Annals of Emerging Technologies in Computing (AETiC)
Vol. 5, No. 5, 2021

Research Article

A Machine Learning Approach for Improving the Performance of Network Intrusion Detection Systems

Adnan Helmi Azizan[1], Salama A. Mostafa[1,*], Aida Mustapha[1], Cik Feresa Mohd Foozy[1], Mohd Helmy Abd Wahab[1], Mazin Abed Mohammed[2] and Bashar Ahmad Khalaf[3,4]

[1]University Tun Hussein Onn Malaysia, 86400 Johor, Malaysia
salama@uthm.edu.my; aidam@uthm.edu.my; feresa@uthm.edu.my; helmy@uthm.edu.my
[2]University of Anbar, Anbar 31001, Iraq
mazinalshujeary@uoanbar.edu.iq
[3]University of Diyala, 32001 Diyala, Iraq
[4]Middle Technical University, 10001 Baghdad, Iraq
basharalzubaidy60@gmail.com
*Correspondence: salama@uthm.edu.my

Received: 17th October 2020; Accepted: 18th December 2020; Published: 20th March 2021

Abstract: Intrusion detection systems (IDS) are used in analyzing huge data and diagnose anomaly traffic such as DDoS attack; thus, an efficient traffic classification method is necessary for the IDS. The IDS models attempt to decrease false alarm and increase true alarm rates in order to improve the performance accuracy of the system. To resolve this concern, three machine learning algorithms have been tested and evaluated in this research which are decision jungle (DJ), random forest (RF) and support vector machine (SVM). The main objective is to propose a ML-based network intrusion detection system (ML-based NIDS) model that compares the performance of the three algorithms based on their accuracy and precision of anomaly traffics. The knowledge discovery in databases (KDD) methodology and intrusion detection evaluation dataset (CIC-IDS2017) are used in the testing which both are considered as a benchmark in the evaluation of IDS. The average accuracy results of the SVM is 98.18%, RF is 96.76% and DJ is 96.50% in which the highest accuracy is achieved by the SVM. The average precision results of the SVM is 98.74, RF is 97.96 and DJ is 97.82 in which the SVM got a higher average precision compared with the other two algorithms. The average recall results of the SVM is 95.63, RF is 97.62 and DJ is 95.77 in which the RF achieves the highest average of recall than SVM and DJ. In overall, the SVM algorithm is found to be the best algorithm that can be used to detect an intrusion in the system.

Keywords: ***Intrusion detection systems; detection rate; false alarms; CIC-IDS2017; random forest; decision jungle; machine; support vector machine***

1. Introduction

Intrusion is a major security problem for the breach in the world of web. It is on the premise that one mistake or intrusion can take over or delete information from your computer and the structure of your system in no time. Failures of security of the system can damage the system. In addition, intrusion can lead to huge financial losses and underlying computer transactions, resulting in poor data in the cyber digital war [1]. In this way, an intrusion detection system and error recognition framework are essential to prevent failures.

Adnan Helmi Azizan, Salama A. Mostafa, Aida Mustapha, Cik Feresa Mohd Foozy, Mohd Helmy Abd Wahab, Mazin Abed Mohammed and Bashar Ahmad Khalaf, "A Machine Learning Approach for Improving the Performance of Network Intrusion Detection Systems", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 201-208, Vol. 5, No. 5, 20th March 2021, Published by International Association of Educators and Researchers (IAER), DOI: 10.33166/AETiC.2021.05.025, Available: http://aetic.theiaer.org/archive/v5/v5n5/p25.html.

An Intrusion Detection System (IDS) is a framework that examines traffic in the internet system to prevent the violation or offensive activity. Some IDSs are capable of taking action when anomalous traffic or malicious activity is discovered, including stop traffic sent from an unsure IP address. Over the last decade, there has been increasing significantly the amount of the network attack. These attacks have been tremendously severe and complex. There are many hacker probes and attack computer networks [2]. To make a defense of these several cyber-attacks and computer virus, there are a lot of security technique that uses for computer been studied in the past decade [3].

