International Journal of Safety and Security Engineering
Vol. 12, No. 2, April, 2022, pp. 239-249
Journal homepage: http://iieta.org/journals/ijsse

Detection of TCP-Based DDoS Attacks with SVM Classification with Different Kernel Functions Using Common Uncorrelated Feature Subsets

Kishore Babu Dasari[1*], Nagaraju Devarakonda[2]

[1] Department of CSE, Acharya Nagarjuna University, Guntur 522510, Andhra Pradesh, India
[2] School of Computer Science and Engineering, VIT-AP University, Amaravati 522237, India

Corresponding Author Email: dasari2kishore@gmail.com

https://doi.org/10.18280/ijsse.120213

Received: 1 February 2022
Accepted: 3 April 2022

Keywords:
CICDDoS2019, DDoS attacks, Kendall correlation, MSSQL, Pearson correlation, spearman correlation, SSDP, support vector machine, Syn flood

ABSTRACT

Distributed Denial of Service (DDoS) is a server-side infrastructure type security attack that aims to prevent legitimate users from accessing server system resources. Huge financial losses, reputation damage and data theft are some of the serious circumstances of DDoS attacks. Available DDoS attack detection methods reduce the severity of the attack's consequences, but they require more data computation, which is more expensive. This research proposed two feature selection methods in order to reduce the data computation for TCP-based DDoS attack detection with Support Vector Machine (SVM) classification algorithm. The first feature selection proposal of this study is to use Pearson, Spearman, and Kendall correlation approaches to select the PSK common uncorrelated feature subset. Use these PSK common uncorrelated feature subsets with SVM classifier with different kernels on TCP-based DDoS attacks and evaluate the classification results. This research, performed operations on Syn flood, MSSQL, SSDP datasets have taken from the CIC-DDoS2019 evaluation dataset. Select TCP-based DDoS attacks common uncorrelated feature subset selected by applying intersection on Syn flood, MSSQL, and SSDP data sets PSK common uncorrelated feature subsets is the second feature selection proposal of this research. Use these TCP-based DDoS attacks common uncorrelated feature subsets with SVM classifier with different kernels on TCP-based DDoS attacks and evaluate the classification results. Results with these two proposed methods also compared in this study. Experiments have been performed with these two approaches on a customized TCP-based DDoS attack that's been developed with Syn flood, MSSQL, and SSDP data sets, and the results have been evaluated. Linear, rbf, poly, sigmoid kernels SVM kernels used in this research. Experiments conclude that SVM with rbf kernel produces better results on TCP-based DDoS attacks.

1. INTRODUCTION

Internet usage increased exponentially during the coronavirus outbreak, it became, the internet has been essential and very important to everyone's daily life. The number of cyber attacks increases proportionately with the number of people using the internet. Reputational damage, data theft, and financial losses are major consequences of cyber attacks. Denial of Service (DoS) is one of the cyber attack, it is a malicious attack which makes system or network unavailable to its intended users. Multiple compromised systems attack a target and cause a denial of service for legitimate users of the targeted resource, such as a server, website, or other network resources is called a Distributed Denial of Service (DDoS) attack [1]. Attackers can use counterfeit traffic from a DDoS attack to overwhelm a business server, causing losses anywhere from $8,000 to $74,000 per hour of downtime, and then steal data while the business is distracted. It shows that the importance of DDoS attack detection. Early detection of DDoS attack is essential to reduce the huge circumstances of the attack. Conventional detection techniques of DDoS attacks require huge computing, it is very expensive. This research proposed two feature selection methods with Support Vector Machine (SVM) to detect TCP-based DDoS attacks. It is computationally cost-effective and detect the attack very fast. Syn flood, MSSQL and SSDP DDoS attacks are TCP-based DDoS attacks.

A syn flood [2] exploits the handshake process of a TCP connection to make a server unavailable to legitimate users by using all available server resources. In order to establish a connection, a TCP connection goes through three distinct stages under normal circumstances. To establish a connection, the client first sends a SYN packet to the server. The server acknowledges the communication by sending a SYN/ACK packet in response to the initial packet. Finally, the client sends an ACK packet to the server to acknowledge receipt of the message. The TCP connection is open and ready to send and receive data when this sequence of packet sending and receiving is completed. An attacker exploits these sequences of handshaking processes to create a denial-of-service attack. The attacker sends a huge number of SYN packets to the targeted server with spoofed IP addresses. The server then acknowledges to each connection request and leaves an open port waiting for the response. The attacker continues to transmit SYN packets while the server waits for the last packet ACK, which never arrives. Each SYN packet leads the server to open a new open port connection for a set amount of time, and after all of the available ports have been used, the server becomes unable to function normally.

An MSSQL attack [3] is made to exploit the Microsoft SQL Server Resolution Protocol when a Microsoft SQL Server responds to a client query or request. When a client requests information from an MS SQL Server, the SQL Resolution Protocol is employed. When a client connects to a database server, the server responds with a list of database instances using the MS SQL Resolution protocol, which helps the client figure out which database instances they're trying to connect to. Attackers can take advantage of SQL servers by running controlled requests from a spoofed IP address that appears to originate from the target server.

A reflection DDoS attack using the Universal Plug and Play (UPnP) network protocols to send an amplified traffic stream to the victim's server is known as a simple service discovery protocol (SSDP) attacks [4]. The SSDP protocol is used to allow UPnP devices to broadcast their existence to other devices on the network in normal conditions. During the typical SSDP attack process, the attacker begins looking for plug-and-play devices that can be used as amplifiers. Identified and created a list of devices capable of responding to queries. The attacker sends a packet to a spoofed target's IP address. The attacker sends packets with misleading queries to every Plug and Play device via a botnet. As a result, each device responds to the designated target, the victim receives a massive amount of traffic from all devices and becomes exhausted and unable to process legitimate traffic.

This section introduces Syn flood, MSSQL and SSDP DDoS attacks. These DDoS attacks have a wide range of consequences, including financial and reputational consequences. The loss is reduced by detecting DDoS attack early stage. It drives DDoS attack detection research. The statistical methods for detecting DDoS attacks are effective, but model creation takes a long time. The machine learning method for detecting DDoS attacks has a high level of accuracy, but it requires more computation. Using feature selection, the data computation is reduced. Related works discussed in section 2 of this paper. Proposed methodologies and SVM classification algorithm explained in section 3 of this paper. Experimental results discussed in section 4. Both PSK and TCP-based common uncorrelated features produces good results with SVM rbf and poly kernels on TCP-based DDoS attacks. Section 5 contains conclusion of this research.

2. RELATED WORK

Statistical measures such as correlation, entropy, and covariance have been used to analyze network traffic in order to detect DDoS attacks. Xiao et al. [5] Proposed correlation-based DDoS attack detection method with KNN machine learning method. Experiments performed on KDD’99 dataset. Bahl and Dahiya [6] Proposed DDoS attacks detection with Random forest and Naive Bayes classifier using Pearson Correlation Feature selection. Experiments performed on NSL_KDD dataset. Wei et al. [7] Proposed DDoS attacks Detection with Spearman correlation. Singh and Shrivastava [8]. Proposed DDoS attack detection method with Kendall correlation method.

