A comparative analysis on traditional wired datasets and the need for wireless datasets for IoT wireless intrusion detection

Teh Boon Seong[1], Vasaki Ponnusamy[2], NZ Jhanjhi[3], Robithoh Annur[4], M N Talib[5]
[1,2,4]Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman Kampar, Malaysia
[3]School of Computer Science & Engineering, SCE, Taylor's University, Malaysia
[5]Papua New Guinea University of Technology, Lae, PNG

Article Info

Article history:

Received May 10, 2020
Revised Mar 9, 2021
Accepted Mar 29, 2021

Keywords:

Internet of things
Intrusion detection system IDS
Machine learning
Support vector machine
Wired dataset
Wireless dataset

ABSTRACT

IoT networks mostly rely on wireless mediums for communication, and due to that, they are very susceptible to intrusions. And due to the tiny nature, processing complexity, and limited storage capacities, IoT networks require very reliable intrusion detection systems (IDS). Although there are many IDS types of research available in the literature, most of these systems are suitable for wired network environments, and the benchmark datasets used for these research works are mostly relying on wired datasets such as KDD Cup'99 and NSL-KDD. IoT and wireless networks are distinct in nature as wireless networks give more emphasis on the data link layer and physical layer. These concerns are not given much attention in traditional wired datasets in the body of knowledge. Therefore, in this research, an IDS system is developed using a newly available IoT wireless dataset (NaBIoT) in the literature with the datasets focusing much on the common IoT related attacks, and related layers are taken into consideration. The IDS system developed is evaluated by comparing with various machine learning algorithms in terms of evaluation metrics such as accuracy, F1 score, false positive, and false negative. Moreover, the IoT wireless dataset is compared against the traditional NSL-KDD datasets to evaluate the need for IoT wireless datasets. The NaBIoT datasets show its effectiveness in detecting wireless intrusions. Besides that, the simulation is performed with different combinations of features to conclude that certain features are primary in detecting attacks, and IDS does not require all the features to perform detection. This can reduce the detection time mainly for machine learning and creating the models. This research results have proposed some of the critically important features to be used and eliminating not such important features.

Corresponding Author:

Noor Zaman Jhanjhi
School of Computer Science and Engineering SCE
Taylor's University, Malaysia
Email: noorzaman.jhanjhi@taylors.edu.my

1. INTRODUCTION

Wireless networks are becoming much more important nowadays due to the popularity of IoT networks. Among that, Wi-Fi and cellular networks are much promising communication mediums for IoT networks [1]. The security measure used in Wi-Fi is IDS, which is a detection system widely used for every network security infrastructure [2]. Wireless networks utilizing Wi-Fi and Cellular communications are one of the main pillars for IoT networks. IoT is a network that consists of physical or digital devices, smart cars, home appliances, and drones. that are embedded with electronics such as sensors. They can send, receive, and share data without the help of a human [3]. IoT devices are very flexible, allowing them to be used in any sector. For instance, sensors can be installed in cities to collect data on road conditions which can be used by the traffic management system to monitor and analyze the traffic flow. The system will be able to adjust the traffic light according to the traffic, reducing the possibility of congestion. Therefore, IoT becomes very popular and has a major impact on the development of society. However, a wireless network is susceptible to malicious activities such as eavesdropping, the man in the middle (MiTM), denial of service (DOS), and malware. According to [4] new types of attacks have been discovered at an astonishing speed. One of the best tools to combat attacks in wireless network is IDS as it is the final line of defence in most of the defense-in-depth systems.

The intrusion detection system can be one of the best tools to mitigate attacks in wireless networks. An intrusion detection system can be designed by using machine learning algorithms to detect possible matched between a known traffic unknown traffics. There are various IDS systems in place, such as signature-based, anomaly-based, and others. Machine learning algorithms such as support vector machine, naïve Bayes, decision tree, and others are deployed to design an IDS in wired and wireless networks.

2. PROBLEM STATEMENT

One of the main research gaps exists in the IDS design for IoT is, although many machine learning-based IDS have been proposed previously, the work relies traditionally on commercially available KDD Cup data and this dataset is not suitable for IDS in IoT as KDD Cup data is designed for traditional wired network attacks and it is very outdated. Moreover, IDS design for IoT has been researched by [5] by applying autoencoders to IoT network traffic for anomaly detection as a complete means of detecting attacks, but the authors could not conclude on the crucial features needed for detecting attacks. There is no unique work employing wireless datasets to design IoT IDS by using a wide range of IoT traffic data. A closer work to this is by [6], but the researchers have been focusing on KDD Cup data, which is not suitable for IoT IDS, as mentioned earlier. Our work uses a wireless dataset simulated using the IoT environment, and through simulation results, we can identify crucial features for detecting attacks and which combination of features gives the best performance.

3. RELATED WORK

A genetic algorithm-based IDS was developed by [7], which primarily relies on KDD datasets again. Another work by [8] has been developed using a support vector machine (SVM) as it generalizes the performances of the IDS using KDD cup datasets. There are many hybrid IDS techniques have been proposed whereby two different classification methods have been used for intrusion detection purpose. A work by [9] uses a combination of SVM and genetic algorithm for intrusion detection and feature selections.

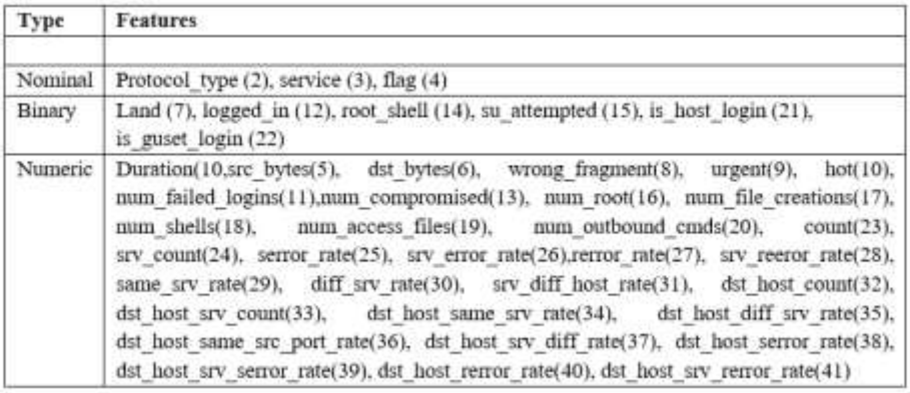
Network traffic is normally obtained either by using a wired or wireless medium. But the nature of these data differ because of because there are different attack vectors are targeted at both wired and wireless medium. According to [10] wired network is composed of an access medium that is well secured whereas wireless network requires airspace monitoring due to open medium of wireless networks. Most of the IDS work in the literature relies on KDD Cup ’99, its improved version of NSL-KDD, and Kyoto 2006+ datasets. Therefore there exists a huge gap in wireless IDS design whereby some wireless parameters are crucial when it comes to classification and detections.

