UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Comparing FTP and SSH Password Brute Force Attack Detection using k-Nearest Neighbour (k-NN) and Decision Tree in Cloud Computing

Muhammad Fakrullah Kamarudin Shah[1], Marina Md-Arshad[2], Adlina Abdul Samad[3] & Fuad A. Ghaleb[4]
Faculty of Computing
Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Malaysia
Email: hyperpakol@gmail.com[1], marinama@utm.my[2], adlina6@graduate.utm.my[3], abdulgaleel@utm.my[4]

Submitted: 22/9/2022. Revised edition: 16/5/2023. Accepted: 16/5/2023. Published online: 30/5/2023
DOI: https://doi.org/10.11113/ijic.v13n1.386

***Abstract*—Cloud computing represents a new epoch in computing. From huge enterprises to individual use, cloud computing always provides an answer. Therefore, cloud computing must be readily accessible and scalable, and customers must pay only for the resources they consume rather than for the entire infrastructure. With such conveniences, come with their own threat especially brute force attacks since the resources are available publicly online for the whole world to see. In a brute force attack, the attacker attempts every possible combination of username and password to obtain access to the system. This study aims to examine the performance of the k-Nearest Neighbours (k-NN) and Decision Tree algorithms by contrasting their precision, recall, and F1 score. This research makes use of the CICIDS2017 dataset, which is a labelled dataset produced by the Canada Institute for Cybersecurity. A signature for the brute force attack is utilised with an Intrusion Detection System (IDS) to detect the attack. This strategy, however, is ineffective when a network is being attacked by a novel or unknown attack or signature. At the conclusion of the study, the performance of both algorithms is evaluated by comparing their precision, recall, and f1 score. The results show that Decision Tree performs slightly better than k-NN at classifying FTP and SSH attacks.**

***Keywords*—Supervised Machine Learning, Cloud Computing, SSH, FTP classification, Feature Selection, Decision Tree, K-Nearest Neighbours**

I. Introduction

Cloud computing is, by definition, a service that can be remotely accessed from any Internet-connected location in the world. Cloud computing can be categorised into three categories, which are Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (IaaS) and Software-as-a-Service (SaaS) [1]. While cloud computing is convenient, it's always exposed to external attacks such as brute force attacks. In its simplest form, a brute force attack is an attempt to fraudulently gain access to a system or device by using several credentials from a wordlist. Normally, the process will be taking a long time, but with the advent of the Graphic Processing Unit (GPU), the time taken will be reduced greatly [4].

To detect two types of common brute force attacks, SSH and FTP, machine learning is used to classify the attack from the dataset. There are three distinct classifications for machine learning which is supervised, unsupervised and reinforcement learning. This research use k-Nearest Neighbours (k-NN) and Decision Tree algorithms because these two algorithms are compatibility towards the dataset in order to have a better understanding in the visualization. By utilizing the visualization, the user can have more focus on mitigating the attack instead of focus on how to diagnose the attack.

This research seeks to test the accuracy, precision and F1 score of k-NN and Decision Tree to classify the brute force assault from a dataset. Precision can be defined as how many positives are from all the projected positives. Recall is a statistic used to identify the real positive from the true positive and false negative. F1 score is used to examine the balance between precision and recall [8].

k-NN is chosen for this research because of its potential for pattern identification and classification, while Decision Tree is picked because of its ability to do binary classification and the process can be viewed [5]. The dataset that will be used in this research is CICIDS17 from the Canadian Institute for Cybersecurity published in July 2017. This dataset comprises of two types of brute force assault, FTP and SSH and benign data. Before classifying the dataset, the dataset will be cleaned using RapidMiner Studio to eliminate superfluous attributes and missing values that cannot be used and are unneeded in the classification process.

The structured of the paper was divided into seven sections. The first and second section is the introduction and literature review of this research. Besides, third and forth section is the dataset and design and implementation of this research. After the design and implementation research was made, the fifth section is result and analysis regarding this research. Suggestions for the improvement was made in sixth section. Finally, the conclusion was concluded in the last section.

