Network Intrusion Detection using Machine Learning Techniques

2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) 978-1-7281-4142-8/20/$31.00 ©2020 IEEE 10.1109/ic-ETITE47903.2020.148

Sumaiya Thaseen I
School of Information Technology and Engineering
VIT University
Vellore, India
sumaiyathaseen@gmail.com

B Poorva
School of Information Technology and Engineering
VIT University
Vellore, India
poorvababu@gmail.com

P Sai Ushasree
School of Information Technology and Engineering
VIT University
Vellore, India
ushasreepamidi1@gmail.com

***Abstract*— Intrusion detection over packets on a network aims at classifying different types of packets without decrypting its contents to detect any intrusions in the network using machine learning. For this, packets are generated and transmitted over a network which are then captured by Wireshark for intrusion detection analysis. The captured data is organized into a dataset with the selected attributes after preprocessing using Weka tool and machine learning algorithms such as Naive Bayes, Support vector machine, Random Forest and KNearestNeighbors are implemented which classifies the data with accuracy 83.63%, 98.23%, 99.81%, 95.13% respectively. The packets are classified as encrypted packets, unencrypted packets, unencrypted malicious packets and encrypted malicious packets with different attributes and features where random forest is identified to be the best classifier with maximum accuracy.**

Keywords*— Intrusion detection, machine learning, Naive Bayes, Support vector machine, Random Forest, KNearestNeighbors.*

I. INTRODUCTION

Network is a group of computers interlinked with each other to share their resources or communicate via various means like cables, telephone lines, optical fibers, satellite, radio waves or infrared light. The types of network are Local Area Network(LAN), Metropolitan Area Network(MAN) and Wide area network(WAN). LAN is a network that is restricted only to a small area while MAN can cover almost an entire city but WAN is covers very huge geographic areas, like across countries or the global, thus making it a global network. Virtual Private Network(VPN) is used for establishing secure communication between networks. Wireless Personal Area Network (WPAN) is used by a personal to connect their personal devices like smart phone, smart tv, Bluetooth, speakers etc. to a network. Transmission of data from the source to destination over a network, involves breaking down of the data into several packets which are then reassembled at the receiver end. Protocols decide how the data is transferred on a network. There are different ways to set up a network connection, out of which, the most frequent forms are bus, ring and star topology. The connection can be wired with the use of ethernet cables or wireless and the physical devices required for establishing connection over a network is routers, modems, hubs, bridges and switches. Bus topology ensures easy installation which require very minimal cabling which is a simple network that delivers the packet to the destination. In a ring network, the data packet travels in one direction and reaches the destination after traversing through all the devices connected to the network whereas in the star network every device is connected individually to the network through hubs or switches, thus providing high speed transmission of data. The communication over a network can be among, client to server, server to client or clients interacting among themselves and sharing resources.

Hypertext transfer protocol (HTTP) is a protocol in the application level which enables the computers to communicate in a standardized manner over a network. It uses default TCP 80 port to deliver data but other ports can also be used. It is a stateless and connection less protocol and is media independent. Hypertext transfer protocol secure (HTTPS) adds a security layer to http where data is transmitted through secure socket layer (SSL) or transfer layer security (TLS) security. In this protocol, every data packet sent from the source is encrypted using TLS or SSL to avoid any attack on the data from the intruders. It thus provides confidentiality of the data to both sender and receiver as its works in collaboration with certificate authorities. The packets are decrypted at the receiver end to obtain the data with the necessary keys.

The packets contain minimal data which may be encrypted or may be not that is assembled at the receiver end to obtain the data but the packet format and features remain the same. A virus code or any other malicious data can be intruded into the network by an attacker and it can reach the destination, and affect the receiver. There is no greater difference between a data packet and a malicious packet, so network cannot identify and remove it automatically, it also just resembles an ordinary data packet [14].Any such intrusion into the network must be detected and resolved without decrypting and viewing the contents of the packet to maintain the authentication and confidentiality of data with the sender and receiver. Any malicious packet transmitted must be detected and removed from the network. Huge number of packets are transmitted over a network and the possibilities of transmitting malicious packets is also high, so in order to detect such intrusion in the network without decrypting [15] and viewing the contents of the packet, the attributes of different types of packets are analyzed, trained and tested for accuracy using machine learning algorithms which can thus detect any intrusion in the network easily without reading the contents of the packet [17].

