7, pp. 98893–98907, 2019.

[11] 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, *arXiv:2112.04698*.

[12] H. Moudoud, S. Cherkaoui, and L. Khoukhi, "An IoT blockchain architecture using oracles and smart contracts: The use-case of a food supply chain," in *Proc. IEEE 30th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, 2019, pp. 1–6.

[13] Z. A. El Houda, "Security enforcement through software defined networks (SDN)," Ph.D. dissertation, Dept. Sci. Eng., Univ. Technol. Troyes, Troyes, France, 2021.

[14] H. Moudoud, L. Khoukhi, and S. Cherkaoui, "Prediction and detection of FDIA and DDoS attacks in 5G enabled IoT," *IEEE Netw.*, vol. 35, no. 2, pp. 194–201, Mar./Apr. 2021.

[15] Z. A. El Houda, A. Hafid, and L. Khoukhi, "Co-IoT: A collaborative DDoS mitigation scheme in IoT environment based on blockchain using SDN," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2019, pp. 1–6.

[16] H. Moudoud, S. Cherkaoui, and L. Khoukhi, "Towards a scalable and trustworthy blockchain: IoT use case," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2021, pp. 1–6.

[17] Z. A. El Houda, L. Khoukhi, and A. Hafid, "ChainSecure—A scalable and proactive solution for protecting blockchain applications using SDN," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 1–6.

[18] H. Moudoud, S. Cherkaoui, and L. Khoukhi, *An Overview of Blockchain and 5G Networks*. Cham, Switzerland: Springer Int., 2022, pp. 1–20.

[19] Z. A. E. Houda, A. Hafid, and L. Khoukhi, "Blockchain meets AMI: Towards secure advanced metering infrastructures," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[20] H. Moudoud, Z. Mlika, L. Khoukhi, and S. Cherkaoui, "Detection and prediction of FDI attacks in IoT systems via hidden Markov model," *IEEE Trans. Netw. Sci. Eng.*, early access, Mar. 23, 2022, doi: 10.1109/TNSE.2022.3161479.

[21] Z. A. E. Houda, A. Hafid, and L. Khoukhi, "Blockchain-based reverse auction for V2V charging in smart grid environment," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2021, pp. 1–6.

[22] H. Moudoud, S. Cherkaoui, and L. Khoukhi, "Towards a secure and reliable federated learning using blockchain," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2021, pp. 01–06.

[23] Z. A. E. Houda, A. Hafid, and L. Khoukhi, "BrainChain—A machine learning approach for protecting blockchain applications using SDN," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[24] B. Brik and A. Ksentini, "Toward optimal MEC resource dimensioning for a vehicle collision avoidance system: A deep learning approach," *IEEE Netw.*, vol. 35, no. 3, pp. 74–80, May/Jun. 2021.

[25] Z. A. El Houda, B. Brik, A. Ksentini, L. Khoukhi, and M. Guizani, "When federated learning meets game theory: A cooperative framework to secure IIoT applications on edge computing," *IEEE Trans. Ind. Informat.*, early access, Apr. 26, 2022, doi: 10.1109/TII.2022.3170347.

[26] B. Brik and A. Ksentini, "On predicting service-oriented network slices performances in 5G: A federated learning approach," in *Proc. IEEE 45th Conf. Local Comput. Netw. (LCN)*, 2020, pp. 164–171.

[27] Z. A. El Houda, A. S. Hafid, and L. Khoukhi, *A Novel Unsupervised Learning Method for Intrusion Detection in Software-Defined Networks*. Cham, Switzerland: Springer Int., 2022, pp. 103–117.

[28] Z. A. El Houda, "Renforcement de la sécurité à travers les réseaux programmables," Ph.D. dissertation, Dept. Comput. Sci. Oper. Res., Université de Montréal, Montreal, QC, Canada, 2021.

[29] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.

[30] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 916–954, 2008.

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.

[32] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proc. Mil. Commun. Inf. Syst. Conf. (MilCIS)*, 2015, pp. 1–6.

[33] S. Mane and D. Rao, "Explaining network intrusion detection system using explainable AI framework," 2021, *arXiv:2103.07110*.

[34] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for intrusion detection systems," *IEEE Access*, vol. 8, pp. 73127–73141, 2020.

[35] S. Wali and I. Khan. "Explainable AI and Random Forest Based Reliable Intrusion Detection System." Dec. 2021. [Online]. Available: https://www.techrxiv.org/articles/preprint/Explainable_AI_and_Random_Forest_Based_Reliable_Intrusion_Detection_system/17169080

[36] B. Brik, N. Lagraa, A. Lakas, and Y. Ghamri-Doudane, "RCS-VC: Renting out and consuming services in vehicular clouds based on LTE-A," in *Proc. Global Inf. Infrastruct. Netw. Symp. (GIIS)*, 2015, pp. 1–6.

[37] N. Tamani, B. Brik, N. Lagraa, and Y. Ghamri-Doudane, "Vehicular cloud service provider selection: A flexible approach," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2017, pp. 1–6.