All proposals are proposed single correlation methods for feature selection to detect DDoS attack detection. This study proposed intersection of multiple correlation methods on dataset to select common uncorrelated feature subset. This study applies Pearson, Spearman and Kendall correlation methods for empirical research on TCP–based DDoS attack detection.

3. METHODOLOGY

3.1 Data set

This study uses CIC-DDoS2019 [9] DDoS attack evaluation data set which contains eleven DDoS attack datasets, each data set contains the corresponding attack and benign target class labels. Each data set contains millions of records. It is collected from the Canadian Institute for Cyber Security. This research focused on TCP-based DDoS attack detection, so collected TCP-based DDoS attack data sets of Syn flood, DDoS_MSSQL, DDoS_SSDP data sets from CIC-DDoS2019 data sets and performed operations on them.

3.2 Preprocessing

Data preprocessing is a set of processing steps to make a data suitable for machine learning algorithms. This study removes Unnamed: 0, Flow ID, Source IP, Source Port, Destination IP, Destination Port, Timestamp, SimillarHTTP features which vary from network to network and remove missing and infinite values records for cleaning the data. Encoding the attack and benign target labels with 0 and 1 respectively. Next apply standardization to the features to increase the efficiency of the classification algorithms on data.

3.3 Feature selection

Feature selection [10] is an essential step for machine learning classification algorithms to reduce the data computation and reduce the model training time. Intrinsic, Wrapper and Filter-based feature selection techniques are available now a days. Variance threshold and correlation filter-based feature selection techniques are applied in this study for feature selection.

3.4 Variance threshold

Variance threshold is a fast processing threshold based feature elimination method. It removes features which vary below a specific threshold value. Variance threshold considers the relationship of the feature itself in all records of the data set. It ignores the features link with target label. By default, it removes constant features which have the variance threshold value equal to zero, it means, that features have the same value in all records of the data.

3.5 Correlation

Correlation [11, 12] describes the relationship between two or more features. The correlation coefficient values vary between -1 to +1 to show the strength of the association between the features. The coefficient value is ± 1 indicates features has the strong correlation between the features. The coefficient value is 0 indicates features have strong uncorrelation. Pearson, Spearman and Kendall correlation techniques [8] are used in this study for finding uncorrelation features.

Pearson correlation coefficient calculated by:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$

where, r - rank of Pearson correlation coefficient;
x_i - independent-feature in a sample;
y_i - target or dependent-feature in a sample;
$\bar{x}$ – mean of the x-feature values;
$\bar{y}$ – mean of the y-feature values.

Spearman correlation coefficient calculated by

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{2}$$

where, ρ — rank of Spearman correlation coefficient;
d_i - difference between the two ranks of each observation;
n - number of observations.

Kendall correlation coefficient calculated by

$$\tau = \frac{N_c - N_d}{\frac{n(n-1)}{2}} \tag{3}$$

where, τ – rank of Kendall rank correlation coefficient;
N_c - number of concordant;
N_d - number of discordant.

3.6 Support vector machine

In Machine Learning, classification [13] is the problem of learning to distinguish records in a dataset that correspond to two or more target labels. The Support Vector Machine (SVM) is a simple and powerful machine-learning algorithm that finds a hyperplane in an N-dimensional space features to classify among label classes. Support Vectors are the data points with the shortest distance to the hyperplane. Due to the kernel functions that turns the input data space into a higher-dimensional space, SVMs are also known as kernelized SVMs [14, 15]. Linear, polynomiall, radial basis function (rbf), and sigmoid are the most common kernel functions.

The linear kernel function can be written as

$$k(x_i, x_j) = x_i * x_j \tag{4}$$

The polynomial kernel function can be written as

$$k(x_i, x_j) = (1 + x_i * x_j)^d \tag{5}$$

The radial basis function (rbf) kernel function can be written as

$$k(x_i, x_j) = \exp\left(-\gamma ||x_i - x_j||^2\right) \tag{6}$$

The sigmoid kernel function can be written as

$$k(x_i, x_j) = \tanh(\propto x^T y + c) \tag{7}$$

4. RESULTS AND DISCUSSION

In this research experiments are performed on data sets of Syn, MSSQL, and SSDP. Experiments also are conducted out on a customized TCP-based DDoS attack data set, which has been created by concatenating 40% of the total of each Syn flood, MSSQL, and SSDP data sets. After pre-processing, remove the constant features from the datasets which have variance threshold=0. Next, remove quasi-constant features from the data sets which have variance threshold=0.01. Table 1 shows the number of constant and quasi-constant features of the data sets. The amount of constant features in each data set, including customized data set, is the same, but the number of quasi-constant features vary.

Table 1. Number of constant and quasi-constant features of the data sets

Datasets	Number of Constant Features	Number of Quasi-constant Features
Syn	12	7
MSSQL	12	5
SSDP	12	8
Customized TCP-Based DDoS attack	12	5

Table 2. Number of correlated features of the data sets

	Correlation Methods		
Data Sets	**Pearson**	**Spearman**	**Kendall**
Syn	41	44	41
MSSQL	39	49	48
SSDP	34	44	43
Customized TCP-Based DDoS attack	39	50	47

Table 3. Number of PSK-common un-correlated features of the data sets

Data Set	Number of PSK-common un-correlated features
Syn	12
MSSQL	9
SSDP	12
Customized TCP-Based DDoS attack	10

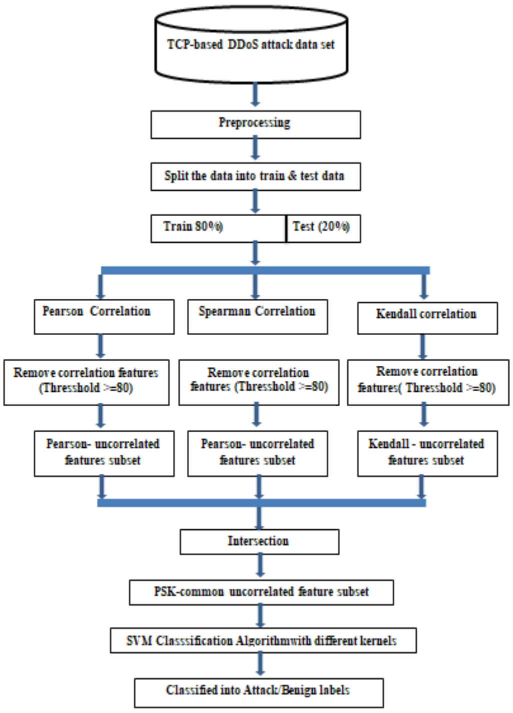
TCP-based DDoS attack data set
Preprocessing
Split the data into train & test data
Train 80%)
Test (20%)
Pearson Correlation
Spearman Correlation
Kendall correlation
Remove correlation features (Thresshold >=80)
Remove correlation features (Thresshold >=80)
Remove correlation features(Thresshold >=80)
Pearson- uncorrelated features subset
Pearson- uncorrelated features subset
Kendall - uncorrelated features subset
Intersection
PSK-common uncorrelated feature subset
SVM Classsification Algorithmwith different kernels
Classified into Attack/Benign labels

