IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 5, 2021, accepted October 2, 2021, date of publication October 6, 2021, date of current version October 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3118361

An Agile Approach to Identify Single and Hybrid Normalization for Enhancing Machine Learning-Based Network Intrusion Detection

MURTAZA AHMED SIDDIQI AND WOOGUIL PAK, (Member, IEEE)
Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, South Korea

Corresponding author: Wooguil Pak (wooguilpak@yu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant NRF-2019R1F1A1062320, and in part by the Ministry of Science and Information Communication Technology (ICT) (MSIT), South Korea, through the Information Technology Research Center (ITRC) support program, supervised by the Institute for Information Communications Technology Planning Evaluation (IITP), under Grant IITP-2021-2016-0-00313.

ABSTRACT Detecting intrusion in network traffic has remained a problematic task for years. Progress in the field of machine learning is paving the way for enhancing intrusion detection systems. Due to this progress intrusion detection has become an integral part of network security. Intrusion detection has achieved high detection accuracy with the help of supervised machine learning methods. A key factor in enhancing the performance of supervised classifiers is how data is augmented for training the classification model. Data in real-world networks or publicly available datasets are not always normally (Gaussian) distributed. Instead, the distributions of variables are more likely to be skewed. To achieve a high detection rate, data normalization or transformation plays an important role for machine learning-based intrusion detection systems. Several methods are available to normalize the attributes of the data before training a classification model. However, opting for the most suitable normalization technique is still a questionable task. In this paper, a statistical method is proposed that can identify the most suitable normalization method for the dataset. The normalization method identified by the proposed approach gives the highest accuracy for an intrusion detection system. To highlight the efficiency of the proposed method, five different datasets were used with two different feature selection methods. The datasets belong to both Internet of things and traditional network environments. The proposed method is also able to identify hybrid normalizations to achieve even improved intrusion detection results.

INDEX TERMS Anomaly detection, Bot-IoT, CIC-IDS 2017, intrusion detection, IoT, ISCX-IDS 2012, normalization, NSL KDD, skewness, scaling, transformation, UNSW-NB15.

I. INTRODUCTION

Studies on machine learning (ML) and deep neural networks (DNN) for intrusion detection systems (IDS) have become prominent due to an increase in knowledge on neural networks [1]. IDS play a significant role in securing a network since they aim to identify and highlight elements that can disrupt network communication. With the efficiency of ML-based IDS, the applications of IDS are no longer limited to traditional networks. The Internet of Things (IoT), which represents a large portion of today's world of interconnected devices represents a unique challenge for security requirements. Due to the limitations of resources in IoT and low-cost production, IoT devices are being targeted by a high number of attacks [2]. However, several researchers have proposed effective methods for IoT security [3]–[5]. A key element of training any ML-based IDS is the pre-processing of training data [6]. Various factors can influence the training model of an ML-based IDS. This is why it is vital to provide the training model with data that is normalized and contains relevant features [7]. Hence, researchers have pursued exploring pre-processing methods, feature selection methods, and data normalization methods to achieve a high anomaly detection rate [8], [9]. Normalization or transformation plays a vital role in network security, as normalization forces integrity which tends to increase the general cleanliness and structure

The associate editor coordinating the review of this manuscript and approving it for publication was Zheng Yan.

for each feature [10]. Data normalization leads to improved representation of data and allows ML-based algorithms to get most of the data, resulting in enhanced predictions. Various publicly available datasets for IDS contain numerous features (variables) whose distribution is outlying from normal (Gaussian) and asymmetric (skewed) [11]. Such factors make things complicated for achieving high detection in an IDS. Normalization helps in improving the interpretation of data, getting insights about the bond between variables in a feature, and meeting norms for statistical inference [12]. However, selecting a normalization or sequence of normalization is quite challenging [13], [14]. The absence of any standard method for evaluating the effects of normalization for the dataset classification results in a selection based on the hit and trial method [15], [16]. Such an approach can be a time-consuming process with no guarantee that the selected normalization is the most suitable for the ML model and dataset. In this paper, a statistical method to identify a suitable normalization method is suggested. The proposed method can be used to identify the most suitable normalization, transformation, or scaling method to achieve a high detection rate in ML-based IDS. The proposed method identifies not only single but also hybrid normalization methods for the dataset. The key contributions of the paper are:

- Identifying a statistical matrix that can assist in finding the most suitable normalization for the data at hand.
- Based on the computed ranks one can identify the most effective single or hybrid normalization for data in hand.
- To prove the validity and generality of the proposed statistical method, five different datasets with both numerical and non-numerical feature attributes were selected. Then, two different feature selection methods were employed for feature selection and three different ML classifiers were implemented to verify the selected normalization method.

The rest of the paper is structured as follows. Section II covers the related work on methods that are used for identifying the normality of the dataset with some prominent data normalization methods. Section III describes the details of the proposed process to identify data normalization. Section IV briefs about the experiments for the proposed process. Section V covers the experiments conducted to evaluate and validate the proposed technique. Section VI represents the results of the evaluation and validation process of the proposed method. Section VII set forth the discussion on the proposed model and a comparison between the proposed model and similar approaches. In the end, section VIII concludes the paper.

II. RELATED WORK

With the rapid expansion of the internet and interconnected devices, network security has come to be increasingly challenging. Network intrusion detection (NIDS) has proven to be an effective method to achieve high accuracy in classifying network anomalies. Most of the supervised classification methods rely on prior normalized datasets to train the model for classification. However, real-world network

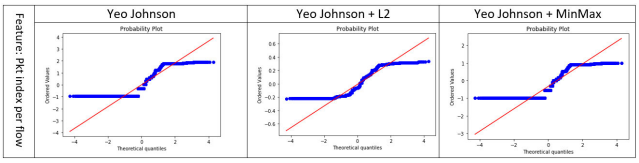


FIGURE 1. QQ-plotting of ISCX 2012 dataset feature 'Pkt index per flow' with normalization (single and hybrid).

data do not contain any normalization pre-process. Suitable normalization for publicly available IDS datasets can easily be established based on available research work. On the other hand, identifying suitable normalization methods for real-world or new datasets remains a concern. Generally, ML methods tend to perform well on a dataset with normal distribution [17], [18]. Distance from normality in a dataset can be illustrated in several different methods; however, the most prominent measures are skewness and kurtosis [19]. The skewness defines the asymmetry of a distribution in a dataset and zero skewness indicates symmetric distribution. Asymmetric distribution with a larger tail on the right has positive skewness and a dataset with a larger left tail has negative skewness. On the other hand, kurtosis deals with both tail heaviness and peakedness of a distribution associated with that of the normal distribution. Therefore, kurtosis is restricted to symmetric distributions [20]. Generally, if the values of skewness and kurtosis significantly diverge from zero and three respectively, it is expected that the dataset in hand may not be normally distributed. However, no official guidelines are specified for the values of kurtosis or skewness to indicate the non-normality of a dataset [21]. Other common methods to check normality before classification of the dataset are histogram, Box plot, QQ (quantile-quantile) plot, Kolmogorov Smirnov test, Lilliefors test, and Shapiro Wilk test [22], [23]. The mentioned methods suffer from diverse limitations. Histogram can be deceptive since changing the graph scale can alter the shape of the distribution and may lead to misperception [24]. Box-plot generates limited information to understand or conclude normality [24]. Similarly, QQ-plot can be a little tricky in identifying the right normalization method as shown in Figure 1. As seen in Figure 1, all three representations of the QQ plot are quite similar yet the classification results are different.

Among the tests Kolmogorov Smirnov, Lilliefors, and Shapiro Wilk, the Shapiro test is the most powerful [23]. The Shapiro-Wilk assessment is based on a random sample from the dataset. The null hypothesis [25] of the Shapiro-Wilk test is that the data is normally distributed. If the p-value [26] of the sample data is lower than 0.05 then the distribution is not normal. However, based on sample size it is possible that the p-value can identify a normally distributed dataset as not normally distributed and vice versa [22]. As a result, the method to identify a fitting normalization method for a dataset before classification is unclear. Transformation and normalization techniques implemented in this paper include

L2-normalization, Yeo-Johnson, Min-Max, Robust scaler, and Standard scaler. The Yeo-Johnson transformation [27] is an extension of the Box-Cox transformation. Mathematically Yeo-Johnson can be represented as Equation 1, where 'y' are the values, 'λ' can be any real number, and $\lambda = 1$ gives the identity transformation [11].

$$y_i^{(\lambda)} = \begin{cases} \left(\frac{(y_1+1)^{\lambda}-1}{\lambda}\right) & \text{if } \lambda \neq 0, y \geq 0 \\ log(y_i+1) & \text{if } \lambda = 0, y \geq 0 \\ \left(\frac{(-[(y_1+1)^2-(\lambda)-1]}{2-\lambda}\right) & \text{if } \lambda \neq 2, y < 0 \\ -log(-y_i+1) & \text{if } \lambda = 2, y < 0 \end{cases} \tag{1}$$

The Min-Max normalization [28] is among the most commonly used normalization [29]–[31]. Min-Max implements linear transformation on the data. Mathematically Min-Max can be represented as Equation 2.

$$x_{scaler} = \frac{x - min(x)}{max(x) - min(x)} \tag{2}$$

The Robust scaler [32] approach to normalizing data is similar to min-max. The only difference is that the Robust scaler scales data based on the quintile range. Equation 3 represents the Robust scaler, where 'x' represent the values while $Q_1 = 25^{th}$ quantile and $Q_3 = 75^{th}$ quantile.

$$x = \frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)} \tag{3}$$

The Standard scaler [33] normalizes the data by removing the mean and scaling the data to unit variance. Mathematically Standard scaler can be represented as Equation 4, where 's' is the standard deviation and 'μ' is the mean.

$$x_{scaler} = \frac{x - \mu_{mean}}{s_{stddiv}} \tag{4}$$

The L2-standardization [34] normalizes the dataset in a way that in each row the sum of the square of each value will be one. Equation 5 represent L2-standardization, where 'x' represent the values of features in the dataset.

$$||x||_2 = (|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2)^{1/2} \tag{5}$$

Data pre-processing plays a vital role in an IDS and for several data mining-related operations. Data normalization is an essential part of data pre-processing, particularly for intrusion detection methods that rely on statistical attributes extracted from the data at hand. As in paper [35], the authors highlighted the importance of data normalization and how it can affect the performance of anomaly detection. The authors implemented four different normalization methods with three different classifiers. This paper was aimed to answer two questions. First, whether attribute normalization is crucial for intrusion detection performance. Second, which technique of attribute normalization is most effective. The authors concluded the paper by providing experimental proof that attribute normalization plays a role in improving anomaly detection. Among the implemented normalization methods, statistical-based normalization resulted in achieving the highest accuracy results. The only downside of this paper was that the authors identified the most suitable normalization based on classification results. This approach of identifying normalization based on classification is not a proficient process. A study by B. Setiawan *et al.* [36], used the information gain method to select the most suitable normalization method. In this research, the authors used information gain on the log normalization, min-max, and z-score normalization schemes. After implementing normalization, attributes were rounded off by 2 to 10 decimals, and Information gain was used on each decimal alteration. Based on the information gain method the quality of attributes was computed. As per results, the highest risk of rounding the normalization was displayed by log normalization and z-score. Yet, the authors implemented log normalization for the intrusion detection system. The authors justified the use of log normalization by stating that it had the three decimal place-safe threshold. Despite the justification, implementing log normalization is a concern as the information gain implemented by authors highlighted that rounding log normalization is not suitable. In a paper by Yu Liping *et al.* [37], they evaluated several normalization methods and concluded that each evaluation purpose requires a different data normalization procedure. Compared to the mentioned papers, this paper presents a more precise statistical approach to identify the most suitable normalization method for the dataset in hand.

III. PROPOSED METHOD

In this article, a statistical method is proposed to identify the most suitable normalization, transformation, or scaling method for the data at hand. The proposed statistical model is not limited to a specific format of a dataset. As the model is validated on datasets that cover both numeric and non-numeric feature attributes. The proposed model is also implanted on IoT-based datasets to further validate the general application of the suggested approach. The flow of the proposed method can be seen in Figure 2.