There may be intrusion when deleting or stealing data from a computer in a limited time. Intrusion is, therefore, one of the most important issues in the web network security system. Defective equipment System hardware is also damaged by the intrusion. Several intrusion strategies are identified; perhaps this is the accuracy of one of the serious questions. The false alarm and detection rate are fundamental in a thorough analysis of accuracy. Intrusion detection needs to be improved to reduce false alarms and increase detection speed [4], [5].

Supervised Machine Learning is used to mechanize the structure of the IDS model building. The system gets trained to decide on choices that find a way to identify patterns. It then in the testing phase analyzes and reviews the provided data to decide on the acceptability of the online requests by using examples without labels [2]. In addition to this approach, unsupervised, semi-supervised and reinforcement learning approaches are also used [6]. Semi-supervised learning and training use fewer labels of data information and a lot of unlabeled information to explain the reasons. Experimental technology is used in reinforcement studies where activities yield better. Order of classification, predictions and regressions are used. Environments, activity and agents are the three essential elements of this type of study. The aim is for the expert to select an activity that exploits the expected and predicted remuneration. With an excellent approach, an agent can quickly achieve a goal [2], [5].

This work is divided into five segments, starting with the introduction. Section 2 portrayed about the most relevant work. In addition, strategies and materials were set out in section 3. Simulation and results were also exhibited in the section. Finally, section 5 completes and concludes the discussion.

2. Related Work

Intrusion detections are an essential part of any security, for example, multipurpose security gadgets and adaptive security appliances, intrusion identification systems and response frameworks, and firewalls. Intrusion detection systems (IDS) attempt to differentiate between intrusions and malfunctions of the computer system framework by gathering and analyzing data information gathered from the system and identifying changes after the attempted attack. Various algorithms are used in IDS, however, which algorithm can manifest the best performance is an issue to be investigated. In related works, a comparison was made from the previous related work of IDS. So in the research that will be done, the machine learning algorithms or techniques that will use are random forest (RF), decision jungle (DJ), and support vector machine (SVM). This algorithm will be tested in the Microsoft Azure to get the accuracy value. These three algorithms will use the same dataset to make this research relevant. The accuracy value will be compared to get more accurate algorithms in the detection of intrusion in the system. SVM is also a sustainable option for a failure detection that can provide continuous detection capability, control very large dimensions of data. SVM plans training vectors and high dimensional component space through non-linear planning and labelling each vector according to its group. Then characterized by defining the number of support vectors, these are members from the set of training inputs, which forms a hyperplane in the feature component space [2].

SVM was introduced in the mid-1990s [7]. Basically, the concept that drives SVM for the intrusion detection is to use the provisioning data only as a representation of the typical normal class or to be known as a non-attack in the intrusion detection framework, thus expecting the rest as unique features and obtaining anomalies [1]. The classification created by the support vector-machine method divides the input information into a limited region where common elements and normal objects are found and the rest of the field hopefully contains inconsistencies. Random Forest is a

classification group used for characterization. Another variant of the group is presented in [8]. Sometimes it acts better than boosting, faster than bagging and accelerating. The first form of irregular forest area can be revealed as a version of bagging, and the basic classification is a random tree. However, it is considered a learning process that uses the selection tree as the basic classification [9].

In addition, the random forest is an algorithm that contains a pool of independent and indistinguishable classification trees, each developed according to random vector. From each tree in the group, one vote is given for the best-known class of input vectors [10]. The significant diversity of the random forest can be obtained by taking a sample of a large number of attributes from the dataset or simply subjectively changing some parameters of the selected decision tree 10. Random trees have two limits that can be offset: the number of vectors that can be incurred in each node regularly established in all nodes and the number of trees that make up the forest. Because RFA models give the impression of high memory; a viable and valid option is a DJ algorithm. The second is an improvement of RFA [2], which returns to the possibility of built-in selective Directed Acyclic Graphics (DAGs) assemblies and needs much a smaller amount of memory than the RFA. The use of DJA is an ongoing area of research focusing on medical characteristic issues: prognosis and diagnosis of infection [11], the expectation of fatal health well-being [12], excellent results for grouping patients and recommending patients; proper diet. Thus, writers and authors must also control the importance and presentation of DJA, which has never been used to detect mechanical and industrial irregularities. Indeed, the impression of memory is seen as a finite commodity, and a legitimate option and valid alternative may be needed for computing.

