Explainable AI and Random Forest Based Reliable Intrusion Detection system

Syed Wali, and Irfan A. Khan, Senior Member, *IEEE*

***Abstract*— Emerging Cyber threats with an increased dependency on vulnerable cyber-networks have jeopardized all stakeholders, making Intrusion Detection Systems (IDS) the essential network security requirement. Several IDS have been proposed in the past decade for preventing systems from cyber-attacks. Machine learning (ML) based IDS have shown remarkable performance on conventional cyber threats. However, the introduction of adversarial attacks in the cyber domain highlights the need to upgrade these IDS because conventional ML-based approaches are vulnerable to adversarial attacks. Therefore, the proposed IDS framework leverages the performance of conventional ML-based IDS and integrates it with Explainable AI (XAI) to deal with adversarial attacks. Global Explanation of AI model, extracted by SHAP (Shapley additive explanation) during the training phase of Primary Random Forest Classifier (RFC), is used to reassess the credibility of predicted outcomes. In other words, an outcome with low credibility is reassessed by secondary classifiers. This SHAP-based approach helps in filtering out all disguised malicious network traffic and can also enhance user trust by adding transparency to the decision-making process. Adversarial robustness of the proposed IDS was assessed by Hop Skip Jump Attack and CICIDS dataset, where IDS showed 98.5% and 100% accuracy, respectively. Furthermore, the performance of the proposed IDS is compared with conventional algorithms using recall, precision, accuracy, and F1-score as evaluation metrics. This comparative analysis and series of experiments endorse the credibility of the proposed scheme, depicting that the integration of XAI with conventional IDS can ensure credibility, integrity, and availability of cyber-networks.**

***Index Terms*—Adversarial attacks, Intrusion Detection System. Explainable AI, cyberattacks, Random Forest Classifier.**

I. Introduction

Technological advancement has increased the dependency on cyber-networks and internet resources, raising serious concern about the security of assets connected with this vulnerable cyber-world. Moreover, with the rapid increment in the number of devices connected to the internet, the number of reported cyberattacks is also increasing drastically, which depicts that ensuring network privacy via the installation of an Intrusion Detection System (IDS) is the utmost requirement of this digitized era as it has been proven effective against internal and external intruders [1,2]. These IDS systems are responsible for monitoring the network traffic to segregate the malicious users from legitimate ones by adopting different strategies including logical operations, data mining, statistical, and machine learning approaches [3]. Among these strategies, the most trusted ones are Machine Learning (ML) based IDS systems that primarily depend on the different types of classifiers such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest Classifier (RFC), and Artificial Neural Networks (ANN). Using any of these adopted strategies, the purpose of all IDS is to maintain confidentiality, integrity, and availability of cyber-networks [2, 4].

Undoubtedly, ML-based systems have shown outclass performance in several domains [5, 6]. However, these ML-based IDS are extremely vulnerable to adversarial attacks [7, 8]. These attacks can easily deceive conventional ML-based IDS by slightly modifying the data that leads toward incorrect predictions [9]. Such adversarial attempts have shown a significant reduction in the performance of IDS systems [10]. Therefore, modern IDS systems must consider these types of attacks during their development phase for ensuring the integrity and reliability of IDS during its critical operation of cyber-network analysis.

Another challenge associated with the ML-based IDS systems is the lack of transparency in the decision-making process. Model biasness or learning of inappropriate weights may lead to serious consequences in many applications. Therefore, feature-based impact analysis or translation of predicted outcomes in terms of features contribution can enhance the trust of stakeholders. Considering these bottlenecks and modern challenges of the cyber domain, this paper presents a novel IDS based on the combination of Random Forest Classifier and Shapley additive explanation (SHAP). It is an explainable AI (XAI) tool that is used to decode the overall model response (global explanation) and each prediction (local explanations) in terms of features contribution.

The proposed SHAP-assisted RFC framework of IDS can assess the credibility of predicted outcomes and ensure a high level of accuracy in detecting modern cyber threats. Furthermore, the adopted strategy makes the final decision after cross-validating the local explanation of predicted outcome with the global explanation of the SHAP framework, and therefore, the proposed IDS shows outclass performance on both test dataset and adversarial samples generated via Hop Skip Jump Attack (HSJA). Thus, the proposed IDS meets all essential requirements of modern cyber-networks by providing a reliable, accurate, trustworthy, transparent intrusion detection

Corresponding author: Irfan Khan. Also, both authors contributed equally and they are co-first authors.

Syed Wali is with the Electrical and Computer Engineering Department, Texas A&M university, College Station, TX 77843 USA (e-mail: syedwali@tamu.edu).

Irfan Khan is with Marine Engineering Technology Department in a joint appointment with the Electrical and Computer Engineering Department, Texas A&M university, College Station, TX 77843 USA (e-mail: irfankhan@tamu.edu).

framework having improved robustness against adversarial attacks.

A brief introduction of the proposed IDS has been presented in this section, whereas the remaining paper is organized as follows. Section II covers the background and relevant research efforts that have been made in this research domain. Section III presents the framework of the proposed IDS and relevant details of its development phase. Section IV shows the performance of the proposed IDS on the test dataset and adversarial samples. Moreover, the comparison of the proposed IDS with other classifiers is also presented in this section, whereas Section V finally concludes the research paper and highlights the potential research domain.

II. RELATED WORK

This research paper focuses on the development of IDS, having more emphasis on the adversarial robustness of IDS and the trustworthiness of explainable AI. Therefore, related work of each aspect is separately presented in this section.

A. XAI and Intrusion Detection System

During the last decade, several IDS systems have been proposed for preventing cyber networks from malicious attacks [11]. Among them, ML-based IDS showed remarkable performance due to their capability of learning millions of parameters. However, these complex models, which are commonly referred as black-box models, are uninterpretable [12], whereas transparent decision-making is the utmost requirement of every IDS. Since a single incorrect prediction of IDS makes the system vulnerable to serious cyber threats, the XAI must be incorporated in conventional IDS for increasing credibility and reliability.

Mane et al. in [13] utilized the NSL-KDD dataset and the DNN-based ML model for detecting network intrusions. In order to add transparency, they utilized five different XAI frameworks for demonstrating the behavior of the trained model. However, they did not utilize the explanation generated by any XAI framework for validating the credibility of predicted outcomes.