4. PROPOSED DETECTION METHOD

The proposed work in unique in that simulated annealing based feature selection is done by having a combination of three random features at each round, and different machine learning algorithms are applied to each combination of features. The purpose of this combination is to identify the most suitable features required for intrusion detection, and this can greatly reduce the training and testing required and subsequently can reduce the detection time. The same method is also applied to traditional NSL-KDD Cup datasets as shown in Table 1, to compare the performance of wireless datasets with traditional wired datasets.

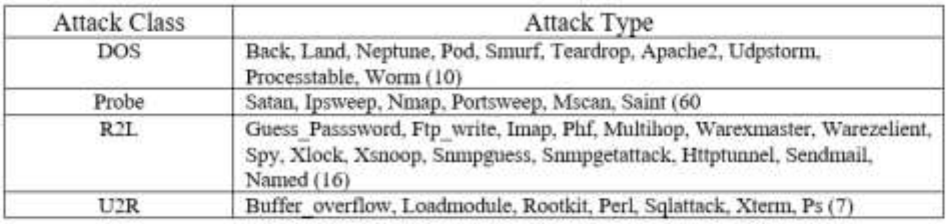
In this experiment, NSL-KDD has been used for training the model since it is suitable for supervised models. As shown in Table 2, there are 4 types of attacks taken into consideration, which are DOS, probing, U2R, and R2L. DOS will utilise all the resources of the victim and would leave the victim unable to respond or process for any legitimate requests. Probing is normally scanning for open port to look for vulnerabilities and attack through these ports. In U2R a local user gets authenticated to super user mode without authorization. For the R2L, a remote device can get grant access to other devices without authorization, and the attacker gains access to the target machine [11].

Table 1. Attribute value type of NSL-KDD [11]



Type	Features
Nominal	Protocol_type (2), service (3), flag (4)
Binary	Land (7), logged_in (12), root_shell (14), su_attempted (15), is_host_login (21), is_guset_login (22)
Numeric	Duration(10,src_bytes(5), dst_bytes(6), wrong_fragment(8), urgent(9), hot(10), num_failed_logins(11),num_compromised(13), num_root(16), num_file_creations(17), num_shells(18), num_access_files(19), num_outbound_cmds(20), count(23), srv_count(24), serror_rate(25), srv_error_rate(26),rerror_rate(27), srv_reeror_rate(28), same_srv_rate(29), diff_srv_rate(30), srv_diff_host_rate(31), dst_host_count(32), dst_host_srv_count(33), dst_host_same_srv_rate(34), dst_host_diff_srv_rate(35), dst_host_same_src_port_rate(36), dst_host_srv_diff_rate(37), dst_host_serror_rate(38), dst_host_srv_serror_rate(39), dst_host_rerror_rate(40), dst_host_srv_rerror_rate(41)

Table 2. Mapping of attack class with attack type [11]



Attack Class	Attack Type
DOS	Back, Land, Neptune, Pod, Smurf, Teardrop, Apache2, Udpstorm, Processtable, Worm (10)
Probe	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint (60
R2L	Guess_Passsword, Ftp_write, Imap, Phf, Multihop, Warexmaster, Warezelient, Spy, Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named (16)
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps (7)

The workflow of the system is shown in Figure 1. In this research, there are two types of dataset used to train the model, which is wired and wireless. The wired dataset, which has 41 features, was used to train the model is KDDTrain+ with testing dataset KDDTest+. Besides, for wireless datasets, it is from the NBaIoT dataset [12] consists of 115 features. For the NSL-KDD dataset, the top 10 features were calculated using ANOVA. The top 10 features were used to train the model, and the precision, recall, f score, and accuracy was used to determine the performance of the model. Besides, 10 groups of random combinations of the top 10 features are used to train the model to get the accuracy score. Each of the group consists of 3 features from the top 10 features. NBaIoT dataset consists of two attacks, which are the Mirai attack and the Bashlite attack, which also known as Gafgyt, Q-Bot, Torlus, LizardStresser, and Lizkebab. Both of this attack is distributed denial of service (DDoS) attack which uses the internet of things (IoT) device as a botnet to perform this attack to a target. It is controlled by a botmaster to launch an attack. There are 5 different files in the Mirai attack, which are scan, ACK, SYN, UDP, and UDP plain. These 5 files are the attack the launch by Mirai attack. The scan is the file which consists of the packet the Mirai attack perform scanning for the vulnerable device in the network. In addition, ACK, SYN, UDP, and UDP plain consist of the packet that used to flood a target device. Bashlite attack is also one of the attacks in IoT, which is also performing DDoS attacks by using an infected botnet. In the Bashlite attack, it consists of 5 files, which are scan, junk, UDP, TCP, and combo. Similar to the Mirai attack, the Bashlite attack also consists of scanning vulnerable devices in the network. The wireless train dataset was created by combining the 5 files in Mirai attack and normal traffic, which is the benign traffic provided in the NaBIoT dataset. For the wireless test dataset, it also consists of 5 files in Mirai attack and benign traffic. On top of that, the wireless test dataset was also combined with 5 files in the Bashlite attack.

The next flow is to have a feature selection of the independent feature. The algorithm that used to calculate the feature score is SelectKBest(score_func=f_classif, k=10). The function will calculate the top 10 feature scores, and the parameter used as the score function is ANOVA (f_classif). The parameter k is the number of features that are needed as a result after the calculation of the feature scores. In this research, the top 10 features are needed. Therefore, the k value will be 10. The relevant features regarding a set of class labels were measured by ANOVA, and it is used for feature selection in various fields. Some works in the literature show ANOVA is a powerful statistic test. It can be used to rank individual feature relevance in supervised classification [13]-[14]. The top 10 feature score was calculated, and the order of the feature score starts from the highest to the lowest. In Table 3, the feature score, which has the highest score, is same_srv_rate, which has a score of 1.64e+05for NSL-KDD, and the feature that has the lowest score is count which has the score of 6.27e+04. Besides, the top 10 scores for NaBIoT dataset was also calculated by SelectKBest(score_func=f_classif, k=10). Same as NSL-KDD feature scoring, the function used is SelectKBest(score_func=f_classif, k=10), the parameter used in this function is ANOVA, which is the algorithm to calculate the top 10 scores and k =10 which is the top 10 feature scores. In Table 4, the feature that has the highest scoring in NaBIoT is MI_dir_L0.1_weight with a score of 1.52e+04, and the lowest score in the top 10 is H_L5_weight which has a scoring of 2.18e+03.