Initially, a dataset is pre-processed by applying basic data cleaning. Details of data pre-processing are included in the experiment section of the paper. After data cleaning, feature selection is applied to the dataset. After feature selection, the normalization, transformation, or scaling methods are applied to the dataset. To find the most suitable normalization approach, two or three most common normalization methods can be applied separately on the dataset. This will result in multiple datasets based on the number of selected normalization methods. The following two steps will be applied to each dataset separately. First, the mean, median, and skewness of each feature of the datasets are computed. Second, the overall average of mean, median, and skewness of the features in the datasets are computed. This will result in a matrix of the average mean, median, and skewness of the features from each dataset. The skewness in the proposed model is taken as an absolute value. The reason is later discussed in the

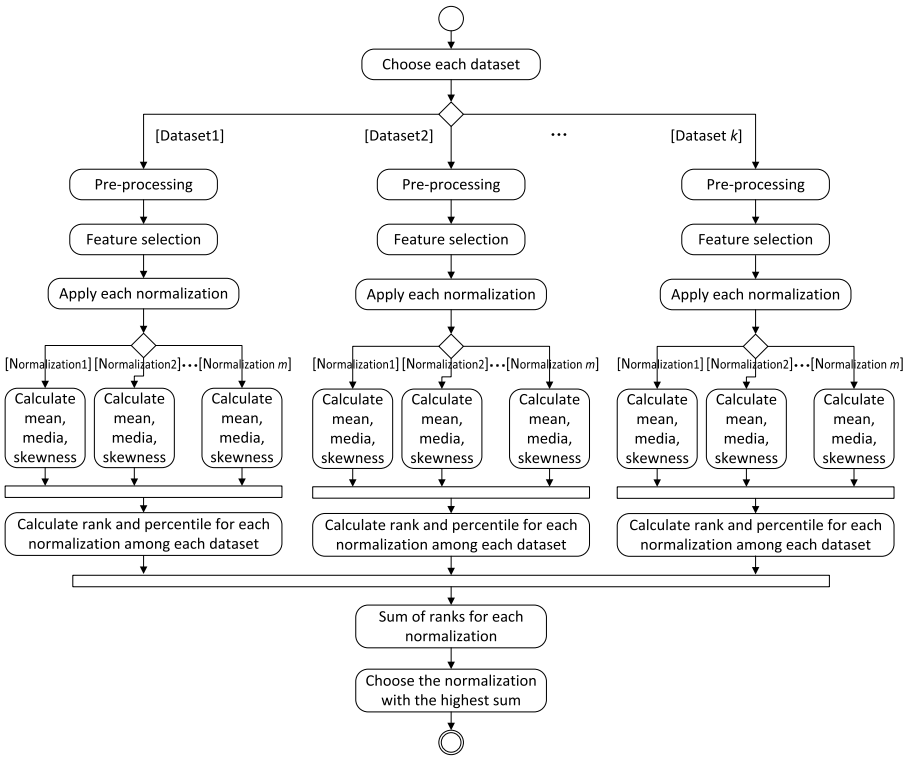


FIGURE 2. Proposed method flow.

sub-section data normalization and encoding. After computing the average mean, median, and skewness of the features of all the datasets the Ranking and Percentile method are applied on the computed matrix. The Rank and Percentile methods assign ranks based on descending order i.e. the highest value in the column will be ranked first. To identify the most suitable normalization method, the sum of the ranks is calculated. As a result, the normalization with the largest sum of ranks is the most suitable normalization for the data at hand. Based on the flow diagram in Figure 2, the proposed statistical model is shown in Algorithm 1.

IV. EXPERIMENTATION

In this paper, we analytically evaluate the effect of different methods of attribute normalization on the performance of ML-based IDS. Three ML algorithms, Support Vector Machine (SVM), Random Forest (RF), and Deep Neural Network (DNN) were employed to validate the proposed statistical method. The reason for implementing three different classifiers is to highlight that the proposed model is not dependent on a particular ML algorithm. As each of the mentioned ML algorithms belongs to a different category of ML classification method. The SVM is associated with spatial regression-based algorithms, RF represents a decision-based classifier, and DNN belongs to supervised deep learning ML algorithms. Attribute normalization schemed used were Yeo-Johnson, Min-Max, Robust scaler, Standard scaler, and L2 normalization. The stated normalization methods are selected because they cover the core techniques for data normalization methods i.e. scaling, clipping, and log-based scaling. The datasets used for the experiment were CIC-IDS 2017 [38], ISCX-IDS 2012 [39], NSL-KDD [40], UNSW-NB 15 [41], and Bot-IoT [42]. All the mentioned datasets are well-known and specifically developed for testing network security-based algorithms. Further, the Bot-IoT dataset was designed particularly for IoT-based security algorithms. Datasets CIC-IDS 2017, ISCX-IDS 2012, and Bot-IoT were created by modeling a real network environment containing both normal and attack traffic. The NSL-KDD dataset is an improved version of the KDD'99 dataset [40]. The KDD'99 dataset suffered from a high number of redundant records, duplicate values, and biased sampling. The PerfectStorm tool created the UNSW-NB 15 dataset. The tool generated a mixture of both normal and attack traffic behaviors to create the UNSW-NB 15 dataset. The details of the attacks simulated in each dataset are shown in Table 1. The motive behind implementing multiple datasets, feature selection methods, and classifiers is to highlight the generality and flexibility of the proposed method. The paper's contribution is twofold. First, the proposed statistical method is implemented to identify the most suitable normalization method for each dataset. Secondly, multiple ML-based IDS are implemented to validate the results of the proposed method. The hardware used for the experiments was an Intel Xeon Gold 32 core (64 threads) processor with 192GB RAM and RTX2080ti GPU. The programming language used for the

Algorithm 1 Proposed Statistical Model

Input: Dataset, "x"
Output: Most suitable normalization method for the dataset.
Where;
$x \triangleq (x_1, x_2, \ldots, x_k)$, $k(\in \mathbb{N})$: datasets.
ith data: $x_i \triangleq (f_1^i, f_2^i, \ldots, f_n^i)$, n is the total number of features.
$N_m = m$-th normalization.
Step 1: Apply Pre-Processing on the dataset.
$x' \leftarrow Pre - Processing(x)$
Step 2: Apply feature selection on dataset.
$x'' \leftarrow FeatureSelection(x')$
Step 3: Apply normalization, transformation or scaling on dataset, N_m.
$x^{(m)} \leftarrow N_m(x'')$, where $x_i^{(m)} \triangleq (f_1^{(m),i}, f_2^{(m),i}, \ldots, f_n^{(m),i})$
Step 4: Compute Mean, Median and, Skewness of each feature.
$mean_j^{(m)} = mean(f_j^{(m),1}, f_j^{(m),2}, \ldots, f_j^{(m),k})$
$median_j^{(m)} = median(f_j^{(m),1}, f_j^{(m),2}, \ldots, f_j^{(m),k})$
$skewness_j^{(m)} = abs(skewness(f_j^{(m),1}), f_j^{(m),2}, \ldots, f_j^{(m),k}))$
Step 5: Compute average Mean, Median and Skewness of the dataset.
$\overline{mean}^{(m)} = mean(mean_1^{(m)}, mean_2^{(m)}, \ldots, mean_n^{(m)})$
$\overline{median}^{(m)} = median(median_1^{(m)}, median_2^{(m)}, \ldots, median_n^{(m)})$
$\overline{skewness}^{(m)} = skewness(skewness_1^{(m)}, skewness_2^{(m)}, \ldots, skewness_n^{(m)})$
Step 6: Get Rank on $\overline{mean}^{(m)}$, $\overline{median}^{(m)}$, and $\overline{skenewss}^{(m)}$ of each N_m, $\forall m$.
Step 7: Sum the ranking of $\overline{mean}^{(m)}$, $\overline{median}^{(m)}$ and, $\overline{skenewss}^{(m)}$ to identify the best normalization (N_{m^*}).
$m^* = argmax_m\{\text{Rank}(\overline{mean}^{(m)}) + \text{Rank}(\overline{median}^{(m)}) + \text{Rank}(\overline{skenewss}^{(m)})\}$

implementation was Python 3.6 on the Ubuntu operating system. The DNN was executed using GPU-enabled TensorFlow 2.3.1 on the Keras framework. The RF and SVM classifiers were implemented using Scikit-Learn 0.23.2 ML library. TensorFlow, Scikit-Learn, and Python are open-source software, which is accessible online for free download. Minitab [43] was used for extracting and analyzing dataset attribute values. MS-Excel data analysis tool 'Rank and Percentile' was used to calculate the ranks of the normalization methods.

A. DATA PRE-PROCESSING

This section covers the pre-process steps applied to the dataset. In this paper, five datasets were used; CIC-IDS 2017, ISCX-IDS 2012, NSL-KDD, UNSW NB-15, and Bot-IoT. First basic data cleaning is applied on all five datasets. Basic cleaning removed missing value samples, duplicate data samples and, infinite data samples from the datasets. Second negative time values are removed from the datasets. In the dataset CIC-IDS 2017, the "BENIGN" class and in dataset ISCX-IDS 2012 "NORMAL" class had a very high number of samples. To avoid biasing, 230124 samples of the "BENIGN" class are used from CIC-IDS 2017 and 25,854 samples of the "NORMAL" class are extracted from ISCX-IDS 2012. In dataset CIC-IDS 2017, two classes "Infiltration" and "Heartbleed" are dropped as they had very few samples. All three classes of web attacks are combined into one class "Web attack" in dataset CIC-IDS 2017. For the UNSW NB-15 dataset the features 'id', 'label', and 'service' were dropped. As the 'service' column contained some missing values. The 'id' column was the index and the 'label' column was a binary representation of attacks, where '0' represented normal and '1' as an attack. In the Bot-IoT dataset, features 'pkSeqID', 'saddr', 'daddr', 'category', and 'attack' are dropped. The categorical features are converted to numbers with the help of label encoding. The 'Data Exfiltration' class is dropped as it had only 6 samples. To avoid any biases the samples of 'TCP', 'UDP', and 'Service_Scan' are reduced. At the end of pre-processing, features with zero variance were dropped for the datasets using zero-variance [44]. After pre-processing, details of the dataset can be seen in Table 1.

To avoid any biasing, the synthetic minority over-sampling technique (SMOTE) is implemented with edited nearest neighbors (ENN) [8] to perform cleaning on the training set of each dataset.

B. FEATURE SELECTION

In this research, two feature selection methods are employed. The filter-based Pearson correlation [45] is applied on CIC-IDS 2017, ISCX-IDS 2012, UNSW NB-15, and Bot-IoT datasets. Whereas, wrapper-based Forward Selection with Decision Tree (FS-DT) [46] is applied on NSL-KDD. Pearson correlation works by computing the correlation between features. Features with a high correlation are more linearly dependent and therefore have nearly the same effect on the dependent variable. Hence, if two features show a high correlation, the Pearson correlation drops one of those features.

TABLE 1. Details of CIC-IDS 2017, ISCX-IDS 2012, and NSL-KDD datasets after pre-processing.

Dataset	CIC-IDS2017	ISCX-IDS2012	NSL-KDD	UNSW NB-15	Bot-IoT
Number of features	79	82	41	42	14
Number of classes	11	5	2	10	7
Number of samples	786,633	73,566	125,973	162,724	74,680
Number of each class	Benign:230,124 Bot:1,956 DDoS:128,025 DoSGoldenEye:10,293 DoSHulk:230,124 DoSSlowhttptest:5,499 DoSslowloris:5,796 FTP-Patator:7,935 PortScan:158,804 SSH-Patator:5,897 Web Attack:2,180	Normal:25,845 BruteForceSSH:10,029 DDoS:25,845 HTTPDoS:3,928 Infiltration:7,919	Normal:67,343 Attack:58,630	Normal:85,720 Analysis:2,032 Backdoor:1,880 DoS:5,494 Exploits:27,424 Fuzzers:20,957 Generic:7,599 Reconnaissance:9,991 Shellcode:1,456 Worms:171	Normal:477 HTTP:2,474 Keylogging:73 OS_Fingerprint:17,914 Service_Scan:17,914 TCP:17,914 UDP:17,914

TABLE 2. Original number of dataset features vs features selected.

	CIC-IDS 2017	ISCX-IDS 2012	NSL-KDD	UNSW NB-15	Bot-IoT
Total Features	79	82	41	42	19
Features Selected	45	42	19	36	14

In this experiment, Pearson correlation was applied using the python library with correlation coefficients 0.95 and −0.95. Features within the defined correlation coefficient limits were selected for the experiment. In datasets, CIC-IDS 2017 and ISCX-IDS 2012 features were compared with each other in the dataset to compute their correlation with each other. As a result, features that have higher correlation such as Average Backward Segment Size, Average Forward Segment Size, Maximum Packet Length, Minimum Packet Length, etc in datasets CIC-IDS 2017 and ISCX-IDS 2012 were dropped by Pearson correlation. Pearson correlation can be computed as shown in Equation 6.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{6}$$

where:

r_{xy} = Pearson Correlation Coefficient Value,

x_i = Individual Sample Points of Each Conditional Attributes,

y_i = Individual Sample Points of the Decision Attribute,

$\bar{x}$ = Average of all Sample Points of each Conditional Attribute,

$\bar{y}$ = Average of all Sample Points of the Decision Attribute.

The wrapper-based FS-DT [47] is a greedy search algorithm that tries to find the "optimum" subset of features by iteratively selecting features based on the classifier performance. Table 2 shows the number of features selected after the feature selection process.

C. DATA NORMALIZATION AND ENCODING

In this experiment, five data normalization methods i.e. Yeo-Johnson, Robust scaler, Min-Max, Standard scaler, and L2 normalization are implemented. The NSL-KDD and UNSW NB-15 datasets contained both numeric and nominal features. Therefore, the one-hot encoding and label-encoding [48] are applied to NSL-KDD and UNSW NB-15 features respectively. Both one-hot encoding and label-encoding were applied using the python libraries. The one-hot encoding works by creating new binary columns to replace the categorical feature in a dataset. For instance, in the NSL-KDD dataset, the categorical feature 'protocol_type' had three attributes tcp, udp, and icmp. So the one-hot encoding encoded the attributes tcp as 001, udp as 010, and icmp as 100 and aggregating one feature column to three feature columns. Due to the one-hot encoding, NSL-KDD features were increased from 19 to 98 On the other hand, the label-encoding on the UNSW NB-15 dataset assigned a unique numeric value to each attribute of the non-numeric feature. For illustration, the feature 'proto' in UNSW NB-15 had non-numeric attributes i.e tcp, udp, igmp, ospf, sctp, etc. the label encoding simply encoded tcp as 0, udp as 1, igmp as 2, and so on. The reason for applying different encoding techniques on NSL-KDD and UNSW NB-15 datasets is due to the difference in the ordinal nature of the categorical attributes. Later, each normalization is applied to all five datasets. Software "Minitab" was then used on the normalized datasets to extract attribute mean, median, and skewness. The average and middle values of the attribute are represented by mean and median respectively. While the skewness as defined earlier is the asymmetry of a distribution in a dataset and zero skewness indicates symmetric distribution. Asymmetric distribution with a larger tail on the right has positive skewness and a dataset with a larger left tail has negative skewness [21]. Equations 7, 8, and 9 were used to compute the mean, median, and skewness respectively.

$$Mean = \frac{\sum_{i-1}^{N} x_1}{N} \tag{7}$$

$$Median = \begin{cases} x\left[\frac{N}{2}\right] & \text{if } N \text{ is even} \\ \frac{\left(x\left[\frac{N-1}{2}\right] + x\left[\frac{N+1}{2}\right]\right)}{2} & \text{if } N \text{ is odd} \end{cases} \tag{8}$$

TABLE 3. Average mean, median, and skewness of each dataset based on different normalization methods.

