Genetic Algorithm-Conditional Mutual Information Maximization based feature selection for Bot Attack Classification in IoT devices

G. Kavitha and N. M. Elango*

School of Information Technology and Engineering, V.I.T. University, Vellore, India
E-mail: kavithagk.vlr@gmail.com; elango.nm@vit.ac.in
**Corresponding Author*

Received 06 May 2021; Accepted 06 July 2021;
Publication 26 August 2021

Abstract

The evolution of computing is increasing in a vast manner that will integrate many physical objects and the internet to generate a new interconnection, such as the Internet of Things (IoT). It is estimated that the number of devices that will be interconnected to the internet will be more than trillions until 2025. Due to the lack of interoperability when these devices are interconnected in a vast heterogeneous network, it is tough to define and apply security mechanisms. The IoT networks have been exposed to many vulnerable attacks that disturb the network. Therefore, designing an intrusion detection system that provides additional security tools specific to IoT is needed to apply security mechanisms to detect the attacks in the network. In this paper, we propose a novel hybrid GA-CMIM machine learning algorithm that improves the efficiency in detecting the botnet intrusions with the set of optimal features that are selected from the dataset using a feature selection method.

Keywords: Internet of Things, Botnet, intrusion detections, machine learning.

Journal of Mobile Multimedia, Vol. 18_1, 119–134.
doi: 10.13052/jmm1550-4646.1816

1 Introduction

The service provided by the n/w to the legitimate users may get affected normally bypassing the traffic to the target system. Those kinds of interpretations are referred to as threats or attacks that are sent to the n/w. The report (2018) states that the danger such as DDoS (Distributed Denial Service of Attack) is excellent every year by year. As we all knew, the Internet Of Things (IoT) interconnects many physical devices that are easy to hack by the BotNet attacks [1]. The term botnet is defined as the network induced by the host computer, with the intervention of executing those activities will affect the regular access to the system that is interconnected to the network [2]. Due to the usage of IoT-based devices, it is highly challenging to provide a secured network to the structure of IoT using existing intrusion detection systems. Syntech has reported that there has been over a 600% increase in attacks against IoT devices [3]. The current I.D.S. for IoT is not highly suitable because it cannot detect the intrusions of many new and different kinds of infrastructure that are created by the attacks [4]. One of the major issues that have is all about the IoT network design where it is integrated with numerous sensors that accomplish such and control tasks [5].

2 Background

IoT is a kind of system which interconnects with a large number of methods that interconnect by more sensors & activations that are interwoven to the internet. Also, it will provide access to gather and share information either through mobiles or web-based application interfaces. Botmasters are the attackers who send the intruders, such as launching distributed cyber-attacks [6]. The primary objective of the intrusion detection system that is designed for detecting and identifying intruders [7].

2.1 Types of IoT Attack

The Internet of Things is a model it can interconnect many physical objects that will be interconnected to the internet. The devices are interconnected from the various heterogeneous landscapes; hence it seems to be difficult in designing the security measures for each service along with their specific security measures [17]. The major drawback identified by most of the researchers [18, 19] is nothing but the intrusion detection system for the traditional network may not suit the complex and heterogeneous IoT systems. The different types of attacks that affect the IoT systems are represented in

Figure 1. The short description of major types of IoT attacks are explained in the following section.

2.1.1 Physical attacks

IoT devices such as sensors, controllers, RFID readers, and RFID tags are highly affected by different physical attacks [19].

2.1.2 Protocol attacks

The data that reside and transmitted in the protocol like TCP, which gets affected during such data transmission is referred to as the protocol attacks [19].

2.1.3 Data attacks

The data that resides in the IoT objects are affected by these types of attacks where those data reside in different IoT devices which are used for communication and connectivity purpose [19].

2.1.4 Software attacks

The IoT application services that include the firmware, operating system applications, and various web services accessed by the user will have a higher chance of getting affected by software attacks such as [19].

This paper describes the detailed study and design of the botnet detection methods on large volumes of network data. The contribution of this paper is to develop a system that can classify the botnet's network data using machine learning. In the forthcoming session, we have discussed different types of botnet and the machine learning method used to those botnets. The term botnet is a kind of network that consists of botmasters who creates malicious activities to launch a DDoS attack, sending spam data to intrude the expected traffic and it can also steal the data by intervening with the actual data which affects normal process [20].

2.1.5 Client-server

Most of the botnets are executed their process as a client-server model. Both server and the client will get infected are awaiting for the commands from the botmasters [20].