3. Methods and Materials

In this section, the testing dataset, the machine learning methods that have been used in this work which are SVM, RF, and DJ and evaluation metrics have been presented. The knowledge discovery in databases (KDD) methodology is used to perform this research and the intrusion detection evaluation dataset (CIC-IDS2017) is utilized to assess the performance of the three classification algorithms. The evaluation metrics of recall, precision and accuracy are utilized for the assessment process.

3.1. Dataset

Data used in the study: [13] developed a new characterization of intruder data detection and intrusion traffic using and the Intrusion Detection Evaluation Dataset (CIC-IDS2017). It has 85 attributes. The data collection period began in 2017, Monday, July 3, at 9 a.m., and ended at 5 p.m. Friday, 2017 July 7, within 5 days. Attacks include DoS, Botnet, Web Attack, Brute Force FTP, Brute Infiltration, DDoS and Force SSH. More details about the dataset can be found in [13].

3.2. Machine Learning Algorithms

Decision jungle (DJ) is a non-parametric model, which displays the limits of non-linear decision boundaries [11], [12]. In addition, the DJ consists of a collection of decision-directed acyclical diagrams (DAGs). They have integrated component selection and configuration and are resilient in the presence of noisy features.

Support vector machine (SVM) is a real learning AI system based on the objective of prediction [14]. Gaussians apply results each time binary, uniformly, and multinomial regression. That is why another Stata is given to order SVM machines. This package is a fine example of LIBSVM [15], which has been widely reported. SVM is a controlled management system and a supervised learning approach to process different types of information through different rules. These materials have been used for both classification design problems and to address non-linear data processing tasks. SVM creates multiple hyperplanes or hyperplane in the upper dimensional. The best hyperplane among which keeps information in different classes and stops at least between classes. The main purpose of this Kernel field function is to intersect between hyperplanes. In recent years, experts have developed a number of innovative additives due to growing motivation for SVMs [16], [17]. SVM is commonly

used in image processing, pattern recognition applications, and video and audio reception applications.

In this work, Random forest (RF) is selected because it protects the overfitting and has shown great results. RF is a team classifier designed to improve accuracy. Irregular or random process of the RF has a low characteristic error, in contrast to some other classification algorithms [18]. There are many random forests to choose from and each of which produces various selected subtrees in the preparation part [19]. In IDS, RFs are classifiers used to organize and retrieve research information related to identity interference and intrusion. RF has high-precision properties even for processing data with noise.

3.3. Evaluation Methods

The confusion matrix of this work is an m x m, where m represents the intrusion classes to be predicted by the classification algorithms. The row of the evaluation matrix represents the target classes while the columns represent the output classes. It is needed to choose a decision threshold to label the instance as positive or negatives [2]. If the probability assigned by the rating, for example, exceeds the limit, it is declared positive, and if the probability is less than the threshold of decision, it is labelled as negative as shown in Table 1.

Table 1. The confusion matrix

	class one	**class two**
	classified	**classified**
class one	TP	FN
class two	FP	TN

wherein Table 1, the Positive (P) observation is known positive e.g., an investigation into positive detection of attack; Negative (N) prediction outcome is not positive, e.g., negative malignancy is detected or severe; True Positive (TP) prediction outcome is positive; False Negative (FN) prediction outcome is known to monitor despite the expected negative; False Positive (FP) prediction outcome is negative but predicted positive and True Negative (TN) prediction outcome is assumed to be negative and is predicted to be negative.

To reflect on the demonstration of each group or classification, the preparations are completed according to all values of 85 attributes. The configuration for the evaluation metrics in SVM, RF, and DJ has examined in more details are as follows.

