2019 IEEE 7th Conference on Systems, Process and Control (ICSPC 2019), 13–14 December 2019, Melaka, Malaysia

Anomaly Analysis for the Classification Purpose of Intrusion Detection System with K-Nearest Neighbors and Deep Neural Network

Kayvan Atefi[1], Habibah Hashim[2], Murizah Kassim[3]

[1,2,3] *Faculty Of Electrical Engineering*
Universiti Teknologi Mara (UiTM)
Shah Alam, Malaysia

[1] atefi@ieee.org, [2] habib350@uitm.edu.my, [3] murizah@uitm.edu.my

***Abstract*—Nowadays, along with network development, due to the threats of unknown sources, information communication is more vulnerable and require more secured information. An Intrusion Detection System (IDS) is important for protecting information with growing of unauthorized activities in-network. Traditional firewall techniques are less capable to protect information against new intrusion. Numerous researches on intrusion detection system have been conducted but old dataset like Kddcup'99 is analyzed. Problem identified that lack of accuracy to detect intrusion with the current available intrusion system. Hence this study aims to anomaly analysis for the classification purpose of the intrusion detection system with the most update dataset named CICIDS-2017 which can be used for the intrusion detection evaluation. This research has conducted the anomaly analysis for the classification purpose based on the K-Nearest Neighbors (KNN) for the machine learning (ML) and Deep Neural Network (DNN) using the Deep Learning (DL) method. One of the results presents a classification performance based on Matthews Correlation Coefficient (MCC) for ML and DL. DNN has performed significantly higher correctness classifier which shows DNN score 0.9293% compared to KNN is at 0.8824%. This research is significant as reference for IDS development which would improve security response for networked systems.**

Keywords*—*Anomaly Analysis, Intrusion Detection System, K-Nearest Neighbors, Deep Neural Network

I. INTRODUCTION

Today, along with network development, due to the threats of unknown sources, information communication is more vulnerable, thus, more secure information is required. Various threats could be supervised by creating effective IDS for supplying security towards the network [5]. Currently, attacking network infrastructure is the major problem for securing the information and network. By growing unauthorized activities in the network, an IDS is required for protecting information as a component of defense since traditional firewall techniques are not capable to protect information against intrusion completely [1]. Based on reference [5], in order to secure communication, an intelligent analysis must be conducted in a large amount of live data which were already secured in network devices. IDS has become one of the most applied methods for securing data against exterior threats.

This research proposes anomaly analysis for the classification purpose of the IDS with the most update dataset named "CICIDS-2017" which can be used for the intrusion detection evaluation. Moreover, this research conducts the classifications of intrusion based on the K-Nearest Neighbors (KNN) for the machine learning (ML) and Deep Neural Network (DNN) for the Deep Learning (DL) method. Moreover, some descriptions within the ML and DL are highlighted in this research to detect intrusion.

II. BACKGROUND OF STUDY

Foundation of IDS was launched in the 1980s after publishing Denning's work about ID. In recent years, many IDSs had been provided as research prototypes and for ensuring commercial procedures. Various methods were used by scalping strategies to identify unpermitted activities. IDS are among fundamental technologies which ensure system dynamic for security. The purpose and aim of ID would be to identify both attacks and non-attacks. IDS could generate a real-time response to the intrusion situations and offensive processes by studying the procedure and signature of intrusion behaviors. [1]. Nowadays various cyber-crimes are committed through attackers of different intentions. Increasing cyber-attacks beside their complexity have caused significant economic, national and international security troubles attracting the attention of experts and researchers of security around the world [10][11]. These threats have been transformed recently into a more organized form to make illegal revenues. Intrusion is among the most effective platforms having a network of sophisticated malware conducting an expanded range of cyber- attacks [15].