Sinclair et al. in [14] extracted rules using Decision Tree and Genetic Algorithm (GA) for improving the performance of the model, whereas Ojugo et al. in [15] worked on the optimization of the IDS model by utilizing GA for extraction of rules. They concluded that instead of having one optimum rule, IDS should be built using a set of rules extracted via ML algorithms. A similar concept of rule-based IDS was also adopted by Dias et al. in [16] for adding transparency in the decision-making process.

Similarly, Mahbooba et al. in [17] also worked on the explanation of each predicted outcome by extracting rules from the decision tree that was trained and evaluated on the KDD dataset. These extracted rules were only used to explain each predicted outcome and overall model response. However, they did not focus on adversarial attacks and the improvement of IDS using explanations provided by XAI tools.

B. Adversarial Robustness and Intrusion Detection System

ML-based IDS have shown outclass performance in enhancing cyber security. However, advanced ML models like Deep Neural Network decreases their capability of classifying normal or malicious network whenever subjected to adversarial attacks. Szegedy et al. in [8] developed the method of generating adversarial samples, commonly known as the Fast Gradient Sign Method (FGSM), by adding small noise to the actual dataset such that the introduced change cannot be assessed manually. As a result, the ML-based system suffers from incorrect classification due to such adversarial attacks. Similar to FGSM, several other methods were also introduced for generating such samples for evaluating the performance of ML-based systems. Researchers in the Cybersecurity domain also started exploring the impact of adversarial attacks on IDS, and developed methodologies to introduce robustness in models to cope up with this modern threat.

Yang et al. in [10] evaluated their DNN-based model on adversarial attacks. Although their model showed 89% accuracy on the actual dataset, it was unable to show promising performance on three different types of adversarial attacks. For example, the performance of their model decreased to approximately 50% on the first two types of selected attacks, Zeroth Order Optimization (ZOO) attack, and Generative Adversarial Nets (GAN) attack, while the model showed compromising the performance of nearly 70% against substitute model attacks implemented by Authors. Similarly, Peng et al. in [18] evaluated their DNN-based IDS and three other ML models against four different types of adversarial attacks. All models showed a significant reduction in accuracy during this evaluation, depicting that such attacks must be considered during the development phase of IDS.

However, most of the research on adversarial samples have been done in image classification and computer vision domain, where different types of defense mechanism are developed by researchers for ensuring the protection against such attacks. Klawikowska et al. in [19] utilized an explanation of the model for improving image classification performance under adversarial attacks. Similarly, [20] also utilized XAI for leveraging the performance and credibility of classifiers and evaluated their approach on SVHN and CIFAR-10 datasets.

Fidel et al. in [21] presented the XAI signatures-based framework for segregating the adversarial samples and normal network traffic. They evaluated their approach on datasets utilized in the image recognition domain and achieved an accuracy of around 97% in detecting the adversarial attack. This type of defense mechanism must also be introduced in cyber-networks. Therefore, this paper proposes a novel XAI-assisted IDS that checks the credibility of ML-model predictions and ensures the outclass performance under normal and adversarial attacks. Furthermore, it adds transparency in the decision-making process, enhancing user trust.

III. PROPOSED FRAMEWORK

The Intrusion Detection System (IDS) proposed in this paper utilizes the explanation provided by SHAP for leveraging the performance and credibility of IDS. There are two phases of this research- the first one is the IDS development stage which is

mainly utilized to develop Classifiers and Credibility Assessment Module (CAM). The second phase depends on the previously developed modules for examining malicious network traffic with high credibility. Furthermore, the proposed IDS development phase, as shown in Fig 1, is dependent on several data processing blocks which are separately explained in this section. In contrast, the IDS evaluation phase presented in this figure is explained in the next section