	CIC-IDS 2017			ISCX-IDS 2012			NSL-KDD		
Normalization	Average Mean	Average Median	Average Skewness	Average Mean	Average Median	Average Skewness	Average Mean	Average Median	Average Skewness
Yeo-Johnson	0	-0.279	5.017	0	-0.338	2.031	0.020	-0.029	32.473
L2 Normalization	0.040	0.015	3.058	0.043	0.020	8.231	0.034	0.016	35.935
Robust Scaler	26581.63	0	15.235	54874	0	12.412	5.061	0.010	38.137
Standard Scaler	0	-0.277	15.235	0	0.064	12.412	0.020	-0.030	38.137
Min-Max	0.193	0.155	9.986	-0.806	-0.939	12.412	0.050	0.026	38.137

	UNSW NB-15			Bot-IoT		
Normalization	Average Mean	Average Median	Average Skewness	Average Mean	Average Median	Average Skewness
Yeo-Johnson	-3.6696	0.0394	0.50142	5.6567	-0.2316	0.31974
L2 Normalization	0.0357	0.0288	32.1929	0.09929	0.1111	74.1810
Robust Scaler	79.1296	7.6941	12.3598	26.7954	4.5935	21.9743
Standard Scaler	-1.3541	-0.1317	12.3598	-2.3284	-0.1956	21.9743
Min-Max	0.1342	0.1469	12.3598	0.3706	0.3226	21.9743

TABLE 4. Average mean, median, and skewness of each dataset based on hybrid normalization methods.

	CIC-IDS 2017 Dataset				ISCX-IDS 2012 Dataset		
Normalization	Average Mean	Average Median	Average Skewness	Normalization	Average Mean	Average Median	Average Skewness
Yeo-Johnson + Min-Max	-0.264	-0.477	5.017	Yeo-Johnson + Min-Max	-0.41	-0.67	2.031
Min-Max + L2 Normalization	0.193	0.155	9.986	Min-Max + L2 Normalization	-0.128	-0.147	6.326
L2 Normalization + Yeo-Johnson	0	-0.4	3.057	L2 Normalization + Yeo-Johnson	0	-0.4	3.057
L2 Normalization+Standard Scaler	0.074	-0.189	2.298	L2 Normalization + Min-Max	-0.86	-0.95	8.321
	NSL-KDD Dataset				UNSW NB-15 Dataset		
Normalization	Average Mean	Average Median	Average Skewness	Normalization	Average Mean	Average Median	Average Skewness
Yeo-Johnson + Standard Scaler	0.02	-0.029	32.473	Min-Max+L2 Normalization	0.0706	0.0729	11.0353
Yeo-Johnson + Robust Scaler	0.045	0.0102	32.473	Yeo-Johnson+Min-Max	0.3744	0.3560	0.5014
Min-Max + L2 Normalization	0.034	0.016	35.935	Yeo-Johnson+Standard Scaler	0.0394	-8.9229	0.5014
Standard Scaler + L2 Normalization	0.015	-0.005	33.781	L2 Normalization+Standard Scaler	-2.2604	-0.0838	44.5423

	Bot-IoT Dataset		
Normalization	Average Mean	Average Median	Average Skewness
Min-Max+L2 Normalization	0.1888	0.1560	21.6463
Yeo-Johnson+Min-Max	0.4502	0.3657	0.31974
Yeo-Johnson+Standard Scaler	5.6185	-0.2316	0.3197
L2 Normalization+Standard Scaler	-2.9353	-0.0169	74.1810

$$Skewness = \frac{N\sum_{i=1}^{N}(x_i - \bar{x})^3}{(N-1)(N-2)s^3} \tag{9}$$

where "N" represents the number of values in the dataset, "x" represents the value in a dataset. Further in Equation 9, "$\bar{x}$" represents the mean, and "s" represents the standard deviation of the dataset. For the proposed statistical method, skewness is taken as an absolute value. Negative skew is generally considered problematic for statistical models [49], [50]. Table 3 represents the average mean, median, and skewness of datasets after each normalization.

In this research, hybrid-normalization methods are also implemented for two reasons. One, to check whether a combination of two normalization methods can further improve IDS performance. Second, to check the flexibility of the proposed model. Based on the five normalization techniques selected for experimentation, a high number of combinations for hybrid normalization are possible. However, only a handful of combinations with high performance were selected as shown in Table 4.

D. IDENTIFYING THE BEST NORMALIZATION

After computing the matrices shown in Tables 3 and 4 percentile ranking is applied to identify the most suitable normalization method. Percentile rank [51] returns a score compares to other scores in the same matrix or set. This method can be used to calculate the relative standing of a value within a matrix or set. In this experiment, the Rank and Percentile data analysis tool from MS-Excel is used. The formula for percentile and ranking can be represented as Equations 10 and 11.

$$P = \frac{x}{N} \times 100 \tag{10}$$

$$r = \frac{P}{100(n+1)} \tag{11}$$

where "P" is the percentile, "x" represents the number of values below the selected value, "N" represents the total number of values, "r" represents the rank, and "n" is the number of values. Ranks are assigned based on descending order. After applying the Rank and Percentile method in Table 3, Table 5 was acquired. The normalization method with the highest sum of rank in Table 5 represents the most suitable normalization for the dataset.

Similarly, the Rank and Percentile method was applied on Table 4 to compute Table 6. The hybrid normalization method with the highest sum of rank in Table 6 represents the most suitable hybrid normalization for the dataset.

Based on the proposed method, normalization methods that achieved the maximum sum of rank in Tables 5 and 6 will achieve the highest accuracy for the respective dataset with ML-based IDS.

V. EVALUATING PROPOSED STATISTICAL MODEL

To evaluate and verify the proposed statistical model, three ML-based IDS are implemented. The implemented IDS models are based on RF, SVM, and DNN. The DNN model for IDS in this paper is the same as our earlier published work [6].

TABLE 5. After applying the rank and percentile on table 3 and computing ranks for each normalization.

	CIC-IDS 2017 Dataset							ISCX-IDS 2012 Dataset						
	Average Mean		Average Median		Average Skewness			Average Mean		Average Median		Average Skewness		
Normalization	Rank	Percentile	Rank	Percentile	Rank	Percentile	∑Rank	Rank	Percentile	Rank	Percentile	Rank	Percentile	∑Rank
Standard Scaler	4	0.00	4	25.00	1	75.00	9	3	25.00	1	100.0	1	75.00	5
Robust Scaler	1	100.0	3	50.00	1	75.00	5	1	100.0	3	50.00	3	50.00	7
Yeo-Johnson	4	0.00	5	0.00	4	25.00	**13**	3	25.00	4	25.00	5	0.00	**12**
Min-Max	2	75.00	1	100.0	3	50.00	6	5	0.00	5	0.00	1	75.00	11
L2 Normalization	3	50.00	2	75.00	5	0.00	10	2	75.00	2	75.00	4	25.00	8
	NSL-KDD Dataset							UNSW NB-15 Dataset						
	Average Mean		Average Median		Average Skewness			Average Mean		Average Median		Average Skewness		
Normalization	Rank	Percentile	Rank	Percentile	Rank	Percentile	∑Rank	Rank	Percentile	Rank	Percentile	Rank	Percentile	∑Rank
Yeo-Johnson	4	0.00	4	25.00	5	0.00	**13**	5	0.00	3	50.00	5	0.00	**13**
L2 Normalization	3	50.00	2	75.00	4	25.00	9	3	50.00	4	25.00	1	100.0	8
Min-Max	2	75.00	1	100.0	1	50.00	4	2	75.00	2	75.00	2	25.00	6
Standard Scaler	4	0.00	5	0.00	1	50.00	10	4	25.00	5	0.00	2	25.00	11
Robust	1	100.0	3	50.00	1	50.00	5	1	100.0	1	100.0	2	25.00	4

	Bot-IoT Dataset						
	Average Mean		Average Median		Average Skewness		
Normalization	Rank	Percentile	Rank	Percentile	Rank	Percentile	∑Rank
Yeo-Johnson	2	75.00	5	0.00	5	0.00	**12**
L2 Normalization	4	25.00	3	50.00	1	100.0	8
Robust Scaler	1	100.0	1	100.0	2	25.00	4
Standard Scaler	5	0.00	4	25.00	2	25.00	11
Min-Max	3	50.00	2	75.00	2	25.00	7

TABLE 6. After applying the rank and percentile on table 4 and computing ranks for each hybrid normalization.

	CIC-IDS 2017 Dataset						
	Average Mean		Average Median		Average Skewness		
Normalization	Rank	Percentile	Rank	Percentile	Rank	Percentile	∑Rank
Yeo-Johnson + Min-Max	4	0.00	4	0.00	2	66.60	**10**
Min-Max + L2 Normalization	1	100.0	1	100.0	1	100.0	3
L2 Normalization + Yeo-Johnson	3	33.30	3	33.30	3	33.30	9
L2 Normalization + Standard Scaler	2	66.60	2	66.60	4	0.00	8
	ISCX-IDS 2012 Dataset						
	Average Mean		Average Median		Average Skewness		
Normalization	Rank	Percentile	Rank	Percentile	Rank	Percentile	∑Rank
Yeo-Johnson + Min-Max	3	33.30	3	33.30	4	0.00	**10**
Min-Max + L2 Normalization	2	66.60	1	100.0	2	66.60	5
L2 Normalization + Yeo-Johnson	1	100.0	2	66.60	3	33.30	6
L2 Normalization + Min-Max	4	0.00	4	0.00	1	100.0	9
	NSL-KDD Dataset						
	Average Mean		Average Median		Average Skewness		
Normalization	Rank	Percentile	Rank	Percentile	Rank	Percentile	∑Rank
Yeo-Johnson + Standard Scaler	3	33.30	4	0.00	3	0.00	**10**
Yeo-Johnson + Robust Scaler	1	100.0	2	66.60	3	0.00	6
Min-Max + L2 Normalization	2	66.60	1	100.0	1	100.0	4
Standard Scaler + L2 Normalization	4	0.00	3	33.30	2	66.60	9
	UNSW NB-15 Dataset						
	Average Mean		Average Median		Average Skewness		
Normalization	Rank	Percentile	Rank	Percentile	Rank	Percentile	∑Rank
Min-Max + L2 Normalization	2	66.60	2	66.60	2	66.60	6
Yeo-Johnson + Min-Max	1	100.0	1	100.0	3	0.00	5
Yeo-Johnson + Standard Scaler	3	33.30	4	0.00	3	0.00	**10**
L2 Normalization + Standard Scaler	4	0.00	3	33.30	1	100.0	8
	Bot-IoT Dataset						
	Average Mean		Average Median		Average Skewness		
Normalization	Rank	Percentile	Rank	Percentile	Rank	Percentile	∑Rank
Min-Max + L2 Normalization	3	33.30	2	66.60	2	66.60	7
Yeo-Johnson + Min-Max	2	66.60	1	100.0	3	33.30	6
Yeo-Johnson + Standard Scaler	1	100.0	4	0.00	4	0.00	**9**
L2 Normalization + Standard Scaler	4	0.00	3	33.30	1	100.0	8

The DNN used in earlier work had four dense layers with 120 nodes in an individual layer; other parameter specifics can be seen in Table 7.

Apart from classification, the Cohen's kappa coefficient [52], and receiver operating characteristics (ROC) [53] were also computed to verify that the normalization selected by the proposed method is the most suitable for the dataset. For the ML classification model accuracy, precision, recall, and the F1-score were calculated using Equations (12)–(15).

$$Accuracy = \frac{TruePositive + TrueNegative}{Total} \tag{12}$$

TABLE 7. DNN parameters for the experiments.

Activation Function	Epochs	Learning rate	Batch size	Layers
Elu	100	0.0001	100	4

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (13)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (14)$$

$$F1 - Score = 2.\frac{Precision.Recall}{Precision + Recall} \quad (15)$$

where true positive is when an attack is correctly identified as an attack and false positive is when normal traffic is incorrectly identified as an attack. True negative is when normal traffic is correctly identified as normal traffic and false negative is when an attack is incorrectly identified as normal traffic. The kappa coefficient score is a very handy measure of an ML model's capability when performing multi-class classification [54]. The kappa coefficient compares the predicted and expected accuracy of an ML algorithm. Mathematically kappa coefficient can be represented as Equation 16.

$$Kappa = \frac{p_0 - p_e}{1 - p_e} \quad (16)$$

where 'p_0' is the overall accuracy of the ML model and 'p_e' represents the agreement between the ML model estimates and the authentic class values as if happening by chance. On the other hand, the receiver operating characteristic (ROC) curve is a graphical representation of the classification model at all classification thresholds [55].