Intrusion detection system has the ability to identify security incidents on a network, any malicious packet injection can be detected and removed. It provides greater visibility across network, and without decrypting the contents provided additional confidentiality to the sender and receiver. By implementing classification of data, different types of packets, whether it is an encrypted or unencrypted packet or malicious encrypted or unencrypted packet are identified and any outlier values and irrelevant data can be considered as intrusions or treats and can be removed from the network before it reaches the destination. By identifying the intrusions, alert mechanisms can be induced, and based on the type of intrusion, high end security can be established, thus securing the data over transmission on a network [16].

Intrusion detection systems are deployed to detect any anomalies in the transmission from host to client or on the network [18]. These systems work by analyzing the digital signature, certificates of the sender and receiver. Any violations of the regulations are monitored and reported to the network administrator alerting them about the possibility of attacks on the network. Sometimes for much more secure transmission, the packets are transmitted after decrypting and viewing its contents [19].

The objective of this system is to detect intrusion in a network without decrypting the contents of a package which enhances the confidentiality and integrity among the sender and receiver. Here, the packets are analyzed and different attributes are considered to be the classification factors [20]. The machine is trained and tested with different data to distinguish a malicious packet either encrypted or unencrypted from normal encrypted or unencrypted packets[22]. Any irrelevant or wrong data for any attribute can be easily identified, and alerted for intrusion. Thus, by classification of data based on the attributes of the packets without decrypting the content enables easy and quicker detection of intrusion, also ensuring confidentiality and security [21].

Wireshark is an open source tool used to analyze the packets transmitted over a network. It provides maximum details about the packets captured in the network. It is widely used by network administrators to detect any troubleshoots, examine security problems, debug protocol implementations and to verify network applications. It can capture live packets transmitted over a network, and the packets can be filtered based on several features. The proposed system uses Wireshark to capture the data packets transmitted over the network to study the details of the packets and organize them into a dataset to apply machine learning on the features to classify the packets as encrypted, encrypted malicious packet, unencrypted packet, unencrypted malicious packet.
Weka, an open source platform, is used for the preprocessing, classification, regression, clustering and framing association rules and visualization of the dataset based on inbuilt tools which is involved in removing null values, outliers from the dataset, and feature selection is done by importing the dataset into weka, listing out the attributes and identifying the key attributes that have major contribution to classification of the data.

Based on the literature analysis performed, various features that are assumed to detect the packet as malicious are identified, and how different networks respond to malicious packets and basic knowledge of how to detect and analyze are explained in detail in the background of this paper.

II. Background

After Dynamic key generation from the multimedia file using a special function is the technique used for symmetric key cryptography for transmission of multimedia files where the key is a dynamic bit pattern selected from the original file itself. Greater the size of the file, greater the no. of keys for encryption. Bit position and pixel depth is the part of bit pattern [1].

Sahmir's Secret sharing threshold scheme, Central IDS, Message forwarding proxies are the approaches used for intrusion detection in encrypted network. All the network traffic is explicitly forwarded to the network intrusion detection system(NIDS) over standard channels. Central IDS carries out traffic analysis and intrusion detection. It operates as separate host in the encrypted network. IDS sensors are located at all end points to ensure that all the network traffic that goes to the receiver is sent through the CIDS first. Each message is split into n messages and sent to a proxy layer. The proxy performs 1 of the 4 predefined functions written on it. CIDS and the receiver gets any number of message and recovers the main message. Attacker will not know which proxy will forward to CIDS/receiver or to both. The same msg will be available in both CIDS and the receiver. So, if any changes are made is easily detected [2].

From the network data, primary features are extracted using NetMate and the additional features are extracted by performing operations on the primary features. Only statistical features of the packets are available if it is encrypted. Feature selection is done by FOS and best first search algorithms. Ngrep to remove similarities in the content of the encrypted file (i.e. in the same visible to others). NetAI is used to identify the statistical features. FOS stops selecting features if the Mean square error (MSE) no more reduces. To validate the accuracy of the feature subsets selected by FOS, KNN classifier is used. Network classification techniques such as port or protocol pairing, Signature analysis, Deep packet inspection, Statistical anomaly analysis have been used for the feature selection in encrypted network [3].