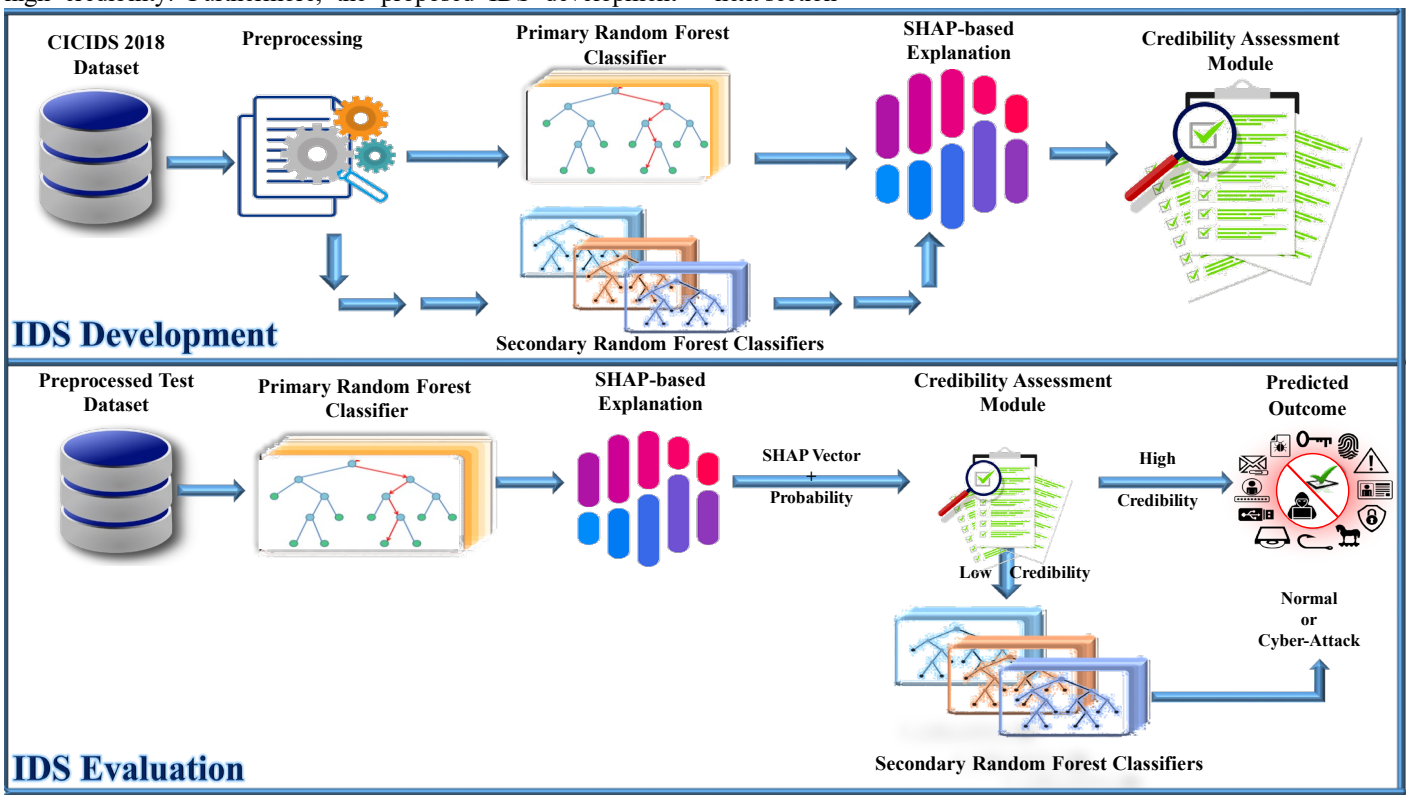


Fig. 1. Development and Evaluation Phase of the proposed IDS Framework. The IDS development phase preprocesses the CICIDS 2018 dataset and splits it into the train, validation, and test datasets. Train and validation split was utilized at this research phase for the development of classifiers and Credibility Assessment Modules. These trained modules are utilized in the IDS evaluation phase for segregating the malicious network traffic.

A. Dataset

The dataset developed by Communication Security Establishment (CSE) and Canadian Institute for Cybersecurity (CIC) in 2018, commonly known as CSE-CIC IDS 2018, has been widely utilized for developing IDS as it encompasses various modern cyberattack vectors [22]. These attack vectors are available in ten different CSV files such that each file contains normal network traffic and categories of cyberattacks. Therefore, the cumulative processing of each attack file was not possible due to the limitation of available computational power. The objective of this research was to develop an IDS capable of making transparent and credible decisions using Explainable AI (XAI), which can also be achieved by processing a few CSV files from CICIDS 2018. Therefore, three different files and following six different attack vectors, as shown in Table I, were selected for this research:

1. SSH-BruteForce: Secure Shell (SSH) is a widely adopted communication protocol for gaining remote access to a machine within a few seconds via an encrypted channel. However, an adversary can make several authentication attempts for gaining this remote access by applying all possible combinations of passwords until access is granted [23].
2. FTP-BruteForce: File Transfer Protocol (FTP) is also vulnerable to brute force attacks, resulting in the loss of valuable information and assets. These attacks can be automated easily without any need of excessive domain knowledge via available brute force attack tools such as Hydra, Rainbow Crack, and John the Ripper [24].
3. DoS attacks HULK: HTTP Unbearable Load King (HULK) attacks are critical attacks as the adversary sends a large amount of virtual traffic to the webserver in order to deprive the legitimate users of the web-portal services [25].
4. DoS attacks SlowHTTPTest: The Denial of Service attack which is specifically associated with the application layer, is known as a slow HTTP DoS attack [26]. There are several methods to create such an adversary for affecting the legitimate users and service by putting an intentional delay in response to the webserver.
5. DDOS attack-HOIC: High Orbit Ion Cannon (HOIC) was a tool developed for testing the network which floods the HTTP server using script files [27]. It can also be utilized for affecting the web server and associated services.
6. DDOS attack-LOIC-UDP: Low Orbit Ion Cannon (LOIC) is also a network testing tool developed by Praetox Technologies, capable of generating intense

network traffic for affecting the server [27]. It can also easily affect the performance of web servers.

TABLE I
DATASET DESCRIPTION

File Name	Network Traffic	Total Count
02-14-2018	Benign FTP-BruteForce SSH-Bruteforce	667626 193360 187589
02-16-2018	DoS attacks-Hulk Benign DoS attacks-SlowHTTPTest	461912 446772 139890
02-21-2018	DDOS attack-HOIC Benign DDOS attack-LOIC-UDP	686012 360833 1730

B. Data Preprocessing

Data Preprocessing is extremely essential before training any ML model as it significantly impacts the performance and the associated computational time. Therefore, CICIDS 2018 dataset was passed through several data preprocessing stages, as shown in Fig 2. Each stage of this process is explained in this section:

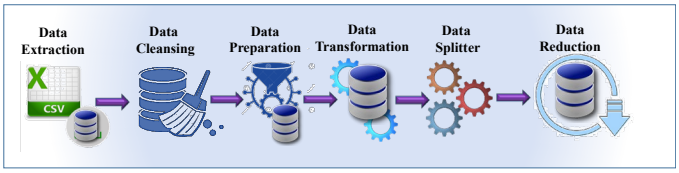


Fig. 2. CICIDS 2018 Dataset Preprocessing Phase. This research phase splits the cleansed dataset into a train-test set with a ratio of 50:50 and a train-validation set with a ratio of 90:10, followed by the dimensionality reduction step.

The first step is the data extraction stage, in which all types of network traffic is extracted from CSV files as Pandas data frame. Then these data frames are passed through the data cleansing stage in which all erroneous data instances having either infinite values or NAN were removed. Moreover, duplicate rows from data were also removed at this stage. This cleansed data is then passed through the data preparation stage in which string labels were converted into numerical values, and the timestamp column was removed so that classifier can decide on behalf of network traffic. After this processing stage, data enters into the transformation stage, where data in each column is normalized. This normalized data is then passed through the data splitter that splits the data into 50:50 train-test ratio and 90:10 train-validation set ratio. After that, the random forest classifier was initialized with default parameters was trained on the training dataset for reducing relatively less important features. Out of 78 features, the top 40 were selected with respect to their relative importance, and the highest priority ones are shown in Fig 3.

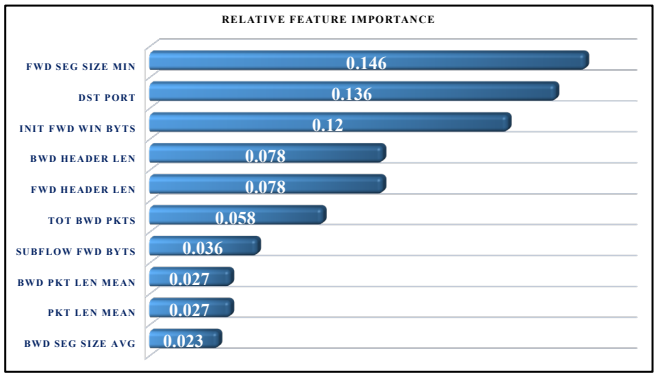


Fig. 3. Relative Feature Importance of CICIDS 2018 dataset.

