Computer Science • 24(1) 2023 https://doi.org/10.7494/csci.2023.24.1.4551

Sagar Dhanraj Pande
Aditya Khamparia

EXPLAINABLE DEEP NEURAL NETWORK-BASED ANALYSIS ON INTRUSION-DETECTION SYSTEMS

Abstract *The research on intrusion-detection systems (IDSs) has been increasing in recent years. Particularly, this research widely utilizes machine-learning concepts, and it has proven that these concepts are effective with IDSs – particularly, deep neural network-based models have enhanced the rates of the detection of IDSs. In the same instance, these models are turning out to be very complex, and users are unable to track down explanations for the decisions that are made; this indicates the necessity of identifying the explanations behind those decisions to ensure the interpretability of the framed model. In this aspect, this article deals with a proposed model that can explain the obtained predictions. The proposed framework is a combination of a conventional IDS with the aid of a deep neural network and the interpretability of the model predictions. The proposed model utilizes Shapley additive explanations (SHAPs) that mixes the local explainability as well as the global explainability for the enhancement of interpretations in the case of IDS. The proposed model was implemented by using popular data sets (NSL-KDD and UNSW-NB15), and the performance of the framework was evaluated by using their accuracy. The framework achieved accuracy levels of 99.99 and 99.96%, respectively. The proposed framework can identify the top-4 features using local explainability and the top-20 features using global explainability.*

Keywords IDS, deep neural network, explainable AI, NSL-KDD, local explainability, global explainability

Citation Computer Science 24(1) 2023: 97–111

1. Introduction

In the domain of IDS, there are a few issues – particularly with system reliability. Cybersecurity specialists now usually settle on IDS guidelines, so the forecasts of the system should be comprehensible. Subsequently, their growing sophistication is a considerable disadvantage given the imposing accuracy levels that are obtained by such systems; they cannot include details about why they make decisions – especially because DNNs are also employed as black-box implementations. Consequently, some details about the causes that underly IDS forecasts must be given, and some clarification regarding the intrusions that are found by cybersecurity professionals must be presented. There are a few studies that describe the new developments in IDSs, yet most of these studies have no effective platform in theory. In this article, a system that is focused on Shapley additive explanations (SHAPs) is proposed in order to overcome these drawbacks and provide a clearer interpretation of IDSs. SHAP has a good theoretical basis for either shallow- or deep-trained models. This structure will not offer some IDS interpretations. This system presents local and international interpretations to enhance the generalizability of all IDSs. Local descriptions in this sense may include knowledge that each function value decrements or increments the anticipated likelihood. In this context, there are two types of global interpretations. Most of all, the second will analyze interactions between the importance of functions and particular forms of threats by extracting essential attributes from each IDS. The NSL-KDD data set was utilized to illustrate the validity of this research method.

The significant contributions of this article can be mentioned as follows:

1) The framework recommends a structure to explain any IDS locally as well as globally. This system leads to a greater comprehension of IDS forecasts and eventually aims to create confidence in IDSs for cyber-users. This system also allows cybersecurity professionals to better recognize cyber assaults (for example, identifying the common aspects of individual threats).

2) The proposed framework is distinctive in the domain of IDSs due to the utilization of SHAP methodologies for the visibility of the IDS enhancement. When compared with other methods, the SHAP methodology has a stronger theoretical basis.

3) The major disparities are discussed among the one-vs.-all classification models and the multi-class classification models. When two kinds of classification models equate the interpretations of similar threats, security professionals can refine the IDS frameworks. The configured framework will then improve the confidence of system interaction in intrusion-detect systems.

This paper is organized into various sections: Section II discusses the related work based on IDSs and threats; Section III discusses the methodologies that are related to explainable AI, deep neural networks, and the proposed model; Section IV discusses the proposed framework that is utilized for enhancing the comprehensibility of any

IDS and illustrates the particulars of the working aspects of the system; Section V discusses the results that were obtained from the implementation of the proposed framework; and Section VI discusses the future aspects as well as the conclusions based on the proposed framework.

2. Related work

As the topic that we have considered is new and unique in itself, few researchers have explored this area using the machine-learning techniques that are discussed in detail below.

In 2018, Holzinger [9] reviewed the process of transitioning from machine-learning concepts to explainable AI in the healthcare domain. The disadvantages of machine learning raised the concept of explainable AI. Machine-learning models act like a black-box kind of working, as they generate predictions but no explanations for them, and explainable AI acts like a glass-box model as it generates predictions along with explanations for them. In 2020, Ignatiev [3] researched the various challenges that were related to explainable AI; this was represented by XAI. An overview of the advancement of explainable AI is based on a rigorous logic-based methodology.