Precision is the ratio of the attacks flows (TP) to the characteristic flows (TP + PF).

$$\text{Precision, P} = \frac{TP}{TP + FP} \tag{1}$$

Recall or Sensitivity: It is a ratio of correctly identified attacks (TP), overall predicted flows (TP+FN).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

Accuracy: The most commonly used metric to judge a model and is not a clear indicator of the performance. The worse happens when classes are imbalanced. Accuracy shows the percentage of true detection over total traffic trace [2].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

4. Modelling and Results

The knowledge discovery in databases (KDD) methodology is used to perform this research and the intrusion detection evaluation dataset (CIC-IDS2017) is utilized to test and assess the performance of the ML algorithms. They are all integrated into a ML-based network intrusion detection system (ML-based NIDS) model as shown in Figure 1.

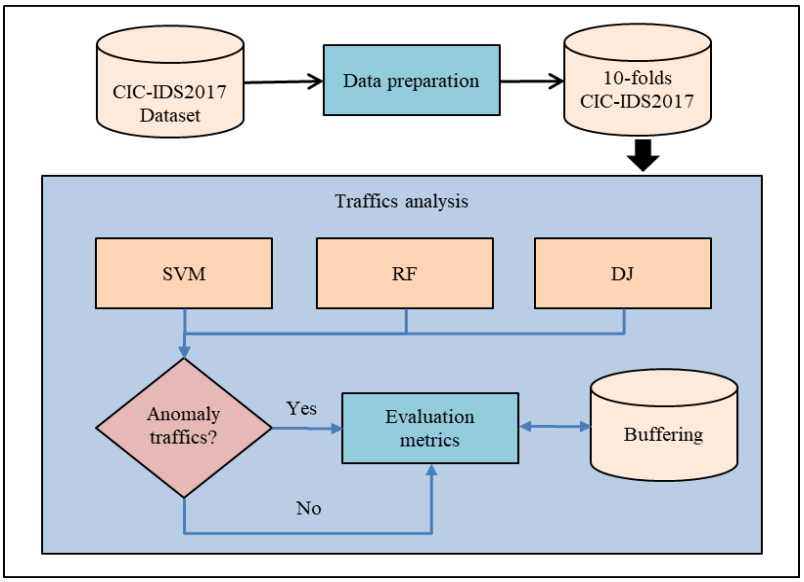

Figure 1. ML-based NIDS model

The KDD methodology, which comprises cleaning, integration, selection, transformation, mining processes to the data is adopted in this work. This methodology aims to perform pattern evaluation and knowledge discovery used for IDS. The following algorithm shows the processing stage of the ML-based NIDS model on the CIC-IDS2017.

1. *Perform data extraction from the collection of CIC-IDS2017 dataset;*
2. *Consolidate conflicting information from multiple sources into a specific resource;*
3. *Select the data a procedure based on the applicable data to the test;*
4. *Convert the data required through mining strategy into an appropriate structure of 10-folds cross-validation;*
5. *Extract patterns potentially helpful for modelling designs in the training phase;*
6. *Perform pattern assessment to the abnormal traffic to enhance the recognition of the model design based on the given evaluation measures of the testing phase;*
7. *Represent the results of the two phases as findings.*

The experiments were performed using the Azure Machine Learning Tool using a validation method for 10-fold training and testing[1]. Moreover, the three algorithms that have been choosing are SVM, RF and DJ. They have been tested to classify the traffics, which imported from the CIC-IDS2017dataset and the following result have been obtained. Table 2 displays the result of the accuracy of these three algorithms. The results using the three algorithms have been carried out 10 test of diverse data allocation of cross-validation folds. The obtained test result shows that the highest value of the accuracy score is recorded by the SVM with an average score of 98.18% and followed by the RF with an average accuracy of 96.76 and the lowest is the DJ with an average accuracy of 96.50%.

