© Universiti Tun Hussein Onn Malaysia Publisher's Office

JSCDM

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/jscdm

Journal of Soft Computing and Data Mining

A Comparative Study between Machine Learning and Deep Learning Algorithm for Network Intrusion Detection

Alya Syazweena Sha'ari[1], Zubaile Abdullah[1]*

[1]Faculty of Computer Science & Information Technology,
Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2022.03.02.005
Received 12 July 2022; Accepted 20 October 2022; Available online 01 November 2022

Abstract: Network Intrusion Detection is a system that can monitor a network system to avoid malicious activities. One of the methods used for intrusion detection systems is using machine learning. Many pieces of research had proved that machine provides good detection in term of accuracy and performance. However, it can only be used with a smaller dataset other than the features can only be determined using human power. So, deep learning is applied to countermeasure the problem as it can form its own features without using human power other than can be tested with a larger dataset. This study aims to conduct a comparative study for network intrusion detection using machine learning and deep learning algorithm. The dataset that will be tested is CSE-CIC-IDS2018 using Support Vector Machine and Convolutional Neural Network.

Keywords: Network intrusion detection, machine learning, deep learning, support vector machine, convolutional neural network

1. Introduction

Network intrusion detection is a new generation system that can monitor network system to avoid malicious activities [1]. Intrusion detection system is important to improve a security level by identifying all malicious and suspicious events or activities that could be observed in computer or network systems [2]. The success of the system depends on the success of the algorithm used and the performance of its method in detecting any attacks [3]. There are a few methods used for intrusion detection system, one of it is using machine learning [2]. Many research had proved that intrusion detection based on machine learning provides better detection in term of accuracy [4] and reported good results and performance [5]. However, machine learning also has its drawbacks such as the problem can only be solved with smaller dataset and also the features in machine learning are determined by the experts [6].

In order to solve the problem, deep learning method is used in the intrusion detection system. Deep learning is also a machine learning approach and consists of multiple-level layers which is capable of running processes at the same time with high-level features developed from the low-level ones, which makes it forming its own features without using human power [7]. A network intrusion detection method based on deep learning was validated using the KDD CUP'99 dataset and resulted in a significant improvement over the traditional machine learning accuracy [8].

The aim of this project is to proposed a comparative study between machine learning and deep learning algorithm for network intrusion detection. The objectives are to study the intrusion detection using machine learning and deep learning algorithm. Next is to design the network intrusion detection model based on machine and deep learning algorithm. Lastly is to evaluate the accuracy of the system based on both machine learning and deep learning algorithm using confusion matrix.

**Corresponding author: zubaile@uthm.edu.my*

In this project, the dataset that will be used as the sample is the CSE-CIC-IDS2018 dataset. This dataset includes seven different attack scenarios including Heartbleed, Brute-force, DoS, DDoS, web attacks, Botnet, and infiltration [9]. The dataset will be tested with one machine learning algorithm, support vector machine, and one deep learning algorithm, convolutional neural network. Then, the result will be evaluated using confusion matrix. The tool that will be used to test the dataset is WEKA. WEKA is a collection of machine learning algorithms to solve real-world data mining problems. It is written in Java language and can run on almost any platforms. There are many features in the tool, including data mining, pre-processing, classification, regression, clustering, and workflow.

2. Related Work

2.1 Network Intrusion Detection

Intrusion detection system (IDS) comes in many shapes and sizes, where some are simply software applications that run on servers or workstations. Their main purpose is to monitor events on systems or networks and notify the security administrators of any events that is determined to be worthy of alert by the sensors. There are several types of IDS that can be used to aid the security administrators as shown in Figure 1. However, this research will be focusing on the network-based.