Wang et al. [27] introduced and implemented a framework that was based on explainable AI in the domain of IDS. This framework was implemented by utilizing the NSL-KDD data set and the SHAP values that were obtained through this model, which led to the explainability of the predictions that were obtained. In 2019, Margaliot and Marcelloni [6] explained the mechanisms of a fuzzy system to develop a system that attains explainable AI. In this explanation of building such a system, four w's were considered: why is explainable AI utilized, when is explainable AI utilized, what is explainable AI, and to where does explainable AI lead.

In 2019, Wang et al. [26] developed and explained the theoretical directing designing of a model based on explainable AI for a user-centered system. In 2020, Scott Lundberg et al. [11] contributed many aspects such as obtaining effective explanations depending on the domain of game theory (a novel kind of explanation that evaluates local attribute interactions) and a modern collection of methods that integrate several local explanations within each estimation to consider the global model structure. In 2018, Yalei Ding and Yuqing Zhai [5] implemented the concept of IDS that was based on a convolutional neural network (CNN) that was based on the NSL-KDD data set. The performance of this system was compared with the various existing machine-learning models such as the random forest, SVM, DBF, and LSTM models. The proposed model improved the efficiency of IDS.

In 2019, Sandeep et al. [7] proposed a custom design model for IDS. It estimated by using machine-learning methodologies such as an autoencoder for the learning of features through unsupervised learning, and logistic classification was utilized for classifying various threats with the aid of the NSL-KDD data set. The proposed framework was evaluated by utilizing various evaluation metrics such as precision, accuracy, and recall; the outcomes look promising.

In 2018, Shone et al. [23] proposed and implemented IDS based on deep-learning concepts. It was modeled using a non-symmetric auto-encoder for learning attributes through unsupervised learning, and it was implemented using the NSL-KDD data set.

In 2018, Danijela et al. [19] reviewed the various data sets that were based on IDS; e.g., KDD CUP 99, NSL-KDD, and Kyoto 2006+. These were discussed in the aspects of various attributes and duplicates in the data set along with categories such as a normal, attack, and unknown.

In 2018, Amarasinghe et al. [1] designed a framework for identifying denial-of-service threats using IDS. The framework was developed based on deep learning along with explanations that were related to why and what a glitch was as well as its confidence. It was implemented by utilizing the NSL-KDD data set; an accuracy of about 97% was obtained through this framework.

In 2020, Barredo et al. [2] discussed the various concepts that were part of explainable AI and their related taxonomies. In addition to these concepts, the chances and confrontations were part of the AI that was responsible for explainable AI.

In 2020, Pande et al. [17] proposed a framework for identifying and classifying DDoS threats that were part of the NSL-KDD data set with the aid of the WEKA tool. The framework was designed based on the random forest methodology and attained an accuracy of 99.76%.

In 2018, Hajimirzaei et al. [8] framed a new framework that was related to an intrusion-detection system. It was designed by utilizing methodologies such as artificial bee colony and multilayer perceptron. The artificial bee colony methodology was utilized for the training aspect, and multilayer perceptron was utilized for classifying the threats by using the NSL-KDD data set and Cloudsim simulator. The evaluation metrics that were utilized for evaluating the proposed framework were MAE, RMSE, and the Kappa statistic.

In 2018, Donghwoon et al. [10] compared a proposed framework that was based on various variations of CNN models for studying the performance across the variations of the CNN models depending on its depth. The proposed framework was implemented by utilizing various kinds of CNN model. This study showed that the performance of the model enhancement did not identify the depth of the CNN improvement. The CNN-based models outperformed the variational auto-encoder, and these models were not very effective when compared to the long short-term memory methodology.

In 2018, Rajesh and Deepa [25] surveyed various methodologies that were based on machine learning while utilizing the NSL-KDD data set. Pande and Khamparia [18] discussed the various methodologies that were related to machine-learning and deep-learning methodologies.

The complete related work can be summarized as follows: the major design of the considered frameworks for the major part of the research in analyzing the DDoS threats utilized NSL-KDD as well as machine-learning and deep-learning methodologies. These models will act like a black-box model that will not provide any explana-

tion for the predicted outcome. Obtaining an explanation of the obtained predicted outcome can only occur through the explainable AI concept.