2.1.6 P2P

The P2P botnet is used to attack in the P2P network which is nothing but a network that avoids a single point of failure [20].

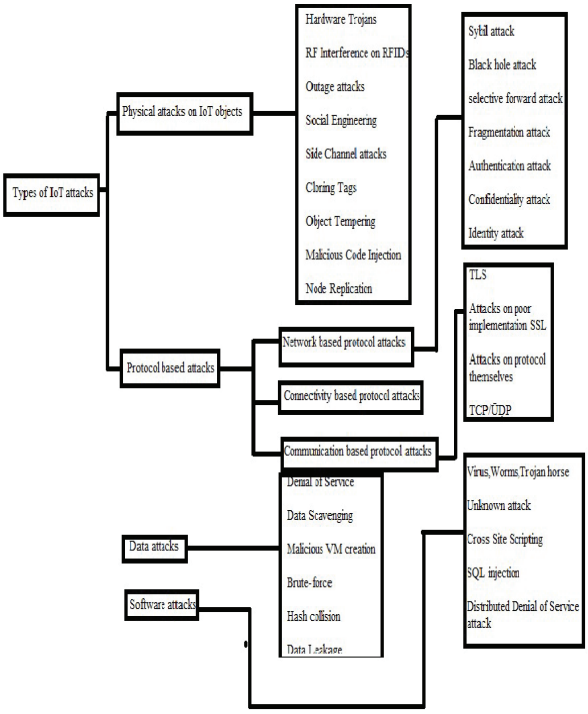

Figure 1 Different types of IoT attacks.

2.1.7 IoT

The emergence of IoT based services, there is an increase in large scale botnet attacks which tends to increase in large attacks [20].

3 Machine Learning for Botnets

The detailed literature review on the machine learning algorithms that are used to detect the botnet with the number of optimal features selected for analysis. Livadas et al. [21] analyzed the traffic data that consists of I.R.C. commands using a combination of three classifier machine learning algorithms C4.5 Tree, Naive Bayes, and Bayesian Network, with 10 features that are selected for analysis. Starter et al. [22] used the same combination of machine learning methods as livadas et al. used for their research purpose but with the 16 sets of features. G. Gu et al. [23] identified the patterns of spammers using behavior analysis using two levels X-means clustering algorithm. Maud et al. [24] used the combination of SVM, C4.5, Naive Bayes,

and boosted decision classifiers to detect the network type botnet with 20 selected features. Husna et al. [25] proposed to detect the P2P bots before it gets launch during the command and control phase where here author evaluated the proposed feature selection method with two different clustering algorithms like K-means and Hierarchal to compare the accuracy of the system with the same set of features. Noh et al. [26] identifying the characteristics of licit or illicit traffic patterns within the network with 7 selected features using the ROCK clustering algorithm. Langin et al. [27] selected eight features of the spam messages within the network. Liao et al. [28] detect the botnet by transforming the raw network traffic flows into multi-dimensional feature streams online. H Choi et al. [29] sees the activities that are get intruded in the DNS traffic using an X-means clustering algorithm with 13 selected features. Saad et al. [30] detection of P2P botnets along with the analysis on the behavior system using SVM, ANN, Gaussian, and Naive Bayes classifiers algorithm with 11 sets of features.

4 System Architecture and Design

The proposed system aims to analyze and detect malicious data intruded on the execution of regular traffic. The dataset that is used for analysis consists of both negative and normal data with equal in ratio using a machine learning method which also uses the hybrid feature selection method to select the optimal number of features for analysis. The forthcoming section describes the dataset used along with the design of the proposed system and the metrics that are used for analysis purposes.

4.1 Workflow of the Proposed System

The designs of the proposed system are been executed in two phases. The first phase of the algorithm tries to select the optimal features from the features present in the dataset and then in the next phase using a machine learning algorithm that detects the intrusions from the standard data. The flow diagram of the proposed model is represented in Figure 2.

4.2 Experimental Design

The process of the system aims to detect the affected data from the model which we built. The proposed method is evaluated and compared with the existing benchmarked methods like and metrics such as accuracy, false alarm rate, detection rate have been used to assess the performance of the system.

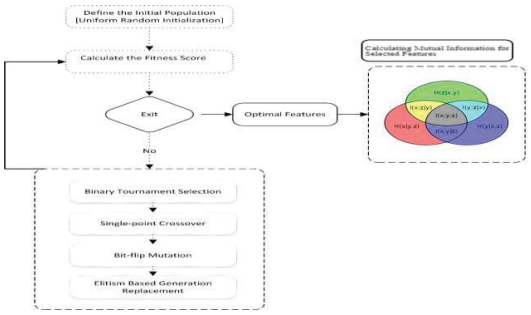

Figure 2 Working process of the proposed system.