To access the gateway to read possible data from the encrypted packet, data size and timing is known without decrypting and is used to identify the type of content and destination URL. Accesses are distinguished based on similarity of information and access frequencies(rules). In SSL /TLS network, IDS used for pattern matching., it reconstructs the http headers and payloads at the client side. The steps involved in the detection of intrusions in web access include feature selection extraction, frequency analysis and attack detection [4].

TTL values of a packet passing through various routers from the source before reaching the destination decreases for every

router. So, if there are any changes in the TTL value like the value being the same or increasing, then the packet is identified to be malicious. Any abnormal time to live values identified in the network, detects the pocket as malicious [5].

Intrusion detection, malware analysis and botnet detection are few approaches to security of big traffic analytics. Intrusion detection is to alert when there is any malicious activity. Role of malware analysis is to identify infected applications while botnet detection identifies if any network of computers is infected with malicious software. The procedure involves
Real time classification with statistical features, Robust classification by recognizing unknown classes and Efficient classification by using correlation [6].

Feature based selection and tools for analysis in a honey net network where the packet sent and received is considered malicious and intrusion detection takes place within the network with the help of Snort, which is a signature based intrusion detection tool, out of the different features available, such as the distributional changes of packets like ip address, port number etc., it identifies and chooses the best one based on the threshold values of entropy level and classifies an activity as malicious or not. Wireshark, net miner is used for analysis, uses various features like source ip address, destination ip address, source port, destination port, indegree, out degree, packet size entropy, application protocol used, origin of ip address for classification of malicious activities [7].

DTrojan model is implemented to analyze the network behavior characteristics to detect any malware. Some of the main features are selected and their attributes are used for the analysis of intrusion in the network packet.
1.Connection establishment,
2.Operating control which includes:

- interactive command detection
- Connecting time detection
- Active time
- Analysis of data flow
- Size detection

3.Connection maintenance

- analysis of the heartbeat of the packets.

4.Abnormal post judgement

- feature suspicious

The above mentioned are the features and attributes considered by a Trojan network to detect any intrusion in the network [8].

III. Materials and Methods

A. Dataset and Attribute Information

The dataset used in this system for the prediction and classification of different types of packets is created by considering the features and attributes of the data packets generated, that are analyzed using the Wireshark tool. The dataset is generated specifically for this study by transmitting different types of packets such as encrypted, unencrypted normal packets, encrypted and unencrypted malicious packets over our college network and capturing those packets based on the ip address from which they are sent and the time using Wireshark tool. The various features of each packet captured is analyzed and the attribute reduction is done to obtain better performance results. Weka tool has been used for the purpose of feature reduction and the attributes considered are given in the table below.
Total number of instances in this dataset :1130.
Number of attributes: 24.
Area: Data packets captured in the network
Attribute characteristics: Integer, decimal and nominal values.

TABLE I. DATASET DESCRIPTION

Sl.No	Attribute	Values
1	Frame size	Integer
2	Epoch time	Decimal
3	Time delta from previous captured packet	Decimal
4	Time delta from previous displayed packet	Decimal
5	Time since reference of first frame	Decimal
6	Frame number	Integer
7	Protocols in frame	Nominal
8	Coloring rule name	Nominal (takes value 1 or 2)
9	Coloring rule string	Nominal (takes value 1 or 2)
10	Sequence	Integer
11	Acknowledgement	Integer
12	Length	Integer
13	File data	Integer
14	Total length	Integer
15	Identification	Integer
16	Time to live	Integer
17	Window size value	Integer
18	Calculated window size	Integer
19	Window scaling factor	Integer
20	Time since first frame in this tcp stream	Decimal
21	Time since previous frame in this tcp stream	Decimal
22	Tcp payload	Integer
23	Tcp segment data	Integer
24	Label	Nominal