C. Random Forest Classifier

Random Forest Classifier (RFC) is an ensemble prediction technique that has been proven efficient in several classification and regression problems because it makes a final prediction based on several decision trees [28]. Moreover, a random selection of data nodes for constructing the decision tree improves the overall performance of the classifier. This classification performance mainly depends on the total number of leaves and trees such that for L number of leaves, the decision tree splits the feature space into L regions denoted as R_L. This feature space is utilized for predicting the final output of a decision tree, which can be mathematically represented as (1) and (2), where the final predicted outcome depends on the majority votes of all trees. Therefore, the total number of leaves and trees are the two most important hyper-parameters of RFC [29] as their optimum selection during the training phase reflects in terms of better performance during evaluation. These parameters must be carefully decided because an excess increase in their values after a certain number results in increased computational complexity [30].

$$f(x) = \sum_{i=1}^{L} constant_L * \prod(x, R_L) \tag{1}$$

$$\prod(x, R_L) = \begin{cases} 1 & if \quad x \in R_L \\ 0 & otherwise \end{cases} \tag{2}$$

The training Phase of RFC deployed in the proposed IDS was carried out in two stages. In the first phase primary RFC, responsible for identifying all types of attacks in the dataset, was developed using a rigorous random search method for selecting the optimum hyper-parameters. 3-Fold cross-validation was utilized to resolve overfitting issues of the training phase. Results obtained via the randomized search of hyper-parameters, as shown in Fig 4, represent that log2 selection criteria of parameters and 40 trees are an optimum classification choice.

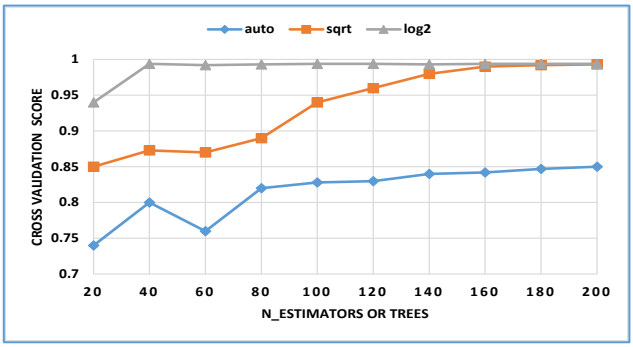


Fig. 4. Randomized Search of RFC Hyperparameter. Three different features selection criteria (auto, sqrt, and log2) were explored on the different number of trees. This exploration revealed that the RFC model shows improved performance with the logarithmic feature selection criteria and 40 trees.

In the second phase, the training dataset was further split into the following three different categories, which were utilized for training secondary Random Forest Classifiers:

1. Category A: Dataset containing Bengin traffic, FTP-BruteForce, and SSH-Bruteforce attacks
2. Category B: Dataset containing Bengin traffic, DoS-Hulk, and DoS-SlowHTTPTest attacks
3. Category C: Dataset containing Bengin traffic, DoS-HOIC, and DOS-LOIC-UDP attacks

Each Secondary Random Forest Classifier was initialized with default parameters of the Scikit-learn library [31] and successfully achieved 99% accuracy on out-of-bag samples. Since each Secondary RFC had to deal with three classes, they achieved better classification performance with few trees as compared to Primary RFC.

D. SHAP-based Explanation

The proposed IDS system utilizes Random Forest as primary and secondary classifiers because tree-based models perform better than neural networks on tabular data in many applications [32]. Moreover, tree-based models are intrinsically interpretable than deep learning models, and this interpretation of predicted outcomes is extremely essential for building user trust and improving the performance of AI-based systems. In order to extract the local and global explanation of the proposed IDS, the concept proposed by Shapley in [33] is adopted, which is purely based on the game theory approach. Conventionally this theory comprises game and players, but in this case of explainable AI, the game is to reproduce the predicted outcome of the pre-trained model, and features of the dataset are players. Therefore, the SHAP-based explanation of AI models measures the impact of all features in the dataset and helps in determining their positive and negative contributions.

The proposed IDS utilizes the Tree Explainer module of SHAP that provides fast and consistent explanations by reducing the complexity of SHAP calculations [32]. During the IDS development phase, the training dataset was passed through the SHAP explainer for extracting Shapley values, and their corresponding summary plots are shown in Fig 5, depicting the behavior of the top 20 key players of each attack class. Furthermore, colors represent the magnitude of features, and the x-axis represents their positive or negative impact on the identification of each individual class. These extracted features and corresponding Shapley values can assess the credibility of predicted outcome.