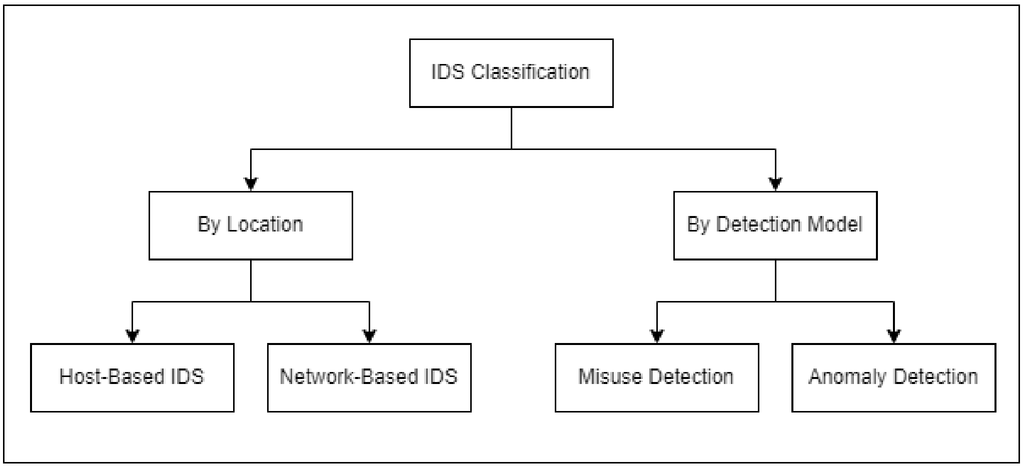


Fig. 1 - IDS classification

Network-based intrusion detection systems (NIDS) are devices that are distributed within networks to monitor the traffic to detect abnormal activities, such as attacks against hosts or servers [10]. At first, NIDS use either statistical measures or computed thresholds on feature sets but are ineffective for present day attacks. It is because they suffer from high rate of false positive and false negative alerts, where the high rate of false positive alerts mean that could unnecessarily alert even when there are no attacks happened, while high rate of false negative alerts mean that NIDS could fail to detect attacks more frequently [11].

Network intrusion detection operated using sets of rule and code signatures created by experts but it is time consuming and can only be created if the attack method that was chosen has been used at least once. In order to solve these issues, machine learning algorithms are being used into NIDS [10]. The existing intrusion detection has been more effective with the development of machine learning recently even though there are still problems with low detection accuracy due to the instability of machine learning algorithm itself [4].

2.2 Machine Learning and Deep Learning

Machine learning is a branch of artificial intelligence that adapted to the new environment which allows programs to find and learn the patterns within data. Machine learning is divided into three sub-domains, which are supervised, unsupervised, and reinforcement learning as shown in Figure 2.

Based on [5], support vector machine (SVM) is a supervised model used for classification, regression, and outlier detection. Various types of SVM classifiers with different capabilities have been introduced. There two main types of SVMs, which are the multi-class classifier and binary classifier which can be classified into another two classifications, called linear and non-linear. On the other hand, the multi-class SVMs can be classified into three categories, which are one against all (OAA), one against one (OAO), and direct acyclic graph-support vector machine (DAGSVM) as shown in Figure 3.