Figure 1. PSK. Common uncorrelated feature subset selection framework

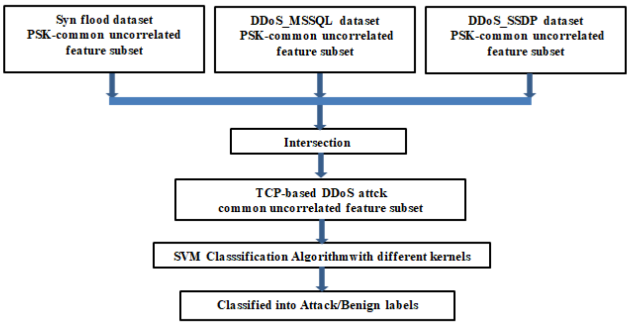

Figure 2. TCP-based DDoS attack common uncorrelated feature subset selection framework

After removing constant features, applied correlation algorithms in order to identify correlated features of the datasets. The data sets in this study are correlated using the Pearson, Spearman, and Kendall algorithms. In this research, correlation features are selected based on a threshold value of >=80. Table 2 shows the number of correlated features of the data sets including customized data set.

Remove correlation features which are selected by the Pearson correlation method for the data set features set, and create a Pearson un-correlation feature subset. Remove correlation features which are selected by the Spearman correlation method for the data set features set, and create a Spearman un-correlation feature subset. Remove features which are selected by the Kendall correlation method for the data set features set, and create a Kendall un-correlation feature subset. Create a PSK-common un-correlated feature subset by intersecting uncorrelated feature subsets of Pearson, Spearman, and Kendall correlation Techniques. Table 3 shows the number of PSK-common un-correlated features of the datasets. Table 4 shows the list of PSK-common un-correlated features of TCP-based DDoS attack data sets. Figure 1 and Figure 2 shows the PSK and TCP common un-correlated features subsets selection Frameworks respectively.

4.1 Evaluation metrics of classification algorithms

4.1.1 Confusion matrix

Confusion matrix is a square matrix is used to evaluate the performance of a classification algorithm. The actual values and predicted values of the target labels are represented by the columns and rows of the confusion matrix respectively. It contains True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values. TP indicates the actual Positive value predicted by the classification algorithm to be Positive. TN indicates the actual Negative value predicted by the classification algorithm to be Negative. FP indicates the actual Positive value predicted by the classification algorithm to be Negative. FN indicates the actual Negative value predicted by the classification algorithm to be Positive. The values of the confusion matrix are used to calculate the classification algorithms evaluation metrics.

4.1.2 Accuracy

The accuracy measure shows how many of the target labels are correctly predicted.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (8)$$

4.1.3 Precision

The precision measure shows the number of target label' predictions are correct out of the total predictions of the label. Precision of positive label is

$$Precision = \frac{TP}{TP + FP} \qquad (9)$$

4.1.4 Recall

The recall measure shows the number of target label' predictions are correct out of the total number of the label. Precision of positive label is

$$Recall = \frac{TP}{TP + FN} \qquad (10)$$

4.1.5 F1-score

F1 Score is harmonic mean of precision and recall.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (11)$$

4.1.6 Specificity

Specificity also called True Negative rate, is True Negative divided by sum of True Negative and False Positive.

$$Specificity = \frac{TN}{TN + FP} \qquad (12)$$

4.1.7 K-Fold cross validation

The K-Fold cross validation process is used to estimate the model's performance on a data set. The procedure is known as K-Fold cross validation because it only has one parameter, K, which specifies how many parts or folds the given data sample should be divided into. It selects one fold as a test set, the remaining folds as training sets, and then evaluates the model. This process is repeated until each fold has been designated as a test fold.

4.1.8 AUC-ROC curve

Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a measurement for evaluating the effectiveness of classification models. It is a probabilistic curve that plots the True Positive Rate against False Positive Rate (FPR) at different threshold values. If the AUC-ROC score is close to 1, the model is referred to as an excellent measure of target class separability. If the AUC-ROC score is close to 0, the model is referred to as the worst measure of target class separability. If the AUC-ROC score is close to 0.5, the model has no capacity for target class separation at all.

4.1.9 Log-loss

One of the most important measurements of error for evaluating the performance of a classification algorithm based on probabilities is log-loss. The log loss in a perfect model would be close to 0. The log loss in a worst model would be close to 1.

$$Log - loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \ln p_i + (1 - y_i)ln(1 - p_i)] \qquad (13)$$

Here, y indicates the actual value, p indicates the prediction probability and N indicates the number of observations.

Table 4. PSK-common un-correlated features lists of the data sets

Sr No	Feature Names			
	Syn Flood attacks	**MSSQL attack**	**SSDP attack**	**Customized Exploitation DDoS attack**
1	Protocol	Protocol	Protocol	Protocol
2	Active Mean	Active Mean	Active Mean	
3	Total Length of Bwd Packets	Total Length of Bwd Packets	Total Length of Bwd Packets	Total Length of Bwd Packets
4	Total Length of Fwd Packets	Total Length of Fwd Packets	Total Length of Fwd Packets	
5	Total Fwd Packets	Total Fwd Packets	Total Fwd Packets	Fwd Packet Length Std
6	Flow Duration	Flow Duration	Flow Duration	Bwd Packet Length Min
7	min_seg_size_forward	Fwd Header Length	Fwd Header Length	Fwd Header Length
8	Total Backward Packets	Total Backward Packets	Bwd Packet Length Min	Total Backward Packets
9	Down/Up Ratio	Fwd Packet Length Std	Fwd Packet Length Std	Down/Up Ratio
10	Flow Packets/s		Fwd Packet Length Max	Inbound
11	Flow IAT Min		Flow IAT Min	Flow IAT Min
12	Active Min		Active Std	min_seg_size_forward

Table 5. Precision, recall, F1-score, specificity and accuracy scores of the SVM classifier kernels with PSK un-correlated features on Syn flood DDoS attack dataset

Kernel	Precision		Recall		F1-Score		Specificity		Accuracy (%)
	Attack	**Benign**	**Attack**	**Benign**	**Attack**	**Benign**	**Attack**	**Benign**	
Linear	1	0.67	1	0.5	1	0.57	05	1	99.95
RBF	1	0	1	0	1	0	0	1	99.94
Poly	1	0.57	1	0.5	1	0.53	0.5	1	99.95
Sigmoid	1	0	1	0	1	0	0	1	99.93

Table 6. Precision, recall, F1-score, specificity and accuracy scores of the SVM classifier kernels with PSK un-correlated features on MSSQL attack