A. *Intrusion Detection System (IDS)*

Undoubtedly, there has been advancements in network security in recent years and several new techniques for network protection has been proposed. IDSs are usually used as countermeasures or security checks to check, detect, and notify any kind of non-permitted application, misuse or abuse of information system or asset's network [8]. Tyler, 2008, argues that network infrastructure attacks will pose hazards to network security and cyber security. Due to increasing non-permitted functions in the network, IDS has become a necessity since traditional firewalls are not able to provide sophisticated security against the intrusion. The IDS is now a significant field in network security research.[7] and in Figure 1,we can get an idea of how IDS is usually employed by a network.

Anomalies can be detected through classification of intrusions based on meta-heuristic approaches and in this paper we compared the performance of a machine learning method (ML) and a deep learning (DL) method.

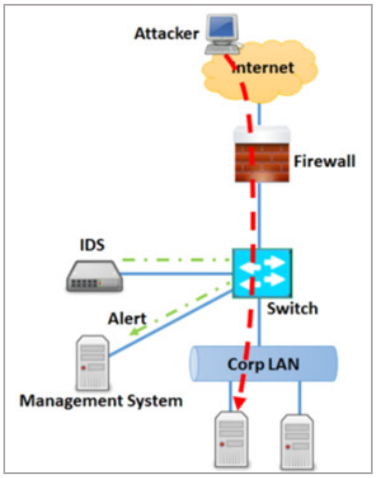


Fig. 1. Intrusion Detection System[25]

B. *K-Nearest Neighbors (K-NN)*

K-NN is a strategy and method applied for regression and classification[8]. In both regression and classification, the input includes the k nearest training instances in the future space. Thus, the output relies on k-NN usage. An object is classified through the vote of its neighbors so that an object is delegated to the most widely used class within k nearest neighbors (k is a positive integer, typically small) [2][16-17][20][23]. Figure 2 presents suggestions for the algorithm of quick explanation to K-Nearest Neighbors.

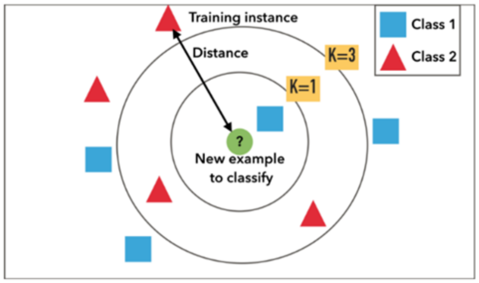


Fig. 2. Quick Explanation to K-Nearest Neighbors Algorithm.[24]

C. *Deep Neural Network (DNN)*

A DNN consists of an ANN and deep learning including layers among the output and input layers[3][12]. It gets the right manipulation of mathematics to transform the input to the output, either as a non-linear or linear relationship. This network moves via the layers while measuring the possibility of outputs.[13]. It provides computational models having multiple processing layers with learning to represent data of multiple abstraction levels. The state of the art in object detection, visual object recognition, speech recognition, as well as many other fields such as genomics and drug discovery have been significantly improved through these methods[14].

D. *Intrusion Detection Evaluation Dataset (CICIDS-2017)*

The references, [18][21][22], proposed that IDSs are the most important instruments of protection against advanced and growing attacks on the network. Because of the apparent lack of reliable testing and valid datasets, intrusion detection strategies based on anomalies suffer from constant and accurate evaluations of performance. Some of the previous datasets suffer from the lack of traffic diversity and volumes, some do not involve the diverse known attacks, while others anonymize packet payload data, thus it is not able to reflect the current trends. CICIDS-2017 dataset is used in this research study [21].