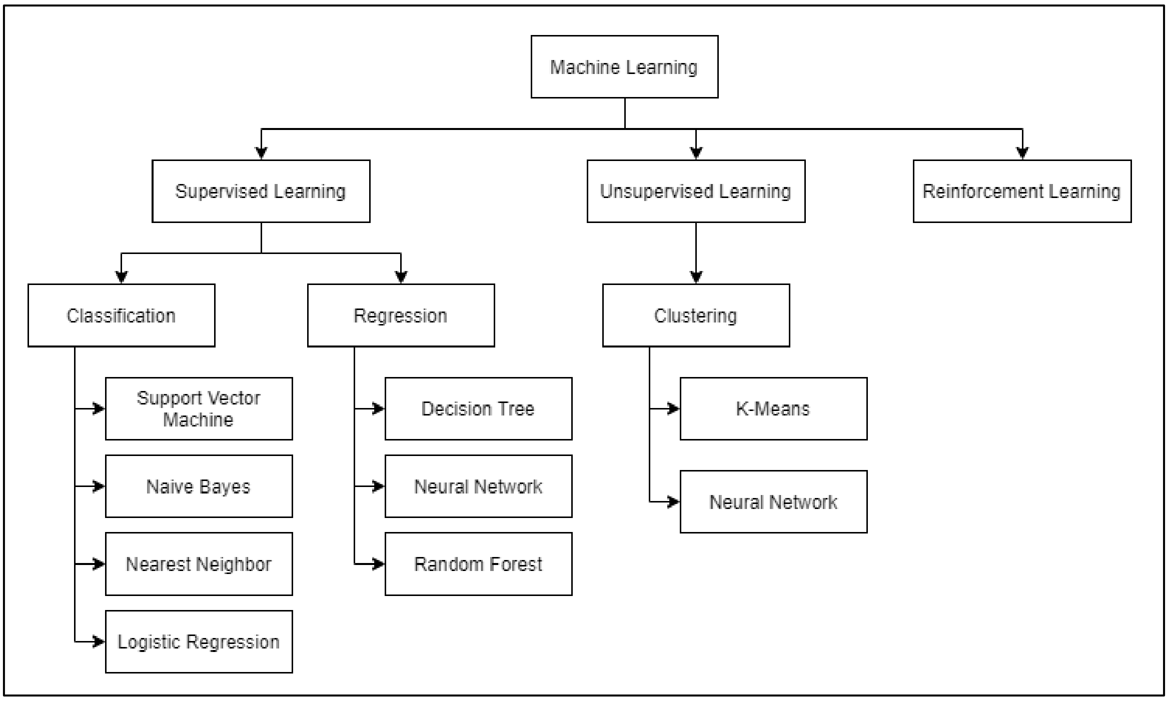


Fig. 2 - Machine learning methods

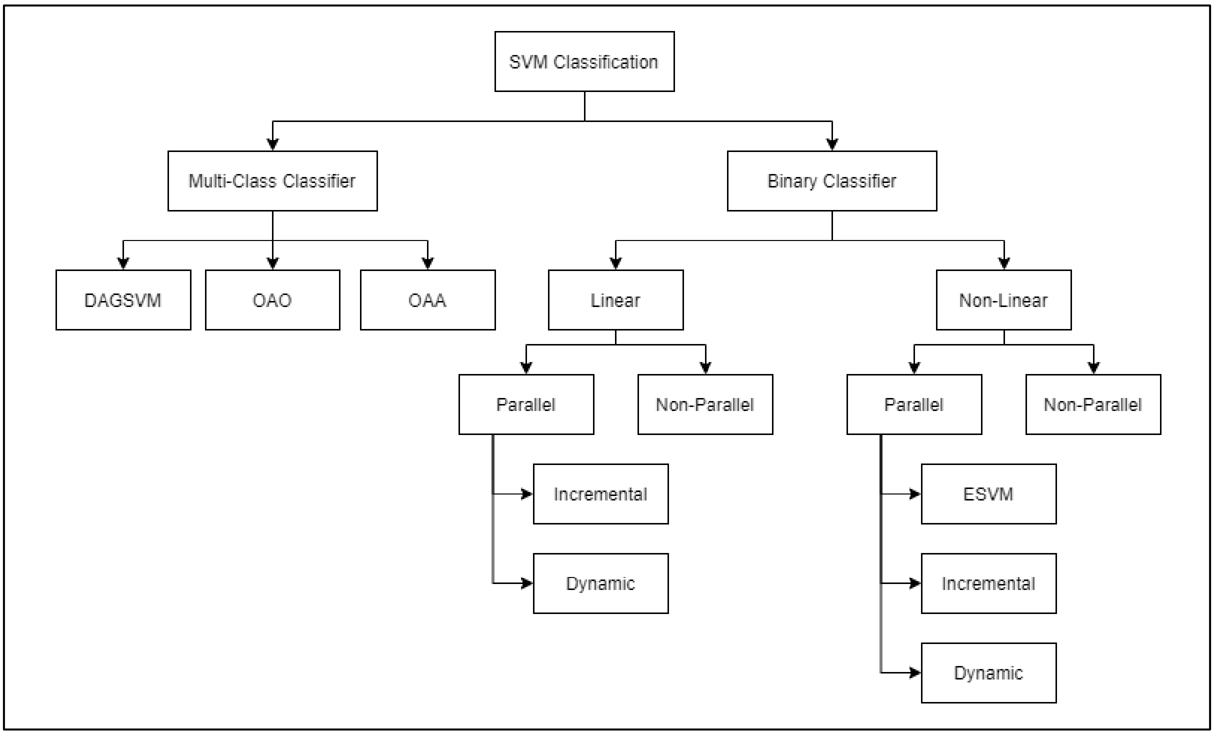


Fig. 3 - SVM classification

According to [7], deep learning is a machine learning approach that consists of multiple-level layers that is capable of running processes simultaneously and high-level features are produced from the low-level ones. Thus, deep learning takes action by forming its own features without using human power. Similar with machine learning, deep learning is also divided into several sub-domains, which are supervised and unsupervised learning as shown in Figure 4.