II. Literature Review

In this section, there will be the elaboration on the current research and threats in cloud computing devices which also included the types of the cloud computing attacks. Besides, the information with the current supervised machine learning algorithm which involve in the attack classification using machine learning algorithms and the nature of the attacks targeting cloud computing devices were also detailed in this section. Lastly, the machine learning techniques in classifying the attacks from network perimeter also explained in this section.

A. Cloud Computing background

According to Oxford Languages, Cloud Computing is a scenario in which data is stored, managed, and processed using remote servers hosted on the internet rather than local servers or personal computers [17]. Cloud computing is also characterised as a model for offering accurate and efficient internet access to a shared pool of programmable grids, storage, servers, software, and amenities that can be rapidly liberated with minimal provider oversight. As shown in Fig. 1, cloud computing consists of three primary service models: PaaS, IaaS, and SaaS [19].

SaaS is defined as any cloud service that allows consumers to access Internet applications from anywhere, regardless of hardware or location. PaaS provides a platform and environment that is typically included in a solution stack, such as an operating system, databases, and middleware. This platform and environment facilitate the creation and development process without the need to construct or maintain the environment via the Internet. IaaS provides highly scalable and budget-friendly computing resources, such as virtual servers, bandwidth, and load balancing, to users [20].

The benefits of SaaS include that the program is always updated, the data is always accessible unless it is removed, and most crucially, the software is easy to access. Google Suite is an example of SaaS because it offers users critical productivity tools such as a word processor, spreadsheet, and presentation. Cloud technologies have transformed the market for computer services by enabling on-demand access to services and resources.

Manufacturers utilised the cloud application service to develop and reinvent their manufacturing process because of the cloud's adaptability and boundless resources [29]. The benefit of this service is that the user is not responsible for building and maintaining the infrastructure. This means that the consumer does not require in-depth knowledge of infrastructures because the cloud provider will manage and construct them [10].

With the rise of cloud computing, security has become of paramount importance. There are numerous risks associated with cloud computing. These risks include data loss and leakage caused by weak keys, an unreliable data centre, insufficient encryption, and insecure credentials. Regardless of the provider, this vulnerability can impact cloud computing. The most prevalent forms of assault are Denial-of-Service (DoS) and brute force. This typically corresponded to the attackers guessing the password, resulting in the flooding of specific services or ports, such as SSH/22, and exhausting the server's resources, hence disabling the service for the genuine user [16]. To alleviate these concerns, consumers should be made aware of the danger and threats posed by the attack.

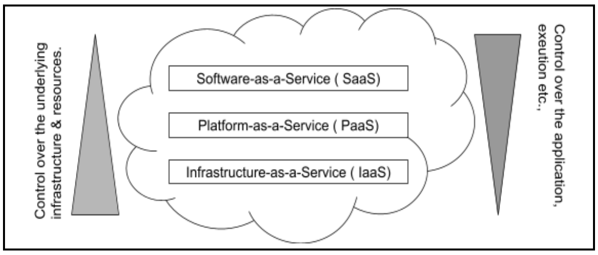

Control over the underlying infrastructure & resources.
Software-as-a-Service (SaaS)
Platform-as-a-Service (PaaS)
Infrastructure-as-a-Service (IaaS)
Control over the application, exeution etc.,

Fig. 1. Cloud computing services [19]

B. Vulnerabilities in cloud computing

Numerous security flaws exist in the cloud computing ecosystem. This paper will briefly examine the most prevalent vulnerabilities, including data breaches, malevolent insiders, denial of service attacks, and insecure systems and APIs [11]. The most prevalent are data breaches. A data breach or data leak is the illegal viewing, accessing, or retrieving of data by an entity. It is a type of security breach in which sensitive information is captured and/or released on an unsecured or unlawful website. This is now one of the greatest threats to consumers. This vulnerability comprises data corruption, which occurs when an attacker deletes or modifies data intending to cause harm to the owner. These vulnerabilities can be avoided by securing the cloud itself with security algorithms [2].