4.3 Datasets

The dataset UNSW-NBIS [31] is considered for the experiment has it reflects network traffic and consists of both real normal and contemporary synthesized attack activities. The dataset consists of a total of 49 features. The various attacks like DDoS, D.O.S., keylogging, and data filtration were presented with this dataset. The 49 features of the dataset are pre-processed, which is used to identify, network-level patterns that devices create. The total number of samples are about 175,341 were used at training level and 82,332 samples were used at testing level analysis.

4.4 Evaluation Metrics for Models

In the experiment, the efficiency of our intrusion detection system is evaluated with the confusion matrix. The evaluation metrics[32][33] such as Accuracy, False Alarm Rate, and Detection Rate are used for analyzing the efficiency of the result.

True Positive: The number of malicious data that are correctly recognized.
True Negative: The number of data that are identified correctly.
False Positive: The number of incorrectly recognized normal code identified as malicious code by the detector.
False Negative: The number of incorrectly recognized malicious code which is identified as normal code by the detector.

The above values are used to evaluate the following metrics are as follows:

$$\text{Accuracy} = \text{TP} + \text{TN/TP} + \text{TN} + \text{FP} + \text{FN}$$

$$\text{False Alarm Rate} = \text{FP/FP} + \text{TN}$$

$$\text{Detection Rate} = \text{TP/TP} + \text{FN}$$
$$\text{Error} = 1 - \text{Accuracy}$$

These measures are used in the system to evaluate the execution of the proposed method the detailed description of the model is described in the forthcoming section.

5 Description of the Proposed System

5.1 First Phase

The feature selection process is done in the first phase of the algorithm using the k-fold cross-validation method. Here the value of k was taken as 10 as it gives the result of a divided dataset with low overfit and then its error rate was computed. The GRFF-FWSVM model, which was proposed in our previous research work has it can capture better feature representation for effective discrimination of samples in multiclass problems but lags in its performance in this binary class model. Nevertheless, it shows good performance in this system compared with other feature selection methods like G.A., CBFS, and R.F.E. The proposed GA-CMIM selects two optimal predictor features to accurately discriminate the samples of binary classes (Normal, Attack) as state and attack_cat with high accuracy than the previous mentioned methods. Table 1 states the optimal features selected by each process where our proposed method determines the optimal set of features for better performance of the system.

5.2 Second Phase

In the previous research, we have proposed a hybrid model GRFF-FWSVM that attempts to captures feature representation for effective discrimination of samples in multiclass problems but lags in its performance in this binary class model. The proposed GA-CMIM selects two optimal predictor features to accurately discriminate the samples of binary classes (Normal, Attack) as state and attack_cat. The figure represents the algorithm steps of the proposed method. The values for the parameters of the genetic algorithm are tabulated in Table 2 and the algorithm steps of the proposed algorithm in Table 3.

Table 1 Number of Features Selected by different methods

CBFS	GA	RFE	GRFF-FWSVM	GA-CMIM
3	9	5	4	2

Table 2 Genetic algorithm parameters

Parameters	Values
Crossover Probability	0.6
Generations	20
Mutation Probability	0.1
Population Size	20
Selection	Binary Tournament

Table 3 Proposed algorithm GA-CMIM

Algorithm: GA-CMIM
Input: α, β, γ, δ, π
Output: S_{opt1},S_{opt2}

generate rand(α)
p $<-$ rand(α)
for i in 1 to δ **do**
  $n_e <- \alpha.\beta$
  $p_1 <-$ best (ne, p)
  $n_{cr} <- (\alpha - n_e)/2$
  for j = 1 to n_{cr} **do**
    rand(S_a, S_b) ϵ p
    generate $cr_o(S_c, S_d)$
    $p_2 <- cr_o(S_c, S_d)$
  end for
  for j = 1 to n_{cr} **do**
    choose $S_j \epsilon p_2$
    $S'_j <- mt(S_j, \gamma)$
    if (S'_j!S_{opt1}) **then**
      update(S_j' $<-$ mod(S'_j))
    end if
    update S_j with S'_j in p_2
  end for
  update $p <- p_1 + p_2$
end for
return $S_{opt1} \epsilon$ p
proc CMIM(S_{opt1}, m_{inf})
for x in 1 to X **do**
  sub[x] $<-$ m_{inf}(x)
  for y in 1 to Y **do**
    num[y] <- $argmax_x$ sub[x]
    for x in X **do**
      sub[x] $<-$ min(sub[x], cond_m_{inf} (x, num[y])
    end for
  end for
end for
return S_{opt2} (sub[x])