Kernel	Precision		Recall		F1-Score		Specificity		Accuracy (%)
	Attack	**Benign**	**Attack**	**Benign**	**Attack**	**Benign**	**Attack**	**Benign**	
Linear	1	0.87	1	0.89	1	0.88	0.89	1	99.31
RBF	1	0.85	1	1	1	0.92	1	1	99.52
Poly	1	0.88	1	0.99	1	0.93	0.99	1	99.60
Sigmoid	1	0.73	0.99	0.86	0.99	0.79	0.86	0.99	98.70

Table 7. Precision, recall, F1-score, specificity and accuracy scores of the SVM classifier kernels with PSK un-correlated features on SSDP attack

Kernel	Precision		Recall		F1-Score		Specificity		Accuracy (%)
	Attack	**Benign**	**Attack**	**Benign**	**Attack**	**Benign**	**Attack**	**Benign**	
Linear	1	0.49	1	0.65	1	0.56	0.65	1	99.82
RBF	1	1	1	0.35	1	0.51	0.35	1	99.89
Poly	1	0.51	1	0.88	1	0.65	0.88	1	99.84
Sigmoid	1	0.5	1	0.35	1	0.41	0.35	1	99.83

Table 8. Precision, recall, F1-score, specificity and accuracy scores of the SVM classifier kernels with PSK un-correlated features on customized dataset

Kernel	Precision		Recall		F1-Score		Specificity		Accuracy (%)
	Attack	**Benign**	**Attack**	**Benign**	**Attack**	**Benign**	**Attack**	**Benign**	
Linear	1	0.89	1	0.83	1	0.86	0.83	1	99.68
RBF	1	0.95	1	0.88	1	0.91	0.88	1	99.8
Poly	1	0.94	1	0.9	1	0.92	0.9	1	99.82
Sigmoid	0.99	0.71	1	0.39	1	0.5	0.39	1	99.1

Table 9. Cross-validation accuracy values with a standard deviation of the SVM classifier kernels with PSK un-correlated features on datasets

Kernel	Syn	MSSQL	SSDP	Customized Dataset
Linear	99.9766% (0.0078%)	99.4542% (0.1342%)	99.9766% (0.0078%)	99.5860% (0.0345%)
RBF	99.9805% (0.0062%)	99.5854% (0.1142%)	99.9220% (0.0151%)	99.7920% (0.0287%)
Poly	99.9786% (0.0095%)	99.9786% (0.0095%)	99.9415% (0.0130%)	99.8003% (0.0274%)
Sigmoid	99.9649% (0.0048%)	98.7511% (0.1031%)	99.9649% (0.0048%)	99.0784% (0.0894%)

4.1.10 Run-time

Run time is the total execution time of the process.

4.2 Results and discussions with PSK common un-correlated feature subsets

Table 5 shows the accuracy, precision, recall, F1-score, and specificity scores of the SVM kernels with a PSK uncorrelated feature subset of Syn dataset. The SVM classifier produces good accuracy with all kernels. For attack classification, SVM gives better precision, recall and F1-scores with all kernels. For benign classification, SVM gives poor precision, recall and F1-scores with 'rbf' and sigmoid kernels, SVM with linear kernel gives better precision, recall and F1-scores and SVM with poly kernel gives good poor precision, recall and F1-scores. For benign classification, an SVM classifier produces best specificity scores with all kernels. For attack classification, SVM with kernels 'rbf' and sigmoid gives poor specificity score, but linear and poly kernels gives a good specificity score. The SVM with linear kernel produces best classification results on Syn DDoS attack data set.

Table 6 shows the accuracy, precision, recall, F1-score, and specificity scores of the SVM kernels with PSK common un-correlated feature subset of MSSQL dataset. The SVM poly kernel gives best accuracy than remaining kernels on MSSQL dataset. The SVM all kernels gives better precision, recall and F1-score values for attack classification. For benign classification, the SVM poly kernel produces best precision and F1-score values, and the SVM rbf kernel produces best recall value. For attack classification, the SVM rbf produces best specificity score. For benign classification, the SVM all kernels produce better specificity scores. The SVM rbf kernel produces better classification results on MSSQL data set.

Table 7 shows accuracy, precision, recall, F1-score, and specificity values of the SVM kernels with PSK common un-correlated feature subset of SSDP dataset. The SVM 'rbf 'kernel produces better accuracy on SSDP dataset than remaining kernels. For attack classification, the SVM all kernels produce better values for precision, recall and F-1 score. For benign classification, the SVM 'rbf' kernel produce the best precision scores and the SVM poly kernel produces best recall and F-1 score values. For benign classification, the SVM produces better specificity value with all kernels. For attack classification, the SVM poly kernel produces best specificity value. Overall results, the SVM poly kernel produces better classification results on SSDP dataset.

Table 8 shows accuracy, precision, recall, F1-score, and specificity scores of the SVM kernels using PSK common un-correlated features-subset of customized dataset. The SVM poly kernel gives the best accuracy on customized dataset. On attack classification, an SVM classifier with all kernels gives better precision, recall and F1-score value. On benign classification, the SVM 'rbf' kernel gives best precision value, and the SVM poly kernel gives best recall and precision score values. The SVM sigmoid kernel gives poor F1-score value. On benign classification, The SVM all kernels give better specificity value. For attack classification, The SVM poly kernel gives best specificity value while, the SVM sigmoid kernel gives poor specificity value. Overall, the SVM poly kernel gives better classification results on the customized dataset with PSK common-uncorrelated feature subset.

Table 9 shows K-fold cross-validation accuracy scores of the SVM kernels. On Syn attack, the SVM rbf kernel gives best K-fold cross validation accuracy value than remaining kernels. On MSSQL, the SVM poly kernel gives best K-fold cross validation accuracy value than remaining kernels. On SSDP, the SVM linear kernel gives best K-fold cross validation accuracy value than remaining kernels. On customized dataset, the SVM poly kernel gives best K-fold cross validation accuracy value than remaining kernels.

Table 10 depicted ROC-AUC scores of the SVM kernels on TCP based DDoS attack data sets. The SVM 'rbf' kernel provides the best ROC-AUC score on all DDoS attack datasets. The SVM linear kernel provides the better ROC-AUC score on all DDoS attack data sets. The SVM poly kernel provides the very poor ROC-AUC score on Syn and MSSQL datasets while it produces good ROC-AUC score on SSDP dataset. The SVM sigmoid kernel provides the good ROC-AUC score on Syn and MSSQL datasets while it produces very poor ROC-AUC score on SSDP dataset. Overall, The SVM rbf kernel gives best ROC-AUC scores on TCP-based DDoS attacks. Figure 3 to Figure 6 shows the ROC curves of the SVM kernels on Syn, MSSQL, SSDP and customized datasets.

Table 11 shows the Log-loss values of the SVM kernels on TCP based DDoS attack data sets. The SVM linear kernel produces best log-loss value on Syn dataset. The SVM r with poly kernel produces best log-loss value of MSSQL dataset. The SVM rbf kernel produces best log-loss value on SSDP data set. The SVM classifier produces poor log-loss values of MSSQL dataset than the remaining two dataset. By observing all log-loss results, the SVM 'rbf' produces better log-loss values on all datasets.