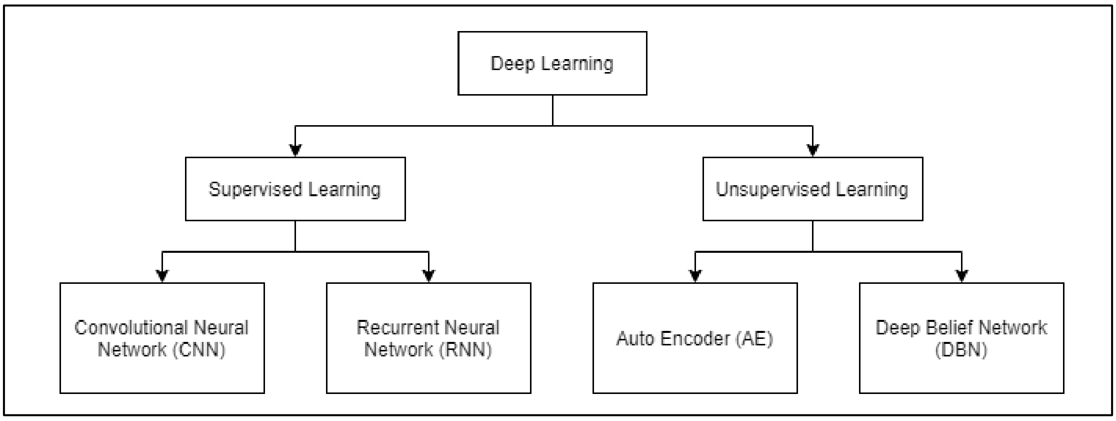

Deep Learning
Supervised Learning
Unsupervised Learning
Convolutional Neural Network (CNN)
Recurrent Neural Network (RNN)
Auto Encoder (AE)
Deep Belief Network (DBN)

Fig. 4 - Deep learning methods

Based on [7], Convolutional Neural Network (CNN) is a special type of artificial neural network (ANN) and also the most frequently used deep learning method. CNN are made up of neurons that capable to learn biases and weights and it also process data that comes in multiple arrays and removes the need for manual feature extraction [5]. There are several past research that used CNN method for intrusion detection, for example [12] used CNN for packet-data anomaly detection in PMU-based state estimator. In the paper, they used a convolutional neural network-based data filter in order to extract features from phasor measurement units. The authors also claimed that the convolutional neural network-based filter has a better performance than other machine learning techniques such as RNN, SVM, bagged and boosted.

2.3 Existing Works

Various researches have been done in network intrusion detection. Table 1 shows a few of the comparison of method in this domain of knowledge.

Table 1 - Comparison of methods in network intrusion detection

Research	Title of the research	Method used	Description
Selected	Research on Network Intrusion Detection Technology Based on Machine Learning [4].	Improved Random Forest-Support Vector Machine	The research focused on Network Intrusion Detection based on an improved algorithm. The proposed method is used on KDDCUP99 dataset.
	Machine Learning and Deep Learning Methods for Intrusion Detection Systems: Recent Developments and Challenges [5].	K-Nearest Neighbour, Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, Recurrent Neural Network, Convolutional Neural Network.	The methods were used on benchmark datasets, such as KDD99, UNSW-NB15, CSE-CIC-IDS2018. The purpose of the research was to proposed a systematic review of machine learning and deep learning in intrusion detection.
	Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study [9].	Recurrent Neural Network, Deep Neural Network, Restricted Boltzmann Machines, Deep Belief Networks, Convolutional Neural Network, Deep Boltzmann Machines, Deep Autoencoder.	The methods were used on two datasets, which are CSE-CIC-IDS2018 and Bot-IoT. The CNN models gave the highest accuracy in both datasets in general.
Proposed	A Comparative Study between Machine Learning and Deep Learning Algorithm for Network Intrusion Detection.	Support Vector Machine, Convolutional Neural Network	The proposed research study focused on the network intrusion detection based on the proposed algorithm.