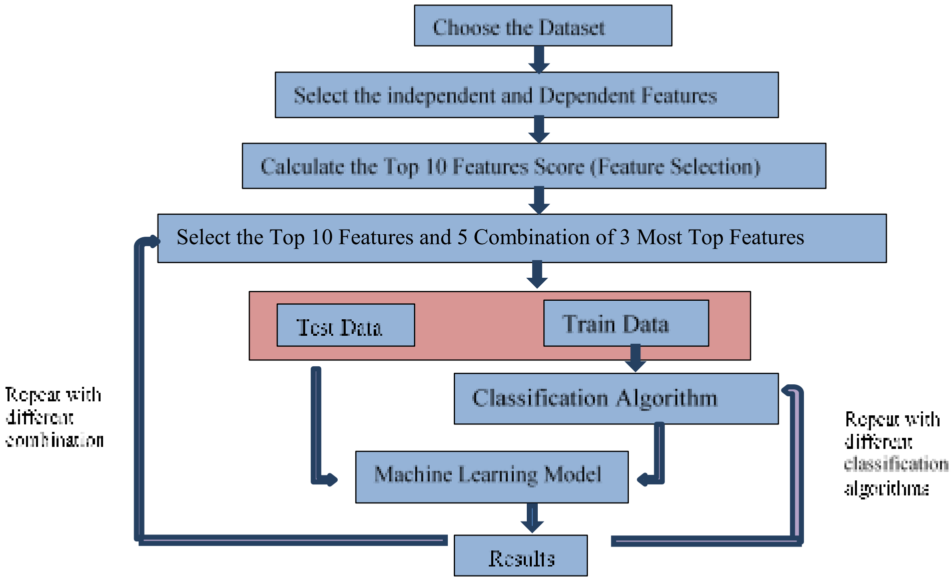

Choose the Dataset
Select the independent and Dependent Features
Calculate the Top 10 Features Score (Feature Selection)
Select the Top 10 Features and 5 Combination of 3 Most Top Features
Test Data
Train Data
Classification Algorithm
Machine Learning Model
Results
Repeat with different combination
Repeat with different classification algorithms

Figure 1. Workflow of this research

The next process for this workflow will be selecting the independent feature after feature scoring. After feature scoring for independent features, the top 10 features were selected to fit into the model in order to train the model. Besides, 5 combinations of 3 from the feature were selected to train the model as sown in Table 5 and Table 6.

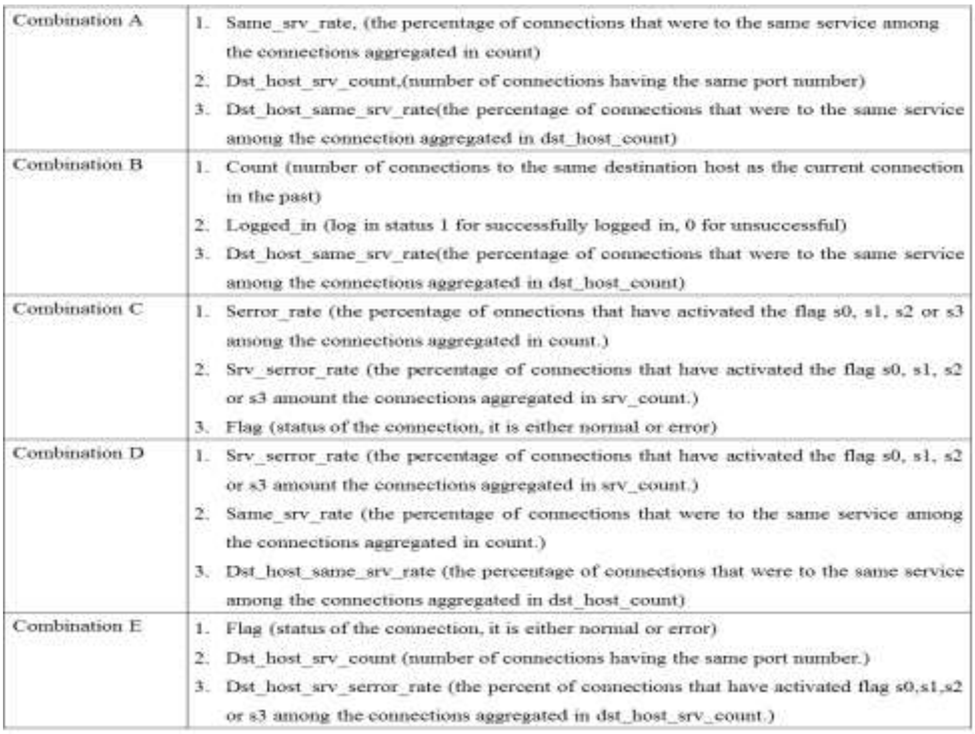
Table 3. Top 10 Feature Scores for NSL-KDD Dataset

Top 10 Features	Score
same_srv_arte	1.64e+05
dst_host_srv_count	1.38e+05
dst_host_same_srv_rate	1.17e+05
logged-in	1.15e+05
dst_host_srv_serror_rate	9.46e+05
dst_host_serror_rate	9.31e+05
serror_rate	9.25e+05
srv_serror_rate	9.13e+05
flag	7.48e+05
count	6.27e+05