The coloring rule name takes values 1 and 2 if it is http and tcp respectively while coloring rule string takes 1 and 2 respectively if it is http or http2 to tcp port 80 and tcp. The

protocols in frame take the nominal values from 1 to 7 based on the different types of protocols that have been used for the transmission of that particular packet. Epoch time is the number of seconds since January 1, 1970.This is stored in .pcap, or .pcapng file. The other time formats in Wireshark are conversions of epoch time. All the timing information obtained on Wireshark is with reference to the epoch time. The labels determine whether the packet is a normal packet(http) or encrypted(ssl/tls) or malicious(http) or encrypted malicious(https). When the packet is encrypted and sent over a network, then there are several tasks performed between the server and client. They are client hello, server hello, certificate exchange between the server and client, encrypted handshake message, certificate status, change cipher spec, client key exchange, server key exchange, new session ticket, server hello done and application data. Similar types of packets are generated for encrypted malicious packets also with an encrypted alert packet added. When it is an unencrypted packet over a network, the types of packets generated are http, http get and http post while for unencrypted malicious packet http ok is also added.
After performing feature reduction, there are 9 attributes in this dataset that are crucial for prediction of the type of packet. They are: Frame size, Epoch time, Time delta from previous displayed packet, Coloring rule name, Length, File data, Total length, Time since previous frame in this tcp stream, Tcp payload.

A simple representation of the dataset with a few attributes which have major contribution in classifying the dataset is given below while the entire attributes in the dataset is described in TABLE I.

TABLE II. ATTRIBUTE DESCRIPTION

Attributes	Normal packet	Encrypted packet	Malicious packet	Encrypted malicious packet
Frame size	342	92	349	1454
Protocols in frame	http	Tls	http	Tls
Coloring rule name	http	Tcp	http	Tcp
Seq	155	1777	1	1401
Ack	1	5623	314	518
Length	288	38	295	1400
File data	27`	0	0	0
Total length	328	78	335	1440
Time to live	51	128	58	88
Tcp payload	288	38	295	1400
Tcp segment data	288	0	0	1277

B. Machine Learning algorithms

Naive Bayes, Random Forest, KNearestNeighbors, and Support vector machine are the four algorithms implemented in this system.

Naive Bayes is the classification algorithm that shows high accuracy when the data has more independent attributes, which is best for classifying categorical data [9].

Random Forest has more advantages of classification over the other algorithms as it is more stable in handling the different types of data and doesn't require scaling of features based on distance, overcomes the issues of overfitting and variance to improve the accuracy [10]. This prediction mechanism involves the selection of dependent attributes from the entire attributes present in the dataset and based on which node is calculated using best split point method. This splitting event is carried on till all the attributes are split into nodes, thus forming several trees. Then all the trees generated are integrated to build the random forest model [11].

Support vector machine is the machine learning technique used for the classification of data based on prior knowledge and retrieval. Certain values are selected for the particular attributes and it is then mapped to the respective values of dependent attributes which produces many boundaries that can be used to separate the attributes while the most optimal one among them is selected. The classification is more accurate if the boundary separates large number of attributes. SVM is not linearly separable if the features in the data are more dependent on each other rather than independent [12].

KNearestNeighbors is a supervised clustering technique which forms clusters of data with more similar values. Based on the different types of data in the dataset, a centroid value is selected based on which the points in the neighborhood are clustered into the number of clusters 'k' that is initialized. Iteration of this algorithm proceeds till all the data in the dataset becomes part of a cluster and no data is left out [13].

C. Performance evaluation metrics

The various evaluation metrics used in this system are accuracy score, confusion matrix, mean absolute error, mean squared error, f1 score, cross validation score, precision, and recall.

Accuracy is the ratio of number of predictions made correctly to the total number of predictions made. The higher the correct number of predictions made by the system, higher is the accuracy.

Confusion matrix is the actual metric to evaluate the performance of the entire model. It involves the identification of false positives, false negatives, true positives and true negatives in the data. The values across the main diagonal is the accuracy value of the model. It forms the basis for the other types of evaluation metrics. The ratio of sum of true positives and false negatives to the total number of values

gives the accuracy of the model. True positive is the number of data values correctly predicted as positive and false positive is the number of values incorrectly predicted as positive. False negative is the number of values correctly predicted as negative and false positive is the number of values incorrectly predicted as positive.
Precision is the fraction of true positive values and total positive values while recall is the fraction of true positive values and correctly classified positive and negative values.

F1 score falls in the range of [0,1] which is the harmonic mean between the precision and recall values. The performance of the model improves with higher f1 score.