3. Methodology

Figure 5 shows the network intrusion detection framework that consists of seven phases, which are raw data, pre-processing of data, features extraction and selection, 10-fold cross validation, Convolutional Neural Network and Support Vector Machine algorithm, and result analysis for both algorithms.

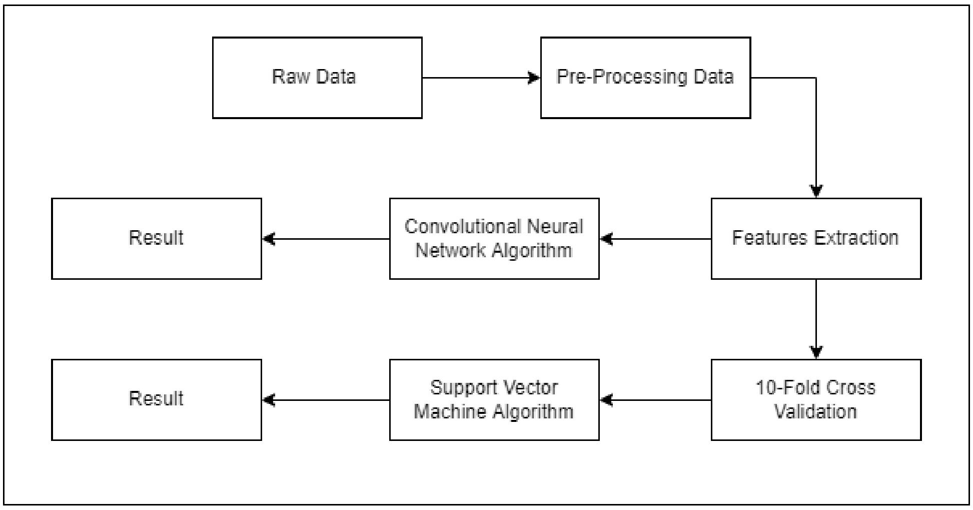


Fig. 5 - Network intrusion detection framework

3.1 Dataset Description

In this research, CSE-CIC-IDS2018 [13] dataset will be used. The dataset consists of multiple attacks on network intrusion detection including Brute Force, DoS, DDoS, Infiltration, and Botnet [13]. This dataset also has been used by [14], [15].

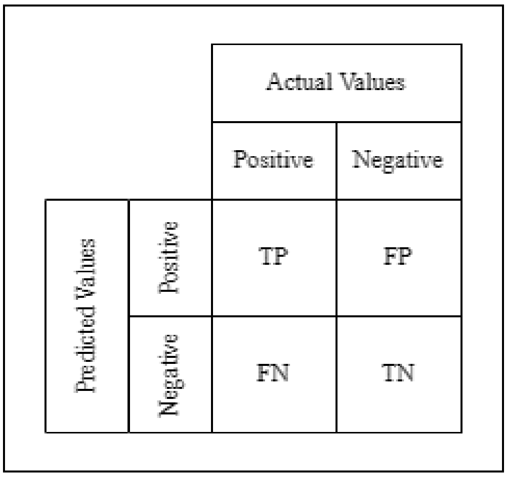
3.2 Classification Algorithms

The following section will be discussing on the two classification algorithms, Support Vector Machine and Convolutional Neural Network. Support Vector Machine (SVM) is a supervised machine learning algorithm that is mostly used for classification problems. SVM is effective in high dimensional cases and uses a subset of training points in the decision function, support vector, which makes its memory efficient. SVM is also an algorithm that takes data as inputs and gives a line that separates data of two classes as outputs, which is called as the hyperplane.

The second algorithm that will be used in Convolutional Neural Network (CNN) which is a deep learning algorithm that is capable of understanding difficult structures and have shown great results in tasks related to image segmentation, object detection, and computer vision applications. CNNs also have been used for both feature extraction and classification in intrusion detection.