Table 4. Top 10 Feature Scores for NaBIoT Dataset

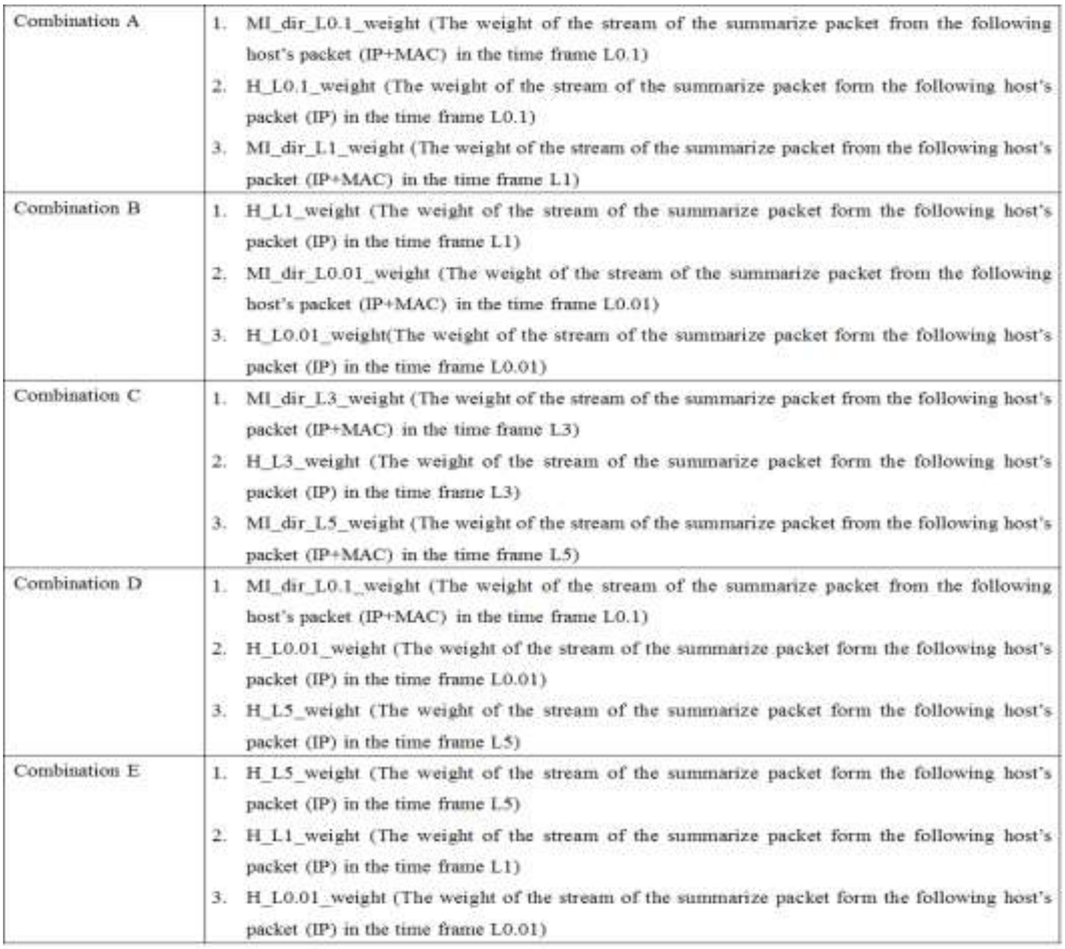
Top 10 Features	Score
MI_dir_L0.1_wieght	1.52e+05
H_L0.1_wieght	1.52e+05
MI_dir_L1_wieght	6.01e+05
H_L.1_wieght	6.01e+05
MI_dir_L0.01_wieght	3.11e+05
MI_L0.01_wieght	3.11e+05
MI_dir_L3_wieght	2.89e+05
H_L3_wieght	2.89e+05
MI_dir_L5_wieght	2.18e+05
H_L5_wieght	2.18e+05

Table 5. Different combinations of features



Combination A	1. Same_srv_rate, (the percentage of connections that were to the same service among the connections aggregated in count) 2. Dst_host_srv_count,(number of connections having the same port number) 3. Dst_host_same_srv_rate(the percentage of connections that were to the same service among the connection aggregated in dst_host_count)
Combination B	1. Count (number of connections to the same destination host as the current connection in the past) 2. Logged_in (log in status 1 for successfully logged in, 0 for unsuccessful) 3. Dst_host_same_srv_rate(the percentage of connections that were to the same service among the connections aggregated in dst_host_count)
Combination C	1. Serror_rate (the percentage of onnections that have activated the flag s0, s1, s2 or s3 among the connections aggregated in count.) 2. Srv_serror_rate (the percentage of connections that have activated the flag s0, s1, s2 or s3 amount the connections aggregated in srv_count.) 3. Flag (status of the connection, it is either normal or error)
Combination D	1. Srv_serror_rate (the percentage of connections that have activated the flag s0, s1, s2 or s3 amount the connections aggregated in srv_count.) 2. Same_srv_rate (the percentage of connections that were to the same service among the connections aggregated in count.) 3. Dst_host_same_srv_rate (the percentage of connections that were to the same service among the connections aggregated in dst_host_count)
Combination E	1. Flag (status of the connection, it is either normal or error) 2. Dst_host_srv_count (number of connections having the same port number.) 3. Dst_host_srv_serror_rate (the percent of connections that have activated flag s0,s1,s2 or s3 among the connections aggregated in dst_host_srv_count.)

Table 6. Combination of NaBIoT dataset



Combination A	1. MI_dir_L0.1_weight (The weight of the stream of the summarize packet from the following host's packet (IP+MAC) in the time frame L0.1) 2. H_L0.1_weight (The weight of the stream of the summarize packet form the following host's packet (IP) in the time frame L0.1) 3. MI_dir_L1_weight (The weight of the stream of the summarize packet from the following host's packet (IP+MAC) in the time frame L1)
Combination B	1. H_L1_weight (The weight of the stream of the summarize packet form the following host's packet (IP) in the time frame L1) 2. MI_dir_L0.01_weight (The weight of the stream of the summarize packet from the following host's packet (IP+MAC) in the time frame L0.01) 3. H_L0.01_weight(The weight of the stream of the summarize packet form the following host's packet (IP) in the time frame L0.01)
Combination C	1. MI_dir_L3_weight (The weight of the stream of the summarize packet from the following host's packet (IP+MAC) in the time frame L3) 2. H_L3_weight (The weight of the stream of the summarize packet form the following host's packet (IP) in the time frame L3) 3. MI_dir_L5_weight (The weight of the stream of the summarize packet from the following host's packet (IP+MAC) in the time frame L5)
Combination D	1. MI_dir_L0.1_weight (The weight of the stream of the summarize packet from the following host's packet (IP+MAC) in the time frame L0.1) 2. H_L0.01_weight (The weight of the stream of the summarize packet form the following host's packet (IP) in the time frame L0.01) 3. H_L5_weight (The weight of the stream of the summarize packet form the following host's packet (IP) in the time frame L5)
Combination E	1. H_L5_weight (The weight of the stream of the summarize packet form the following host's packet (IP) in the time frame L5) 2. H_L1_weight (The weight of the stream of the summarize packet form the following host's packet (IP) in the time frame L1) 3. H_L0.01_weight (The weight of the stream of the summarize packet form the following host's packet (IP) in the time frame L0.01)

5. SIMULATION & RESULTS

In this research, the input for the model is numerical input, and the output of the model will be categorical output. Therefore, ANOVA is the feature selection algorithm that used to measure the top 10 feature scores in this research. ANOVA is an algorithm that is based on a f-test that has a f distribution under the null hypothesis. The existence variance among numerous population means can be determined by ANOVA, which is a statistical technique. It is used to differentiate whether it has a difference in two or more datasets that are statistically important. The model uses numerical variables as the input, which is either integer variables or floating-point variables, and the output of the model is categorical variables, which are Boolean variables. The output of the model is either 1 or 0, which is an attack or normal. Several advantages can achieve in feature selection. It reduces overfitting to the model. In other words, there is lesser chances that when the model is making a decision it is based on the redundant data or noise. Besides, feature selection also increases the accuracy of the model if it is able to find the optimum number of features to fit into the model. This is due to there is less noise in the data, which has the possibility that it will mislead the prediction. The last advantage that is able to obtain by feature selection is reducing training time. For example, in this research, the machine learning algorithm that takes the longest time is the support vector machine. If all of the features are fitted into the machine learning model, it will be taking too long to train the model. After feature selection is done, the training time for support vector machine decreases as compared to all of the features fit into the model. The combination of 3 is randomly selected from the top 10 features of the dataset.