Table 2. Accuracy results of the SVM, RF and DJ

Test	Split data	SVM	RF	DJ
1	90:10	100	100	100
2	80:20	99.86	98.76	98.45
3	70:30	99.45	97.43	97.67
4	60:40	99.51	97.02	96.44
5	50:50	99.52	96.23	95.87
6	40:60	99.45	95.96	95.23
7	30:70	98.67	95.56	95.04
8	20:80	98.23	94.98	94.98
9	10:90	97.56	94.45	94.35
10	66:34	99.56	97.45	96.97
	Average	**98.18**	**96.76**	**96.50**

[1] Azure Machine Learning Studio, link, https://studio.azureml.net/, accessed on: 2019.

Figure 2 shows that the accuracy result of the SVM, RF and DJ. When split data to (90:10), the accuracy of these algorithms are the same where the accuracy is 100%. However, when the split data to (20:80), the accuracy of SVM (98.23%) are higher than the other two algorithms. While the accuracy of RF and DJ is the same, which is 94.98%. At split data (66.34), the accuracy for SVM is 99.56%, RF is 97.45% and DJ is 96.97%. Its means that the highest accuracy (100%) for all algorithm is when the data is split into 90% training and 10% testing. While the lowest accuracy of DJ is (94.35%) when the data is split into 10% training and 90% testing.

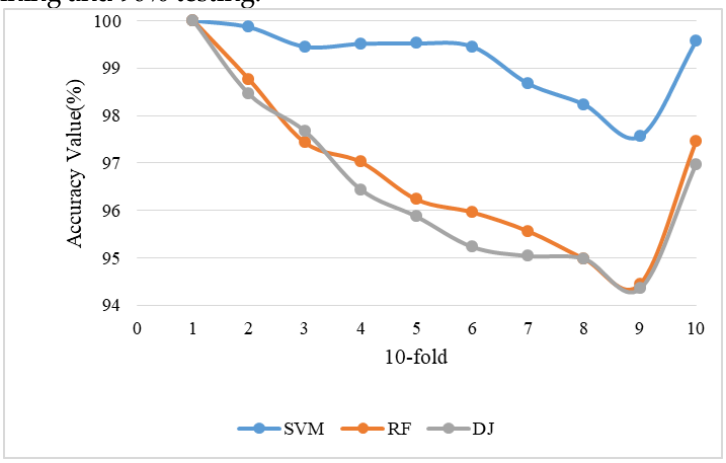

Figure 2. The Accuracy results of the SVM, RF and DJ

Table 3 shows that the test results of precision for the SVM, RF and DJ. The higher average precision value is SVM with an average score of 98.74% and follows by RF with an average score of 97.96% and the DJ with an average score of 97.82% makes the lowest average precision value.

Table 3. The Precision results of the SVM, DF and GBM

Test	Split data	SVM	RF	DJ
1	90:10	100	100	100
2	80:20	96.64	98.32	97.40
3	70:30	98.34	97.65	98.23
4	60:40	99.51	97.23	98.76
5	50:50	100	97.65	97.23
6	40:60	99.45	98.32	98.32
7	30:70	98.76	96.45	96.84
8	20:80	98.24	97.96	96.54
9	10:90	97.72	98.56	96.03
10	66:34	98.69	97.45	98.56
Average		**98.74**	**97.96**	**97.82**

Figure 3 shows the precision results for the SVM, RF and DJ. When the split data to (90:10), the precision of SVM, RF and DJ show the best result of 100% precision. When the split data to (40:60), the precision of SVM (99.45) are the best, while RF (98.32%) and DJ (98.32%) has the same precision score. DJ has the worst precision score at split data (10:90) which score 96.03 than SVM (97.72) and RF (98.56%).