Mean absolute error is calculated as the average difference between the original and predicted values and mean squared error is almost similar to it while it takes the average of square of the difference between the values.

Cross validation is an efficient technique to measure the effectiveness of the model, which involves holding out certain portion of data and training the remaining data to the model and then testing it with the holded data.

IV. Proposed System

Step 1: Initially several types of packets are generated like http, https, tls etc. These packets are generated with different contents like text data, audio, video, image and document attachments. Also packets with malicious content is also generated for the purpose of classification of different types of packets. The virus packet is then encrypted and sent over the network; Thus, the dataset contains details of packets that are encrypted, normal packets with different kinds of attachments, encrypted malicious ,and unencrypted malicious packets.

Step 2:The details of these packets are captured using Wireshark. The features and attributes of the packets are visible in the captured Wireshark data when transmitted over a network from a source to destination address.

Step 3: The several different attributes determined in the packets are interpreted and certain features with very minimal role in classification of different types of packets is removed, and dataset with the other features is created.

Step 4: Feature selection is done on the dataset to select the important features that determine the classification of packets. It is done by using Weka tool.

Step 5: Then over the created dataset, different machine learning algorithms like Naive Bayes, Random Forest, Support vector machine and KNearestNeighbors is implemented to check the accuracy of classification to distinguish different type of packets. The dataset is divided in a ratio of 80% and 20% for training and testing set respectively. Then the algorithms are evaluated for their performance based on their accuracy value.

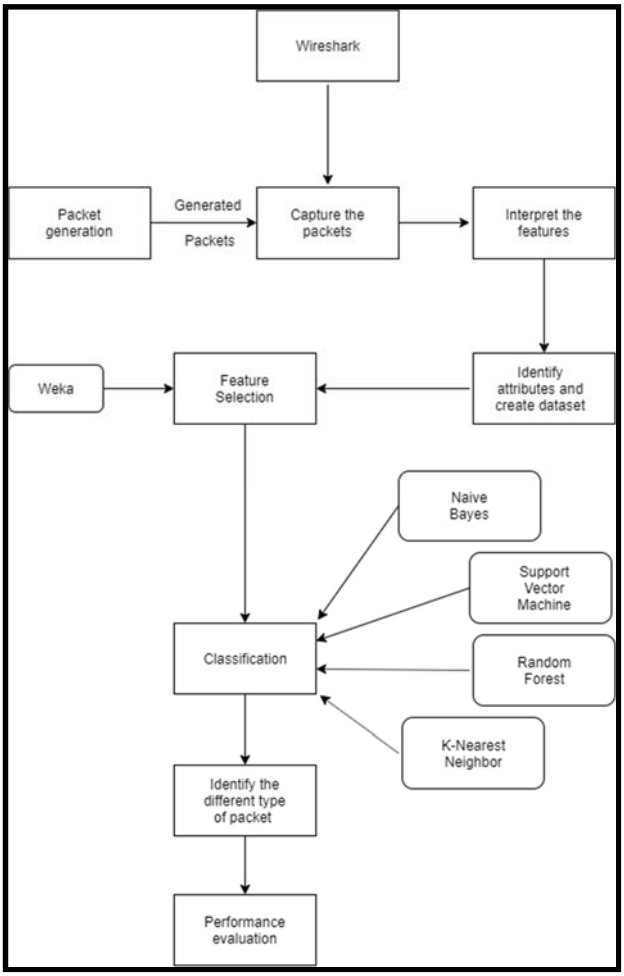

Fig. 1. Architecture of the proposed system

V. Results And Discussions

A. Results

Naive Bayes, Random Forest, KNearestNeighbors, and Support vector machine are the four algorithms implemented on the dataset for prediction and classification of data into different types of packets like normal packet, encrypted packet, packet with attachment, normal malicious packet, encrypted malicious packet. These algorithms are implemented on python Jupiter notebook platform.