The CICIDS2017 dataset includes secure and the most up-to-date prevalent attacks resembling the true actual-world data (PCAPs). It also contains the outcomes of the network traffic associated by labeled flows according to the time stamp, source, and destination IPs, source and destination ports, protocols and attack. Creating real background traffic was the main purpose of making this dataset. This dataset arranged the abstract behavior according to the HTTP, HTTPS, FTP, SSH, and email protocols. Eleven criteria were defined through evaluation in the mentioned dataset which is required for a reliable dataset. None of the previous IDS datasets was able to cover everything of the eleven criteria. For example, considering the labeled dataset, it has secure and attacks labels for each day and the feature set. More than 80 network flow features were derived from the generated network traffic. Furthermore, this dataset contained the most prevalent attacks according to report of McAfee, including Web-based, Brute force, DoS, DDoS, Infiltration, Heart-bleed, Bot, and Scan which had been covered in this dataset as attack diversity [14][18][21-22].

III. METHODOLOGY

The methodology discusses on the procedure applied and operations in meeting the objectives of the research. Results are visualized where MATLAB software is used in the experimental process and study. DL and ML techniques are used in the research method.

Figure 3 shows the complete system block diagram and the overall of the experimental view. Figure shows stage for classification with machine learning and deep learning. As can see in this phase the classification is based on the K-NN and DL methods. In this section, the researcher used the whole set of datasets for the classification purpose. After pre-processing of the dataset and select the number set of datasets, the machine is categorized for machine learning and deep learning methods. The machine learning method has used the K-NN technique. The procedure presents the first phase of the dataset, after pre-processing is done then KNN is derived. The next process is labeling which in the next, the results shows the outcome based on the metrics of accuracy, time and error.

After finishing this phase, the results showed the accuracy, time and other predefined metrics for each stage as the overall view in Figure 3. Each stage is compared, and results are visualized on charts and graph. Next, evaluation and validation process of the model is done based on the defined metrics, after the model implementation and results are produced. Figure 4 indicates the overall and the experimental model of study.

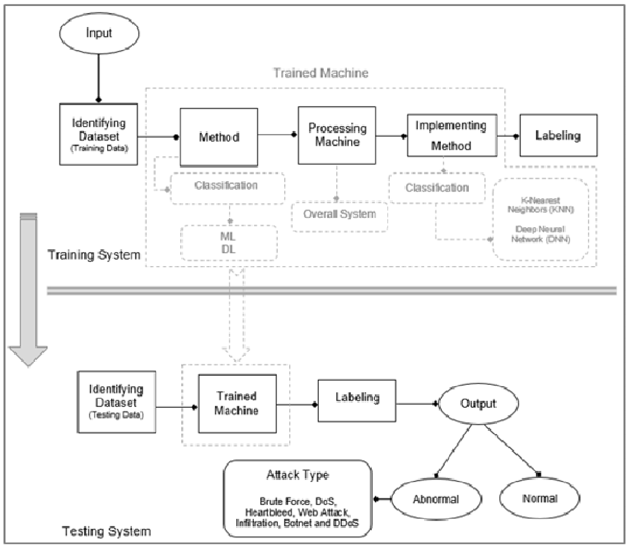


Fig. 3. Complete System Block Diagram

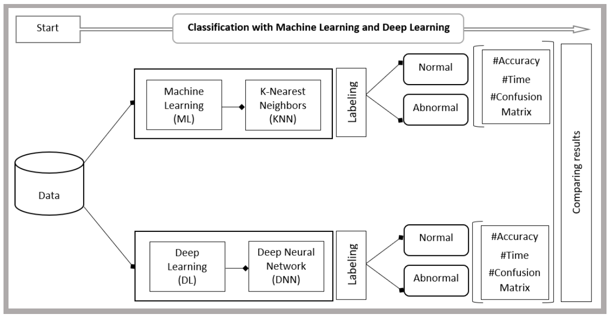


Fig. 4. The Overall and Experimental View of System

IV. Analysis and Results

All the classification techniques are implemented in the MATLAB and the experimental sets are done on windows 10 operating system with a 1.99 GHz CPU and 24 GB memory. The figures and tables show the evaluation of classification methods to be achieved by the average of 100 runs to evaluate the efficiency of intrusion detection. Thus, 80% of the data has been used for training each iteration, whilst the remaining 20% was used for testing in a process. This stage is the outcome based on the ML and DL and shows the comparison of a perfect platform of classification for intrusion detection. Based on the results, it shows which DL is a better platform for implantation. Stage one is a combination of two levels and lastly, each level had their own results based on the important metrics like accuracy, time and error.