Table 12 depicted the execution times of the SVM kernels on datasets. The SVM poly kernel contain best run time on Syn data set. The SVM rbf kernel contain best run time on MSSQL dataset. The SVM linear kernel contain best run time on SSDP dataset. By observing all execution time results, an SVM classifier with rbf contain best values on all datasets.

4.3 Results and discussions with TCP-based DDoS attacks common uncorrelated feature subsets

Accuracy, precision, recall, F1-score, and specificity scores of the SVM kernels on Syn dataset using the TCP common un-correlated feature subset showed in Table 13. On Syn flood attack data set, SVM produces better accuracy values with all kernels. For attack classification, SVM produces best evaluation values. For benign classification, SVM linear and poly kernels give good and SVM with rbf and sigmoid kernels give very low precision, recall, F1-scores and specificity values.

Accuracy, precision, recall, F1-score, and specificity scores of the SVM kernels on MSSQL attack data set using the TCP common un-correlated feature subsets depicted in Table 14. The SVM 'rbf' kernel gives better accuracy on MSSQL dataset. For attack classification, SVM 'linear', 'rbf' and 'poly' kernels gives the better precision, recall and F1-score values. The SVM 'sigmoid' kernel gives the better precision value, but gives lower values than other kernels recall and F1 score value for attack classification. For benign classification, the SVM 'linear' and 'rbf' give better precision, recall and F1-score values. The SVM 'sigmoid' kernel gives better recall value, but it gives poor precision and F1-score values compare to other kernels. The SVM 'rbf' kernel gives best specificity values for attack classification. For benign classification, the SVM classifier gives good results with all kernels. From overall these results, SVM with 'rbf' produces better classification results on MSSQL dataset with TCP common uncorrelation feature subset.

Table 10. ROC-AUC scores of the SVM classifier kernels with PSK un-correlated features on datasets

Kernel	Syn	MSSQL	SSDP	Customized Dataset
Linear	0.875117004680187	0.9013935183456373	0.976684398495299	0.9943542082601688
Rbf	0.999926872074883	0.9988391406068616	0.986826353304181	0.953029610564549
Poly	0.000175507020280	0.0082899972886221	0.922360637344032	0.946619500637433
Sigmoid	0.621948127925117	0.8684299823192183	0.853126361819465	0.668447069011137

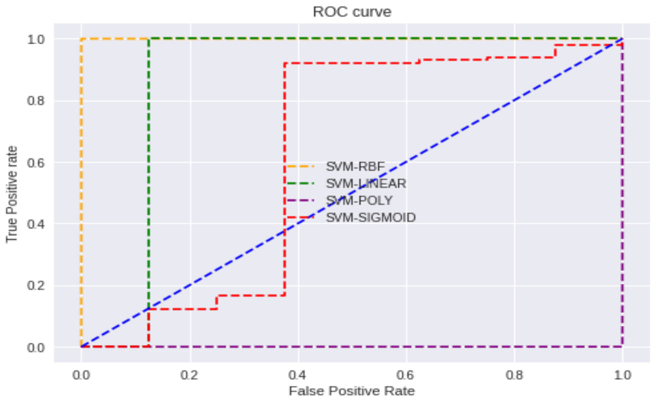

Figure 3. ROC curves of the SVM kernels using PSK common un-correlated feature subset on Syn flood attack

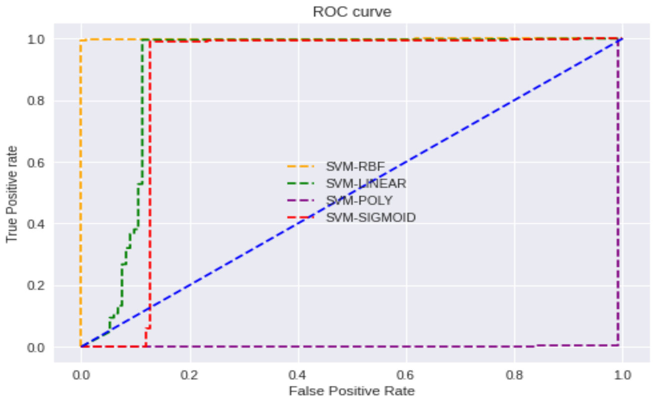

Figure 4. ROC curves of the SVM kernels using PSK common un-correlated feature subset on MSSQL attack

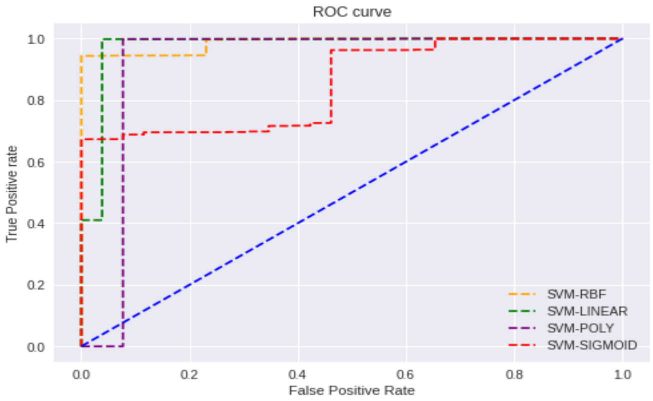

Figure 5. ROC curves of the SVM kernels using PSK common un-correlated feature subset on SSDP attack

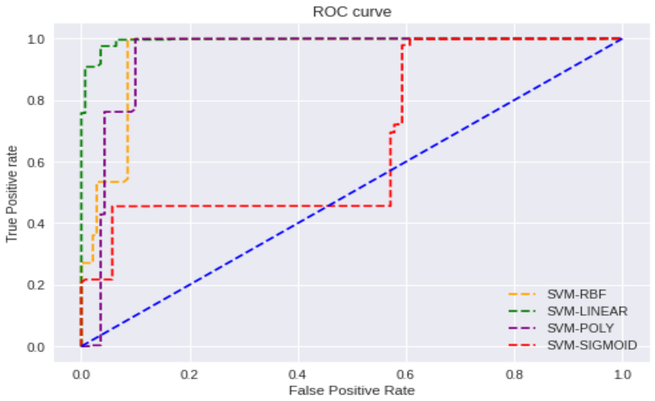

Figure 6. ROC curves of the SVM kernels using PSK common un-correlated feature subset on customized TCP-based DDoS attack

Table 11. Log-loss values of the SVM classifier kernels with PSK un-correlated features on datasets

Kernel	Syn	MSSQL	SSDP	Customized Dataset
Linear	0.016154962329217152	0.23925096872239923	0.060622385688063156	0.11208449573471901
RBF	0.02154011599148268	0.16674892046241613	0.03817023934666906	0.0689752227187703
Poly	0.018847414495954815	0.1377492760496931	0.056131561377174344	0.06322718214778413
Sigmoid	0.024232568158220336	0.44950027893275407	0.05837754530481742	0.3103890677676761