A comparison between NSL-KDD and NaBIoT was conducted. First, the accuracy between different combinations of features for NSL-KDD was compared. Figure 2 shows the accuracy of a different combinations of features for NSL-KDD dataset. The highest accuracy is combination C from support vector machine classifier. The algorithm SVM has the highest accuracy among all of the combinations and classifier, it has an accuracy of 0.748. The combination contains 3 important features, which are serror_rate, srv_error_rate, and flag. Serror_rate is the percentage of connections that activated the flag s0, s1, s2, and s3 among the connection in the count. Srv_error_rate is the percentage of connections that activated the flag s0, s1, s2, and s3 among the connections in srv_count. The flag is the connection status, it is either normal or error. The combination that has the lowest detection accuracy is combination C from naïve Bayes. It has only a detection accuracy of 0.574.

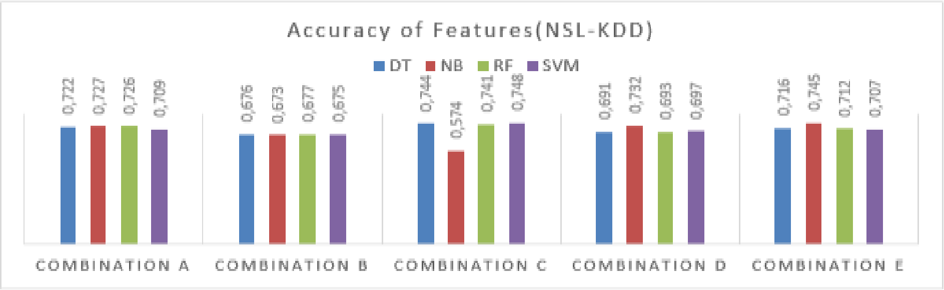


Figure 2. Accuracy of different combination of features (NSL-KDD)

Figure 3 shows the accuracy of different combinations of features for the NaBIoT dataset. The highest accuracy in NaBIoT dataset is 0.959, which is a combination B from the support vector machine classifier. The combination has 3 important features that contribute to such high accuracy. The involved features are H_L1_weight, MI_dir_L0.01_weight, and H_L0.01_weight. H_L1_weight is the weight of the stream of summarizing the packet from the following host's packet (IP) in the time frame L1. MI_dir_L0.01_weight is the weight of the stream of summarising packet from the following host's packet (IP+MAC) in the time frame L0.01. H_L0.01_weight is the stream of summarizing packet from the flowing host's packet (IP) in the time frame L0.01. The combination that gives the lowest detection accuracy is combination B from the support vector machine. Therefore, feature selections give a certain degree of contribution in order to increase the accuracy of the model. In addition, the accuracy of the support vector machine from NaBIoT was compared with a related work which is done by [15]. The datasets used in the related work is from the Cyber Range Lab of the Australian Centre for Cyber Security. The detection accuracy for support vector based detection in the related work is 0.880 while the accuracy for support vector machine from NaBIoT dataset in this research is 0.959. A different dataset is used to train the support vector machine classifier.

In addition, the performance matrix of the combinations and classifier for the NSL-KDD dataset was measured in a false positive rate, as shown in Figure 4 and the true positive rate shown in Figure 5. The lowest false positive rate among all of the combinations is combination C from naïve Bayes, which has a score of 0.00536. The highest true positive rate among all of the combinations is combination A from naïve Bayes, which has a score of 0.579. Although both of this combination from naïve Bayes gives the best result in false positive rate and true positive rate. But it does not give the best result in accuracy, which has been discussed. In terms of overall, combination C from the support vector machine has the most balanced score. For combination C from naïve Bayes, it does gives the best result in a false-positive rate but the performance for true positive rate is only 0.256 and with an accuracy score of 0.574. For combination A from naïve Bayes, it does give the best result in the true positive rate, but the performance for false positive is only 0.0601. With the comparison in terms of overall in accuracy, false-positive rate, and true positive rate. The best combinations that perform the best is combination C from support vector machine, which has a score of a false positive rate of 0.00834, a true positive rate of 0.563, and an accuracy of 0.748.

The performance matrix for combination and classifier for the NaBIoT dataset was measured in a false positive rate shown in Figure 6, and the true positive rate shows in Figure 7. The lowest false positive rate among all of the combinations of features is combination C from naïve Bayes classifier, which has a rate of 0.059. The highest false-positive rate positive rate among all of the combinations is combination E from the support vector machine, which has a score of 0.295. The classifier that performs the best among all of the combinations is a support vector machine classifier, which has a score of 0.965. Among all of the combinations, combination B from the decision tree gives the lowest true positive rate, which is 0.891.

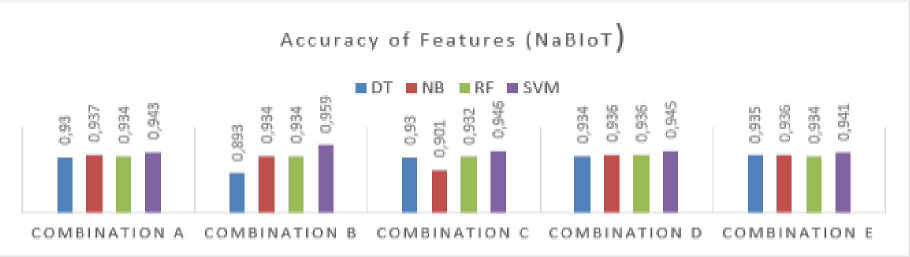


Figure 3. Accuracy of different combination of features (NaBIoT)

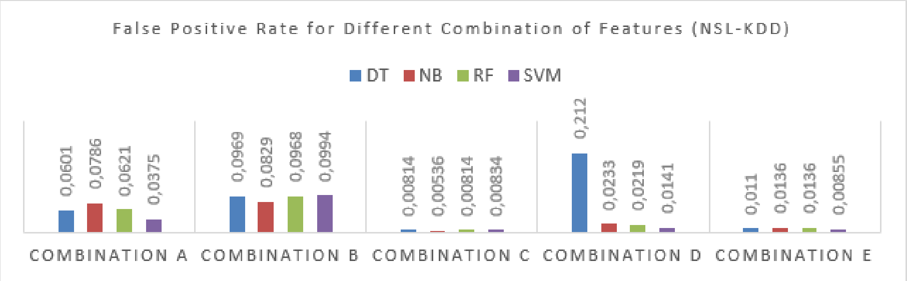


Figure 4. False positive rate for different combination of features (NSL-KDD)