3. Methodology discussion

The major methodologies that have been utilized so far detect and classify various DDoS threats by using the NSL-KDD data set with the aid of the machine-learning methodologies of deep-learning concepts. The major drawback with this aspect is that no user will be able to identify what is the major reason for attaining a particular predicted value; hence, it is considered black-box modeling. So, this raises a situation that needs some logical reason behind that expected value from the trained model. This argument brought about a new concept called explainable AI, which is considered to be glass-box modeling due to the logical reasoning behind the predicted value.

3.1. Explainable AI

As mentioned earlier, explainable AI generates the interpretability of a model. Interpretability can be categorized into two classes: locally concentrated interpretation, and globally concentrated interpretation. Locally concentrated interpretation can explain the logical reason for an obtained output for a corresponding input that is given to a model. Globally concentrated interpretation can understand the structure of a model by looking at its overall structure. The concept of SHAP [12] plays a vital role in enhancing the interpretability of IDSs. This concept of a methodology with locally concentrated and globally concentrated interpretations in the same instance has strong theoretical and mathematical support when compared with other methodologies. The concept of SHAP [24] links the concept of LIME [21] and Shapley values [14]. LIME (local interpretable model-agnostic explanation) [21] concentrates more on learning a local replacement model in order to evaluate its individual forecasts. LIME produces a new modified data set that is composed of permutated samples and also determines the accompanying forecasts of the black-box model; then, an interpretable model will be trained on the new modified data. Certain machine-learning techniques such as linear regression, logistic regression, decision tree, and random forest are utilized as interpretable models. A good local solution to black box framework forecasts should be a local surrogate framework; this can be evaluated and represented as follows

$$\Psi(a) = \begin{pmatrix} argmin \\ j \epsilon J \end{pmatrix} \left\{ \zeta(h, j, k^a) + \Phi(j) \right\} \tag{1}$$

In the notation of Equation (1), j signifies the model of explanation for a sample of a, J signifies the possible set of explanations, $\zeta()$ signifies the loss function, h represents the original model, and k^a signifies the weight aspect between the sampled and original data. If the correlation between the sampled and original data is higher, this indicates that the weight will also be higher (and vice versa), and $\Phi(j)$ signifies

the complexity of function j. As per Equation (1), the LIME model trains interpretable local surrogate framework j on the newly obtained data set by decrementing the loss function and then explores the prediction of a sample by interpreting local framework $\Psi(a)$. Shapley explained the evaluation methodology of obtaining Shapley values [14]; this methodology is utilized in game theory to ascertain the proportion of each individual of a game. This process can be more understandable by utilizing the concepts that are related to the predictions of machine-learning methodologies. The mean offerings of an attribute value to the prediction in all possible combinations can be referred to as Shapley values.

$$\xi_i(g, y') = \sum_{x' \subseteq (y_1', y_2' \ldots, y_n') \setminus (y_i')} \frac{\mid x' \mid !(N - \mid x' \mid -1)!}{N!} \cdot [g(x' \cup y_i') - g(x')] \qquad (2)$$

In the notations in Equation (2), x' represents the subset of attributes that are utilized in the model, y' represents the attribute values that have a vector (and the instances of this are explained through y_i'), N represents the number of attributes that are considered, and $g(x')$ represents the predictions for the attribute values in x' (the evaluation of this prediction value involves masking out the i^{th} attribute). By drawing the random instances through simulation or the i^{th} attribute′s random values from the data set are selected. The three properties that abide by the Shapley values are the symmetry, dummy, and additivity properties; these properties are represented in Equations (3) through (5), respectively:

$$f(x' \cup y_i') = f(x' \cup y_j'), f(x') \subseteq (y_1', y_2', \ldots, y_n') \setminus (y_i', y_j') \qquad (3)$$

$$f(x' \cup y_i') = f(x'), f(x') \subseteq (y_1', y_2', \ldots, y_n') \setminus (y_i') \qquad (4)$$

$$f(x' \cup y_i') = f^1(x' \cup y_i') + f^2(x' \cup y_i'), then \xi_i(g, y') = \xi_i(f^1, y') + \xi_i(f^2, y') \qquad (5)$$

High computational time is essential for evaluating Shapley values due to the number of possible combinations of attribute values of 2^k. How SHAP is designed is suggested by S. Lundberg [12]. This illustrates a case x estimation by the estimation of each feature's relationship to the estimation (Fig. 1).