Table 12. Execution times (in seconds) of the SVM classifier kernels with PSK un-correlated features on datasets

Kernel	Syn	MSSQL	SSDP	Customized Dataset
Linear	26.6 s	3.93 s	4.7 s	1min 15s
RBF	4.22 s	2.15 s	7.88 s	8.54 s
Poly	1.78 s	16.7 s	16.5 s	37.1 s
Sigmoid	1.67 s	3.14 s	6.13 s	17.3 s

Table 13. Precision, Recall, F1-score, Specificity and Accuracy scores of the SVM classifier kernels with TCP un-correlated features on Syndata set

Kernels	Precision		Recall		F1-Score		Specificity		Accuracy (%)
	Attack	**Benign**	**Attack**	**Benign**	**Attack**	**Benign**	**Attack**	**Benign**	
Linear	1	0.6	1	0.38	1	0.46	0.38	1	99.95
RBF	1	0	1	0	1	0	0	1	99.94
Poly	1	0.6	1	0.38	1	0.46	0.38	1	99.95
Sigmoid	1	0	1	0	1	0	0	1	99.94

Table 14. Precision, Recall, F1-score, Specificity and Accuracy scores of the SVM classifier kernels with TCP un-correlated features on MSSQL dataset

Kernels	Precision		Recall		F1-Score		Specificity		Accuracy (%)
	Attack	Benign	Attack	Benign	Attack	Benign	Attack	Benign	
Linear	1	0.87	1	0.89	1	0.88	0.89	1	99.33
RBF	1	0.86	1	1	1	0.92	1	1	99.54
Poly	0.99	0.81	1	0.62	0.99	0.7	0.62	1	98.53
Sigmoid	1	0.14	0.82	1	0.9	0.25	0.82	1	82.83

Table 15. Precision, Recall, F1-score, Specificity and Accuracy scores of the SVM classifier kernels with TCP un-correlated features on SSDP dataset.

Kernels	Precision		Recall		F1-Score		Specificity		Accuracy (%)
	Attack	Benign	Attack	Benign	Attack	Benign	Attack	Benign	
Linear	1	0.6	1	0.12	1	0.19	0.12	1	99.84
RBF	1	1	1	0.38	1	0.56	0.38	1	99.9
Poly	1	0.4	1	0.65	1	0.5	0.65	1	99.78
Sigmoid	1	0.5	1	0.12	1	0.19	0.12	1	99.83

Table 16. Precision, Recall, F1-score, Specificity and Accuracy scores of the SVM classifier kernels with TCP un-correlated features on customized dataset

Kernels	Precision		Recall		F1-Score		Specificity		Accuracy (%)
	Attack	Benign	Attack	Benign	Attack	Benign	Attack	Benign	
Linear	0.99	0.69	1	0.39	1	0.5	0.39	1	99.08
RBF	0.99	0.76	1	0.49	1	0.6	0.49	1	99.23
Poly	0.99	0.95	1	0.13	0.99	0.23	0.13	1	98.98
Sigmoid	0.99	0.41	1	0.08	1	0.13	0.08	1	98.79

Table 17. K-fold cross-validation accuracy scores (with a standard deviation) in % of the SVM classifier with different SVM kernels at gamma='auto' on Syn attack data set with common features of common uncorrelated feature subsets of. TCP-based DDoS attack data sets

Kernels	Syn flood attack	MSSQL attack	SSDP attack	Customized Exploitation DDoS attack
Linear	99.9825% (0.0039%)	99.4437% (0.1213%)	99.4437% (0.1213%)	99.2074% (0.0981%)
RBF	99.9805% (0.0062%)	99.6117% (0.1015%)	99.8879% (0.0174%)	99.3072% (0.0788%)
Poly	98.9786% (0.0413%)	98.7826% (0.3742%)	99.6117% (0.1015%)	98.9786% (0.0413%)
Sigmoid	99.9630% (0.0073%)	99.3335% (0.1911%)	99.7855% (0.0407%)	99.1387% (0.0363%)

Table 18. ROC-AUC scores of the SVM classifier kernels with TCP un-correlated features

Kernels	Syn	MSSQL	SSDP	Customized Dataset
linear	0.5000780031201248	0.9115668679364141	0.97216754073562	0.804878166117432
Rbf	0.9998147425897035	0.9985558259717529	0.97746705336078	0.9166135472542274
poly	0.01884747682815236	0.009709817623306864	0.96056421276190	0.9602001298919971
sigmoid	0.02154011599148268	0.9659640896670526	0.22380673308588	0.7238688571909653

Table 19. Log-loss values of the SVM classifier kernels with TCP un-correlated features

Kernels	Syn	MSSQL	SSDP	Customized Dataset
Linear	0.018847476828152368	0.23200084781720595	0.05613260096299011	0.3161369087387047
Rbf	0.02154011599148268	0.15949896739883287	0.035924931149806244	0.267279329018493
Poly	0.018847476828152368	0.5075051065313313	0.05837785718056215	0.41672705319774966
Sigmoid	0.02154011599148268	5.930461606011113	0.05613260096299011	0.3161369087387047

Table 20. Execution times (in seconds) of the SVM classifier kernels with TCP un-correlated features

Kernels	Syn	MSSQL	SSDP	Customized Dataset
Linear	10.8 s	2.78 s	12.8 s	17.3 s
Rbf	1.9 s	2.76 s	8.27 s	18.4 s
Poly	3.07 s	8.7 s	4min 38s	24.1 s
Sigmoid	1.41 s	2.86 s	7.93 s	26.9 s

Accuracy, precision, recall, F1-score, and specificity scores of the SVM kernels on SSDP dataset using the TCP-common un-correlated feature subsets depicted in Table 15. An SVM classifier gives better accuracy values with all kernels, but it gives the best accuracy with 'rbf' kernel. An SVM classifier gives better precision, recall and F1-score for attack classification. For benign classification, an SVM classifier with all kernels gives better specificity value. An SVM with 'rbf' kernel gives better precision and F1-score value and SVM 'poly' kernel gives better recall value for benign classification. An SVM with 'poly' kernel gives better specificity value for attack classification. From these overall results, the SVM 'rbf' produces better classification results on SSDP dataset with TCP common uncorrelation features sub set.

Accuracy, Precision, Recall, F1-score, and specificity scores of the SVM kernels with TCP common feature subset depicted in Table 16. SVM rbf kernel gives best accuracy score while SVM remaining kernels also gives better accuracy results. For attack classification, SVM all kernels gives better precision, recall and F1-score values. SVM with poly kernel gives best precision value, the SVM rbf kernel gives good recall and F1-score values than others. For benign classification for benign classification, the SVM poly and sigmoid kernels produce poor recall and F1-score values. For benign classification, SVM all kernels produce the best specificity score. The SVM rbf kernel gives a good specificity score while SVM with poly and sigmoid kernel produce poor specificity scores for attack classification. From overall results show SVM rbf kernel produce better classification results on customized dataset.