The classification in this part is based on normal and abnormal data of different type attacks. The goal of this work is to comparison deep learning and machine learning to show the performance analysis in terms of anomaly detection and shows its performance results.

A. Analysis Classification Performance based on Accuracy, Recall and Precision Measurement for ML and DL

Analysis Classification Performance based on Accuracy, Recall and Precision Measurement for ML and DL are collected as normal and abnormal had been carried out. Different ML and DL, which for the ML implement the KNN and for the DL implement the DNN. Figure 5 shows the results for accuracy, precision, and recall. The accuracy is identified as the percentage of overall objects which is correctly classified. The recall which also called true positive rate is indicating the amounts of items out of the total correctly identified as positive true positive. Lastly, precision presented the amounts of items properly recognized as positive out of the complete positive items. The presented graph shows the accuracy, precision, and recall within the ML and DL classification methods.

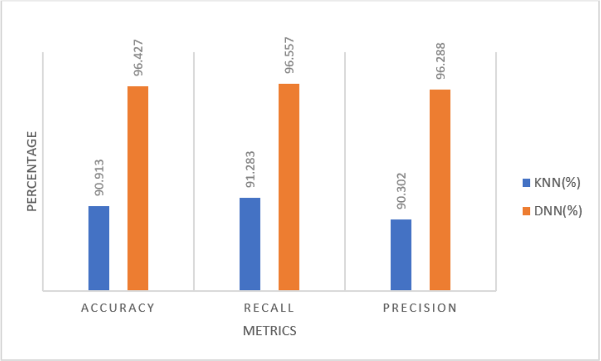


Fig. 5. Reported the results for accuracy, precision, and recall within the ML and DL classification methods

Two given platforms namely, KNN and DNN with three metrics namely accuracy, recall and precision are compared and measured. KNN and DNN scored 90.913% and 96.427% respectively in accuracy. KNN and DNN scored 91.283% and 96.557% respectively for recall and KNN and DNN scored 90.302% and 96.288% respectively for precision. DNN shows the highest scores in all two metrics under this experiment. The KNN scored the lowest percentages in all two metrics under this experiment.

The difference between the highest and lowest is 5.514% which shows the validity of selecting DNN as the good platform as classification. According to the results, we can conclude that DNN not only scored the highest percentages throughout the test of different metrics but also show consistency in results.

Therefore, based on the theses results we can ascertain that DNN is the best suitable platform to execute this experiment in term of classification of anomaly detection. The lowest results are achieved for classification, 90.913 % accuracy, 91.283% recall and 90.302% precision which belongs to KNN. Based on the results from Figure 5 it can be realized that the achieved results of DNN are greater than other algorithms.

B. Analysis Classification Performance based on , CPU Time and Elapsed time Measurement for ML and DL

Performance is an important aspect of IDS that is based on the techniques and methods which had been for classification. If the number iteration and procedure within the classification methods are high, then the complexity of classification will increase in terms of the elapsed time and CPU time. The observations showed that KNN techniques need more time for the classification process than DNN techniques.

Thus, the processing time for classification and clustering normal and abnormal data is so important [19]. Figure 6 and Figure 7 show the CPU time and elapsed time consumed by ML and DL based on the KNN and DNN.