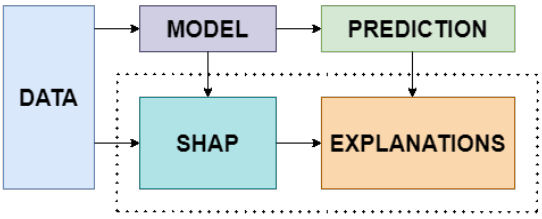

Figure 1. Generalized SHAP System

In connection with Shapley values, LIME methodologies can be viewed with the explanation that can be signified by the linear model; it links the two methodologies

(LIME as well as the Shapley values). The contribution of each feature of the model can be explained with the help the SHAP values to classify them into positive or negative classes. The main advantages of utilizing SHAP values are that these values are evaluated for any model with just an effortless linear model and that each set of data records will have its corresponding set of SHAP values. An instance of a data set can be explained using a specific SHAP value by using the following equation as mentioned:

$$f(C') = \xi_0 + \sum_{i=1}^{N} \xi_i C_i' \tag{6}$$

In the notations in Equation (6), f is known as the explanation model, $C's$ is known as the coalition vector with values of 0 and 1 for each of the instances of the data, 1 indicates that the instances in the new data set are the same as those of the original data set, 0 indicates that the instances in the new data set are different from those of the original data set, N indicates the size of the maximum coalition, and ξ_i is the feature contribution for attribute i for an instance of the data set and x_i is known as Shapley value.

3.2. Deep neural network

In this article, a deep neural network is utilized for the model to predict the various anomalies in the KDD-NSL data set for identifying DoS attacks. An input instance can be represented by X (which is of the form R^n), and each instance i of the data set that is related to an attribute can be represented as x_i; thus, the data set can be represented as $X = \{x_i\}_{i=1}^{n}$, and the corresponding labels are represented by Y. Classification function-based deep-neural-network mapping can be represented by $g : R^n \implies R^+$. Multiple layers are involved in the deep-learning network model, including the input, output, and multiple hidden layers (with multiple neurons in each of these layers). The neurons in each of these hidden layers can be activated as is represented mathematically in Equation (7):

$$h_i^{k+1} = f\left(\sum_j h_j^{(k)} w_{ji}^{(k,k+1)} + b_i^{(k+1)}\right) \tag{7}$$

In the notations in Equation (7), $h_i^{k+1}$is the activation of the $(k+1)^{th}$ layer of the i^{th} neuron, $w_{ji}^{(k,k+1)}$ is the weight of the connection between the j^{th} neuron of the k^{th} layer and the i^{th} neuron of the $(k+1)^{th}$ layers, $b_i^{(k+1)}$ is the bias of the i^{th} neuron of the $(k+1)^{th}$ layer, and $f(.)$ is the activation function. In this framework, the activation function that is utilized is ReLu; it can be represented as shown in Equation (8):

$$f(z) = Max(0, z) \tag{8}$$

The Softmax function is utilized for the output layer for the classification aspects that are related to the provided inputs; this activation function can be represented as shown in Equation (9):

$$Prob(Y = y_i \mid X) = \frac{e^{(h_i)}}{\sum_k e^{(h_k)}} \tag{9}$$

In the notations in Equation (7), h_i is the obtained value from the above-mentioned activation function $f(.)$. Depending on the predicted class classification, the SHAP values and their corresponding explanations will be generated. The whole process of explainable AI and a deep neural network is a learning model that is combined to form an explainable AI model.

3.3. Proposed model

This section discusses the proposed framework along with its flowchart implementation for obtaining the better interpretability of an IDS framework. This interpretable IDS framework is essential for any user – along with the accuracy of the framework. Consequently, an IDS framework can be generated along with transparency, which is essential at this moment in time. Figure 2 represents the flowchart of the proposed framework in this article.