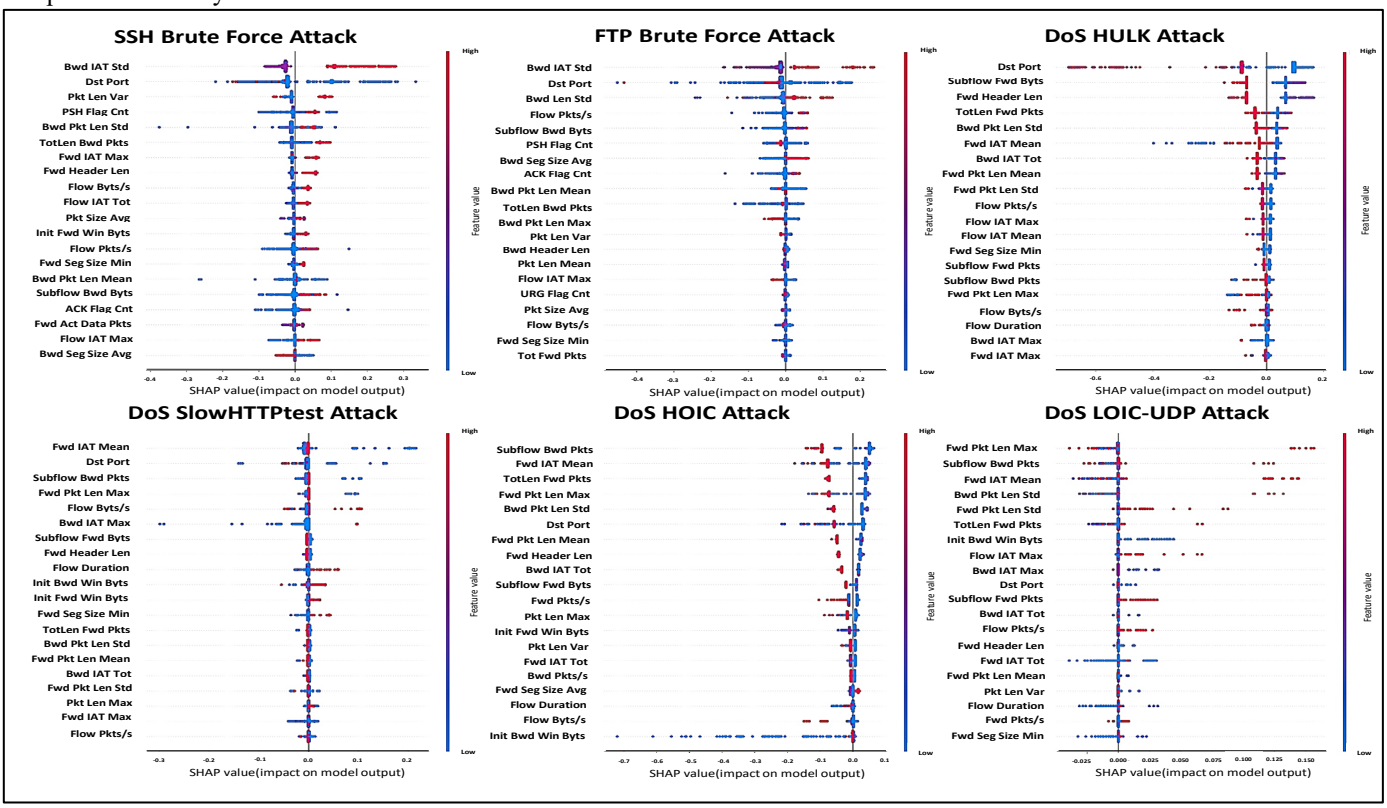


Fig. 5. SHAP Global Explanation of SSH BruteForce attack, FTP BruteForce attack, DoS HULK attack, DoS SlowHTTP attack, DoS HOIC attack and DoS LOIC-UDP attack. Top 20 features of each summary plot depict the overall behavior of the model in identifying different types of cyberattacks. These extracted Shapley values from the training dataset can be utilized for the credibility assessment of predicted outcomes.

E. Credibility Assessment Module

The Credibility Assessment Module (CAM) is one of the most important parts of the proposed IDS, ensuring the credibility and transparency of predicted outcomes. The development of the CAM module requires two inputs from previous modules of the proposed framework. The first one is the predicted probability vector and the second is the array of Shapley values extracted from the SHAP module. For simplicity, the algorithm utilized in the CAM module is shown in Fig 6, depicting that the correlation was used for measuring the confidence in the predicted outcome. This confidence array flags the predicted outcome whenever the correlation of predicted Shapley values is not stronger than the correlation of the second high probability outcome. Furthermore, three different correlation methods, including Pearson, Spearman, and Kendall, were investigated for measuring the confidence array on the validation dataset. The result of their corresponding confidence array and accuracies is shown in Table II, representing that the Pearson correlation is the optimum choice for assessing the predicted outcome as the flagged outputs are only 2.3%. Reassessment of these flagged outputs via secondary classifiers can further improve the performance of IDS.

Credibility Assessment Module Pseudocode

```
Class0_Shap ←— Training Dataset Bengin Shapley Values
Class1_Shap ←— Training Dataset SSH BruteForce Shapley Values
Class2_Shap ←— Training Dataset FTP BruteForce Shapley Values
Class3_Shap ←— Training Dataset DoS HULK Shapley Values
Class4_Shap ←— Training Dataset DoS SlowHTTPtest Shapley Values
Class5_Shap ←— Training Dataset DoS HOIC Shapley Values
Class6_Shap ←— Training Dataset DoS LOIC Shapley Values
Classes_Shap ←— [Class0_Shap,Class1_Shap … Class6_Shap]
X_Validation ←— Validation Dataset Features
Y_Validation ←— Validation Dataset Labels
Initialize Confidence Array
Initialize CAM_Prediction Array

For x in X_Validation:
    y_pred ←———— RandomForest Predict Function on x
    y_shap ←———— Store Predicted class Shapley Values
    P_vect ←———— RandomForest Probability outputs
    C_2nd ←———— Second Highest Probability Class from P_vect
    y_shap2←———— Second Highest Probability Class Shapley values

    Coef1 ←———— Calculate Correlation between y_shap and Classes_Shap[y_pred]
    Coef2 ←———— Calculate Correlation between y_shap2 and Classes_Shap[C_2nd]
    if (Coef1>Coef2):
        Confidence.append(1);
        CAM_Prediction.append(y_pred);
    else:
        Confidence.append(0);
        CAM_Prediction.append(C_2nd);
end
```

Fig. 6. Pseudocode of Algorithm utilized in Credibility Assessment Module

TABLE II
CREDIBILITY ASSESSMENT MODULE PERFORMANCE ON THE VALIDATION DATASET

Correlation Method	Confident Data Instances	Flagged Data Instances	CAM Accuracy
Pearson	97.3%	2.7%	96.5%
Spearman	78.5%	21.5%	83.7%
Kendall	64.2%	35.8%	74.3%

IV. EXPERIMENTS AND RESULTS

Experimental evaluation of the proposed framework, as shown in Fig 1, depends on the pre-trained classifiers and CAM module. Whenever the predicted outcome is flagged by the CAM module, it is passed through secondary classifiers for further evaluation. Among three classifiers, one or two secondary-classifiers will evaluate the suspicious outcome based on the results of the primary RFC and CAM module. For instance, if primary RFC predicts FTP BruteForce class and CAM module predicts SSH BruteForce class as an output, then only secondary classifier A will reassess the data because both classes belong to this category. However, if the results of the CAM module and primary RFC belong to the two different categories of secondary classifiers then both will reassess the data and the final prediction will be based on the highest predicted probability.

In order to assess the proposed framework, performance evaluation was carried out on the test dataset and adversarial samples. Moreover, they were also compared with two renowned classifiers. Results of this evaluation phase are described in this section.

A. Performance Evaluation on Test Dataset

Test dataset was first passed through the primary RFC for detecting the intrusions in the system, and its resulting confusion matrix is shown in Fig 7. This confusion matrix depicts high accuracy in five classes. However, FTP BruteForce class and DoS-SlowHTTP class results are not reliable. These results of RFC are passed through a SHAP-based explainer for producing corresponding Shapley values. Then CAM module assesses the credibility of predicted outcomes using SHAP and RFC results. Consequently, it flags only 4.2% data instances of the complete test dataset and sends them for reassessment. The final confusion matrix after reassessment is also shown in Fig 7, depicting the improved performance of IDS in all classes with an overall accuracy of 100%.