[38] K. Amarasinghe and M. Manic, "Improving user trust on deep neural networks based intrusion detection systems," in *Proc. 44th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, 2018, pp. 3262–3268.

[39] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable AI in intrusion detection systems," 2018, *arXiv:1811.11705*.

[40] H. Li, F. Wei, and H. Hu, "Enabling dynamic network access control with anomaly-based IDS and SDN," in *Proc. ACM Int. Workshop Security Softw. Defined Netw. Netw. Funct. Virtualization*, 2019, pp. 13–16.

[41] I. A. Khan, N. Moustafa, D. Pi, K. M. Sallam, A. Y. Zomaya, and B. Li, "A new explainable deep learning framework for cyber threat discovery in industrial IoT networks," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 11604–11613, Jul. 2022.

[42] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symp. Comput. Intell. Security Defense Appl.*, 2009, pp. 1–6.

[43] S. Z. Lin, Y. Shi, and Z. Xue, "Character-level intrusion detection based on convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2018, pp. 1–8.

[44] Z. Li, Z. Qin, K. Huang, X. Yang, and S. Ye, "Intrusion detection using convolutional neural networks for representation learning," in *Neural Information Processing*. Cham, Switzerland: Springer Int., 2017, pp. 858–866.

[45] K. Doshi, Y. Yilmaz, and S. Uludag, "Timely detection and mitigation of stealthy DDoS attacks via IoT networks," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 5, pp. 2164–2176, Sep./Oct. 2021.

[46] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, and M. Ghogho, "Deep learning approach for network intrusion detection in software defined networking," in *Proc. Int. Conf. Wireless Netw. Mobile Commun. (WINCOM)*, 2016, pp. 258–263.

[47] K. Wu, Z. Chen, and W. Li, "A novel intrusion detection model for a massive network using convolutional neural networks," *IEEE Access*, vol. 6, pp. 50850–50859, 2018.

[48] M. A. Ambusaidi, X. He, P. Nanda, and Z. Tan, "Building an intrusion detection system using a filter-based feature selection algorithm," *IEEE Trans. Comput.*, vol. 65, no. 10, pp. 2986–2998, Oct. 2016.

[49] K. Singh, S. C. Guntuku, A. Thakur, and C. Hota, "Big data analytics framework for peer-to-peer botnet detection using random forests," *Inf. Sci.*, vol. 278, pp. 488–497, Sep. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025514003570

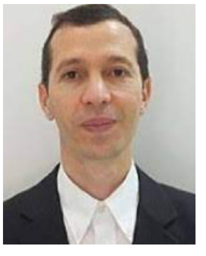
[50] N. Moustafa, G. Misra, and J. Slay, "Generalized outlier Gaussian mixture technique based on automated association features for simulating and detecting Web application attacks," *IEEE Trans. Sustain. Comput.*, vol. 6, no. 2, pp. 245–256, Apr.–Jun. 2021.

[51] N. Moustafa, E. Adi, B. Turnbull, and J. Hu, "A new threat intelligence scheme for safeguarding industry 4.0 systems," *IEEE Access*, vol. 6, pp. 32910–32924, 2018.

[52] "SHAP (SHapley Additive exPlanations) Library." 2018. [Online]. Available: https://shap.readthedocs.io/en/latest/index.html

ZAKARIA ABOU EL HOUDA (Member, IEEE) received the B.Eng. degree in computer science from the National School of Applied sciences, Marrakech, Morocco, the M.Sc. degree in computer networks from Paul Sabatier University, Toulouse, France, in 2016 and 2017, respectively, the first Ph.D. degree in computer science from the University of Montreal, Canada, and the second Ph.D. degree in computer engineering from the University of Technology of Troyes, France, in 2021. His current research interests include applied machine/deep learning for intrusion detection systems, security of distributed machine learning, and blockchain for network security.

BOUZIANE BRIK (Member, IEEE) received the Engineering degree (First Class) in computer science and the Magister degree from the University of Laghouat, Algeria, in 2010 and 2013, respectively, and the Ph.D. degree from the University of Laghouat, France, and the University of La Rochelle, France, in 2017. He is currently working as an Associate Professor with Burgundy (Bourgogne) University and DRIVE Laboratory. Before joining Burgundy University, he was a Postdoctoral Fellow with the University of Troyes, CESI School, and Eurecom School. He has been working on network slicing in the context of H2020 European projects on 5G, including MonB5G and 5GDrones. His research interests also include the Internet of Things (IoT), the IoT in industrial systems, smart grid, and vehicular networks. He also acted or still acts as a Reviewer of many IFIP, ACM, and IEEE conferences and journals, such as the IEEE transactions, letters, and magazines.



LYES KHOUKHI (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Sherbrooke, Canada, in 2006. From 2007 to 2008, he was a Postdoctoral Researcher with the Department of Computer Science and Operations Research, University of Montreal. Since September 2020, he has been a Full Professor with the ENSICAEN, GREYC Laboratory, Caen, France. Previously, he was an Associate Professor with the University of Technology of Troyes. His current research interests include network security, blockchain, and advanced network.