VI. RESULTS

In this section results of the ML classifiers on CIC-IDS 2017, ISCX-IDS 2012, NSL-KDD, UNSW NB-15, and Bot-IoT datasets are presented. The reason behind using five different datasets and multiple ML classifiers is to highlight the flexibility and generality of the proposed method. Table 8 represents the most suitable normalization method based on the proposed method ranking computation as shown in Table 5.

A. RANDOM FOREST BASED IDS MODEL

As a part of verifying the proposed statistical model, RF was implemented for classification on all five datasets as shown in Table 9. The normalization methods highlighted by the proposed method (i.e. Table 8) achieved the highest accuracy. For the single normalization method, Yeo-Johnson achieved the highest accuracy in all five datasets. While for the hybrid normalization method, the combination of Yeo-Johnson and Min-Max achieved the highest accuracy for CIC-IDS 2017 and ISCX-IDS 2012 datasets. While the combination of Yeo-Johnson + Standard was able to achieve the highest accuracy on NSL-KDD.

Based on Equation 16, the Kappa coefficient score was computed for all five datasets. Table 10 represents the Kappa score of the normalization method which achieved the highest accuracy based on the RF-based IDS model.

TABLE 8. Most suitable normalization method for all three datasets based on the proposed method.

Dataset	Normalization	Hybrid Normalization
CIC-IDS 2017	Yeo-Johnson	Yeo-Johnson + Min-Max
ISCX-IDS 2012	Yeo-Johnson	Yeo-Johnson + Min-Max
NSL-KDD	Yeo-Johnson	Yeo-Johnson + Standard Scaler
UNSW NB-15	Yeo-Johnson	Yeo-Johnson + Standard Scaler
Bot-IoT	Yeo-Johnson	Yeo-Johnson + Standard Scaler

To visualize the classification performance of RF-based IDS, Figure 3 represents the classification matrix of each dataset with both single and hybrid transformations. Although achieving high accuracy is not the core purpose of this research but, the RF-based IDS was able to achieve good classification results on each dataset excluding the UNSW NB-15 dataset. Figures 3 (a), (b), (c), (d), and (e) represents the classification matrix of the datasets CIC-IDS 2017, ISCX-IDS 2012, NSL-KDD, UNSW NB-15, and Bot-IoT based on Yeo-Johnson normalization. Figures 3 (f) and (g) represent the classification matrix of the datasets CIC-IDS 2017, ISCX-IDS 2012 based on Yeo-Johnson + Min-Max normalizations respectively. Figures 3 (h), (i), and (j) represent the classification matrix of the NSL-KDD, UNSW NB-15, and Bot-IoT datasets based on Yeo-Johnson + Standard scaler normalizations respectively.

The ROC curves for each dataset classified by RF-based IDS are presented in Figure 4. Each colored line in the ROC graph represents a class in a dataset. Figures 4 (a), (b), (c), (d), and (e) represent the ROC of the datasets CIC-IDS 2017, ISCX-IDS 2012, NSL-KDD, UNSW NB-15, and Bot-IoT based on Yeo-Johnson normalization. Figure 4 (f) and (g) represent the ROC of the datasets CIC-IDS 2017, ISCX-IDS 2012 based on Yeo-Johnson + Min-Max normalizations respectively. Figures 4 (h), (i), and (j) represent the ROC of the NSL-KDD, UNSW NB-15, and Bot-IoT datasets based on Yeo-Johnson + Standard scaler normalizations respectively.

B. SUPPORT VECTOR MACHINE BASED IDS MODEL

In this research, the SVM-based IDS model is executed for 1000 epochs. As the main goal of performing classification is to verify the proposed statistical model and not to achieve high accuracy results. Based on Table 11, Yeo-Johnson achieved the highest accuracy as highlighted by the proposed method. However, in hybrid normalization for CIC-IDS 2017 and BoT-IoT datasets the highlighted normalization method (i.e. Table 8) did not achieved the best classification results. For datasets, ISCX-IDS 2012, NSL-KDD, and UNSW NB-15 the highest accuracy was achieved by the normalization method identified by the proposed statistical method as shown in Table 8.

Based on Equation 16, the Kappa coefficient score is computed for all five datasets. Table 12 represents the Kappa score of the normalization method which achieved the highest accuracy based on the SVM-based IDS.

To visualize the classification performance of SVM-based IDS, Figure 5 represents the classification matrix of each

TABLE 9. Classification results of the RF-based IDS on all three datasets and normalizations.

Dataset	CIC-IDS 2017 dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	Min-Max+L2	L2+Yeo-Johnson	Yeo-Johnson+Min-Max	L2+Standard
Accuracy	99.64%	99.71%	99.75%	99.68%	99.87%	99.71%	99.71%	99.88%	99.65%
F1-Score	97.19%	97.28%	97.41 %	97.73%	99.86%	98.61%	97.80%	99.13%	97.36%
Precision	95.58%	95.70%	95.97%	96.33%	98.54%	97.81%	96.60%	98.67%	95.84%
Recall	99.66%	99.62%	99.66%	99.48%	99.61%	99.54%	99.24%	99.62%	99.27%
Dataset	ISCX-IDS 2012 dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	Min-Max+L2	L2+Yeo-Johnson	Yeo-Johnson+Min-Max	L2+Min-Max
Accuracy	95.72%	95.18%	95.75%	94.34%	95.89%	95.16%	95.05%	95.77%	94.92%
F1-Score	94.29%	93.02%	94.01 %	91.45%	94.27%	92.99%	92.68%	94.06%	95.20%
Precision	94.39%	92.73%	93.82%	91.08%	94.35%	92.53%	92.23%	93.98%	92.12%
Recall	94.22%	93.42%	94.22%	92.69%	94.22%	93.65%	93.37%	94.14%	93.17%
Dataset	NSL-KDD dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	Min-Max+L2	Yeo-Johnson+Standard	Yeo-Johnson+Robust	Standard+L2
Accuracy	99.73%	99.87%	99.86%	99.82%	99.87%	99.71%	99.88%	99.86%	99.62%
F1-Score	99.72%	99.87%	99.86%	99.82%	99.87%	99.70%	99.88%	99.86%	99.62%
Precision	99.73%	99.87%	99.86%	99.83%	99.87%	99.71%	99.88%	99.87%	99.63%
Recall	99.72%	99.86%	99.85%	99.82%	99.86%	99.70%	99.88%	99.86%	99.61%
Dataset	UNSW NB-15 dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	MinMax+L2	Yeo-Johnson+MinMax	Yeo-Jonson+Standard	L2+Standard
Accuracy	75.79%	76.66%	75.53%	36.87%	76.92%	74.87%	76.57%	77.04%	63.25%
F1-Score	55.53%	56.44%	55.73%	31.66%	56.81%	54.40%	56.69%	56.87%	45.80%
Precision	54.20%	54.77%	54.32%	47.18%	54.93%	53.44%	54.79%	54.97%	48.31%
Recall	67.61%	67.93%	67.68%	42.92%	68.15%	64.72%	68.21%	68.20%	55.77%
Dataset	Bot-IoT dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	MinMax+L2	Yeo-Johnson+MinMax	Yeo-Jonson+Standard	L2+ Standard
Accuracy	99.41%	98.73%	99.29%	98.33%	99.42%	98.57%	99.17%	99.25%	98.90%
F1-Score	98.25%	97.71%	96.28%	94.29%	98.34%	96.93%	96.36%	96.33%	95.50%
Precision	97.98%	97.33%	94.34%	91.48%	98.12%	95.80%	94.51%	94.42%	93.28%
Recall	98.67%	98.29%	98.60%	98.05%	98.68%	98.19%	98.56%	98.60%	98.37%

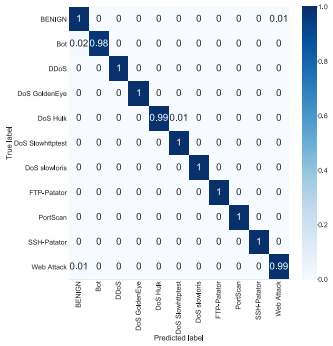
TABLE 10. Kappa coefficient score of the RF-based IDS on highest accuracy normalizations.

Datasets	Normalization	Kappa Score	Hybrid Normalization	Kappa Score
CIC-IDS 2017	Yeo-Johnson	0.9983	Yeo-Johnson + Min-Max	0.9984
ISCX-IDS 2012	Yeo-Johnson	0.9429	Yeo-Johnson + Min-Max	0.9386
NSL-KDD	Yeo-Johnson	0.9976	Yeo-Johnson + Standard Scaler	0.9976
UNSW NB-15	Yeo-Johnson	0.7213	Yeo-Johnson + Standard Scaler	0.7228
Bot-IoT	Yeo-Johnson	0.9925	Yeo-Johnson + Standard Scaler	0.9902

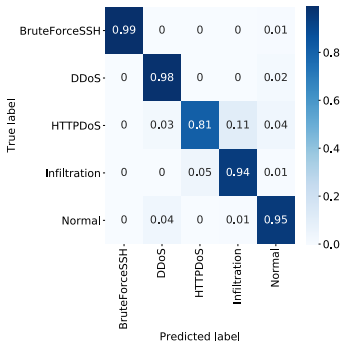
TABLE 11. Classification results of the SVM-based IDS on all three datasets and normalizations.

Dataset	CIC-IDS 2017 dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	Min-Max+L2	L2+Yeo-Johnson	Yeo-Johnson+Min-Max	L2+Standard
Accuracy	34.25%	20.73%	43.99%	46.22%	75.73%	49.65%	71.51%	71.30%	68.69%
F1-Score	28.45%	03.91%	31.92 %	25.17%	68.02%	38.52%	52.69%	56.59%	43.00%
Precision	38.65%	08.49%	36.93%	31.61%	67.27%	45.21%	56.95%	54.74%	44.29%
Recall	49.57%	09.74%	52.11%	41.67%	75.99%	56.66%	61.78%	64.09%	52.21%
Dataset	ISCX-IDS 2012 dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	Min-Max+L2	L2+Yeo-Johnson	Yeo-Johnson+Min-Max	L2+Min-Max
Accuracy	86.33%	18.93%	89.73%	90.22%	93.49%	86.45%	92.95%	93.57%	91.82%
F1-Score	83.67%	19.00%	88.55%	86.76%	90.05%	83.85%	89.60%	90.47%	89.05%
Precision	83.83%	48.39%	88.65%	85.90%	89.16%	84.26%	88.63%	89.97%	88.53%
Recall	84.21%	29.40%	88.68%	87.71%	91.04%	84.09%	90.78%	91.03%	89.60%
Dataset	NSL-KDD dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	Min-Max+L2	Yeo-Johnson+Standard	Yeo-Johnson+Robust	Standard+L2
Accuracy	98.06%	98.06%	98.08%	98.78%	99.38%	98.09%	99.38%	99.25%	98.20%
F1-Score	97.99%	97.99%	98.07%	98.77%	99.38%	98.08%	99.38%	99.25%	98.19%
Precision	98.05%	98.05%	98.12%	98.80%	99.38%	98.14%	99.38%	99.24%	98.25%
Recall	97.95%	97.95%	98.03%	98.74%	99.39%	98.04%	99.39%	99.25%	98.15%
Dataset	UNSW NB-15 dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	MinMax+L2	Yeo-Johnson+MinMax	Yeo-Jonson+Standard	L2+ Standard
Accuracy	53.49%	03.86%	59.76%	24.11%	70.35%	53.47%	62.20%	70.32%	27.36%
F1-Score	30.36%	04.44%	38.44%	07.68%	49.95%	30.45%	39.55%	49.95%	12.59%
Precision	42.55%	25.28%	42.41%	08.89%	49.60%	37.61%	44.84%	49.60%	27.50%
Recall	33.66%	07.92%	48.68%	15.54%	63.04%	34.11%	49.49%	63.04%	26.28%
Dataset	Bot-IoT dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	MinMax+L2	Yeo-Johnson+MinMax	Yeo-Jonson+Standard	L2+ Standard
Accuracy	72.89%	18.80%	72.73%	39.52%	72.92%	67.09%	77.00%	70.92%	44.39%
F1-Score	64.17%	09.34%	63.80%	25.14%	65.56%	68.62%	67.84%	65.56%	19.61%
Precision	59.11%	37.48%	59.02%	31.41%	63.38%	63.02%	65.08%	63.38%	24.03%
Recall	80.57%	12.61%	79.62%	37.90%	81.56%	79.20%	85.21%	81.56%	27.13%

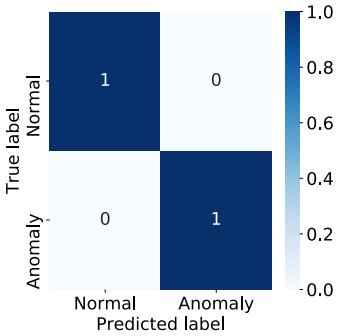
dataset with both single and hybrid transformations. Figure 5 (a), (b), (c), (d), and (e) represent the classification matrix of the datasets CIC-IDS 2017, ISCX-IDS 2012, NSL-KDD, UNSW NB-15, and Bot-IoT based on Yeo-Johnson normal-



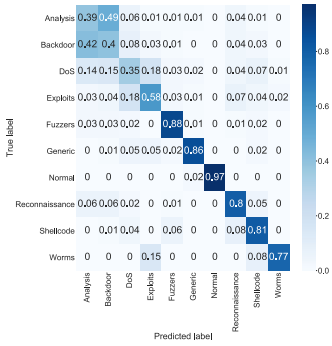
(a) CIC-IDS 2017 dataset with Yeo-Johnson