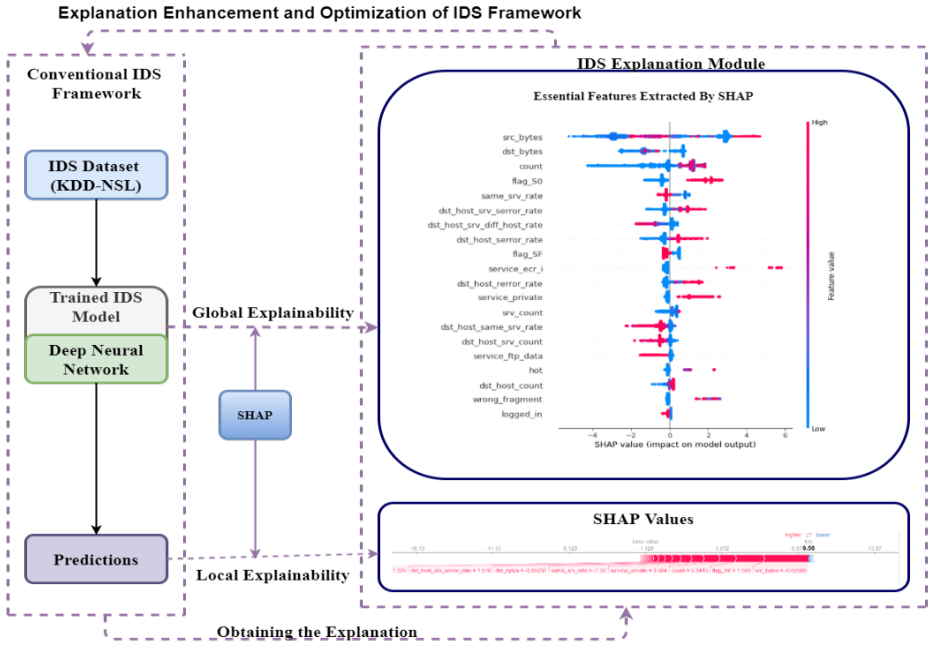

Figure 2. Proposed Framework Flowchart Overview

This flowchart can be categorized into two segments: the left segment is the conventional the IDS framework, and the right segment is utilized to obtain the interpretability that corresponds to the prediction that is obtained by the conventional IDS framework. A deep neural network model is utilized in the conventional IDS framework; this is used for training as well as prediction using the KDD-NSL data set. The predicted classifiers are utilized to compare the prediction with the explanation results, which can provide guidance as well as a reference for experts who are dealing with IDS. The concentration of the proposed work is on obtaining greater explainability for the predictions of the IDS framework. Hence, local explainability and global explainability provide a proper explanation for the obtained predictions from the IDS framework. The global explainability is generated by utilizing two methodologies; this is the first methodology to analyze the essential features of IDS. The second methodology provides the relationship between the feature values and their impacts on the obtained prediction. The local explainability generates the explanation for the output that is generated by the IDS framework, and it provides the significance of the input features for the predictions that are obtained from the IDS framework.

As mentioned in Figure 2, the proposed framework can be utilized for obtaining a considerable improvement in the IDS framework's transparency. Experts who work on this framework will be able to validate the predictions that are obtained from the IDS framework with the aid of local explainability as well as global explainability. Moreover, the deep neural network is utilized in this proposed framework. By considering the differences between the explanation and the classifiers that are obtained, experts can adjust the parameters of the model that is utilized in the IDS framework to obtain an optimized prediction and a favorable explanation.

4. Result analysis

This section mainly deals with a discussion that is related to the data set and the performance of the framed IDS. A discussion of the obtained results then follows from the perspective of the proposed model in terms of local and global explainability. This proposed framework aims to obtain an explanation that corresponds to the predicted class from IDS.

4.1. Data set discussion

The data set that was utilized for implementing the proposed framework was the NSL-KDD data set [22]. Before this data set, the KDD′99 data set was preferred for various research aspects of IDS; however, it had various aspects to worry about (e.g., unbalanced distribution, and redundancy in the data set). Due to this, NSL-KDD is preferred, as it is comparable with KDD′99 with a well-structured form. The data set is divided into training and testing data sets (KDDTrain+ and KDDTest+, respectively). The number of classes in the data set was 5, the number of attacks in the training data set was 21, and the number of attacks in the testing data set

was 37 [4,20]; this represents that the number of additional novel attacks in the testing data set was 16 [16, 28]. There were about 41 attributes for each record of the data set; the comprehensive report of the attributes are mentioned in [13, 15].

4.2. Performance evaluation

The evaluation metrics that were considered for the performance of the proposed IDS framework were accuracy, precision, recall, and F1 score. Accuracy can be defined as the ratio of those instances that were identified correctly to the complete test set. Precision can be defined as the ratio of those instances that were identified as attacks to all of the instances that were classified as attacks. Recall can be defined as the ratio of the instances that were identified as attacks to all of the instances that were of the class of attacks. The F1 score was measured by considering both the precision and recall. The proposed framework training model utilized a deep neural network with a learning rate of 0.001, epochs of 20, and a batch size of 36. This framework achieved an accuracy of 99.99%, a precision of 94.28%, a recall of 100%, and an F1 score of 97.05%.