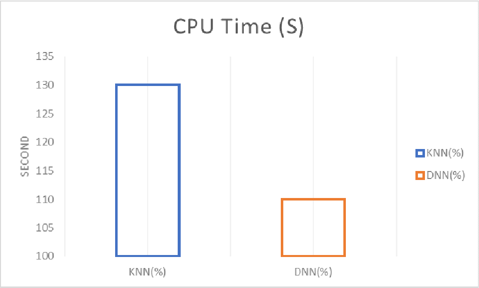


Fig. 6. Show the CPU time consumed by ML and DL based on the KNN and DNN

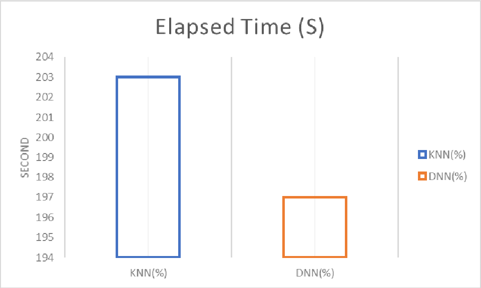


Fig. 7. Show the elapsed time consumed by ML and DL based on the KNN and DNN

The performance assessment is implemented to take into account how a different method can increase CPU time and elapsed time during classification and detection process. The results showed that the value of CPU time and elapsed time increase with using KNN.

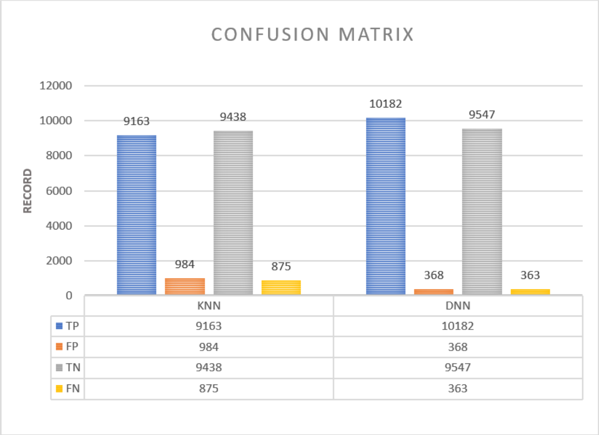
The worst and the best values of CPU time are 130 (s) and 110 (s) for KNN and DNN in the defined dataset respectively, the worst and the best values of elapsed time are 203 (s) and 197 (s) for KNN and DNN in the defined dataset respectively.

C. Analysis Classification Performance based on Confusion Matrix for ML and DL

Figure 8 shows a comparison result of a confusion matrix for the classification performance of ML and DL. TP is the percentage of positive instances that have been recognized accurately. In this research, it considers as the amounts of attacks which are properly predicted as an offense. In this experiment, the highest rate of TP is belonging to DNN with a rate of 10182 records and the minimum rate is for KNN which is 9163records. This study presents that FP number of records of negative instances is wrongly classified as positive which it is referred to the amounts of benign events foreseen as attacks. Based on the outcome, the max and min rate is for KNN and DNN with the rates of 984 and 368 records which the lowest rate shows the better classification which has less incorrectly classified portion.

Furthermore, TN is described as the number of records of negative instances that have been correctly categorized. In this case study, it shows the number of benign occurrences effectively marked as normal. According to the results of this study, the sublime rate of TN is for DNN and the minority rate is for KNN which are 9547 and 9438 records.

However, FN is the number of positive instances classified as negative wrongly and for the outcome of this research. It illustrates the amount of attack which is wrongly predicted as normal. The utmost and minimalist portion of FN in this research respectively is 875 and 363 records for KNN and DNN. It shows the highest rate is for KNN and the min is belonging to DNN.



	KNN	DNN
TP	9163	10182
FP	984	368
TN	9438	9547
FN	875	363

Fig. 8. Indicated the Outcomes of Confusion Matrix Analysis for the Classification Performance of ML and DL

D. Classification Performance based on Matthews Correlation Coefficient (MCC) for ML and DL

The MCC has a value of -1 to 1 in which -1 show an incorrect classification, while 1 implies a right classification. Using the MCC enables you to measure how well your model/function of classification works [7]. MCC requires true and false positives and negatives into account and is widely considered a balanced measure that can be used for various class sizes. Figure 9 illustrates the results of the Matthews Correlation Coefficient analysis for the classification performance of ML and DL.