Table 17 shows K-fold cross-validation accuracy values of SVM kernels on Syn attack data set with TCP common uncorrelated feature subset depicted in Table 17. On MSSQL DDoS attack data set, an SVM classifier with 'poly' kernel gives the best K-fold cross-validation accuracy value and SVM with 'rbf' kernel gives better K-fold cross validation value. On customized data set, an SVM classifier with 'poly' kernel gives the best K-fold accuracy value and SVM 'rbf' kernel gives better K-fold cross validation value.

ROC-AUC scores of the SVM classifier with different SVM kernels on Syn attack data set with TCP common uncorrelated feature subset shows Table 18. An SVM classifier with 'rbf' kernel gives the best ROC-AUC scores on SYN, MSSQL, and SSDP TCP-based DDoS attack data sets, but it gives better ROC-AUC score value of customized data set with TCP common uncorrelated feature subset. While SVM classifier with 'poly' kernel gives the best ROC-AUC score on customized data set and better score on SSDP dataset, but it gives poor ROC-AUC score values on Syn and MSSQL attacks datasets. In terms ROC_AUC score, SVM with 'sigmoid' kernel gives better score on MSSQL attack data set, good score on customized data set and it gives poor values on Syn and MSSQL attack data sets. Conclusion of these results, SVM classifier with 'rbf' produces best ROC-AUC scores on TCP-based DDoS attacks data sets with TCP common uncorrelated feature subset. Figure 7 to Figure 10 shows ROC curves of the SVM classifier with different kernels on Syn flood, MSSQL, SSDP and customized datasets.

Table 19 shows the SVM kernels log-values on the Syn attack data set with TCP common uncorrelated feature subset. On Syn flood DDoS attack data set, SVM with 'linear' and 'poly' kernels gives better log-loss score and SVM classifier with 'rbf' and 'sigmoid' kernels gives good log-loss values. An SVM classifier with 'rbf' kernel gives better log-loss values on MSSQL, SSDP and customized data set than others. An SVM classifier with 'sigmoid' kernel gives very poor log-loss value on MSSQL DDoS attack data set. From these results, an SVM classifier with 'rbf' gives better log values on TCP based DDoS attacks with a common uncorrelated feature subset of TCP based DDoS attacks.

Table 20 shows the execution values of the SVM classifier with different SVM kernels on a Syn attack dataset with TCP-common uncorrelated feature subset. An SVM classifier with 'sigmoid' and 'rbf' kernels contain better execution time on Syn dataset. An SVM classifier with 'rbf' and 'linear' kernels contain better execution time on MSSQL DDoS attack data set. An SVM classifier with 'poly' kernel contains better execution time on SSDP dataset. An SVM classifier with 'linear' and 'rbf' kernels contain better execution time on customized dataset. From these results, SVM classifier with 'rbf' contain better execution time on TCP common uncorrelated feature subsets.

4.4 Compare results with common uncorrelated feature subsets of PSK and TCP based DDoS attack

SVM classification gives better classification results with both PSK and TCP common feature subsets on TCP based DDoS attacks. An SVM classifier with 'linear' kernel produces the best accuracy, specificity and classification results consists of precision, recall, and F1-score on a Syn attack dataset with PSK common uncorrelated feature subset. An SVM classifier with both 'linear' and 'poly' kernels produce better accuracy, specificity and classification results consist of precision, recall, and F1-score on Syn attack data set with TCP based DDoS attacks common uncorrelated feature subset. An SVM classifier with 'linear' and 'poly' kernels give accuracy value 99.95% with both PSK and TCP-based DDoS attacks common uncorrelated feature subset. SVM classification algorithm with 'rbf' kernel produces better accuracy, specificity and classification results on MSSQL DDoS attack data set with both PSK and TCP-based DDoS attacks common uncorrelated feature subset. The SVM classification algorithm gives 99.52% accuracy with 'rbf' kernel on MSSQL dataset with PSK common uncorrelated feature subset. The SVM classification algorithm gives 99.54% accuracy with 'rbf' kernel with TCP-based DDoS attacks common uncorrelated feature subset on MSSQL DDoS attack data set. SVM with poly kernel produces better classification results on SSDP attack dataset with PSK common uncorrelation feature subset. SVM with 'rbf' produces better classification results on SSDP DDoS attack data set with a common uncorrelation feature subset of TCP-based DDoS attacks. The SVM classification algorithm gives 99.89% accuracy with 'rbf' kernel on SSDP DDoS attack data set with PSK common uncorrelated feature subset. The SVM classification algorithm gives 99.90% accuracy with 'rbf' kernel with TCP-based DDoS attacks common uncorrelated feature subset on SSDP dataset. SVM with 'rbf' kernel produce better classification results on a customized dataset with a common uncorrelation feature subset of TCP based DDoS attacks. The SVM classification algorithm gives 99.80% accuracy with 'rbf' kernel with PSK common feature subset on customized dataset. The SVM classification algorithm gives 99.23% accuracy with 'rbf' kernel with TCP-based DDoS attacks common uncorrelated feature subset on customized data set.

SVM classifier kernels gives same classification results with both PSK and TCP common uncorrelated feature subsets in terms of K-fold cross validation, ROC-AUC score and log-loss values. SVM classification with 'rbf' kernel gives better results in terms of in terms of K-fold cross validation, ROC-AUC score and log-loss values with both PSK and TCP common uncorrelated feature subsets on TCP based DDoS attacks.

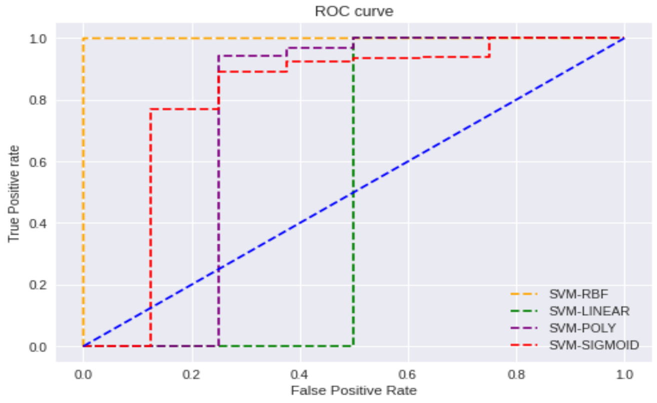

Figure 7. ROC curves of the SVM classifier kernels with TCP un-correlated features on Syn flood dataset

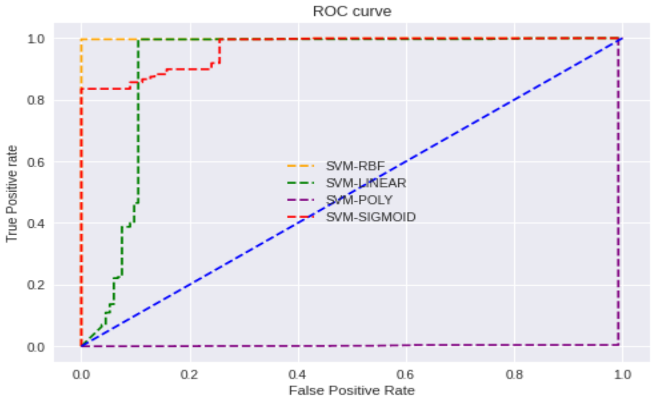

Figure 8. ROC curves of the SVM classifier kernels with TCP un-correlated features on MSSQL dataset