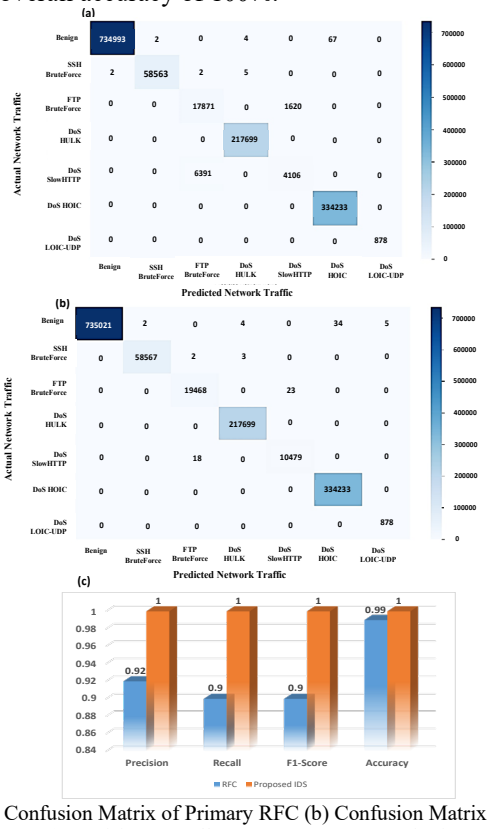


Fig. 7. (a) Confusion Matrix of Primary RFC (b) Confusion Matrix of Proposed IDS (c) Precision, Recall, F1-Score, Accuracy of Primary RFC and Proposed IDS on Test Dataset

B. Comparison with other Classifiers

The effectiveness of the proposed IDS was measured using precision, recall, f1-score, and accuracy. These four evaluation metrics, adopted for measuring the performance on the CICIDS dataset, can be mathematically represented as (3), (4), (5), and (6). These metrics were calculated on the predictions made by baseline Primary RFC, Proposed IDS, K-Nearest Neighbors, and Support Vector Machine. Comparative analysis, as shown in Table III represents that the proposed IDS outperforms all the classifiers.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (6)$$

Here TP is Total Positive, TN is Total Negative, FP is False Positive and FN is False Negative.

TABLE III
PERFORMANCE EVALUATION OF KNN, SVM, RFC, AND PROPOSED IDS ON TEST DATASET

Network Traffic	KNN				SVM				RFC				Proposed IDS			
	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
Benign	0.99	1	1	1	0.999	1	1	1	0.999	1	1	1	1	1	1	1
SSH Brute Force	1	1	1	1	0.999	1	1	1	1	1	1	1	1	1	1	1
FTP Brute Force	0.99	0.78	0.77	0.78	0.957	0.67	0.96	0.79	0.99	0.74	0.92	0.82	1	1	1	1
DoS HULK	0.99	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DoS SlowHTTP	0.99	0.59	0.61	0.6	0.1077	0.58	0.11	0.18	0.99	0.72	0.39	0.51	1	1	1	1
DoS HOIC	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DoS LOIC-UDP	1	0.99	1	1	0.992	1	0.99	1	1	1	1	1	1	0.99	1	1

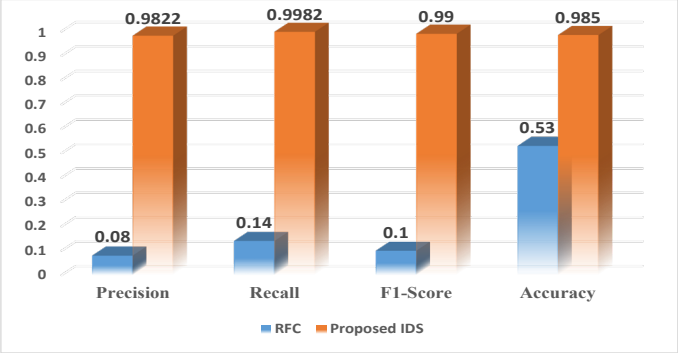
C. Performance Evaluation on Adversarial Attacks

In cyber-networks, the process of bypassing a security device for any malicious purpose, such that the attack remains unidentified until an attacker achieves their objective, is termed as evasion. Adversarial attacks come under this category, where malicious samples try to disguise as legitimate network traffic and become a serious threat for cyber-networks and interconnected devices. Therefore, security devices must have the capability to deal with such modern threats and should be evaluated on such attacks before deployment.

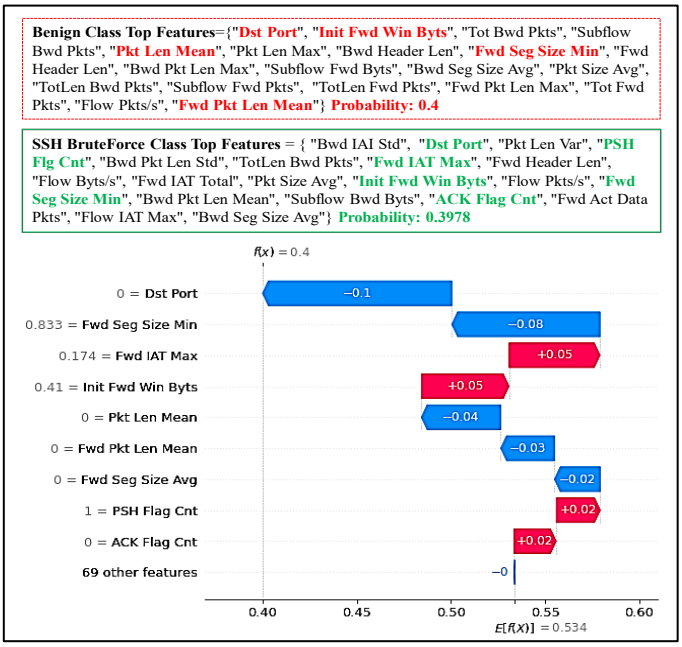
There are two types of adversarial attacks. One is white-box attacks, where an attacker has complete information about the implemented model, including its weight and structural parameters, which is usually rare. In contrast, the second one is known as a black-box attack, where an attacker has limited information and is more practical. Based on available information, black-box attacks are further categorized into score-based attacks and decision-based attacks. In score-based attacks, an attacker has access to the output layer, whereas decision-based attacks only require predicted labels of an ML model.