As per the given assumptions of MCC, our results show that DNN is much closer to 1 which means higher correct classifier. KNN has also shown good results but lesser than DNN. According to results KNN and DNN score 0.82261% and 0.92870% respectively. Therefore, based on these results we can ascertain that DNN has performed significantly higher correctness classifiers as compare to others.

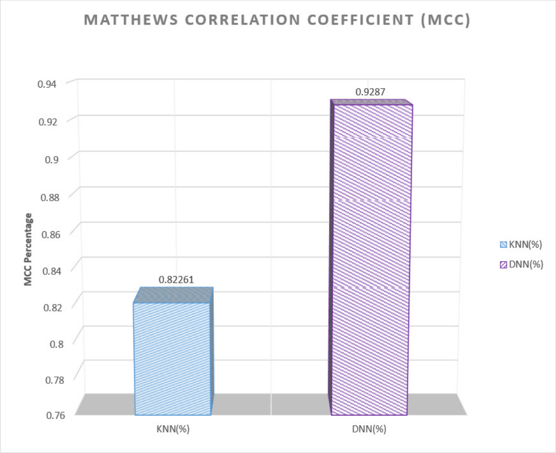


Fig. 9. Illustrate the Matthews Correlation Coefficient percentage for the classification of ML and DL

E. *Analysis Classification Performance based on , F-Measure for ML and DL*

The F-Measure is the rhythmic mean measurement based on precision-recall equilibrium. In reality, another exam is taken to ascertain in detail the accuracy of the detection scheme. The F1 score is regarded as a measure of the subject matter dataset's accurate classification or detection. Generally, the F1 score can be as high as to 1 which is the nearer the rate is to 1, the more accurate is regarded the classifier [6][19]. Figure 10 shows the max F-Measure rate is belonging to the DNN with the rate of 96.534% which the lowest is for KNN with a rate of 90.790%. The performance of classification focused is on F-Measure for ML and DL.

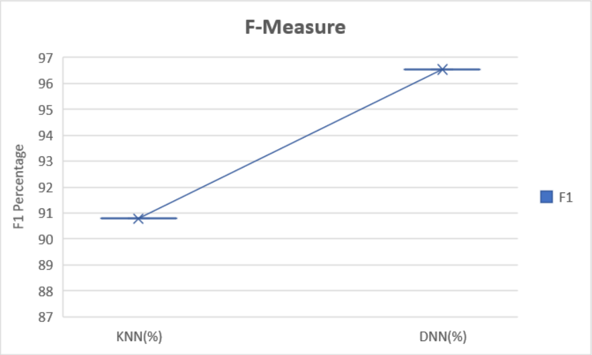


Fig. 10. Shows the Performance of Classification Focused on F-Measure for ML and DL

V. Conclusion

In this research, the authors design and implement the model to accomplish ML and DL anomaly analysis for the classification of intrusions in an IDS with the most updated dataset named "CICIDS-2017" which can be used for the intrusion detection evaluation. Moreover, this research conducts the anomaly analysis for the classification purpose based on the K-Nearest Neighbors (KNN) for the machine learning (ML) and Deep Neural Network (DNN) for the Deep Learning (DL) method. Further, the outcomes illustrated the performance in detecting intrusion based on defined metrics and the results based on the accuracy, precision, recall, time, confusion matrix and MCC for the number of events that being predicted as attacks correctly have been shown in the achieved result which is shown in the clear figures and table. In the achievement results, DNN classification has higher performance than other classification of KNN based on the defined dataset. Therefore, based on the theses results we can ascertain that DNN is the best suitable platform to execute the experiment in term of classification of anomaly detection.

Acknowledgment

The authors would like to thank the Ministry of Education for funding the research through 600-IRMI/FRGS 5/3 (150/2019) and the Faculty of Electrical Engineering, Universiti Teknologi MARA for their support.