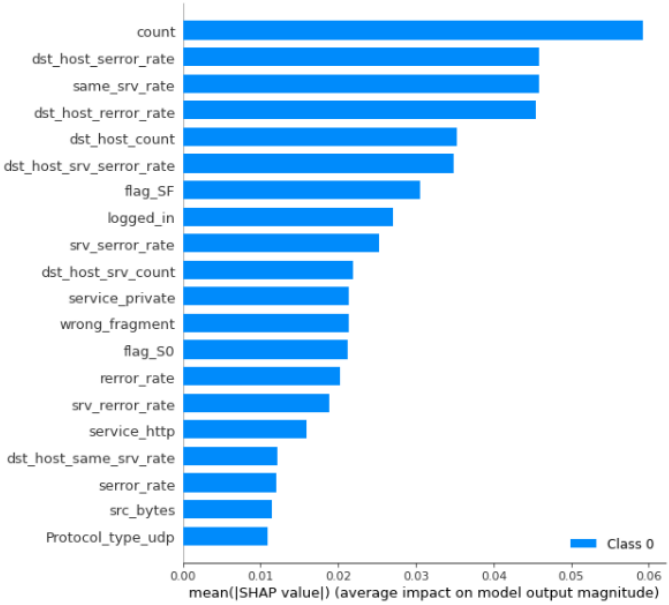

Figure 3. Distribution of SHAP Values

The summary of the SHAP values that were obtained through the proposed IDS framework can be represented as mentioned in Figure 3. The interpretation of the obtained results can be explained based on the obtained graph that can be found in Figure 4. The interpretation of the obtained results can be explained based on the

obtained graph that can be found in Figure 4 (this graph helps identify the essential features; the identified top-four essential features were *src_bytes*, *flag_S0*, count, and *service_private*).

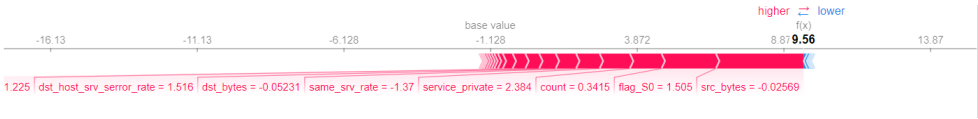

Figure 4. Interpretation of Deep Neural Network Classifier

The global explanation for the obtained results can be represented as mentioned in Figure 5 (this figure gives information regarding the top-20 essential attributes that were identified for DoS threat and their corresponding attribute values). The colors represent the attribute values from low to high depending on the Shapley values. In Figure 5, the Shapley values are considered on the x-axis, and attributes are considered on the y-axis. The attribute value increases as the red color's intensity increases; on the other hand, the attribute value decreases as the blue color's intensity increases. In the y-axis direction, the overlap points are pulsated; this represents the distribution by the function of the Shapley values. The attributes are organized accordingly.