6 Results

In this section, we have presented the results of our experiment and prove the efficiency of our proposed method by comparing the development of the proposed model with the existing two standard methods used to detect intrusions in the botnet network. As we discussed earlier, the values of evaluation metrics such as accuracy, False Alarm Rate, Detection rate, and the Error have been tabulated in Table 4, which also listed down the metrics values of the other two existing methods. The result states the accuracy value of the proposed method is 99.870% high than the current two methods, 99.647%, and 99.678%.

The classification methods like SVM, R.F., L.D.S., and BPNN on the selected features in the dataset using the proposed method GA-CMIM are tabulated in Table 5.

The performance of classifiers under different metrics such as accuracy, sensitivity, and specificity is shown below in Figure 3.

The accuracy value that is attained by different classifiers on the selected features by other feature selection techniques are represented in Figure 4. It states that the proposed method determines the optimal features that improve the accuracy attained by the classifiers used for differentiating the malicious and regular attacks.

Table 4 Comparison of existing methods with the proposed system in (%)

Literatures	Accuracy	False Alarm Rate	Detection Rate	Error
Al-Jarrah et al.	99.647	0.0012	96.752	0.353
Mai et al.	99.678	0.0015	95.756	0.322
Proposed (GA-CMIM-LDA)	99.870	0.08013	99.91	0.130

Table 5 Performance of the classification algorithms on GA-CMIM

Algorithms	Accuracy	False Alarm Rate	Detection Rate	Error
SVM	99.30	0.01648	99.78	0.70
RF	99.00	0.01942	99.04	1.00
LDA	99.87	0.08013	99.91	0.13
BPNN	99.03	0.01754	99.23	0.97

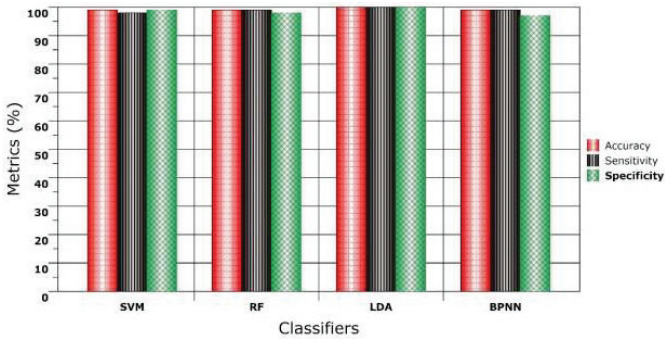

Figure 3 Performance of classifiers under various metrics with GA-CMIM identified features.

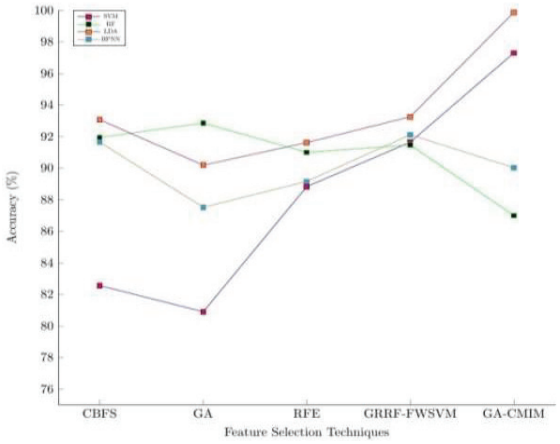

Figure 4 Accuracy attained by various classifiers under different feature selection techniques.

7 Conclusion

The significant aspect of designing an I.D.S. is to detect, identify, and track the attackers. In general, most of the I.D.S. have used ML algorithms. Still, due to its inherent nature, complexity, and also the cost of computational requirements will grow indefinitely when they are applied to the larger datasets. In this paper, we proposed two phases of botnet intrusion detection system for large-scale network datasets. To do a better analysis of a large-scale dataset, it is highly recommended to eliminate redundant and irrelevant features, we proposed and used a novel hybrid feature selection GA-CMIM

method to detect botnet instructions from the normal network data. The results show that the proposed method had achieved an accuracy value of 99.84% and reduced the error rate which also reduces the computational time and cost. As the number of devices connected to the network is increasing gradually year by year than providing security to the devices that are interconnected within the web will face a critical challenge to the global connectivity and accessibility of the Internet of Things (IoT). The computing power and memory usage of the IoT-based devices will be high. Thus our future work aims by using the learning techniques and detect the network's intrusion using online mode and the natural way of solving the large and complex problems by using the learning algorithms in the distribution environment.