The second weakness is the presence of malicious insiders. This vulnerability is the most dangerous because it cannot be identified by IDS, which is likely configured to protect against external threats. This vulnerability might manifest in numerous ways. Whether it's a former employee, sysadmin, contractor, or even a business partner, you should always treat everyone with respect [11]. For instance, in December 2019, one of the most renowned IT businesses, Microsoft, suffered a database leak due to employee negligence, exposing up to 250 million records including support cases and details, emails, IP addresses, geolocation, and notes from Microsoft support agents. This has led Microsoft to pay a fine of $750 for each victim of the breach [7].

Denial of Service (DoS) is arguably the most infamous form of cyberattack. A Denial of Service (DoS) is initiated by repeatedly overwhelming a service or the cloud. This assault may involve several computers. The greater the number, the higher the attack's success rate. Typically, the computer used to launch an attack is a "zombie" or a machine that was previously pawned or controlled by the attacker. Denial of Service (DoS) attacks may target a specific service or port, the cloud's bandwidth, or the cloud's processing resources. Denial of Service (DoS) is capable of evading detection. Even security solutions cannot identify the attack because its packet cannot be separated from typical user packets. A DoS attack often has two goals. First is overwhelming the target's resources or network connection with numbers. The second step involves sending malicious packets to exploit a security flaw [3]. Fig. 2 depicts the attacker's Denial of Service (DoS) attempt.

Another form of attack is brute force. System vulnerabilities and Application Programming Interface (API). APIs are accessible to all users. By exploiting API vulnerabilities, attackers can gain extensive access to cloud resources, which are typically databases, by abusing the create, read, update, and delete (CRUD) process that is typically allowed to the API. Moreover, Platform-as-a-Service (PaaS) operating systems may contain vulnerabilities that can be abused to provide attackers complete access to the system and cause additional harm to the cloud and services provided.

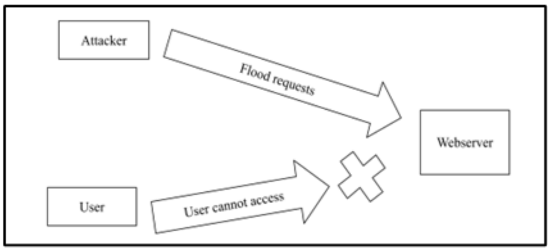


Fig. 2. DoS Attack

C. Related Studies

There are numerous machine learning algorithms available for classifying brute-force attacks. To ensure correct classification, the dataset utilised must be carefully selected to ensure that only relevant information is accessible for the classification algorithm to use in order to improve the algorithm's accuracy.

A study has been undertaken in which Random Forest is utilised to characterise the attack as either an SSH brute force attack or a UDP and HTTP flood [13].

Long Short-Term Memories is applied to the CICIDS2017 dataset to detect brute force attacks with high precision, however, the model is susceptible to overfitting and may not perform well when applied to new data [24]. A study determined that brute force attacks contain fewer packets and bytes than successful SSH logins. Based on aggregated NetFlow data, the study utilised C4.5D Decision Tree to identify SSH brute force attacks. In general, the decision tree is susceptible to overfitting [9].

The dataset was analysed using principal component analysis prior to training. Before training the Naive Bayes classifier utilising WeKa tools, this is performed to reduce the dataset's dimensionality and ensure that the model concentrates on features that have a significant impact on data classification. The constraint is that Nave Bayes implies that the features are independent, hence the dataset must be composed of independent variables [22]. Collecting packet data using the "tcpdump" tool on a honeypot server to record SSH assault traffic is another classification experiment. The packets are then categorised using machine learning methods and the WeKa tool. One of the techniques is the Decision Tree, which is susceptible to overfitting, susceptible to noise in the dataset, and unsuitable for large datasets [23]. An experiment has been conducted in which brute force attacks are identified using a model based on processing the network log to obtain information by reading certain alerts in the network log, such as "SSH user failed to login from IP," to determine that the IP is executing a brute force attack [12]. Another experiment classifies the brute-force attack using deep learning. Before training the classifier, the data will be converted into a two-dimensional image matrix and then taught using a Convolutional Neural Network (CNN) to categorise the attack, which has a high level of accuracy in picture recognition but requires a large amount of training data [18].