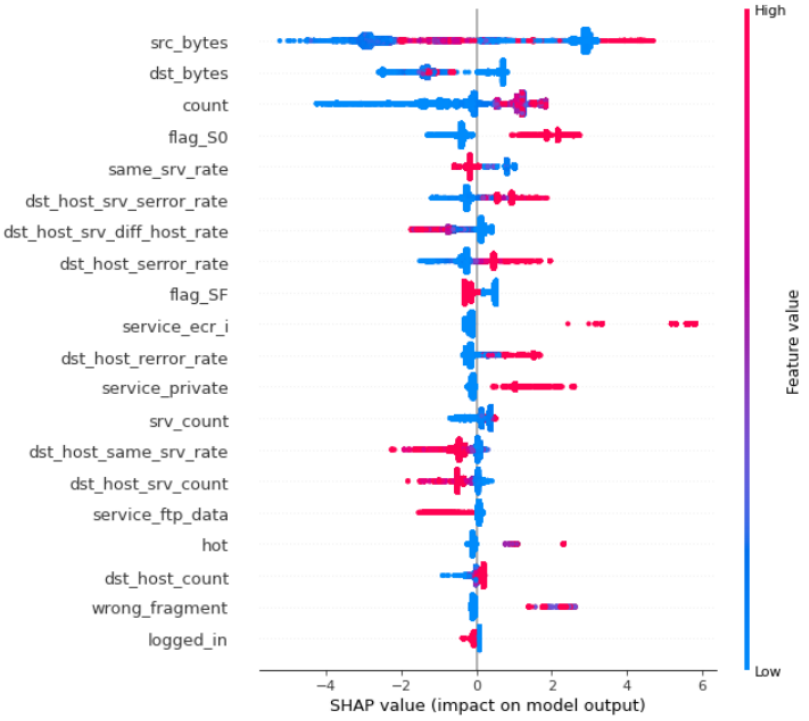

Figure 5. Top-20 attributes of DoS attack

To justify the significance of the proposed technique, the model was implemented on a real-time IOT-based UNSW 2015 data set. This data set was comprised of a sum of 175,341 rows and 45 attributes. In the data set preprocessing step, any missing values were dropped at first. Because of this, the data set was changed over to close to half of its size. Furthermore, label-encoding methods were used in order to deal with the textual values. For performing the scaling, the standardization method was later applied. For creating the effective aftereffects of the proposed model, the execution was performed with a high setup that contained an AMD RYZEN 9 processor with 8 cores, the 64-digit Windows 10 operating system, 16 GB of RAM, and a 6 GB GTX 1660 TI GPU. Specifically, four different types of attacks were available in this data set (R2L, U2L, Probe, and DDOS). In this paper, we specifically focused on the distributed denial of service attacks. In Table 1, the results that were obtained by using various existing models along with proposed model are depicted (for which the NSL-KDD data set was used), whereas Table 2 depicts the results that were obtained using the UNSW-NB-15 data set.

Table 1
Results obtained using various deep-learning algorithms on NSL-KDD data set

Sr. No.	Algorithm	Accuracy
1	XGboost Classifier	0.963
2	Adaboost Classifier	0.924
3	Extra Trees Classifier	0.927
4	Random Forest Classifier	0.978
5	**Proposed Methodology**	**0.999**

Table 2
Results obtained using various deep-learning algorithms on UNSW-NB15 data set

Sr. No.	Algorithm	Accuracy [%]
1	Convolutional Neural Network	92.71
2	Deep Neural Network	94.08
3	Gated Recurrent Neural Networks + Recurrent Neural Network	96.92
4	Artificial Neural Network	99.76
5	**Proposed Methodology**	**99.96**

5. Conclusion and future work

When researchers analyze and predict various aspects based on a considered data set but there is no reason that exists for that particular prediction to be made based on all the features of the dataset. It is a similar case while dealing with machine learning, computer vision, and natural language processing. It is essential to identify the reason for a particular prediction that enhances the explainability of a model. It is very

crucial while dealing with security-related data aspects. Considering this context, the model that was proposed based on the NSL-KDD data set enhances the explainability of the prediction. The performance of the model was evaluated by using various evaluation metrics such as accuracy, precision, recall, and F1 score. The accuracy of the proposed model is promising compared to the existing techniques. Besides the prediction, the proposed model enhanced the interpretability of the obtained predictions by using local explainability as well as global explainability; this will be helpful for experts who deal with IDS. The analysis that is currently ongoing can be strengthened. First of all, further data sets could be utilized to show the feasibility of the model for network IDSs. Second of all, while SHAP has quick calculations to explicitly translate the machine-learning model relative to the Shapley estimate, it still is not feasible to operate in real-time. Finally, more complicated threats such as advanced persistent threats (APTs) can be investigated by the SHAP process. This research gives useful insight into IDS interpretability. In future aspects, additional analysis will concentrate on testing with additional databases, operating on a system in real-time, and describing advanced threats.

References

[1] Amarasinghe K., Kenney K., Manic M.: Toward explainable deep neural network based anomaly detection. In: *2018 11th International Conference on Human System Interaction (HSI)*, pp. 311–317, IEEE, 2018.

[2] Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., García S., *et al.*: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*, vol. 58, pp. 82–115, 2020.

[3] Bessière C. (ed.): *IJCAI'20: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020.

[4] Dhanabal L., Shantharajah S.: A study on NSL-KDD dataset for intrusion detection system based on classification algorithms, *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4(6), pp. 446–452, 2015.

[5] Ding Y., Zhai Y.: Intrusion detection system for NSL-KDD dataset using convolutional neural networks. In: *CSAI'18: Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence*, pp. 81–85, 2018.

[6] Fernandez A., Herrera F., Cordon O., del Jesus M.J., Marcelloni F.: Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?, *IEEE Computational Intelligence Magazine*, vol. 14(1), pp. 69–81, 2019.

[7] Gurung S., Ghose M.K., Subedi A.: Deep learning approach on network intrusion detection system using NSL-KDD dataset, *International Journal of Computer Network and Information Security*, vol. 11(3), pp. 8–14, 2019.