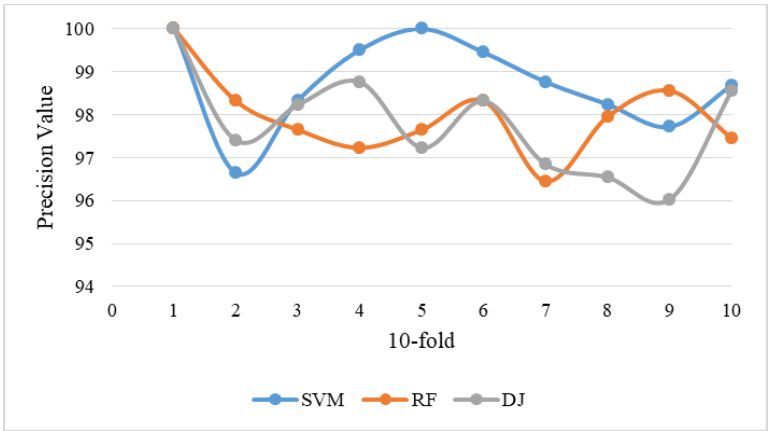

Figure 3. The Precision results of the SVM, RF and DJ

Table 4 shows the test results of recall score that found in the SVM, RF and DJ. The average recall for SVM is 95.63 while for RF is 97.62 and for DJ is 95.77. SVM got the best average for recall than RF and DJ.

Table 4. The recall results of the SVM, RF and DJ

Test	Split data	SVM	RF	DJ
1	90:10	93.00	93.00	93.00
2	80:20	100	98.95	97.43
3	70:30	98.45	100	98.35
4	60:40	95.45	99.45	97.67
5	50:50	93.00	95.56	96.98
6	40:60	94.43	97.23	96.45
7	30:70	94.12	98.90	95.12
8	20:80	94.01	97.02	92.34
9	10:90	97.63	96.34	92.12
10	66:34	96.54	99.74	98.23
Average		**95.63**	**97.62**	**95.77**

Subsequently, Figure 4 shows the result of recall in which when split the data to (90:10), the recall results are the same for these three algorithms which are 93.00%. However, when the split data to (70:30), RF have the highest recall score of 100% while SVM 98.45% and DJ 98.35%. While when splitting the data to (10:90), the recall for DJ is 92.12%, SVM 97.63% and RF is 96.34. Hence, the overall results confirm that the SVM provides the best performance in the ML-based NIDS model among the three algorithms in detecting abnormal patterns of DDoS attacks.

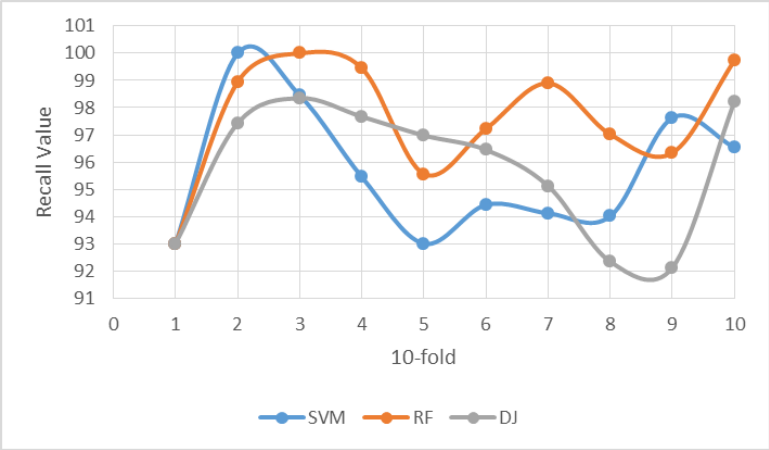

Figure 4. The recall results of the SVM, RF and DJ

5. Conclusion

This research is about determining the classification algorithm that can give the best detect performance to intrusion in the IDSs. It presented an analysis of intrusion detection systems (IDS) using three popular classification algorithms, which are random forest (RF), decision jungle (DJ) and support vector machine (SVM), The aim is to apply the proposed algorithms into a ML-based Network Intrusion Detection System (ML-based NIDS) model. The ML-based NIDS is implemented and tested using the knowledge discovery in databases (KDD) methodology and the intrusion detection evaluation dataset (CIC-IDS2017). The average accuracy results of the SVM is 98.18% while the RF is 96.76% and the DJ is 96.50%. The average precision of the SVM is 98.74 while the RF is 97.96 and the DJ is 97.82. The average recall of the SVM is 95.63, the RF is 97.62 and the DJ is 95.77. So that the SVM has the best overall results that can be best used to detect an intrusion in IDSs. In future research, we will explore more ML algorithms of features selection and classification along with new datasets.