3.3 Confusion Matrix

Confusion matrix is a table that is used to evaluate the performance of a classification model on a set of data like shown in Figure 6. In this research, confusion matrix is used to measure the performance of network intrusion detection model based on Support Vector Machine and Convolutional Neural Network. This matrix considers four different values, which are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These values will then be used to calculate the Accuracy, Precision, and Recall.



		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

Fig. 6 - Example of a confusion matrix

True Positive (TP) is where the predicted positive value matches the actual values, meanwhile True Negative (TN) is where the predicted negative value matches the actual values. False Positive (FP), also known as Type 1 error, is when the value is wrongly predicted where the actual value is negative but the model predicted a positive value. False Negative (FN), also known as Type 2 error, is also an error in predicting the value where the actual one is positive but the model predicted a negative value.

Accuracy of the intrusion detection model is the measurement used to determine the trueness with systematic error and precision with random errors. Eq. 1 defined the probability of the total of TP and TN to the total number of the outcome. The accuracy percentage shown how many networks intrusion is predicted correctly by the algorithm.

$$Accuracy = \ (TP + TN)/(TP + FP + TN + FN) \tag{1}$$

Precision determines on how many of the correctly predicted cases actually turns out to be positive. Eq. 2 defined the probability of TP to the total of TP and FP. The precision percentage shown how many of the network intrusion that is predicted correctly turned out to be true cases.

$$Precision = \ TP/(TP + FP) \tag{2}$$

Recall also determines the actual positive cases that are able to correctly predict with the model. Eq. 3 defined the probability of TP to the total of TP and FN. The recall percentage shows how many of the positive cases that are predict successfully by the model.

$$Recall = \ TP/(TP + FN) \tag{3}$$

3.4 Hardware and Software Requirements

To ensure that the experiment can be run smoothly, hardware is the most important tool to be used to achieve the objectives and expected result. The hardware used to run the experiment is as show in Table 2.

Table 2 - Hardware requirements

Hardware	Description	
DELL Inspiron 5490	Windows Edition	System
	Windows 11 Pro © 2021 Microsoft Corporation. All right reserved	Processor: Intel(R) Core^{TM} i5-10210U CPU @ 1.60GHz 2.11 GHz Installed Memory (RAM): 8.00 GB System Type: 64-bit operating system, x64-based processor

For software requirement, Waikato Environment for Knowledge Analysis (WEKA) will be used. WEKA is a free software licensed under the GNU General Public License that was developed at the University of Waikato, New Zealand. WEKA is written in Java programming language and contains a collection of machine learning algorithms for data mining tasks. This software also contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

4. Results and Discussion

4.1 Result

This section discusses the results that obtained from WEKA by comparing the performance of the selected classification techniques on the dataset. The performance of classifiers is based on two selected classifiers, which are Support Vector Machine and Convolutional Neural Network, and all models are tested on one dataset with two sets of features. One of the sets consist of the top 25 features of dataset and the other one consists of the top 50 features of the dataset. Before running the dataset on WEKA, the dataset is cleaned by removing all Nan and Infinity values and long decimal digits were reduced to one decimal digit to save time and memory. Then, the dataset is loaded into WEKA to do feature selection. Table 3 shows the top 50 features that were ranked by WEKA using Ranker Search Method.