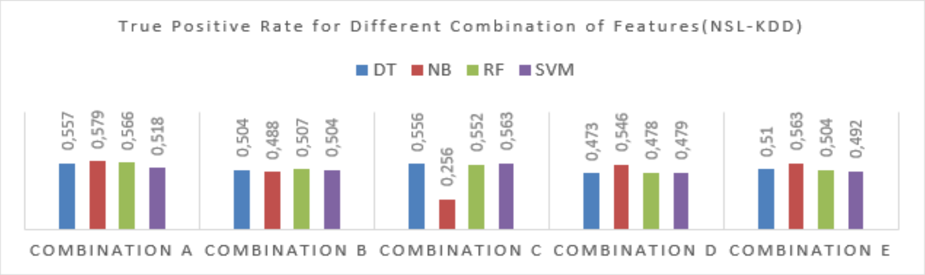


Figure 5. True positive rate for different combination of features (NSL-KDD)

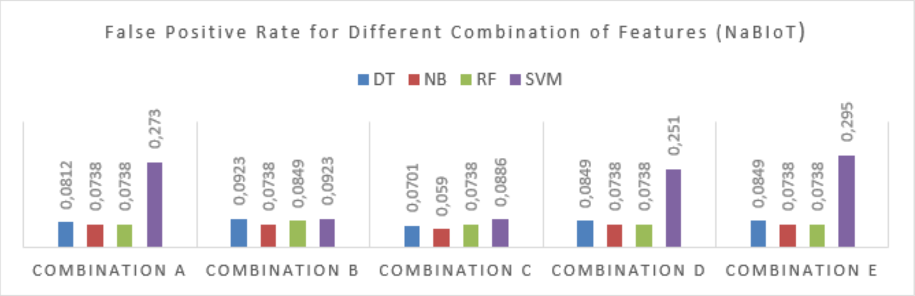


Figure 6. False positive rate for different combination of features (NaBIoT)

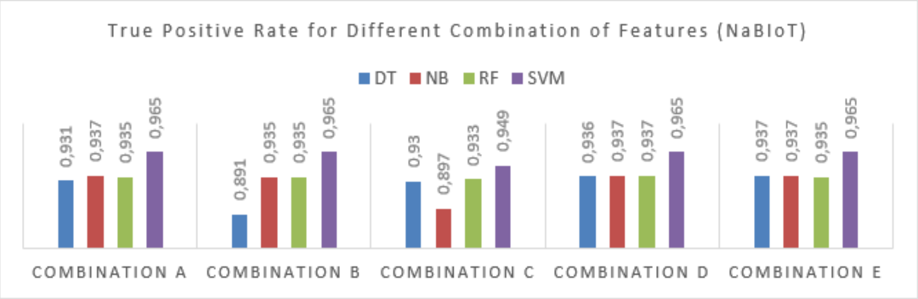


Figure 7. True positive rate for different combination of features (NaBIoT)

The f1 score is compared in both of the datasets, as shown in Figure 8.

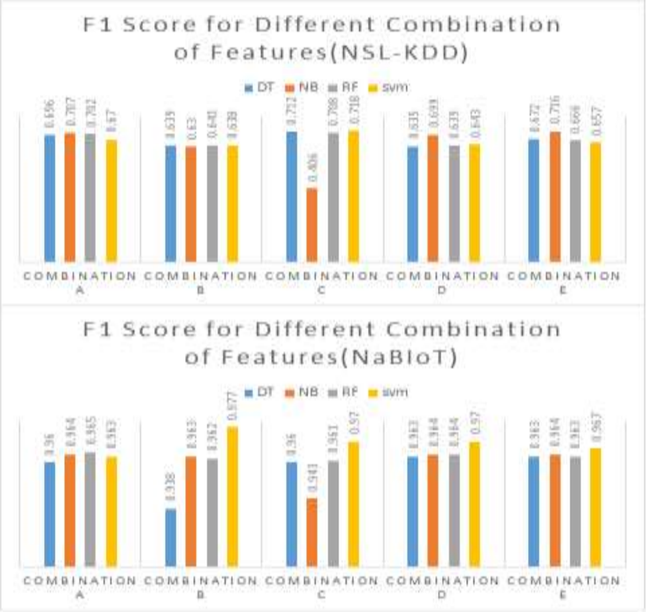


Figure 8. Comparison of F1 score

In NSL-KDD datasets, the combination and classifier algorithm that gives the highest f1 score is combination C from the support vector machine, which has a score of 0.718. The combinations of features that contribute are serror_rate, which is the percentage of connections that have activated the flag s0,s1,s2, or s3 among the connections aggregated in the count. The second feature in the combination is srv_serror_rate, which is the percentage of connections that have activated the flag s0,s1,s2, or s3 among the connections aggregated in srv_count. The last feature in the combination is a flag which is the status of the connection. It is either normal or error. The lowest among all of the combination is combination C from naïve Bayes. It only has a score of 0.406. In the NaBIoT dataset, the highest f1 score among all of the combinations is combination B from the support vector machine, which is 0.977. The first feature that contributes is H_L1_weight, which is the weight of the stream of the summarize packer from the following host's packet (IP) in the time frame L1. The second feature is MI_dir_L0.01_weight, which is the weight of the stream of the summarize packet from the following host's packet (IP+MAC) in the time frame L0.01. The last feature in the combination is H_L0.01_weight, which is the weight of the stream of the summarize packet from the following host's packet (IP) in the time frame L0.01.

6. DISCUSSION

After the F1 score of the compared, the chosen combination and classifier algorithm from both datasets is a support vector machine, as shown in Figure 9. The false-positive rate in support vector machine from the NSL-KDD dataset is 0.00834, which is lower than the support vector machine from NaBIoT, which is 0.0923. The true positive rate for support vector machine NaBIoT has a score of 0.959, which is significantly higher than NSL-KDD. The accuracy score for the support vector machine from NaBIoT is higher than the support vector machine from NSL-KDD. The accuracy score of the support vector machine from NaBIoT is 0.959. The third performance matrix to compare with recalls score. The recall score for the support vector machine from NaBIoT is significantly higher than NSL-KDD. It has a score of 0.965. The precision score for NaBIoT is higher than NSL-KDD. The highest precision score is 0.990, which is higher than 0.989. The F1 score is the final performance matrix that measured in this research. The f1 score for NaBIoT is higher than NSL-KDD, which is 0.977. Among all of this comparison, the support vector machine from NaBIoT has a performance matrix that is higher than NSL-KDD. In addition, these datasets are used in a different perspective for several other researches, few of them are such as [16]-[21]. Most of the researches use for the Internet of Things or its closely related application of wireless. Furthermore, the wireless sensor network and IoT datasets were being used by the [22]-[24]. The few of the older datasets were used by the wireless sensor networks [25]. The researcher in review [26] elaborate more IoT related IDS datasets issues and challenges.