Since decision-based attacks are practically more viable, an attacker can easily target an IDS using such adversarial attempts. In order to assess the performance of proposed IDS on adversarial attacks, an advanced version of black-box attacks by [34] was utilized. One thousand HSJA samples were generated via the Adversarial Robustness Toolbox [35] and were passed through the primary RFC model.

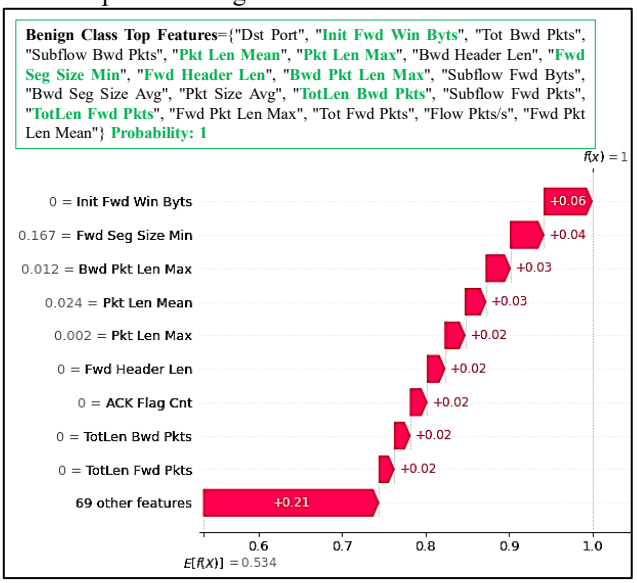
Although the primary RFC model depicted reliable performance on the test dataset, the same model failed to detect attack classes in thousand adversarial samples because each attack was successfully disguised as benign network traffic. However, when the proposed IDS correlates the Shapely results of each individual adversarial sample (local explanation) with the Shapley values of the second high probability class, it flags the predicted outcome and sends for reassessment using secondary RFCs, this reassessment successfully filters out the malicious network traffic. Macro-Average accuracy, precision, recall, and F1 score of proposed IDS and Primary RFC on thousand adversarial samples is shown in Fig. 8, proving that the proposed IDS has intrinsic adversarial robustness and does not require any additional training for detecting such attacks.



Moreover, top-twenty features extracted from the global explanation of the SHAP framework during the IDS development phase may also assess predicted class credibility. Fig. 9 shows the SHAP waterfall plot of an instance where the primary RFC classifier incorrectly classifies the SSH Brute Force attack as benign network traffic and Shapely features of two high probability classes. Since more waterfall plot features appear in the class other than the predicted one, the predicted outcome has low credibility. This manual assessment of feature space can be beneficial, but Shapley values correlation is a more reliable approach for filtering out this disguised attack.



The proposed IDS also helps in the validation of correctly predicted outcomes, as shown in Fig. 10, where the predicted outcome is benign network traffic with 100% probability and almost all features of the waterfall plot are present in the SHAP features space of benign class.



V. Conclusion

There is no doubt that technological development has transformed the concept of globalization into reality. However, this increased dependency on the vulnerable cyber-world has brought several challenges. Among them, cyber-attacks are a serious security concern. Moreover, the utilization of modern AI-based techniques by adversaries for their malicious purposes has highlighted the need for intelligent IDS systems. Considering this important aspect, this paper presented an IDS capable of identifying all types of malicious content in network traffic by utilizing the global explanations developed via the SHAP framework. Furthermore, this IDS presents the transparent decision-making approach by assessing model explanations developed during the development and evaluation phase for enhancing user trust and maintaining operational integrity. Moreover, the presented research work exhibited 98.5% and 100% accurate results on the adversarial samples and test dataset, respectively. It depicts that the promising research domain of integrated XAI and ML-based systems must be explored further by analyzing more adversarial attacks and other AI models for devising more robust and reliable IDS.

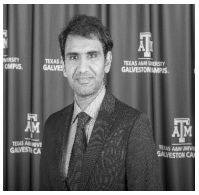
References

[1] H. Sedjelmaci, S. M. Senouci, and N. Ansari, "A hierarchical detection and response system to enhance security against lethal cyber-attacks in UAV networks," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 48, no. 9, pp. 1594–1606, 2018.
[2] B. Hu, C. Zhou, Y.-C. Tian, Y. Qin, and X. Junping, "A collaborative intrusion detection approach using blockchain for Multimicrogrid Systems," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 49, no. 8, pp. 1720–1730, 2019.
[3] A. F. AlEroud and G. Karabatis, "Queryable semantics to detect cyber-attacks: A flow-based detection approach," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 48, no. 2, pp. 207–223, 2018.
[4] V. Mohammadi, A. M. Rahmani, A. M. Darwesh, and A. Sahafi, "Trust-based recommendation systems in Internet of Things: a systematic literature review," Human-centric Computing and Information Sciences, vol. 9, article no. 21, 2019. https://doi.org/10.1186/s13673-019-0183-8
[5] A. Thakkar and R. Lohiya, "A review on machine learning and deep learning perspectives of IDS for IoT: recent updates, security issues, and challenges," Archives of Computational Methods in Engineering, vol. 28, no. 4, pp. 3211-3243, 2021.
[6] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," IEEE Access, vol. 6, pp. 33789-33795, 2018.
[7] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: reliable attacks against black-box machine learning models," 2017 [Online]. Available: https://arxiv.org/abs/1712.04248.
[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016, pp. 2818-2826.
[9] H. Xu, Y. Ma, H. C. Liu, D. Deb, H. Liu, J. L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: a review," International Journal of Automation and Computing, vol. 17, no. 2, pp. 151-178, 2020.
[10] K. Yang, J. Liu, C. Zhang, and Y. Fang, "Adversarial examples against the deep learning based network intrusion detection systems," in Proceedings of 2018 IEEE Military Communications Conference (MILCOM), Los Angeles, CA, 2018, pp. 559-564.
[11]A. A. Salih and A. M. Abdulazeez, ''Evaluation of classification algorithms for intrusion detection system: A review,'' J. Soft Comput. Data Mining, vol. 2, no. 1, pp. 31–40, Apr. 2021.
[12]A. Rai, "Explainable AI: from black box to glass box," Journal of the Academy of Marketing Science, vol. 48, no. 1, pp. 137-141, 2020.
[13]S. Mane and D. Rao, "Explaining network intrusion detection system using explainable ai framework," arXiv preprint arXiv: 2103.07110 , 2021.
[14]C. Sinclair, L. Pierce, and S. Matzner, "An application of machine learning to network intrusion detection," Proceedings - Annual Computer Security