Table 3 - Top 50 features selected using ranker search method

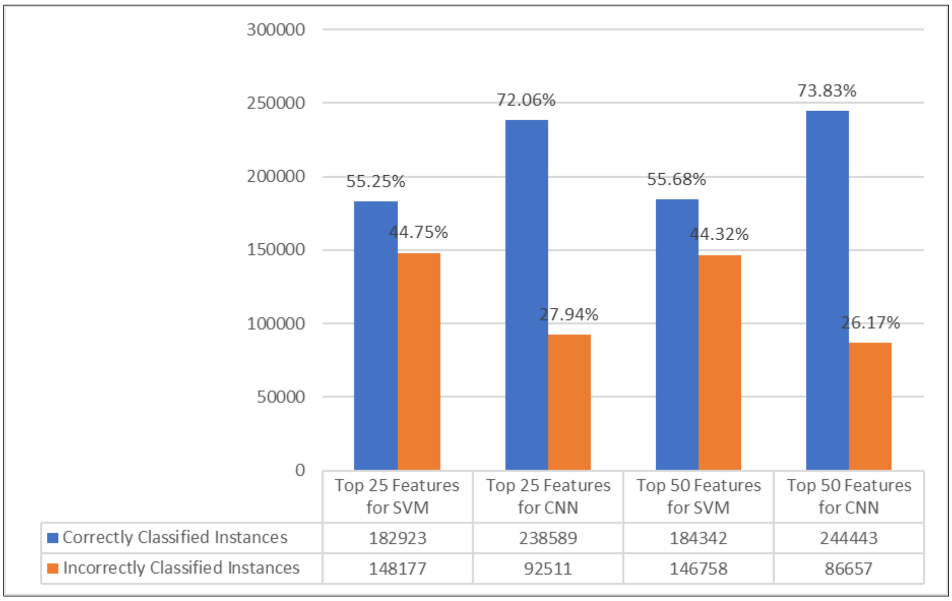
Label Index	Features	Label Index	Features
14	Flow Byts/s	66	Idle Std
1	Flow Duration	57	Init Fwd Win Byts
16	Flow IAT Mean	64	Active Min
15	Flow Pkts/s	32	Fwd Header Len
34	Fwd Pkts/s	62	Active Std
35	Bwd Pkts/s	60	Fwd Seg Size Min
21	Fwd IAT Mean	68	Idle Min
17	Flow IAT Std	52	Bwd Seg Size Avg
18	Flow IAT Max	12	Bwd Pkt Len Mean
20	Fwd IAT Tot	51	Fwd Seg Size Avg
23	Fwd IAT Max	8	Fwd Pkt Len Mean
26	Bwd IAT Mean	9	Fwd Pkt Len Std
22	Fwd IAT Std	13	Bwd Pkt Len Std
27	Bwd IAT Std	65	Idle Mean
25	Bwd IAT Tot	58	Init Bwd Win Byts
28	Bwd IAT Max	56	Subflow Bwd Byts
29	Bwd IAT Min	5	TotLen Bwd Pkts
24	Fwd IAT Min	4	TotLen Fwd Pkts
61	Active Mean	54	Subflow Fwd Byts
63	Active Max	33	Bwd Header Len
19	Flow IAT Min	67	Idle Max
40	Pkt Len Var	6	Fwd Pkt Len Max
39	Pkt Len Std	37	Pkt Len Max
50	Pkt Size Avg	10	Bwd Pkt Len Max
38	Pkt Len Mean	53	Subflow Fwd Pkts

4.2 Discussions

The result of the overall performance for all models on the dataset have been generated using WEKA. The performance results will show which models give the best accuracy. The accuracy for each of the experiments is shown as in Table 4. Figure 7 shows the graphic chart for the accuracy of the experiments.

Table 4 - Results of the experiments

	Accuracy	Precision	Recall	Time Taken
Top 25 Features using SVM	0.5525	0.594	0.552	240 seconds
Top 50 Features using SVM	0.5568	0.635	0.557	149 seconds
Top 25 Features using CNN	0.7206	0.686	0.721	2401 seconds
Top 50 Features using CNN	0.7383	0.724	0.738	1555 seconds



	Top 25 Features for SVM	Top 25 Features for CNN	Top 50 Features for SVM	Top 50 Features for CNN
■ Correctly Classified Instances	182923	238589	184342	244443
■ Incorrectly Classified Instances	148177	92511	146758	86657

Fig. 7 - Accuracy of each experiment

5. Conclusion

In conclusion, the is done using the Support Vector Machine and Convolutional Neural Network on the CSE-CIC-IDS2018 dataset. The study is run using the WEKA tool as it contains a collection of machine learning algorithms for data mining tasks. This paper compares and analyses the performance of two data mining techniques on CSE-CIC-IDS2018 dataset with top 50 and top 25 features. The result shows that CNN with top 50 features had performed the best as compared to others. During this research, there are a few problems that had been identified such as: (a) lack of memory on the computer to run the experiments. This had caused the time taken to be longer to build the models for each experiment. (b) finding the most accurate dataset is a barrier to have a great result because the dataset also plays a big role in the research. However, there are several suggestions for future improvements that are related to this research in getting the best result which are: (a) increase the RAM memory on computer to improve the time taken to build the model. (b) the use of multiple datasets may help in getting a more accurate result.