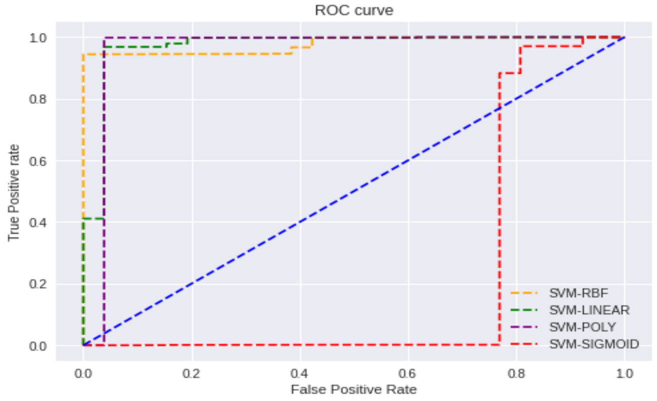

Figure 9. ROC curves of the SVM classifier kernels with TCP un-correlated features on SSDP dataset

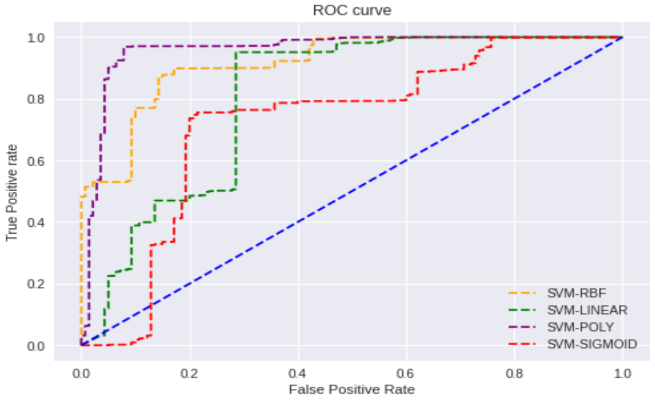

Figure 10. ROC curves of the SVM classifier kernels with TCP un-correlated features on customized dataset

5. CONCLUSION

This research proposed two feature selection method of PSK and TCP common uncorrelated feature subsets by using Pearson, Spearman and Kendall correlation methods. SVM classifier with linear, poly, rbf and sigmoid kernels produces good classification results on Syn flood, MSSQL, SSDP and customized TCP-based DDoS attack datasets with both proposed feature selection method. Although the TCP-based DDoS attacks common uncorrelated feature subset has less features than the PSK common uncorrelated feature subset, it yields comparable results. SVM classifier with rbf kernel produces better results with both feature selection methods on datasets. Correlation based feature selection ignores the interaction with the target class. So, in future, will do research on DDoS attack detection with Wrapper based feature selection methods which are selected the feature subsets based on the target class.

REFERENCES

[1] Dasari, K.B., Nagaraju, D. (2018). Distributed denial of service attacks, tools and defense mechanisms. International Journal of Pure and Applied Mathematics, 120(6): 3423-3437. https://acadpubl.eu/hub/2018-120-6/3/247.pdf.

[2] Ramkumar, B.N., Subbulakshmi, T. (2021). TCP Syn flood attack detection and prevention system using adaptive thresholding method. ITM Web of Conferences, 37: 01016. https://doi.org/10.1051/itmconf/20213701016

[3] Kwang, P. (2017). A countermeasure technique for attack of reflection SSDP in Home IoT. Journal of Convergence for Information Technology. https://doi.org/10.22156/CS4SMB.2017.7.2.001

[4] Moubayed, A., Aqeeli, E., Shami, A. (2020). Ensemble-based feature selection and classification model for DNS typo-squatting detection. IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1-6. https://doi.org/10.1109/CCECE47787.2020.9255697

[5] Xiao, P., Qu, W., Qi, H., Li, Z. (2015). Detecting DDoS attacks against data center with correlation analysis. Computer Communications, 67: 66-74. https://doi.org/10.1016/j.comcom.2015.06.012

[6] Bahl, S., Dahiya, D. (2016). Enhanced intrusion detection system for detecting rare class attacks using correlation based dimensionality reduction technique. Indian Journal of Science and Technology, 9(11): 1-10. https://doi.org/10.17485/ijst/2016/v9i11/84277

[7] Wei, W., Chen, F., Xia, Y., Jin, G. (2013). A rank correlation based detection against distributed reflection DoS attacks. IEEE Communication Letter, 17(1): 173-175. https://doi.org/10.1109/LCOMM.2012.121912.122257

[8] Singh, S., Shrivastava, S.K. (2018). Multilevel and multi-class support vector machine based on affinity propagation clustering for intrusion detection.

International Journal of Innovative Research in Computer and Communication Engineering, 6(12): 9159-9171. https://doi.org/10.15680/IJIRCCE.2018.0612016

[9] Dasari, K.B., Devarakonda, N. (2021). Detection of different DDoS attacks using machine learning classification algorithms. Ingénierie des Systèmes d'Information, 26(5): 461-468. https://doi.org/10.18280/isi.260505

[10] Mekala, S., Padmaja Rani, B. (2020). Kernel PCA based dimensionality reduction techniques for preprocessing of Telugu text documents for cluster analysis. IJERT, 8(12): 1337-1352.

[11] Li, Z.L., Hu, G.M., Yang, D. (2008). Global abnormal correlation analysis for DDoS attack detection. 2008 IEEE Symposium on Computers and Communications, pp. 310-315. https://doi.org/10.1109/ISCC.2008.4625614

[12] Dasari, K.B., Devarakonda, N. (2021). Detection of different DDoS attacks using machine learning classification algorithms. Ingénierie des Systèmes d'Information, 26(5): 461-468. https://doi.org/10.18280/isi.260505

[13] Sahoo, K.S., Tripathy, B.K., Naik, K., Ramasubbareddy, S., Balusamy, B., Khari, M., Burgos, D. (2020). An evolutionary SVM model for DDOS attack detection in software defined networks. IEEE Access, 8: 132502-132513. https://doi.org/10.1109/ACCESS.2020.3009733

[14] Subbulakshmi, T., BalaKrishnan, K., Shalinie, S.M., AnandKumar, D., GanapathiSubramanian, V., Kannathal, K. (2011). Detection of DDoS attacks using Enhanced Support Vector Machines with real time generated dataset. 2011 Third International Conference on Advanced Computing, pp. 17-22. https://doi.org/10.1109/ICoAC.2011.6165212

[15] Srinoy, S. (2007). Intrusion detection model based on particle swarm optimization and support vector machine. 2007 IEEE Symposium on Computational Intelligence in Security and Defense Applications, pp. 186-192. https://doi.org/10.1109/CISDA.2007.368152