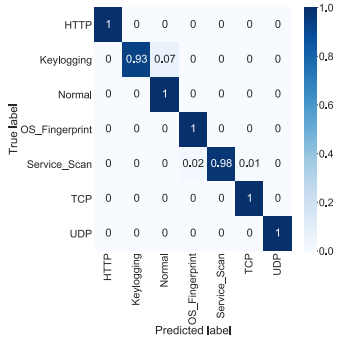
(b) ISCX-IDS 2012 dataset with Yeo-Johnson



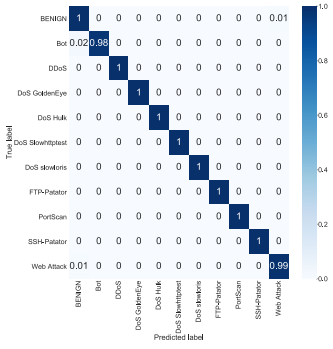
(c) NSL-KDD dataset with Yeo-Johnson



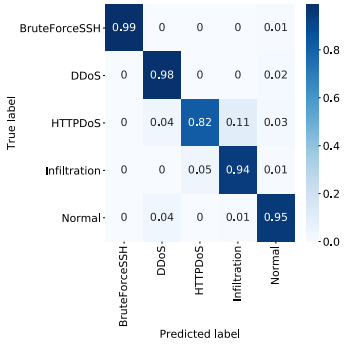
(d) UNSW NB-15 dataset with Yeo-Johnson



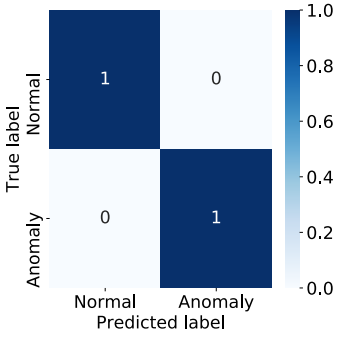
(e) Bot-IoT with Yeo-Johnson



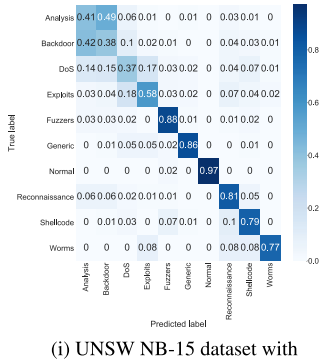
(f) CIC-IDS 2017 dataset with Yeo-Johnson + Min-Max



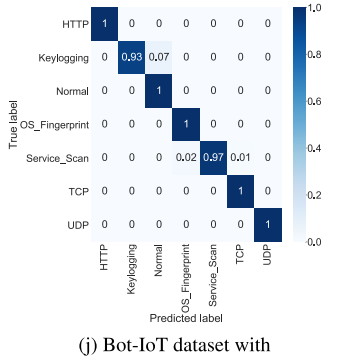
(g) ISCX-IDS 2012 dataset with Yeo-Johnson + Min-Max



(h) NSL-KDD dataset with Yeo-Johnson + Standard Scaler

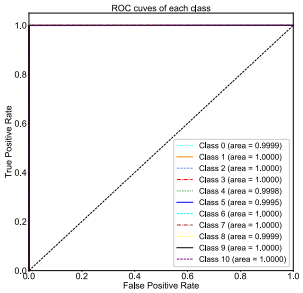


(i) UNSW NB-15 dataset with Yeo-Johnson + Standard Scaler

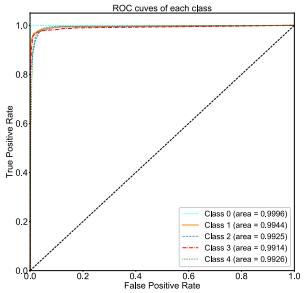


(j) Bot-IoT dataset with Yeo-Johnson + Standard Scaler

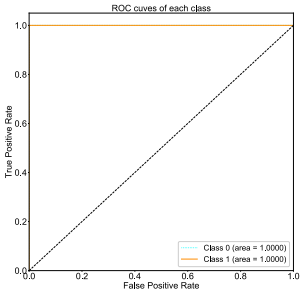
FIGURE 3. Classification matrix of highest accuracy normalization methods based on the RF-based IDS model.



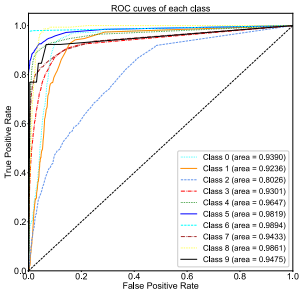
(a) CIC-IDS 2017 dataset with Yeo-Johnson



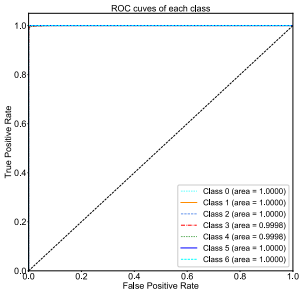
(b) ISCX-IDS 2012 dataset with Yeo-Johnson



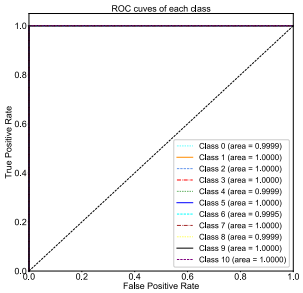
(c) NSL-KDD dataset with Yeo-Johnson



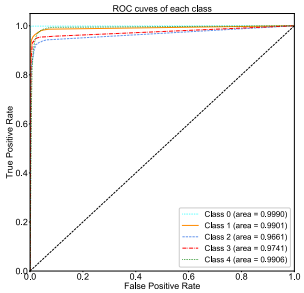
(d) UNSW NB-15 dataset with Yeo-Johnson



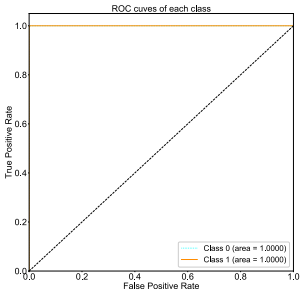
(e) Bot-IoT dataset with Yeo-Johnson



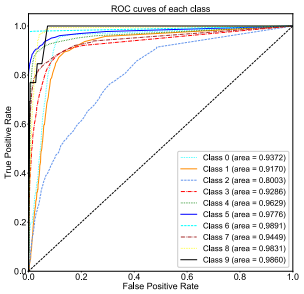
(f) CIC-IDS 2017 dataset with Yeo-Johnson + Min-Max



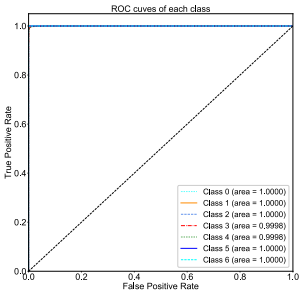
(g) ISCX-IDS 2012 dataset with Yeo-Johnson + Min-Max



(h) NSL-KDD dataset with Yeo-Johnson + Standard Scaler



(i) UNSW NB-15 dataset with Yeo-Johnson + Standard Scaler



(j) Bot-IoT dataset with Yeo-Johnson + Standard Scaler

FIGURE 4. ROC (receiver operating characteristic curve) of highest accuracy normalization methods based on the RF-based IDS model.

TABLE 12. Kappa coefficient score of the SVM-based IDS on highest accuracy normalizations.

Datasets	Normalization	Kappa Score	Hybrid Normalization	Kappa Score
CIC-IDS 2017	Yeo-Johnson	0.6831	Yeo-Johnson + Min-Max	0.6234
ISCX-IDS 2012	Yeo-Johnson	0.9099	Yeo-Johnson + Min-Max	0.9109
NSL-KDD	Yeo-Johnson	0.9991	Yeo-Johnson + Standard Scaler	0.9877
UNSW NB-15	Yeo-Johnson	0.6451	Yeo-Johnson + Standard Scaler	0.6451
Bot-IoT	Yeo-Johnson	0.6477	Yeo-Johnson + Standard Scaler	0.6277

TABLE 13. Classification results of the DNN-based IDS on all three datasets and normalizations.

Dataset	CIC-IDS 2017 dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	Min-Max+L2	L2+Yeo-Johnson	Yeo-Johnson+Min-Max	L2+Standard
Accuracy	98.90%	94.36%	98.76%	98.18%	99.75%	98.69%	99.58%	99.57%	98.87%
F1-Score	94.65%	81.25%	93.05%	90.73%	97.80%	94.12%	97.25%	96.29%	93.12%
Precision	91.54%	85.38%	89.50%	85.88%	96.28%	90.75%	95.36%	93.84%	89.61%
Recall	99.34%	86.94%	98.85%	98.47%	99.64%	98.85%	99.62%	99.65%	98.82%
Dataset	ISCX-IDS 2012 dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	Min-Max+L2	L2+Yeo-Johnson	Yeo-Johnson+Min-Max	L2+Min-Max
Accuracy	94.35%	90.47%	94.46%	91.68%	95.37%	93.11%	94.13%	95.16%	93.10%
F1-Score	92.34%	86.57%	92.30%	87.69%	93.42%	90.42%	91.44%	92.84%	90.79%
Precision	92.52%	84.92%	91.68%	87.70%	93.27%	91.05%	90.33%	92.15%	90.40%
Recall	92.24%	90.51%	93.16%	89.12%	93.62%	90.12%	92.72%	93.60%	91.20%
Dataset	NSL-KDD dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	Min-Max+L2	Yeo-Johnson+Standard	Yeo-Johnson+Robust	Standard+L2
Accuracy	98.95%	99.15%	99.36%	99.07%	99.75%	98.86%	99.74%	99.72%	99.09%
F1-Score	98.94%	99.15%	99.36%	99.06%	99.75%	98.85%	99.74%	99.72%	99.09%
Precision	98.93%	99.13%	99.36%	99.10%	99.75%	98.85%	99.74%	99.72%	99.07%
Recall	98.96%	99.16%	99.36%	99.03%	99.74%	98.86%	99.73%	99.71%	99.12%
Dataset	UNSW NB-15 dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	Min-Max+L2	Yeo+Min-Max	Yeo+Standard	L2+Standard
Accuracy	71.21%	69.23%	70.64%	26.94%	75.24%	69.17%	73.43%	75.20%	42.98%
F1-Score	52.04%	51.53%	51.32%	19.44%	54.64%	50.69%	53.89%	54.90%	31.51%
Precision	52.32%	52.04%	50.97%	27.12%	52.99%	51.40%	52.81%	53.69%	40.39%
Recall	67.53%	64.93%	62.45%	35.54%	64.79%	68.40%	69.82%	64.88%	43.88%
Dataset	Bot-IoT dataset								
Normalization	Min-Max	Robust	Standard	L2	Yeo-Johnson	Min-Max+L2	Yeo+Min-Max	Yeo+Standard	L2+Standard
Accuracy	96.51%	97.55%	97.27%	66.12%	98.57%	96.86%	97.77%	98.67%	95.88%
F1-Score	92.77%	92.42%	95.49%	66.58%	94.11%	93.35%	92.54%	94.46%	91.47%
Precision	90.07%	89.00%	93.89%	63.82%	91.15%	90.53%	89.26%	91.73%	88.51%
Recall	96.92%	97.56%	97.37%	78.89%	98.17%	96.98%	97.69%	98.06%	95.44%

TABLE 14. Kappa coefficient score of the DNN-based IDS on highest accuracy normalizations.

Datasets	Normalization	Kappa Score	Hybrid Normalization	Kappa Score
CIC-IDS 2017	Yeo-Johnson	0.9967	Yeo-Johnson + Min-Max	0.9930
ISCX-IDS 2012	Yeo-Johnson	0.9357	Yeo-Johnson + Min-Max	0.9329
NSL-KDD	Yeo-Johnson	0.9954	Yeo-Johnson + Standard Scaler	0.9948
UNSW NB-15	Yeo-Johnson	0.7020	Yeo-Johnson + Standard Scaler	0.7016
Bot-IoT	Yeo-Johnson	0.9814	Yeo-Johnson + Standard Scaler	0.9827

ization. Figure 5 (f) and (g) represent the classification matrix of the datasets CIC-IDS 2017 and ISCX-IDS 2012 based on Yeo-Johnson + Min-Max normalizations respectively. Figure 5 (h), (i), and (j) represent the classification matrix of the NSL-KDD, UNSW NB-15, and BoT-IoT dataset based on Yeo-Johnson + Standard scaler normalizations respectively.

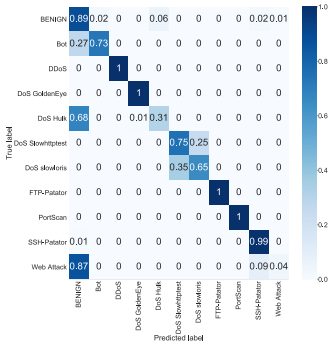
The ROC curves for each dataset classified by SVM-based IDS are presented in Figure 6. Each colored line in the ROC graph represents a class in a dataset. Figures 6 (a), (b), (c), (d), and (e) represent the ROC of the datasets CIC-IDS 2017, ISCX-IDS 2012, NSL-KDD, UNSW NB-15, and Bot-IoT based on Yeo-Johnson normalization. Figure 6 (f) and (g) represent the ROC of the datasets CIC-IDS 2017 and ISCX-IDS 2012 based on Yeo-Johnson + Min-Max normalizations respectively. Figures 6 (h), (i), and (j) represent the ROC of the NSL-KDD, UNSW NB-15, and Bot-IoT dataset based on Yeo-Johnson + Standard scaler normalizations respectively.