Using the neural network's method Multi-Layer Perception (MLP), a study has classified the attack data and benign data from the CICIDS2017 dataset. It is concluded that the accuracy of MLP increases linearly with the number of packets utilised in the training phase, but that the training method is resource-intensive and time-consuming [15]. A study recommends using Diffie-Hellman key exchange for SSH before creating a connection, with a three-time restriction before the key pair is updated. If an assault is occurring, the Diffie-Hellman key exchange will be more than usual, which can be used to differentiate between a normal login and an SSH brute force attack. Support Vector Machine (SVM) is one of the techniques utilised in this study; it has a high memory footprint and requires a considerable amount of time to train the model [28]. A study was undertaken to detect a brute-force attack that attempted to evade detection by halting the attack for a predetermined amount of time. This threat detection utilises flows received from the backbone network that provides information about several victims of a single IP address, which will be utilised as a sign of a brute-force attack. This study employed an Ada-boosted Decision Tree, which required less parameter tuning and enhanced accuracy, but required high-quality learning data and was extremely sensitive to noise and outliers [25].

Among all supervised learning algorithms, Decision Tree and k-Nearest Neighbours (k-NN) have been chosen for this study. Decision Tree is selected due to its unique properties for classifying data, its process's ability to be displayed, and its use of conditions to classify data, which is suited for the dataset utilised in this study, which contains a large number of numerical values. This research utilises the k-Nearest Neighbours (k-NN) technique because it exploits the similarity present in the dataset to appropriately classify the data to its category. K-Nearest Neighbours (k-NN) output is also simpler to comprehend, and the execution time of the algorithm is less than that of other algorithms.

D. k-Nearest Neighbours

k-Nearest Neighbours is a supervised learning algorithm that can be used for both regression and classification.

$$(x^{[i]}, y^{[i]}) \in D \; (|D| = n) \tag{1}$$

This algorithm will store all available data and will use the similarity in the data to predict and will classify the data as the same categories [6]. k-NN is a lazy learning algorithm that does not use training data to make any generalization which means that the training phase is fast [27]. Besides, k-NN also suitable for multiclass classification [29] ,which means that it is suitable to be used for classifying the FTP and SSH brute force attack

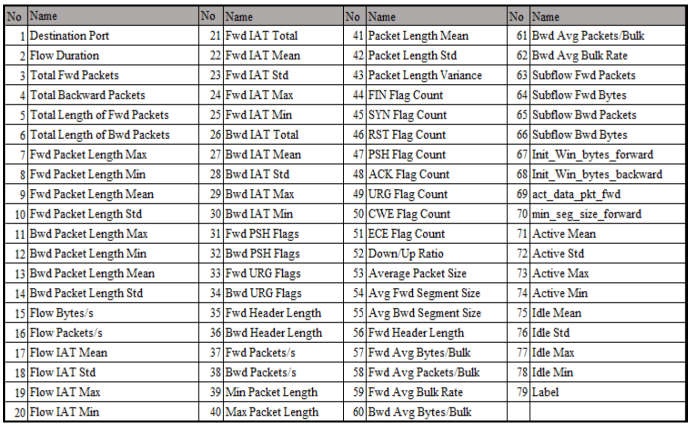
E. Decision Tree

Decision Tree is one of the algorithms that is categorised under supervised learning which is a tree-based technique. To understand the decision tree better, it consists of decision, which is yes, and no. A Decision Tree can also be visualised to increase the understanding of how the algorithm works. Decision Tree also can be used to classify multiclass dataset [30].