Acknowledgement:

This research is fully supported by Universiti Tun Hussein Onn Malaysia (UTHM) under Tier 1 Grant Scheme Vot H237.

References

[1] Richariya, V., Singh, U. P., and Mishra, R. (2012). Distributed approach of intrusion detection system: Survey. *International Journal of Advanced Computer Research*, *2*(4), 358.

[2] Khalaf, B. A., Mostafa, S. A., Mustapha, A., Mohammed, M. A., and Abduallah, W. M. (2019). Comprehensive review of artificial intelligence and statistical approaches in distributed denial of service attack and defense methods. *IEEE Access*, *7*, 51691-51713.

[3] Ahmad, I., Basheri, M., Iqbal, M. J., and Rahim, A. (2018). Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE Access, 6, 33789-33795.

[4] Primartha, R., and Tama, B. A. (2017, November). Anomaly detection using random forest: A performance revisited. In *2017 International conference on data and software engineering (ICoDSE)* (pp. 1-6). IEEE.

[5] Mostafa, S. A., Mustapha, A., Shamala, P., Obaid, O. I., and Khalaf, B. A. (2020). Social networking mobile apps framework for organizing and facilitating charitable and voluntary activities in Malaysia. *Bulletin of Electrical Engineering and Informatics*, *9*(2), 827-833.

[6] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018, January). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *ICISSP* (pp. 108-116). Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152).

[7] Farnaaz, N., and Jabbar, M. A. (2016). Random forest modeling for network intrusion detection system. *Procedia Computer Science*, *89*(1), 213-217.

[8] Fadel, H., Hameed, R. S., Hasoon, J. N., & Mostafa, S. A. (2020). A Light-weight ESalsa20 Ciphering based on 1D Logistic and Chebyshev Chaotic Maps. *Solid State Technology*, *63*(1), 1078-1093.

[9] Li, X., Chen, W., Zhang, Q., and Wu, L. (2020). Building Auto-Encoder Intrusion Detection System Based on Random Forest Feature Selection. *Computers & Security*, 101851.

[10] Gunarathne, W. H. S. D., Perera, K. D. M., & Kahandawaarachchi, K. A. D. C. P. (2017, October). Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 291-296). IEEE.

[11] Akbulut, A., Ertugrul, E., and Topcu, V. (2018). Fetal health status prediction based on maternal clinical history using machine learning techniques. *Computer methods and programs in biomedicine*, *163*, 87-100.

[12] UNP, Canadian Institute for Cybersecurity, Available, https://www.unb.ca/cic/datasets/ids-2017.html.

[13] Guenther, N., and Schonlau, M. (2016). Support vector machines. The Stata Journal, 16(4), 917-937.

[14] Chang, C. C., and Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), 1-27.

[15] Bamakan, S. M. H., Wang, H., Yingjie, T., and Shi, Y. (2016). An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization. Neurocomputing, 199, 90-102.

[16] Khalaf, B. A., Mostafaa, S. A., Mustapha, A., Ismaila, A., Mahmoudb, M. A., Jubaira, M. A., and Hassana, M. H. (2019). A Simulation Study of Syn Flood Attack in Cloud Computing Environment. *AUS Journal*, 1-10.

[17] Farnaaz, N., and Jabbar, M. A. (2016). Random forest modeling for network intrusion detection system. Procedia Computer Science, 89, 213-217.

[18] Liu, Y., Wang, Y., and Zhang, J. (2012, September). New machine learning algorithm: Random forest. In *International Conference on Information Computing and Applications* (pp. 246-252). Springer, Berlin, Heidelberg.

[19] Khalaf, B. A., Mostafa, S. A., Mustapha, A., and Abdullah, N. (2018, August). An adaptive model for detection and prevention of DDoS and flash crowd flooding attacks. In *2018 International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR)* (pp. 1-6). IEEE.