C. DEEP NEURAL NETWORK-BASED IDS MODEL

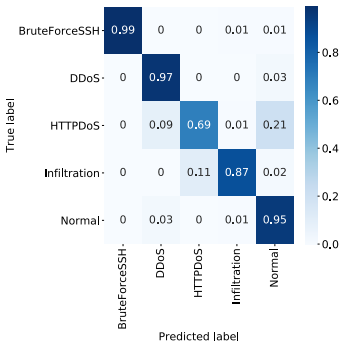
The DNN-based IDS model implemented in this paper is based on our earlier work [8]. Table 13 represents the classification results achieved by the DNN-based IDS model. The DNN-based IDS model validated the Yeo-Johnson normalization as the most suitable normalization method for all five datasets as predicted by the proposed model (i.e. Table 8). On the other hand, in hybrid normalization for CIC-IDS 2017 dataset, L2 normalization + Yeo-Johnson achieved the highest accuracy rather than Yeo-Johnson + Min-Max. For datasets, ISCX-IDS 2012, NSL-KDD, UNSW NB-15, and Bot-IoT the highest accuracy was achieved by the normalization method identified by the proposed statistical method as highlighted in Table 8.

Based on Equation 16, the Kappa coefficient score was computed for all five datasets. Table 14 represents the Kappa score of the normalization method which achieved the highest accuracy based on the DNN-based IDS.

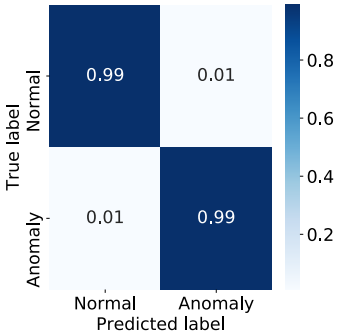
To visualize the classification performance of DNN-based IDS, Figure 7 represents the classification matrix of each dataset with both single and hybrid transformations. Even though achieving high accuracy is not the main purpose of this research but, the DNN-based IDS was able to perform with good accuracy on each dataset excluding the UNSW



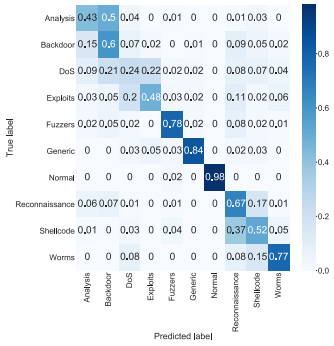
(a) CIC-IDS 2017 dataset with Yeo-Johnson



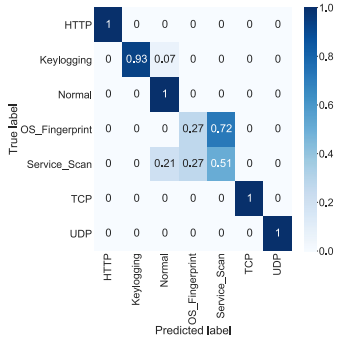
(b) ISCX-IDS 2012 dataset with Yeo-Johnson



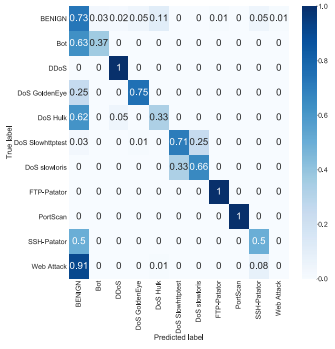
(c) NSL-KDD dataset with Yeo-Johnson



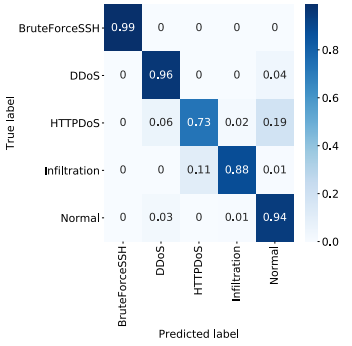
(d) UNSW NB-15 dataset with Yeo-Johnson



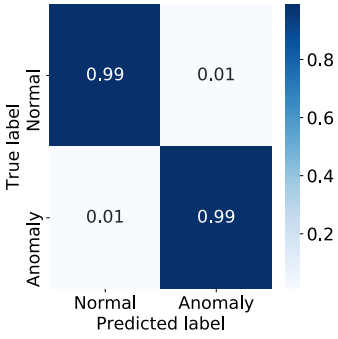
(e) Bot-IoT dataset with Yeo-Johnson



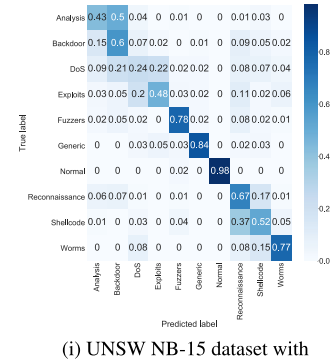
(f) CIC-IDS 2017 dataset with Yeo-Johnson + Min-Max



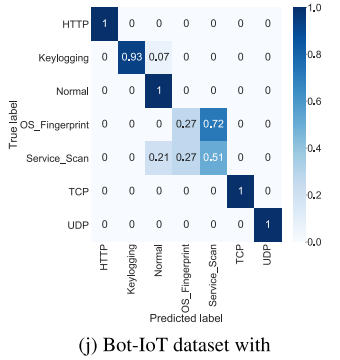
(g) ISCX-IDS 2012 dataset with Yeo-Johnson + Min-Max



(h) NSL-KDD dataset with Yeo-Johnson + Standard Scaler

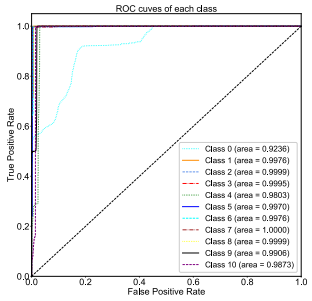


(i) UNSW NB-15 dataset with Yeo-Johnson + Standard Scaler

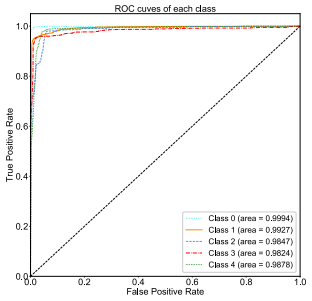


(j) Bot-IoT dataset with Yeo-Johnson + Standard Scaler

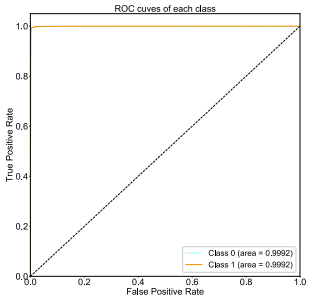
FIGURE 5. Classification matrix of highest accuracy normalization methods based on the SVM-based IDS model.



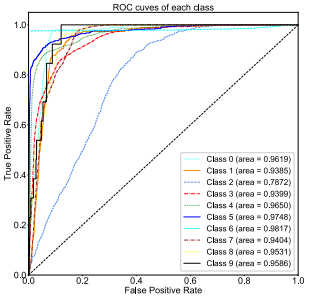
(a) CIC-IDS 2017 dataset with Yeo-Johnson



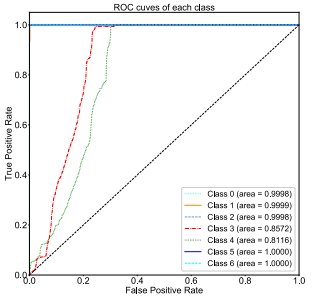
(b) ISCX-IDS 2012 dataset with Yeo-Johnson



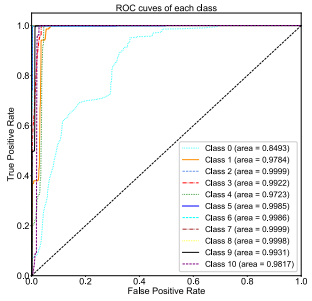
(c) NSL-KDD dataset with Yeo-Johnson



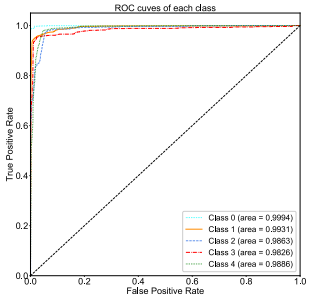
(d) UNSW NB-15 dataset with Yeo-Johnson



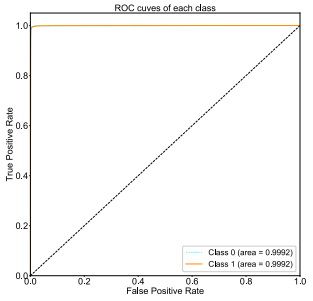
(e) Bot-IoT dataset with Yeo-Johnson



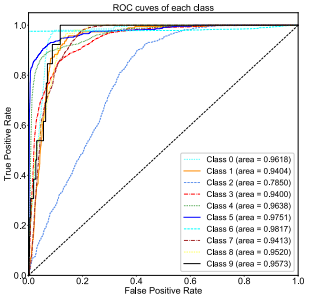
(f) CIC-IDS 2017 dataset with Yeo-Johnson + Min-Max



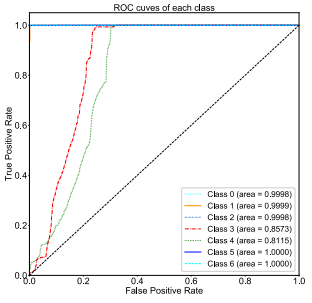
(g) ISCX-IDS 2012 dataset with Yeo-Johnson + Min-Max



(h) NSL-KDD dataset with Yeo-Johnson + Standard Scaler

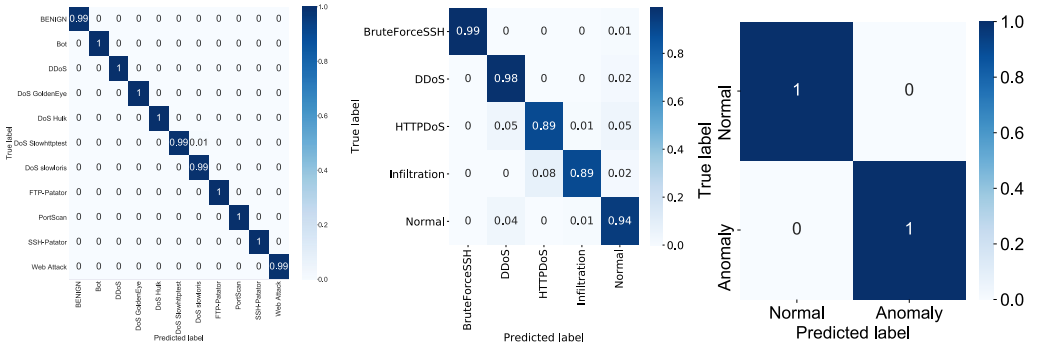


(i) UNSW NB-15 dataset with Yeo-Johnson + Standard Scaler



(j) Bot-IoT dataset with Yeo-Johnson + Standard Scaler

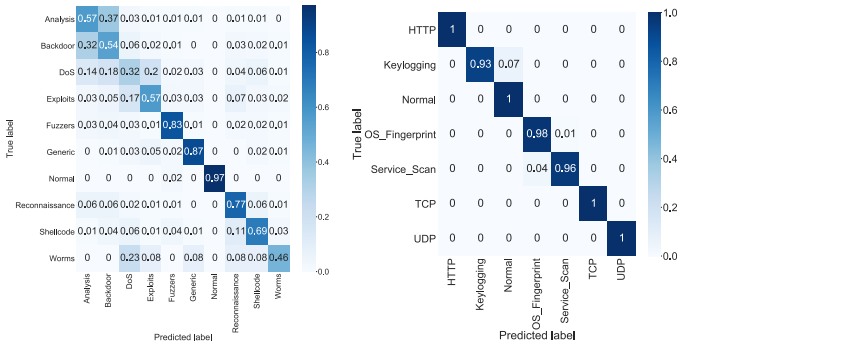
FIGURE 6. **ROC (receiver operating characteristic curve) of highest accuracy normalization methods based on SVM-based IDS model.**



(a) CIC-IDS 2017 dataset with Yeo-Johnson

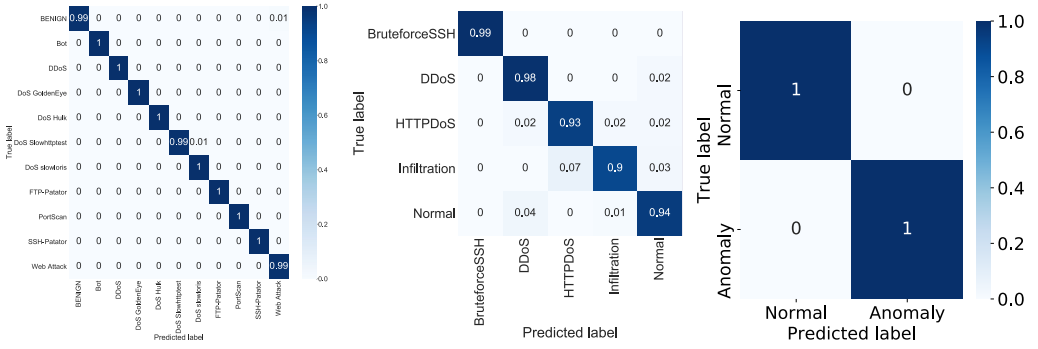
(b) ISCX-IDS 2012 dataset with Yeo-Johnson