- Naive Bayes code is executed by importing the necessary libraries and evaluated for the accuracy and validation score of the dataset which are 83.63% and 77.88% respectively when the testing is applied on 40% of the data (i.e. test size is 40%).
- Support Vector Machine (SVM) algorithm in machine learning is applied on the captured packet dataset and an accuracy of 98.23% is obtained when training and testing data is split as 70% and 30% respectively, while precision is 98.33%, recall 98.23%, f1 score 98.19%.
- Random Forest algorithm is applied on the dataset with 10-fold cross validation and training and testing data to be 60% and 40% respectively. Confusion matrix, mean absolute error, mean squared error, root mean squared

error and accuracy are obtained as the measures of this algorithm. The accuracy of random forest is obtained as 99.81% and mean absolute error (MAE) value is 0.714, mean squared error (MSE) is 9.24 and root mean squared error (RMSE) is 3.04.

- KNearestNeighbors is a clustering technique where data (i.e. packets in this case) are grouped together based on their similarities. In this technique the dataset is split as 80% and 20% for training and testing data respectively. The number of neighbors to be formed id determined by finding the square root of length of prediction dataset and subtracting 1 from it. Other evaluation metrics such as confusion matrix and f2 score are also determined. The accuracy of this algorithm is 95.13%.

The comparative measures of the algorithms used is given in the table below:

TABLE III. EVALUATION METRICS OF ALGORITHMS USED

Sl. No	Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1	Naive Bayes	83.63	90.83	83.62	84.78
2	Random Forest	99.81	99.81	99.81	99.76
3	Support Vector machine	98.23	98.23	98.23	98.19
4	KNearestNeighbors	95.13	94.54	95.13	94.46

B. Discussion

- For normal packet the protocol is http and it is sent through port 80 while it is tcp for encrypted packets and the port cannot be found for this.
- The window scaling factor is 256 for encrypted packets or 128. It is -1 for simple http get and post packets or 8 or 16 or 512 based on the attachments.
- The frame size of http get and post packets, client hello, server hello, handshake messages, cipher suite is same for different packets. The length or tcp payload of the transmission packet varies each time but the client hello, server hello, change cipher spec of the https packet is the same for whatever packet sent i.e.: 517,1364 and 1364 respectively. For http and http post packets it is 419 and 422 respectively. The length of certificate exchanges, handshake exchanges, key exchanges will differ for every packet.
- The acknowledgement number is sent by the tcp server, indicating that is has received data has several packets and is prepared to accept the next packet. This number is 1 for all http post packets and 423 for all http packets. 518 for all Server hello and Certificate, server key exchange, server hello done packets, Server key exchange, server hello done and other server related packets. For all encrypted handshake messages and change cipher spec packets, it is 644. For server key exchanges and handshake message packets the acknowledgement number differs variably.
- File data information is obtained only for http packets and not for encrypted packets.
- The total length of the http and http get, post packet with attachment is same irrespective of audio, video and image as 459 and 462 respectively. While for packets without attachment it varies. The length of client hello packet is the same for all https packets and the value is 557, for server hello packet it is 1404, for certificate exchange it is 1205, 298 for new session ticket, 166 for client key exchange, and other packet lengths vary for every transmission.
- The identification attribute is unique for every packet on the network.
- The time to live is 128 for http get and post, client hello and Client key exchange, change cipher spec, encrypted handshake message packets, other values changes.
- Window size value and calculated window size differs for every packet.
- Tcp segment data is visible for certain packets and is not same for all the packets, for few packets the value matches (server hello, change cipher spec packet value is 1231).
- For https packets the tcp segment data is available for all server hello, server certificate exchange and status, key exchange and server responding packets.

VI. CONCLUSION

The comparative study of machine learning algorithms like Random Forest, Naive Bayes, Support Vector Machine and KNearestNeighbors follows proving Random Forest algorithm to be more accurate compared to the other algorithms used in this system for the prediction and classification of different types of data packets as whether it is normal packet or encrypted packet or normal malicious or encrypted malicious packet. There are many inferences obtained from analyzing the attributes of the data packets of different types of data mentioned in the discussion section above. The further approach to this study involves the using of deep learning algorithms to improve the performance and accuracy of recognition and classification of different types of packets transmitted over a network.

REFERENCES

[1]. Verma, V., & Kumar, R. (2014). A Unique approach to multimedia based dynamic symmetric key cryptography. International Journal of Computer Scienceand Mobile Computing, 3(5), 1119-1128.

[2]. Goh, V. T., Zimmermann, J., & Looi, M. (2009, March). Towards intrusion detection for encrypted networks. In 2009 International Conference on Availability, Reliability and Security (pp. 540-545). IEEE.