References

[1] Yuebin Bail, Hidetsune Kobayashil (2003). Detection Systems: Technology and Development. IEEE.

[2] Kilian Q. Weinberger , L. K. S. (2009). "Distance Metric Learning for Large Margin Nearest Neighbor Classification." Journal of Machine Learning Research.

[3] Bengio, Y. (2009). ""Learning Deep Architectures for AI"." Foundations and Trends in Machine Learning. 2 (1): 1–127.

[4] N. S. Altman Biometrics Unit, C. U. (2012). "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression." Journal The American Statistician 14853 , USA: Pages 175-185.

[5] S. Aneetha, T.S. Indhu, S. Bose. (2012). Hybrid Network Intrusion Detection System Using Expert Rule Based Approach. Ccseit.

[6] Nadarajan, G. K. V. a. R. A. (2012). "HTTP botnet detection using adaptive learning rate multilayer feed-forward neural network." Springer in Information Security Theory and Practice. Security, Privacy and Trust in Computing Systems and Ambient Intelligent Ecosystems: 38-48.

[7] Manjusha, K., et al. (2013). "An Efficient Method of Spam Classification by Multiclass Support Vector Machine Classifier."

[8] Atefi, K., et al. (2013). A hybrid intrusion detection system based on different machine learning algorithms. Proceedings of the 4th International Conference on Computing and Informatics, ICOCI.

[9] Butun, I., et al. (2014). "A survey of intrusion detection systems in wireless sensor networks." IEEE Communications Surveys & Tutorials 16(1): 266-282.

[10] J. Armin, B. T., D. Ariu, G. Giacinto, F. Roli, and P. Kijewski (2015). ""2020 Cybercrime Economic Costs: No Measure No Solution"." Proceedings of the Availability, Reliability and Security (ARES), 2015 10th International Conference: 701-710.

[11] Me, G. B. a. G. (2015). "A Survey on Financial Botnets Threat." Global Security, Safety and Sustainability: Tomorrow's Challenges of Cyber Security Springer: 172-181.

[12] Schmidhuber, J. (2015). ""Deep Learning in Neural Networks: An Overview"." Neural Networks. 61: 85–117.

[13] Yann LeCun, Y. B. G. H. (2015). "Deep learning." Springer Nature Publishing AG. 521(Nature volume): 436–444

[14] Gharib, A., Sharafaldin, I., Habibi Lashkari, A., and Ghorbani, A. A. (2016). "An evaluation framework for intrusion detection dataset. ." International Conference on Information Science and Security (ICISS): 1–6.

[15] P. Farina, E. C., G. Papaleo, and M. Aiello (2016). ""Are mobile botnets a possible threat? The case of SlowBot Net"." Computers & Security 58: 268-283.

[16] Kumar, P., Tiwari, Arvind (2017). "Ubiquitous Machine Learning and Its Applications." IGI Global.

[17] Mohamed, A. E. (2017). "Comparative Study of Four Supervised Machine Learning Techniques for Classification." International Journal of Applied Science and Technology 7(2).

[18] Lashkari, A. H., A. Seo, G. D. Gil and A. Ghorbani (2017). "CIC-AB: Online ad blocker for browsers. ." International Carnahan Conference on Security Technology (ICCST).

[19] Lettier, D. (2017). "You need to know about the Matthews Correlation Coefficient." github.io.

[20] Srivastava, T. (2018). "Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm(with implementation in Python & R)." analyticsvidhya.com.

[21] Sharafaldin, I., A. H. Lashkari and A. A. Ghorbani (2018). "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization." ICISSP.

[22] unb.ca (2019). "Intrusion Detection Evaluation Dataset." University of New Brunswick.

[23] Guo-Feng Fan, Y.-H. G., Jia-Mei Zheng and Wei-Chiang Hong (2019). "Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting." Intelligent Optimization Modelling in Energy Forecasting.