(c) NSL-KDD dataset with Yeo-Johnson



(d) UNSW NB-15 dataset with Yeo-Johnson

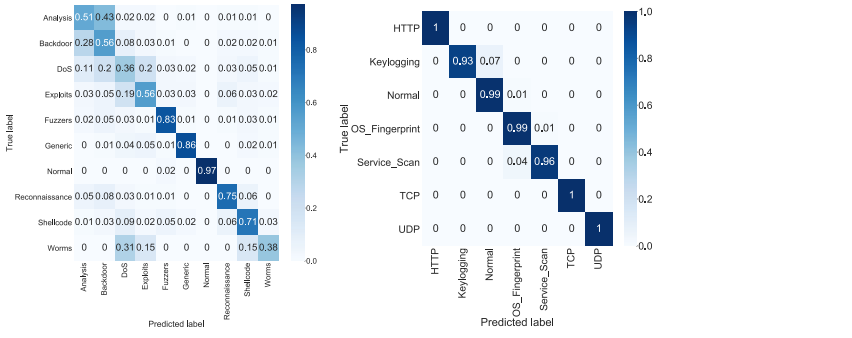
(e) Bot-IoT dataset with Yeo-Johnson



(f) CIC-IDS 2017 dataset with Yeo-Johnson + Min-Max

(g) ISCX-IDS 2012 dataset with Yeo-Johnson + Min-Max

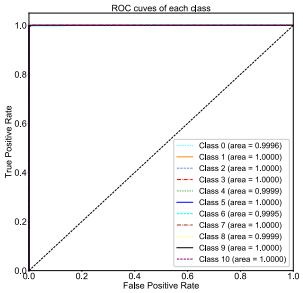
(h) NSL-KDD dataset with Yeo-Johnson + Standard Scaler



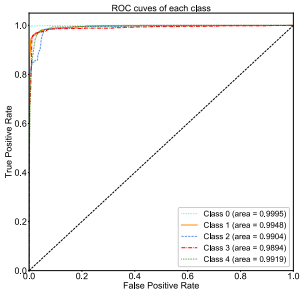
(i) UNSW NB-15 dataset with Yeo-Johnson + Standard Scaler

(j) Bot-IoT dataset with Yeo-Johnson + Standard Scaler

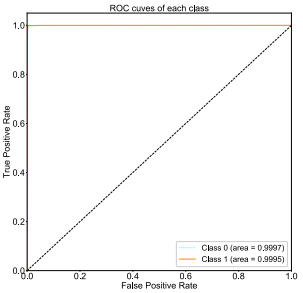
FIGURE 7. Classification matrix of highest accuracy normalization methods based on the DNN-based IDS model.



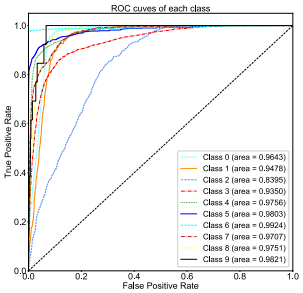
(a) CIC-IDS 2017 dataset with Yeo-Johnson



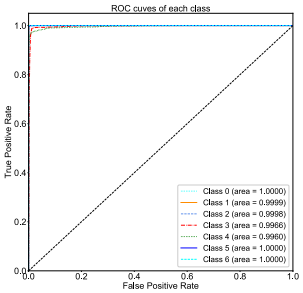
(b) ISCX-IDS 2012 dataset with Yeo-Johnson



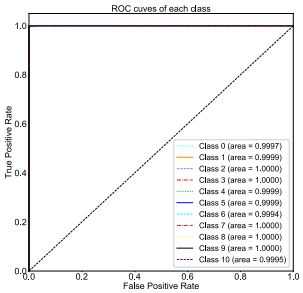
(c) NSL-KDD dataset with Yeo-Johnson



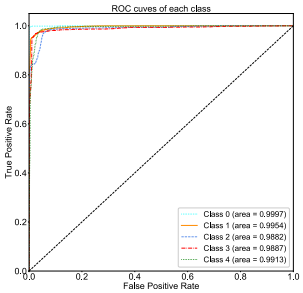
(d) UNSW NB-15 dataset with Yeo-Johnson



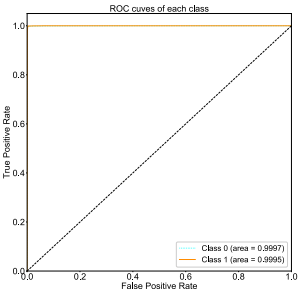
(e) Bot-IoT dataset with Yeo-Johnson



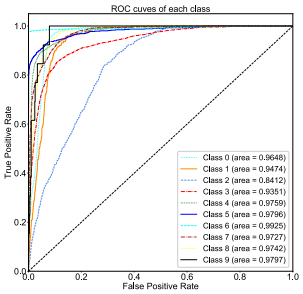
(f) CIC-IDS 2017 dataset with Yeo-Johnson + Min-Max



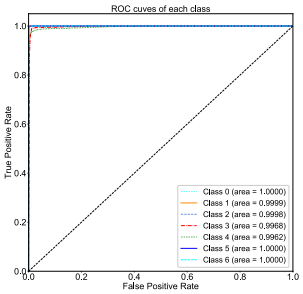
(g) ISCX-IDS 2012 dataset with Yeo-Johnson + Min-Max



(h) NSL-KDD dataset with Yeo-Johnson + Standard Scaler



(i) UNSW NB-15 dataset with Yeo-Johnson + Standard Scaler



(j) Bot-IoT dataset with Yeo-Johnson + Standard Scaler

FIGURE 8. **ROC (receiver operating characteristic curve) of highest accuracy normalization methods based on the DNN-based IDS model.**

TABLE 15. Comparison and discussion on papers adopting different methods for identifying dataset normalization.

Reference	Method to identify Normalization	Normalization Methods	Comparison
[11]	Visual graphs i.e. QQ-plot, heatmap, and graphs	A combination of the objective function, the rectified Box-Cox and Yeo-Johnson transforms	Concern: Missing Validation method i.e. classification, regression, etc. In this article, multiple classifiers and datasets are implemented to validate the identified normalization method.
[20]	Statistics i.e. Jarque-Bera	No normalization applied to the self-generated dataset	Concern: The selected method was not able to identify normality in presence of outlying attributes. The method was yet to be tested with a small dataset. On the other hand in this article, the proposed method is validated on datasets containing outlying attributes and a comparatively small dataset i.e. NSL-KDD.
[58]	Applied all normalization methods one by one	Normalization function, z-score, Categorical transformation	Concern: Suitable normalization was identified based on classification results. Such an approach is quite extensive, as each normalization is applied and classification is performed to identify the suitable normalization method. However, in this article, the suggested method identifies the most suitable normalization without performing classification.
[59]	Applied all normalization methods one by one	MinMax (0,1), MinMax (-1,1), Standardization (0,1), Standardization (-1,1)	Concern: Suitable normalization was identified based on classification results. Such an approach is quite extensive, as each normalization is applied and classification is performed to identify the suitable normalization method. Experimentation was done on one dataset. In contrast, the proposed method in this article was applied on five different datasets with some well-known normalization methods. The proposed model was also able to identify the most suitable normalization without performing classification.
[60]	Jarque-Bera (JB), Robust Jarque-Bera (RJB), Shapiro–Wilk (SW), Kolmogorov–Smirnov (KS), Lilliefors (LF), Anderson–Darling (AD)	Self-modification	Concern: Use of only one dataset, self-modification to introduce non-normalization in the dataset, and the absence of a validation method i.e. classification, regression, etc. While in this article, the proposed method is implemented on five different datasets with a comprehensive validation method.

NB-15 dataset. Figure 7 (a), (b), (c), (d), and (e) represent the classification matrix of the datasets CIC-IDS 2017, ISCX-IDS 2012, NSL-KDD, UNSW NB-15, and Bot-IoT based on Yeo-Johnson normalization. Figure 7 (f) and (g) represent the classification matrix of the datasets CIC-IDS 2017 and ISCX-IDS 2012 based on Yeo-Johnson + Min-Max normalizations respectively. Figure 7 (h), (i), and (j) represent the classification matrix of the NSL-KDD, UNSW NB-15, and Bot-IoT datasets based on Yeo-Johnson + Standard scaler normalization.

The ROC curves for each dataset classified by DNN-based IDS are presented in Figure 8. Each colored line in the ROC graph represents a class in a dataset. Figure 8 (a), (b), (c), (d), and (e) represent the ROC of the datasets CIC-IDS 2017, ISCX-IDS 2012, NSL-KDD, UNSW NB-15, and Bot-IoT based on Yeo-Johnson normalization. Figure 8 (f) and (g) represent the ROC of the datasets CIC-IDS 2017 and ISCX-IDS 2012 based on Yeo-Johnson + Min-Max normalizations respectively. Figure 8 (h), (i), and (j) represent the ROC of the NSL-KDD, UNSW NB-15, and Bot-IoT dataset based on Yeo-Johnson + Standard scaler normalizations respectively.

VII. DISCUSSION

Conventionally the Min-Max and standardization are considered as the most common normalization methods [56]. Even though the general application of the mentioned method can be questioned. In comparison to the existing methods highlighted in related work and Table 15, the study conducted in this paper is much more comprehensive and generally applicable for improving an ML-based IDS. To corroborate the generalization of the proposed model, five different datasets with two different feature selection methods were used. The datasets contained both numeric and nominal features. After applying the proposed statistical method, the most suitable normalization method for the datasets was highlighted in the form of ranks with the help of the Rank and Percentile method. To validate the proposed model, three different ML-based IDS were implemented. Based on the validation procedure, the normalization methods identified by the proposed statistical model achieved higher accuracy as compared to the other normalization methods. On the other hand, not all hybrid normalizations were identified successfully. The proposed model successfully identified eighteen hybrid normalizations out of the twenty hybrid normalizations. However, it is possible to further improve hybrid normalization detection by testing a few more potential combinations of normalizations. Some of the hybrid normalizations were even able to achieve improved classification results as compared to the single normalization methods. The reason behind implementing the hybrid normalization method was to check the ability of the proposed model to identify non-standard normalization methods. Therefore, researchers can also compare newly proposed normalization methods with existing standardized methods. As the proposed method deals with a very specific aspect of the ML pre-processing chain, it can also be used to improve domains other than security. Such areas may include network traffic classification using ML [57], ML for low power devices [2], etc. as the proposed model is computationally efficient. The computational complexity

of our proposed algorithm can be presented as follow:

$$\Theta(k \cdot n \cdot m \log(m)) \tag{17}$$

where k, n, m are dataset size, feature number, and the number of attributes normalized. Noting that $m < n < k$, we can say that the complexity is defined by 'k' only. Which means that the proposed algorithm is very efficient in term of computation complexity as it is 1st order of the polynomial, i.e. $\Theta(k)$. In contrast, other normality testing algorithms such as Q-Q plot, D'Agostino's K-squared test, Jarque-Bera test, etc, cannot have a computational complexity lower than $\Theta(k)$. Table 15 represents a comprehensive comparison between the existing methods to identify suitable normalization methods and the proposed algorithm.

VIII. CONCLUSION

The rising rate of complex attacks on networks has truly tested the limitations of network security. The ML-based IDS can play an integral part in providing enhanced security measures. Normalization, which is a part of pre-processing, plays an important part in improving ML-Based IDS. While identifying which normalization method is suitable for the data at hand is quite challenging. In this study, a statistical model to identify the most suitable normalization method to enhance the performance of ML-based IDS is proposed. The proposed model is agile and does not require high computation. The proposed model used a matrix of mean, median, and skewness with the Rank and Percentile method to identify the most suitable normalization method for the selected dataset. To validate the proposed statistical model three classifiers are also implemented. Based on the validation results the proposed model was able to identify the most suitable normalization method with high accuracy. Such statistical methods open opportunities for researchers to further improve existing methods to assist pre-processing methods to improve ML-based IDS. For future research, we are looking to further improve the algorithm's ability to identifying hybrid normalizations and to identify the most suitable combination of normalization methods that can improve the performance of an ML-based IDS.

REFERENCES

[1] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 1, pp. 41–50, Feb. 2018.

[2] G. Bovenzi, G. Aceto, D. Ciuonzo, V. Persico, and A. Pescape, "A hierarchical hybrid intrusion detection approach in IoT scenarios," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, Dec. 2020, pp. 1–7.

[3] A. Chopra, S. Behal, and V. Sharma, "Evaluating machine learning algorithms to detect and classify DDoS attacks in IoT," in *Proc. 8th Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, New Delhi, India, Mar. 2021, pp. 517–521.

[4] K. Albulayhi and F. T. Sheldon, "An adaptive deep-ensemble anomaly-based intrusion detection system for the Internet of Things," in *Proc. IEEE World AI IoT Congr. (AIIoT)*, Seattle, WA, USA, May 2021.

[5] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An ensemble of autoencoders for online network intrusion detection," 2018, *arXiv:1802.09089*. [Online]. Available: https://arxiv.org/abs/1802.09089

[6] M. A. Siddiqi and W. Pak, "Efficient filter based feature selection flow for intrusion detection system," in *Proc. Int. Workshop Emerg. (ICT)*, Gyeongsan, South Korea, 2020, doi: 10.3390/electronics9122114.