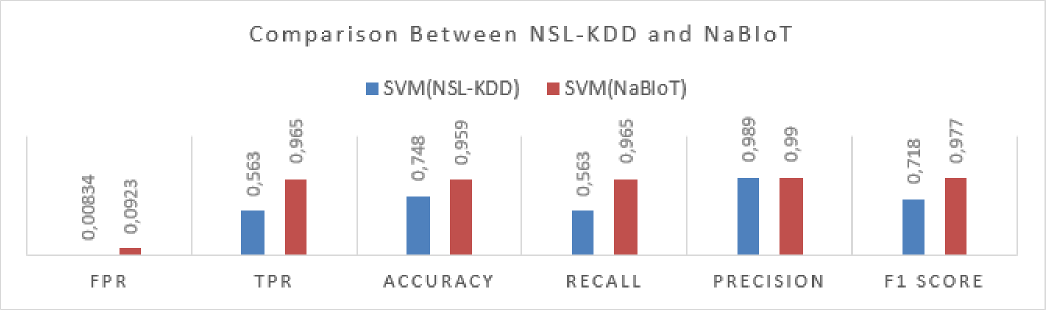


Figure 9. Comparison between NSL-KDD and NaBIoT

7. CONCLUSION

The main focus of this research is to identify the best datasets for wireless IoT IDS. A comparison was made between NSL-KDD and NaBIoT dataset with various machine learning algorithms. Besides that, with the simulated annealing approach with different combinations of features, the most prominent features necessary for IDS are identified. The dataset that performs the best in detecting normal traffic and attack traffic is NaBIoT. The highest f1 score among all of the combinations and classifiers is combination B from the support vector machine classifier. It has a f1 score of 0.977, an accuracy of 0.959, false positive rate of 0.0923, and a true positive rate of 0.965. The feature that is involved in the training of the model is H_L1_weight, MI_dir_L0.01_weight, and H_L0.01_weight. In the future, researchers in a similar domain can consider using NaBIoT datasets for IDS design in wireless networks and focus on the purposed best features to reduce the detection time whilst improving the detection accuracy.

REFERENCES

[1] E. Rattagan, “Wi-Fi usage monitoring and power management policy for smartphone background applications," In *2016 Management and Innovation Technology International Conference (MITicon)*, IEEE, pp. MIT-171, doi: 10.1109/MITICON.2016.8025223.

[2] M. E. Aminanto, R. Choi, H. C. Tanuwidjaja, P. D. Yoo, and K, Kim, “Deep abstraction and weighted feature selection for Wi-Fi impersonation detection,” *2017 IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 621-636, doi: 10.1109/TIFS.2017.2762828.

[3] M. Rouse, S. Shea, and M/ Haughn, M, “IoT devices (internet of things devices),” 2018, Dosegljivo: https://internetofthingsagenda. techtarget. com/definition/IoT-device.

[4] Calyptix, “Top 8 Network Attacks by type in 2017,” pp. 1, 2017, [Online]. Available from :<https://www.calyptix.com/top-threats/top-8-network-attacks-type-2017/> [Accessed on 12 Nov. 2018].

[5] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, “N-baiot-network-based detection of iot botnet attacks using deep autoencoders,” *2018 IEEE Pervasive Computing*, vol. 17 ,no. 3, pp. 12-22, doi: 10.1109/MPRV.2018.03367731.

[6] S. Chawla, “Deep learning based intrusion detection system for Internet of Things,” *2017 (Doctoral dissertation).*

[7] M. S. Hoque, M. Mukit, M. Bikasand, and A. Naser, “An implementation of intrusion detection system using genetic algorithm,” 2012, arXiv preprint arXiv:1204.1336.

[8] J. Jha, and L Ragha, “Intrusion detection system using support vector machine,” *International Journal of Applied Information Systems (IJAIS)*, vol. 3, pp. 25-30, 2013, doi: 10.1109/SACI.2014.6840052.

[9] B. M., Aslahi-Shahri, R. Rahmani, M. Chizari, A. Maralani, M. Eslami, M. J. Golkar, and A. Ebrahimi, A, “A hybrid method consisting of GA and SVM for intrusion detection system,” *2016 Neural computing and applications*, vol. 27, no. 6, pp. 1669-1676, doi: 10.1007/s00521-015-1964-2.

[10] Z. M. Fadlullah, H. Nishiyama, N. Kato and M. M. Fouda, "Intrusion detection system (IDS) for combating attacks against cognitive radio networks," in *IEEE Network*, vol. 27, no. 3, pp. 51-56, May-June 2013, doi: 10.1109/MNET.2013.6523809.

[11] L. Dhanabal, and S. P. Shantharajah, “A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms”, *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, no. 6, 2015.

[12] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, “N-baiot-network-based detection of iot botnet attacks using deep autoencoders,” *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12-22, 2018, doi: 10.1109/MPRV.2018.03367731.

[13] M. Peña *et al*., "ANOVA and Cluster Distance Based Contributions for Feature Empirical Analysis to Fault Diagnosis in Rotating Machinery," *2017 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC)*, Shanghai, pp. 69-74, 2017, doi: 10.1109/SDPC.2017.23.

[14] Z. A. Bakar, D. I. Ispawi, N. F. Ibrahim, and N. M. Tahir, “Classification of Parkinson’s disease based on Multilayer Perceptrons (MLPs) Neural Network and ANOVA as a feature extraction,” *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*, Melaka, pp. 63-67, 2012, doi: 10.1109/CSPA.2012.6194692.

[15] M. N., Chowdhury, K. Ferens, and M., Ferens, “Network intrusion detection using machine learning,” In *Proceedings of the International Conference on Security and Management (SAM)*, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), , p. 30, 2016.

[16] Zahrah A. Almusaylim, Abdulaziz Alhumam, N. Z. Jhanjhi, |Proposing a Secure RPL based Internet of Things Routing Protocol: A Review,” Ad Hoc Networks, Volume 101, 2020, 102096, ISSN 1570-8705, doi: 10.1016/j.adhoc.2020.102096.

[17] Diro, H. Reda, N. Chilamkurti, A. Mahmood, N. Zaman, and Y. Nam, "Lightweight Authenticated-Encryption Scheme for Internet of Things Based on Publish-Subscribe Communication," in *IEEE Access*, vol. 8, pp. 60539-60551, 2020, doi: 10.1109/ACCESS.2020.2983117.