Acknowledgement

The authors would also like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its support and encouragement throughout the process of conducting this study.

References

[1] Borkar, A., Donode, A. & Kumari, A. (2017). A survey on Intrusion Detection System (IDS) and Internal Intrusion Detection and protection system (IIDPS), 2017 International Conference on Inventive Computing and Informatics (ICICI), pp. 949–953, doi: 10.1109/ICICI.2017.8365277.

[2] Jahwar, A. F., & Ameen, S. Y. (2021). A Review on Cybersecurity based on Machine Learning and Deep Learning Algorithms. *Journal of Soft Computing and Data Mining*, *2*(2), 14-25.

[3] Innovations in Intelligent Systems and Applications (INISTA), pp. 1–4, doi: 10.1109/INISTA.2018.8466271.

[4] Kunang, Y.N., Nurmaini, S., Stiawan, D. & Suprapto, B.Y. (2021). Attack classification of an intrusion detection system using deep learning and hyperparameter optimization. J. Inf. Secur. Appl. 2021, vol. 58, pp. 102804. doi: https://doi.org/10.1016/j.jisa.2021.102804.

[5] Wu, F., Li, T., Wu, Z., Wu, S. & Xiao, C. R (2021). Research on Network Intrusion Detection Technology Based on Machine Learning, Int. J. Wirel. Inf. Networks, vol. 28, no. 3, pp. 262–275, 2021, doi: 10.1007/s10776-021-00520-z.

[6] Kocher, G. & Kumar, G. (2021). Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges. Soft Comput 25, 9731–9763. https://doi.org/10.1007/s00500-021-05893-0

[7] Karatas, G., Demir, O. & Koray Sahingoz, O. (2018). Deep Learning in Intrusion Detection Systems, 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), pp. 113–116, doi: 10.1109/IBIGDELFT.2018.8625278.

[8] Osken, S., Yildirim, E. N., Karatas, G. & Cuhaci, L. (2019). Intrusion Detection Systems with Deep Learning: A Systematic Mapping Study, 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), pp. 1–4, doi: 10.1109/EBBT.2019.8742081.

[9] Peng, W., Kong, X., Peng, G., Li, X. & Wang, Z. (2019). Network Intrusion Detection Based on Deep Learning, 2019 International Conference on Communications, Information System and Computer Engineering (CISCE), pp. 431–435, doi: 10.1109/CISCE.2019.00102.

[10] Ferrag, M. A., Maglaras, L., Moschoyiannis, S. & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study, J. Inf. Secur. Appl., vol. 50, p. 102419, doi: https://doi.org/10.1016/j.jisa.2019.102419.

[11] Alhajjar, E., Maxwell, P. & Bastian, N. (2021). Adversarial machine learning in Network Intrusion Detection Systems,” Expert Syst. Appl., vol. 186, p. 115782, 2021, doi: https://doi.org/10.1016/j.eswa.2021.115782.

[12] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A. & Venkatraman, S. (2019). Deep Learning Approach for Intelligent Intrusion Detection System, IEEE Access, vol. 7, pp. 41525–41550, doi: 10.1109/ACCESS.2019.2895334.

[13] Basumallik, S., Ma, R., & Eftekharnejad, S. (2019). Packet-data anomaly detection in PMU-based state estimator using convolutional neural network, Int. J. Electr. Power Energy Syst., vol. 107, pp. 690–702, doi: https://doi.org/10.1016/j.ijepes.2018.11.013.

[14] A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018). https://registry.opendata.aws/cse-cic-ids2018.

[15] Kilincer, I. F., Ertam, F. & Sengur, A. (2021). Machine learning methods for cyber security intrusion detection: Datasets and comparative study, Comput. Networks, vol. 188, p. 107840, doi: https://doi.org/10.1016/j.comnet.2021.107840.

[16] D’hooge, L., Wauters, T., Volckaert, B. & De Turck, F. (2020). Inter-dataset generalization strength of supervised machine learning methods for intrusion detection, J. Inf. Secur. Appl., vol. 54, p. 102564, doi: https://doi.org/10.1016/j.jisa.2020.102564.