[7] S. Sarvari, N. F. M. Sani, Z. M. Hanapi, and M. T. Abdullah, "An efficient anomaly intrusion detection method with feature selection and evolutionary neural network," *IEEE Access*, vol. 8, pp. 70651–70663, 2020.

[8] M. A. Siddiqi and W. Pak, "Optimizing filter-based feature selection method flow for intrusion detection system," *Electronics*, vol. 9, no. 12, p. 2114, Dec. 2020.

[9] X. Larriva-Novo, M. Vega-Barbas, V. A. Villagrá, D. Rivera, M. Álvarez-Campana, and J. Berrocal, "Efficient distributed preprocessing model for machine learning-based anomaly detection over large-scale cybersecurity datasets," *Appl. Sci.*, vol. 10, no. 10, p. 3430, May 2020.

[10] D. Suprina. (Jan. 7, 2013). *The Importance of Data Normalization in IPS*. Help Net Security. Accessed: May 3, 2021. [Online]. Available: https://www.helpnetsecurity.com/2013/01/07/the-importance-of-data-normalization-in-ips/

[11] J. Raymaekers and P. J. Rousseeuw, "Transforming variables to central normality," *Mach. Learn.*, Mar. 2021, doi: 10.1007/s10994-021-05960-5.

[12] T. M. Schendzielorz. (Jan. 16, 2020). *A guide to Data Transformation*. [Onli Available:. Accessed: May 3, 2021. [Online]. Available: https://medium.com/analytics-vidhya/a-guide-to-data-transformation-9e5fa9ae1ca3

[13] J. Brownlee. (Jul. 15, 2020). *How to Choose Data Preparation Methods for Machine Learning*. Machine Learning Mastery. Accessed: May 3, 2021. [Online]. Available: https://machinelearningmastery.com/choose-data-preparation-methods-for-machine-learning/

[14] M. Kuhn and K. Johnson, "Engineering numeric predictors," in *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1st ed. Oxfordshire, U.K.: Chapman, 2019.

[15] P. Huilgol. (Jul. 28, 2020). *Feature Transformation and Scaling Techniques to Boost Your Model Performance*. Accessed: Apr. 12, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/07/types-of-feature-transformation-and-scaling/

[16] S. Raschka. (Jul. 11, 2014). *About Feature Scaling and Normalization*. Accessed: Apr. 12, 2021. [Online]. Available: https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

[17] A. A. Salih and M. B. Abdulrazaq, "Combining best features selection using three classifiers in intrusion detection system," in *Proc. Int. Conf. Adv. Sci. Eng. (ICOASE)*, Zakho - Duhok, Iraq, Apr. 2019, pp. 94–99.

[18] J. M. Jo, "Effectiveness of normalization pre-processing of big data," *Korea Inst. Electron. Commun. Sci.*, vol. 14, no. 3, pp. 547–552, 2019.

[19] T. U. Islam, "Ranking of normality tests: An appraisal through skewed alternative space," *Symmetry*, vol. 11, no. 7, p. 872, Jul. 2019.

[20] G. Brys, M. Hubert, and A. Struyf, "A robustification of the Jarque-Bera test of normality," in *Proc. COMPSTAT*, Prague, Czech Republic, 2004, pp. 753–760.

[21] S. Simon. (Aug. 14, 2018). *Testing Normality Including Skewness and Kurtosis*. University of Cambridge. Accessed: Mar. 9, 2021. [Online]. Available: https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/Simon

[22] R. R. DeFilippi. (May 6, 2018). *Testing for Normality-Applications With Python*. Accessed: Mar. 3, 2021. [Online]. Available: https://medium.com/@rrfd/testing-for-normality-applications-with-python-6bf06ed646a9.

[23] J. Korstanje. (Dec. 13, 2019). *6 Ways to Test for a Normal Distribution-Which One to Use?*. Towards Data Science. Accessed: Mar. 3, 2021. [Online]. Available: https://towardsdatascience.com/6-ways-to-test-for-a-normal-distribution-which-one-to-use-9dcf47d8fa93.

[24] P. Nandakishore. (Nov. 19, 2019). *Normality Testing: The Graphical Way*. Towards Data Science. Accessed: Mar. 3, 2021. [Online]. Available: https://towardsdatascience.com/normality-testing-the-graphical-way-20902abd8543

[25] S. K. Haldar, "Chapter 9—Statistical and geostatistical applications in geology," in *Mineral Exploration*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2018, pp. 167–194.

[26] R. Bevans. (Jan. 7, 2021). *The p-Value Explained*. Accessed: Mar. 3, 2021. [Online]. Available: https://www.scribbr.com/statistics/p-value/

[27] S. Galli, "Performing Yeo-Johnson transformation on numerical variables," in *Python Feature Engineering Cookbook*. Birmingham, U.K.: Packt Publishing, 2020.

[28] G. Ciaburro, *Min-Max Normalization Regression Analysis*. Birmingham, U.K.: Packt Publishing, 2018.

[29] A. Davis, S. Gill, R. Wong, and S. Tayeb, "Feature selection for deep neural networks in cyber security applications," in *Proc. IEEE Int. IoT, Electron. Mechatronics Conf. (IEMTRONICS)*, Vancouver, BC, Canada, Sep. 2020, pp. 1–7.

[30] M. S. Yadav and R. Kalpana, "Data preprocessing for intrusion detection system using encoding and normalization approaches," in *Proc. 11th Int. Conf. Adv. Comput. (ICoAC)*, Chennai, India, Dec. 2019, pp. 265–269.

[31] Z. Khokhar and M. A. Siddiqi, "Machine learning based indoor localization using Wi-Fi and smartphone," *J. Independ. Stud. Res. Comput.*, vol. 18, no. 2, 2021, doi: 10.31645/06.

[32] J. Brownlee. (Aug. 28, 2020). *How to Scale Data With Outliers for Machine Learning*. Accessed: Apr. 8, 2021. [Online]. Available: https://machinelearningmastery.com/robust-scaler-transforms-for-machine-learning/

[33] J. Hale. (Mar. 4, 2019). *Scale, Standardize, or Normalize With Scikit-Learn*. Accessed: Apr. 8, 2021. [Online]. Available: https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02.

[34] R. Karim. (Dec. 27, 2018). *Intuitions on L1 and L2 Regularisation*. Accessed: Apr. 9, 2021. [Online]. Available: https://towardsdatascience.com/intuitions-on-l1-and-l2-regularisation-235f2db4c261#f810

[35] W. Wang, X. Zhang, S. Gombault, and S. J. Knapskog, "Attribute normalization in network intrusion detection," in *Proc. 10th Int. Symp. Pervas. Syst., Algorithms, Netw.*, Kaoshiung, Taiwan, Dec. 2009, pp. 448–453.

[36] B. Setiawan, I. T. S. Nopember, S. Djanali, T. Ahmad, I. T. S. Nopember, and I. T. S. Nopember, "Increasing accuracy and completeness of intrusion detection model using fusion of normalization, feature selection method and support vector machine," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 4, pp. 378–389, Aug. 2019.

[37] L. Yu, Y. Pan, and Y. Wu, "Research on data normalization methods in multi-attribute evaluation," in *Proc. Int. Conf. Comput. Intell. Softw. Eng.*, Wuhan, China, Dec. 2009, pp. 1–5.

[38] D. Kurniabudi, D. Stiawan, M. Y. B. Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 dataset feature analysis with information gain for anomaly detection," *IEEE Access*, vol. 8, pp. 132911–132921, 2020.

[39] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *Comput. Secur.*, vol. 31, no. 3, pp. 357–374, 2012.

[40] 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. Secur. Defense Appl.*, Ottawa, ON, Canada, Jul. 2009, pp. 1–6.

[41] S. Moualla, K. Khorzom, and A. Jafar, "Improving the performance of machine learning-based network intrusion detection systems on the UNSW-NB15 dataset," in *Computational Intelligence and Neuroscience*. Hindawi, 2021, p. 13, doi: 10.1155/2021/5557577.

[42] M. Shafiq, Z. Tian, A. K. Bashir, X. Du, and M. Guizani, "CorrAUC: A malicious bot-IoT traffic detection method in IoT network using machine-learning techniques," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3242–3254, Mar. 2021.

[43] Minitab. *Minitab Statistical Software*. Minitab. Accessed: Apr. 27, 2021. [Online]. Available: https://www.minitab.com/en-us/

[44] C. Lesmeister, "Zero and near-zero variance features," in *Mastering Machine Learning With R* 3rd ed. Birmingham, U.K.: Packt, 2019.

[45] E. Saccenti, M. H. W. B. Hendriks, and A. K. Smilde, "Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models," *Sci. Rep.*, vol. 10, no. 1, pp. 1–9, Dec. 2020.

[46] J. Tan. (Oct. 24, 2020). *Feature Selection for Machine Learning in Python-Wrapper Methods*. Accessed: Apr. 26, 2021. [Online]. Available: https://towardsdatascience.com/feature-selection-for-machine-learning-in-python-wrapper-methods-2b5e27d2db31

[47] A. Kumar. (Jul. 30, 2020). *Sequential Forward Selection—Python Example*. Accessed: Apr. 30, 2021. [Online]. Available: https://vitalflux.com/sequential-forward-selection-python-example/

[48] D. Yadav. (Dec. 7, 2019). *Categorical Encoding Using Label-Encoding and One-Hot-Encoder*. Towards Data Science. Accessed: Apr. 27, 2021. [Online]. Available: https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd.

[49] F. Ahemad. (Mar. 26, 2019). *Selecting the Right Metric for Skewed Classification Problems*. Toward Data Science. Accessed: Apr. 27, 2021. [Online]. Available: https://towardsdatascience.com/selecting-the-right-metric-for-skewed-classification-problems-6e0a4a6167a7

[50] M. F. M. Chowdhury and A. Lavelli, "Impact of less skewed distributions on efficiency and effectiveness of biomedical relation extraction," in *Proc. COLING*, Mumbai, India, 2012, pp. 205–216.

[51] M. Thakur. (2019). *Percentile Rank Formula*. Accessed: Apr. 28, 2021. [Online]. Available: https://www.educba.com/percentile-rank-formula/

[52] K. Pykes. (Feb. 27, 2020). *Cohen's Kappa: Understanding Cohen's Kappa coefficient*. Towards Data Science. Accessed: Apr. 29, 2021. [Online]. Available: https://towardsdatascience.com/cohens-kappa-9786ceceab58

[53] P. N. Tan, "Receiver operating characteristic," in *Encyclopedia of Database Systems*. Boston, MA, USA: Springer, 2009.

[54] S. M. Vieira, U. Kaymak, and J. M. C. Sousa, "Cohen's kappa coefficient as a performance measure for feature selection," in *Proc. Int. Conf. Fuzzy Syst.*, Barcelona, Spain, Jul. 2010, pp. 1–8.

[55] K. Feng, H. Hong, K. Tang, and J. Wang, "Decision making with machine learning and ROC curves," in *Proc. SSRN*, 2019, doi: 10.2139/ssrn.3382962.

[56] L. Ruff, R. K. Jacob, A. V. Robert, G. Montavon, W. Samek, M. Kloft, G. D. Thomas, and K. R. Müller, "A unifying review of deep and shallow anomaly detection," in *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, Feb. 2021.

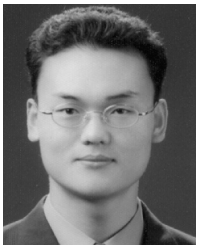
[57] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, "Mobile encrypted traffic classification using deep learning: Experimental evaluation, lessons learned, and challenges," *IEEE Trans. Netw. Service Manag.*, vol. 16, no. 2, pp. 445–458, Feb. 2019.

[58] X. Larriva-Novo, V. A. Villagrá, M. Vega-Barbas, D. Rivera, and M. Sanz Rodrigo, "An IoT-focused intrusion detection system approach based on preprocessing characterization for cybersecurity datasets," *Sensors*, vol. 21, no. 2, p. 656, Jan. 2021.

[59] V. Jan and H. Martin, "Evaluation of data preprocessing techniques for anomaly detection systems in industrial control system," in *Proc. 30th DAAAM Int.*, Vienna, Austria, 2019, pp. 1–8.

[60] S. Rana, N. N. Eshita, and A. S. M. A. Mamun, "Robust normality test in the presence of outliers," in *Proc. Int. Conf. Math., Statist. Data Sci. (ICMSDS)*, Bogor, Indonesia, 2021, doi: 10.1088/1742-6596/1863/1/012009.

MURTAZA AHMED SIDDIQI received the B.S. degree from Greenwich University, Pakistan, and the M.S. degree from Mohammad Ali Jinnah University, Pakistan. He is currently pursuing the Ph.D. degree with the Advance Intelligent and Secure System (AISS) Laboratory, Department of Information and Communication Engineering, Yeungnam University, Republic of Korea. He is also serving as an Assistant Professor with the Computer Science Department, Sukkur IBA University, Pakistan. His research interests include network and system security, network intrusion detection based on machine learning, and data analysis.



WOOGUIL PAK (Member, IEEE) received B.S. and M.S. degrees in electrical engineering and the Ph.D. degree in electrical engineering and computer science from Seoul National University, in 1999, 2001, and 2009, respectively. In 2010, he joined Jangwee Research Institute for National Defence, as a Research Professor, and Keimyung University, Daegu, South Korea, in 2013. Since 2019, he has been an Associate Professor with Yeungnam University, Gyeongsan, South Korea. His current research interests include network and system security, blockchain, and real-time network intrusion prevention based on machine learning for over 1Tbps networks.