References

[1] Zarpelao, B.B., Miani, R.S., Kawakani, C.T., de Alvarenga, S.C. A survey of intrusion detection in Internet of Things. J. Netw. Comput. Appl. **2017**, 84, 25–37.

[2] Christ, A., Gondal, I., Vamplew, P., Kamruzzaman, J. Survey of intrusion detection systems: Techniques, datasets, and challenges. Cybersecurity **2019**, 2, 20.

[3] Jing Liu & Yang Xiao et al., Botnet: Classification, Attacks, Detection, Tracing, and Preventive Measures, EURASIP Journal on Wireless Communications and Networking, Vol. 2009, Article ID 692654, 2009.

[4] Panda, M., & Patra, M. R. (2007). Network intrusion detection using naive Bayes. International Journal of Computer Science and Network Security, 7 (12), 258–263.

[5] Rafael A. Rodriguez-Gomez & Gabriel Macia-Fernandez, Pedro Garcia-Teodoro, Survey and Taxonomy of Botnet Research trough Life-Cycle, Journal of A.C.M. Computing Survey, Vol. 45 Issue 4, August 2013, Article No. 45.

[6] C. Kolias, A. Stavrou, J. Voas, I. Bojanova, R. Kuhn, Learning Internet-of-things security "Hands-on", IEEE Security and Privacy Jan/Feb 20 (February) (2016) pp. 2–11. doi:10.1109/MSP.2016.4.

[7] Pragati Chandhankhede, Autonomous Botnets Detection, Journal of Information Engineering and Applications, ISSN 2224-5782 (print) ISSN 2225-0506 (online) Vol. 3, No. 13, 2013.

[8] C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, M. Rajarajan, A survey of intrusion detection techniques in Cloud, Journal of Network and Computer Applications 36(1) (2013) 42–57.

[9] Sergio S.C Silva, Rodrigo M.P. Silva, Raquel C.G. Pinto, Ronaldo M. Salles, Botnets: A Survey, Journal of Computer Networks 57, pp. 378–403, 2012.

[10] Ashfaq, R.A.R., Wang, X.Z., Huang, J.Z., Abbas, H., He, Y.L. Fuzziness based semi-supervised learning approach for the intrusion detection system. Inf. Sci. **2017**, 378, 484–497.

[11] Roshna R.S. & Vinodh Ewards, Botnets Detection Using Adaptive Neuro-Fuzzy Inference System, International Journal of Engineering Research and Applications, Vol. 3, Issue 2, March–April 2013, pp. 1440–1445.

[12] D. Singh, G. Tripathi, A. J. Jara, A survey of Internet-of-things: Future vision, architecture, challenges, and services, in Internet of Things (WF-IoT), 2014 IEEE World Forum on, IEEE, 2014, pp. 287–292.

[13] García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusiondetection: Techniques, systems and challenges. Computers & Security, 28 (1–2), 18–28. doi:10.1016/j.cose.2008.08.003

[14] Pragati Chandhankhede, Autonomous Botnets Detection, Journal of Information Engineering and Applications, ISSN 2224-5782 (print) ISSN 2225-0506 (online) Vol. 3, No. 13, 2013.

[15] Son T. Vuong & Mohammed S. Alam, Advanced Methods for Botnet Intrusion Detection Systems, Chapter in Book: Intrusion Detection Systems, ISBN: 978-953-307-167-1, 2011.

[16] Sagar A. Yeshwantrao and Prof. Vilas J. Jadhav, Threats of Botnet to Internet Security and Respective Defense Strategies, International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 1, January 2014.

[17] Liao, H.J.; Lin, C.H.R.; Lin, Y.C.; Tung, K.Y. Intrusion detection system: A comprehensive review. J. Netw.Comput. Appl. **2013**, 36, 16–24.

[18] Jayveer Singh & Manisha J. Nene, A Survey on Machine Learning Techniques for Intrusion Detection System, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 11, November 2013.

[19] Christ, A.; Gondal, I.; Vamplew, P. An Anomaly Intrusion Detection System Using C5 Decision Tree Classifier. In Trends and Applications in Knowledge Discovery and Data Mining; Springer International Publishing: Cham, Switzerland, 2018; pp. 149–155.