[18] Almusaylim, Z., Jhanjhi, N, “Comprehensive Review: Privacy Protection of User in Location-Aware Services of Mobile Cloud Computing,” *Wireless Pers Commun.*, vol. 111, pp. 541-564, 2020, doi: 10.1007/s11277-019-06872-3.

[19] Fatima-tuz-Zahra, N. Jhanjhi, S. N. Brohi and N. A. Malik, "Proposing a Rank and Wormhole Attack Detection Framework using Machine Learning," *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, Karachi, Pakistan, 2019, pp. 1-9, doi: 10.1109/MACS48846.2019.9024821.

[20] Almusaylim, Z. A., Zaman, N, “A review on smart home present state and challenges: linked to context-awareness internet of things (IoT),” *Wireless Netw*, vol. 25, pp. 3193-3204, 2019, doi: 10.1007/s11276-018-1712-5.

[21] K. Hussain, S. J. Hussain, N. Jhanjhi, and M. Humayun, "SYN Flood Attack Detection based on Bayes Estimator (SFADBE) For MANET," *2019 International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, pp. 1-4, 2019, doi: 10.1109/ICCISci.2019.8716416.

[22] S. M. Muzammal, R. K. Murugesan and N. Z. Jhanjhi, "A Comprehensive Review on Secure Routing in Internet of Things: Mitigation Methods and Trust-Based Approaches," in *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4186-4210, 15 March15, 2021, doi: 10.1109/JIOT.2020.3031162.

[23] N. Zaman, T. J. Low, and T. Alghamdi, "Enhancing routing energy efficiency of Wireless Sensor Networks," *2015 17th International Conference on Advanced Communication Technology (ICACT)*, PyeongChang, Korea (South), pp. 587-595, 2015, doi: 10.1109/ICACT.2015.7224928.

[24] Kok, S. H., Abdullah, A., Jhanjhi, N. Z, and Supramaniam, M, “A review of intrusion detection system using machine learning approach,” *International Journal of Engineering Research and Technology,* vol. 12, no. 1, pp. 8-15, 2019.

[25] Zaman, N., Khan, A. R, and Salih, M, “Designing of energy efficient routing protocol for wireless sensor network (wsn) using location aware (la) algorithm,” *J. Inform. Commun. Technol,* vol. 3, no. 2, pp. 56-70, 2009.

[26] Dhuha Khalid Alferidah, N. Z. Jhanjhi, “A Review on Security and Privacy Issues and Challenges in Internet of Things,” in *International Journal of Computer Science and Network Security (IJCSNS),* vol 20, no. 4, pp. 263-286, 2020.

BIOGRAPHIES OF AUTHORS

Teh Boon Seong is a final year student, currently pursuing a degree in Communications and Networking in University Tunku Abdul Rahman, Kampar. He shows a high interest in cybersecurity and has planned to get certified as an ethical hacker once he graduates in 2020. His research interest is mainly on Intrusion Detection Systems in Wireless and IoT networks. He is self-motivated in research domains and has high passion in cybersecurity fields.

Dr. Vasaki Ponnusamy is an Assistant Professor at Universiti Tunku Abdul Rahman, Malaysia. She is the Head of the Department for the Department of Computer and Communication Technology at the Faculty of Information and Communication Technology. She obtained her Bachelor of Computer Science and MSc (Computer Science) from Science University of Malaysia and her Ph.D. in IT from Universiti Teknologi PETRONAS (UTP), Malaysia (2013). She is currently working on cybersecurity, IoT security trends, and digital governance. She has great International exposure as Keynote Speaker, Visiting Professor, and several Fellowships. She has edited/authored several research books with international reputed publishers, earned several research grants and supervising postgraduate students. She is also a Human Resource Development Fund Certified Trainer and Master Trainer for Computational Thinking and Computer Science Teaching. Her career in academia started in 1999. She has been teaching information and network security, wireless security, and data communication and networking. She is specialized in handling Cisco devices for routing and security. It is her passion to share her technical knowledge with the community. Her area of specializations are Networking, Communication, Penetrating Testing, and Cybersecurity. She has worked on several projects on IoT Intrusion Detection Systems, Cybersecurity Governance, Social Engineering Attacks Mitigation/Awareness and Human Behavior based Authentication Systems.

Dr. Noor Zaman ***(NZ Jhanjhi)*** is currently working as Associate Professor at School of Computer Science and Engineering, Taylor’s University, Malaysia. Besides of teaching, research, and administrative responsibilities, he is heading the Cybersecurity research Cluster, where he is supervising a great number of Postgraduate students mainly in the cybersecurity for Data Science. Cybersecurity research cluster has extensive research collaboration globally with several professional. He has introduced courses related to the cybersecurity specialization including, Wireless Networks and Security, Secured Software Systems, Computer Forensic, Intrusion Detection Systems. He is the link between industry partners and academia as a cybersecurity specialist for Data Science to represent the Taylor’s. A good number of postgraduate students are graduated under his supervision. Dr NZ Jhanjhi is Associate Editor and Editorial Assistant Board for several reputable journals including IEEE Access Journal, PC member for several IEEE conferences worldwide, and guest editor for the reputed indexed journals. Active reviewer for a series of Q1 journals, has been awarded globally as a top 1% reviewer by Publons (Web of Science). He has high indexed publications in WoS/ISI/SCI/Scopus, and his collective research Impact factor is more than 139 points. He has international Patents on his account, edited/authored more than 21 research books published by world-class publishers. He has a great experience of supervising and co-supervising postgraduate students, an ample number of PhD and Master students graduated under his supervision. He is an external PhD/Master thesis examiner/evaluator for several universities globally. He has completed more than 20 international funded research grants successfully. He has served as Keynote speaker for several international conferences, presented several Webinars worldwide, chaired international conference sessions. He has achieved consecutively higher performance award.

Dr. Robithoh Annur is currently an Assistant professor in the Department of Computer and Communication Technology (DCCT), Faculty of Information and Communication Technology (FICT), UTAR. She received the B. Eng. and M. Eng. degrees from Gadjah Mada University, Yogyakarta, Indonesia, and the National University of Singapore, Singapore, respectively. She was a tutor in the Diploma of Electrical engineering at Gadjah Mada University before eventually completing her doctoral degree at Chulalongkorn University, Thailand. She has already published a series of papers both in international conferences and journals that some of the acknowledged as the best paper award, with a book chapter in a book entitled "Advanced Trends in Wireless Communications, 2011. Currently, she is focusing on the algorithms for tag identification in RFID system, which is simple to implement and performs better than existing known algorithms

Dr. Muhammad Nabeel Talib is a full-time Lecturer II at PNG University of Technology, Lae Papua New Guinea. His interests revolve around Cloud infrastructure towards energy efficient data centers and service-oriented Trust.