Applications Conference, ACSAC, vol. Part F1334, no. 0293, pp. 371–377, 1999, doi: 10.1109/CSAC.1999.816048.
[15]A. A. Ojugo, A. O. Eboka, O. E. Okonta, R. E. Yoro, and F. O. Aghware, "Genetic Algorithm Rule-Based Intrusion Detection System (GAIDS)," Journal of Emerging Trends In Computing Information Systems, vol. 3, no. 8, pp. 1182–1194, 2012, [Online]. Available: http://www.cisjournal.org.
[16]T. Dias, N. Oliveira, N. Sousa, I. Praca, O. Sousa, " A Hybrid Approach for an Interpretable and Explainable Intrusion Detection System," arXiv preprint arXiv: :2111.10280, 2021
[17]B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model," Complexity, vol. 2021, Article ID 6634811, 11 pages, 2021.
[18]Y. Peng, J. Su, X. Shi, and B. Zhao, "Evaluating deep learning based network intrusion detection system in adversarial environment," in Proceedings of 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 2019, pp. 61-66.
[19] Z. Klawikowska, A. Mikołajczyk, and M. Grochowski, "Explainable AI for inspecting adversarial attacks on deep neural networks," in Artificial Intelligence and Soft Computing. Cham, Switzerland: Springer, 2020, pp. 134-146.
[20] O. Amosy and G. Chechik, "Using explainabilty to detect adversarial attacks," 2019 [Online]. Available: https://openreview.net/forum?id=B1xu6yStPH.
[21] G. Fidel, R. Bitton, and A. Shabtai, "When explainability meets adversarial learning: detecting adversarial examples using shap signatures," in Proceedings of 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-8.
[22]M. Ghurab, G. Gaphari, F. Alshami, R. Alshamy, and S. Othman,"A Detailed Analysis of Benchmark Datasets for Network Intrusion Detection System," Asian Journal of Research in Computer Science, vol. 7, pp. 14-33, 2021.
[23]S. K. Wanjau, G. M. Wambugu, and G. N. Kamau, "SSH-Brute Force Attack Detection Model based on Deep Learning," International Journal of Computer Applications Technology and Research, vol. 10, no. 01, pp. 42–50, Jan. 2021.
[24]M. D. Hossain, H. Ochiai, F. Doudou, and Y. Kadobayashi, "SSH and FTP brute-force attacks detection in Computer Networks: LSTM and Machine Learning Approaches," 2020 5th International Conference on Computer and Communication Systems (ICCCS), 2020.
[25]O. Hassen and H. Ibrahim, "Preventive Approach against HULK attacks in Network Environment," International Journal of Computing and Business Research, vol. 7, no. 3, Oct. 2017.
[26]A. Dhanapal and P. Nithyanandam, "The slow HTTP distributed denial of service attack detection in cloud," Scalable Computing: Practice and Experience, vol. 20, no. 2, pp. 285–298, 2019.
[27]R. Papadie and I. Apostol, ''Analyzing websites protection mechanisms against DDoS attacks,'' in Proceeding 9th International Conference Electronic, Computer Artificial Intelligence (ECAI), June 2017, pp. 1–6.
[28] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," Journal of Big Data, vol. 7, no. 1, 2020
[29] D. Ali and S. Frimpong, "Deepimpact: A deep learning model for whole body vibration control using impact force monitoring," Neural Computing and Applications, vol. 33, no. 8, pp. 3521–3544, 2020.
[30] D. Ali, M. B. Hayat, L. Alagha, and O. K. Molatlhegi, "An evaluation of machine learning and artificial intelligence models for predicting the flotation behavior of Fine High-Ash Coal," Advanced Powder Technology, vol. 29, no. 12, pp. 3493–3506, 2018.
[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[32] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.
[33] L. S. Shapley, "A value for n-person games," Contributions to the Theory of Games, vol. 2, no. 28, pp. 307–317, 1953.
[34] J. Chen, M. I. Jordan, and M. J. Wainwright, "HopSkipJumpAttack: A query-efficient decision-based attack," 2020 IEEE Symposium on Security and Privacy (SP), 2020.
[35] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba,V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. M. Molloy, and B. Edwards, ''Adversarial robustness toolbox v1.0.0,'' 2018, arXiv:1807.01069.

Syed Wali received the Bachelor of Engineering degree in Electrical Engineering from NED University of Engineering and Technology, Karachi, Pakistan, in 2018. He received the Master of Engineering degree in Electrical Power System from NED University of Engineering and Technology, Karachi, Pakistan in 2021. He is currently pursuing a Ph.D. degree in Electrical Engineering from Texas A&M University, Texas, USA and working as a Research Assistant at Clean And Resilient Energy Systems (CARES) Lab, Texas A&M, USA.

Syed worked as an Electrical Power Engineer in the largest refinery of Pakistan and served the power sector of the country for more than 2 years. He also worked as a Research Assistant in Neurocomputation Lab under the National Centre of Artificial Intelligence. His research areas include Machine learning, Neural network, Cybersecurity, digital signal processing, and Electrical Power System.



Irfan Khan (S'14, M' 18, SM' 20) is an Assistant Professor at the Department of Marine Engineering Technology with a joint appointment with the Electrical and Computer Engineering at Texas A&M College Station. He is the director of the Clean And Resilient Energy Systems (CARES) Lab that focuses on the reliability, sustainability, and security of the electric energy supply on marine vessels. He has been fortunate to receive several grants from multiple funding agencies to work on marine electric distribution systems, electric vehicle fast charging, and electric microgrids. Dr. Khan is an affiliate faculty member with the TAMU Energy Institute and the TEES Smart Grid Center. Before joining TAMU in 2018, Dr. Khan received a Ph.D. in Electrical and Computer Engineering from Carnegie Mellon University USA. He has published more than 90 refereed reputed journal and peer-reviewed conference papers in the smart energy systems-related areas.

Dr. Khan is a registered Professional Engineer (P.E.) with the State of Texas, USA. He is the Vice-Chair for the IEEE Galveston Bay Section (GBS) of Region 5. He has organized several special sessions at various international conferences. Further, Dr. Khan is a regular reviewer of more than 30 reputed journals and conferences, wherein the year 2020, he reviewed more than 230 articles. He is also helping with editorial responsibilities at various journals, e.g., IEEE Transactions on Industry Applications, IEEE Access, Electronics MDPI, Frontiers in Energy Research, etc.