[20] Matija Stevanovic & Jens Myrup Pedersen, On the Use of Machine Learning for Identifying Botnet Network Traffic, Journal of Cyber Security, Vol. 4, pp. 1–32, 2016.

[21] C. Livadas, R. Walsh, D. Lapsley, W. Strayer, Using machine learning techniques to identify botnet traffic, in Local Computer Networks, Proceedings 2006 31st IEEE Conference on, 2006, pp. 967–974. doi:10.1109/LCN.2006.322210.

[22] M. Masud, T. Al-khateeb, L. Khan, B. Thuraisingham, K. Hamlen, Flow-based identification of botnet traffic by mining multiple log files, in Distributed Framework and Applications, 2008. D.F.A. 2008. First International Conference on, 2008, pp. 200–206. doi: 10.1109/ICDFMA. 2008.4784437.

[23] H. Husna, S. Phithakkitnukoon, S. Palla, R. Dantu, Behavior analysis of spam botnets, in Communication Systems Software and Middleware workshops, 2008. COMSWARE 2008. 3rd International Conference on, 2008, pp. 246–253. doi: 10.1109/COMSWA.2008.4554418.

[24] W. T. Strayer, D. Lapsley, R. Walsh, C. Livadas, Botnet detection based on network behavior, in W. Lee, C. Wang, D. Dagon (Eds.), Botnet Detection, Vol. 36 of Advances in Information Security, Springer, 2008, pp. 1–24.

[25] Ramamoorthy, S., Prabu, M., & Balajee, J. M. (2021). Design and Evaluation of Wi-Fi Offloading Mechanism in Heterogeneous Networks. *International Journal of e-Collaboration (IJeC)*, *17*(1), 60–70.

[26] S.-K. Noh, J.-H. Oh, J.-S. Lee, B.-N. Noh, H.-C. Jeong, Detecting p2p botnets using a multi-phased flow model, in Digital Society, 2009. ICDS'09. Third International Conference on, 2009, pp. 247–253. doi:10.1109/ICDS.2009.37.

[27] C. Langin, H. Zhou, S. Rahimi, B. Gupta, M. Zargham, M. Sayeh, A self-organizing map and its modeling for discovering malignant network traffic, in Computational Intelligence in Cyber Security,2009. CICS '09. IEEE Symposium on, 2009, pp. 122–129. doi:10.1109/CICYBS.2009.4925099.

[28] W.-H. Liao, C.-C. Chang, Peer to peer botnet detection using data mining scheme, Internet Technology and Applications, 2010 International Conference on, 2010, pp. 1–4. doi:10.1109/ITAPP.2010.5566407.

[29] H. Choi, H. Lee, Identifying botnets by capturing group activities in DNS traffic, Journal of Computer Networks 56 (2011) 20–33.

[30] Vinoth Kumar, V., Karthikeyan, T., Praveen Sundar, P. V., Magesh, G., & Balajee, J. M. (2020). A Quantum Approach in LiFi Security using Quantum Key Distribution. *International Journal of Advanced Science and Technology*, *29*, 2345–2354.

[31] Moustafa, N., & Slay, J. (2014, May) UNSW NB15 DataSet for Network Intrusion Detection Systems. Retrieved from http://www.cybersecurity.unsw.adfa.edu.au/ADFA%20NB15%20Datasets

[32] Hossin, Mohammad, and M. N. Sulaiman. "A review of evaluation metrics for data classification evaluations." *International Journal of Data Mining & Knowledge Management Process* 5.2 (2015): 1.

[33] Kumar, G. (2014). Evaluation metrics for intrusion detection systems-A study. *Evaluation*, *2*(11), 11–7.

Biographies

G. Kavitha received her B.Sc. from V.I.T. university, Vellore, and M.C.A. from Arunai Engineering College (Affiliated with Anna University), Tiruvannamalai. She is doing a Ph.D. at V.I.T. University, Vellore. She has over nine years of teaching experience as an Assistant Professor (junior) at V.I.T. University. Her areas of interest are data analytics, Big Data adoptions, and large-scale data analysis concerning machine data.

N. M. Elango holds a Ph.D. in Computer Applications from SASTRA university, Thanjavur. He has over 33 years of experience in research and teaching. His areas of interest are image processing, enterprise modernization, and machine learning. Having published papers in many international conferences and refereed journals of repute, he is currently is working as Associate Professor, School of Information Technology and Engineering, V.I.T. University, Vellore, and mentors research students in various fields of I.T. He is a well-known academician and researcher in the academic and software industry.