[8] Hajimirzaei B., Navimipour N.J.: Intrusion detection for cloud computing using neural networks and artificial bee colony optimization algorithm, *ICT Express*, vol. 5(1), pp. 56–59, 2019.

[9] Holzinger A.: From machine learning to explainable AI. In: *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pp. 55–66, IEEE, 2018.

[10] Kwon D., Natarajan K., Suh S.C., Kim H., Kim J.: An empirical study on network anomaly detection using convolutional neural networks. In: *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pp. 1595–1598, IEEE, 2018.

[11] Lundberg S.M., Erion G., Chen H., DeGrave A., Prutkin J.M., Nair B., Katz R., Himmelfarb J., Bansal N., Lee S.I.: From local explanations to global understanding with explainable AI for trees, *Nature Machine Intelligence*, vol. 2(1), pp. 56–67, 2020.

[12] Lundberg S.M., Lee S.I.: A unified approach to interpreting model predictions. In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017.

[13] Luo Y., Cao X., Chen J., Gu J., Yu H., Sun J., Zou J.: Platelet-derived growth factor-functionalized scaffolds for the recruitment of synovial mesenchymal stem cells for osteochondral repair, *Stem Cells International*, vol. 2022, 2022.

[14] Pande S., Gadicha A.B.: Prevention mechanism on DDOS attacks by using multilevel filtering of distributed firewalls, *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3(3), pp. 1005–1008, 2015.

[15] Pande S., Kamparia A., Gupta D.: Recommendations for DDOS attack-based intrusion detection system through data analysis. In: *Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021*, pp. 899–909, Springer, 2022.

[16] Pande S., Kamparia A., Gupta D.: Recommendations for DDOS Threats Using Tableau. In: *Proceedings of Data Analytics and Management: ICDAM 2021, Volume 2*, pp. 73–84, Springer, 2022.

[17] Pande S., Khamparia A., Gupta D., Thanh D.N.H.: DDOS detection using machine learning technique. In: *Recent Studies on Computational Intelligence: Doctoral Symposium on Computational Intelligence (DoSCI 2020)*, pp. 59–68, Springer, 2021.

[18] Pande S.D., Khamparia A.: A review on detection of DDOS attack using machine learning and deep learning techniques, *Think India Journal*, vol. 22(16), pp. 2035–2043, 2019.

[19] Protić D.D.: Review of KDD Cup '99, NSL-KDD and Kyoto 2006+ datasets, *Vojnotehnički glasnik/Military Technical Courier*, vol. 66(3), pp. 580–596, 2018.

[20] Revathi S., Malathi A.: A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection, *International Journal of Engineering Research & Technology (IJERT)*, vol. 2(12), pp. 1848–1853, 2013.

[21] Ribeiro M.T., Singh S., Guestrin C.: "Why should i trust you?". Explaining the predictions of any classifier. In: *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[22] Shapley L.S.: A Value for n-Person Games. In: H.W. Kuhn, A.W. Tucker (eds.), *Contributions to the Theory of Games (AM-28),* Volume II, pp. 307–318, Princeton University Press, Princeton, 1953. doi: 10.1515/9781400881970-018.

[23] Shone N., Ngoc T.N., Phai V.D., Shi Q.: A deep learning approach to network intrusion detection, *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2(1), pp. 41–50, 2018.

[24] Tavallaee M., Bagheri E., Lu W., Ghorbani A.A.: A detailed analysis of the KDD CUP 99 data set. In: *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–6, IEEE, 2009.

[25] Thomas R., Pavithran D.: A survey of intrusion detection models based on NSL-KDD data set, *2018 Fifth HCT Information Technology Trends (ITT)*, pp. 286–291, 2018.

[26] Wang D., Yang Q., Abdul A., Lim B.Y.: Designing theory-driven user-centric explainable AI. In: *CHI'19: Proceedings of the 2019 CHI Conference on Human Factors in Computing System*, pp. 1–15, 2019.

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

[28] Yadav N., Pande S., Khamparia A., Gupta D.: Intrusion detection system on IoT with 5G network using deep learning, *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–13, 2022.

Affiliations

Sagar Dhanraj Pande
VIT-AP University, School of Computer Science and Engineering (SCOPE), Amaravati, AP, India, sagarpande30@gmail.com

Aditya Khamparia
Bhimrao Ambedkar University, Department of Computer Science, Babasaheb Satellite Center, Amethi, Tikarmafi, UP, India, aditya.khamparia88@gmail.com

Received: 13.11.2021
Revised: 26.06.2022
Accepted: 19.01.2023