[3]. McGaughey, D., Semeniuk, T., Smith, R., & Knight, S. (2018, April). A systematic approach of feature selection for encrypted network traffic classification. In 2018 Annual IEEE International Systems Conference (SysCon) (pp. 1-8). IEEE.

[4]. Yamada, A., Miyake, Y., Takemori, K., Studer, A., & Perrig, A. (2007, May). Intrusion detection for encrypted web accesses. In 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07) (Vol. 1, pp. 569-576). IEEE.

[5]. Yamada, R., & Goto, S. (2013). Using abnormal TTL values to detect malicious IP packets. Proceedings of the Asia-Pacific Advanced Network, 34, 27-34.

[6]. Miao, Y., Ruan, Z., Pan, L., Wang, Y., Zhang, J., & Xiang, Y. (2018). Automated Big Traffic Analytics for Cyber Security. *arXiv preprint arXiv:1804.09023*.

[7]. Sqalli, M. H., Firdous, S. N., Salah, K., & Abu- Amara, M. (2013). Classifying malicious activities in Honeynets using entropy and volume- based thresholds. *Security and Communication Networks*, *6*(5), 567-583.

[8]. Xue, L., & Sun, G. (2015). Design and implementation of a malware detection system based on network behavior. *Security and Communication Networks*, *8*(3), 459-470.

[9]. Bhosale, K. S., Nenova, M., & Iliev, G. (2018, December). Modified Naive Bayes Intrusion Detection System (MNBIDS). In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)* (pp. 291-296). IEEE.

[10]. Chang, Y., Li, W., & Yang, Z. (2017, July). Network intrusion detection based on random forest and support vector machine. In *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)* (Vol. 1, pp. 635-638). IEEE.

[11]. Zhang, J., Zulkernine, M., & Haque, A. (2008). Random-forests-based network intrusion detection systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(5), 649-659.

[12]. Bao, X., Xu, T., & Hou, H. (2009, September). Network intrusion detection based on support vector machine. In *2009 International Conference on Management and Service Science* (pp. 1-4). IEEE.

[13]. Ma, Z., & Kaban, A. (2013, September). K-Nearest-Neighbours with a novel similarity measure for intrusion detection. In *2013 13th UK Workshop on Computational Intelligence (UKCI)* (pp. 266-271). IEEE.

[14]. Wright, C. V., Monrose, F., & Masson, G. M. (2006). On inferring application protocol behaviors in encrypted network traffic. *Journal of Machine Learning Research*, *7*(Dec), 2745-2769.

[15]. Berthier, R., Urbina, D. I., Cárdenas, A. A., Guerrero, M., Herberg, U., Jetcheva, J. G., ... & Bobba, R. B. (2014, November). On the practicality of detecting anomalies with encrypted traffic in AMI. In *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)* (pp. 890-895). IEEE.

[16]. Clarke, N., Li, F., & Furnell, S. (2017). A novel privacy preserving user identification approach for network traffic. *computers & security*, *70*, 335-350.

[17]. Mirza, A. H. (2018, May). Computer network intrusion detection using various classifiers and ensemble learning. In *2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

[18]. Haripriya, L., & Jabbar, M. A. (2018, March). Role of Machine Learning in Intrusion Detection System. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 925-929). IEEE.

[19]. Azwar, H., Murtaz, M., Siddique, M., & Rehman, S. (2018, November). Intrusion Detection in secure network for Cybersecurity systems using Machine Learning and Data Mining. In *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS)* (pp. 1-9). IEEE.

[20]. Almseidin, M., Alzubi, M., Kovacs, S., & Alkasassbeh, M. (2017, September). Evaluation of machine learning algorithms for intrusion detection system. In *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)* (pp. 000277-000282). IEEE.

[21]. Patgiri, R., Varshney, U., Akutota, T., & Kunde, R. (2018, November). An Investigation on Intrusion Detection System Using Machine Learning. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1684-1691). IEEE.

[22]. Mishra, P., Varadharajan, V., Tupakula, U., & Pilli, E. S. (2018). A detailed investigation and analysis of using machine learning techniques for intrusion detection. *IEEE Communications Surveys & Tutorials*, *21*(1), 686-728.