III. Dataset

CICIDS2017 is a dataset made available with the link https://www.unb.ca/cic/datasets/ids-2017.html to the public by the Canadian Institute for Cybersecurity [26]. This dataset seeks to emulate a real-world scenario in which both an attacker and a regular user are utilising the services. This dataset employs a complete network topology and includes all required networking equipment, including routers, switches, hubs, and modems. The network also has a variety of operating systems, including Microsoft Windows, Ubuntu Linux, and Apple Mac OS X. This dataset utilises a mirror port to capture all network traffic and store it on a storage server to obtain a complete capture of the packet. This dataset is publicly accessible on the website of the Canadian Institute for Cybersecurity. The dataset contains a variety of attack vectors based on the day it was collected. In this study, the Tuesday dataset is selected. Tuesday's datasets include three types of network packets: FTP-Patator, SSH-Patator, with 79 columns and 445910 rows. Table I shows the list of features from the dataset.

TABLE I. FEATURE OF CICIDS2017



No	Name	No	Name	No	Name	No	Name
1	Destination Port	21	Fwd IAT Total	41	Packet Length Mean	61	Bwd Avg Packets/Bulk
2	Flow Duration	22	Fwd IAT Mean	42	Packet Length Std	62	Bwd Avg Bulk Rate
3	Total Fwd Packets	23	Fwd IAT Std	43	Packet Length Variance	63	Subflow Fwd Packets
4	Total Backward Packets	24	Fwd IAT Max	44	FIN Flag Count	64	Subflow Fwd Bytes
5	Total Length of Fwd Packets	25	Fwd IAT Min	45	SYN Flag Count	65	Subflow Bwd Packets
6	Total Length of Bwd Packets	26	Bwd IAT Total	46	RST Flag Count	66	Subflow Bwd Bytes
7	Fwd Packet Length Max	27	Bwd IAT Mean	47	PSH Flag Count	67	Init_Win_bytes_forward
8	Fwd Packet Length Min	28	Bwd IAT Std	48	ACK Flag Count	68	Init_Win_bytes_backward
9	Fwd Packet Length Mean	29	Bwd IAT Max	49	URG Flag Count	69	act_data_pkt_fwd
10	Fwd Packet Length Std	30	Bwd IAT Min	50	CWE Flag Count	70	min_seg_size_forward
11	Bwd Packet Length Max	31	Fwd PSH Flags	51	ECE Flag Count	71	Active Mean
12	Bwd Packet Length Min	32	Bwd PSH Flags	52	Down/Up Ratio	72	Active Std
13	Bwd Packet Length Mean	33	Fwd URG Flags	53	Average Packet Size	73	Active Max
14	Bwd Packet Length Std	34	Bwd URG Flags	54	Avg Fwd Segment Size	74	Active Min
15	Flow Bytes/s	35	Fwd Header Length	55	Avg Bwd Segment Size	75	Idle Mean
16	Flow Packets/s	36	Bwd Header Length	56	Fwd Header Length	76	Idle Std
17	Flow IAT Mean	37	Fwd Packets/s	57	Fwd Avg Bytes/Bulk	77	Idle Max
18	Flow IAT Std	38	Bwd Packets/s	58	Fwd Avg Packets/Bulk	78	Idle Min
19	Flow IAT Max	39	Min Packet Length	59	Fwd Avg Bulk Rate	79	Label
20	Flow IAT Min	40	Max Packet Length	60	Bwd Avg Bytes/Bulk		

IV. Design and Implementation

The setup for the brute force attack classification experiment was discussed in this section. The selected dataset was trained using the k-Nearest Neighbours (kNN) and Decision Tree methods. Feature selection is utilised to locate data relevant to this research. The data is divided into 70% training and 30% test sets, and the algorithms are trained using the training sets. This process will be repeated until a satisfactory result is achieved. The evaluation of the performance of the two classifiers, performance metrics was used to measure the performance by using the confusion matrix, precision, recall and f1-score.

A. Data Cleaning

To prepare the dataset for model training, RapidMiner Studio is used to clean the dataset. First, the superfluous attribute 'Fwd. Header Length' is deleted prior to importing the dataset into RapidMiner. Next, in order to further clean the dataset, these trials eliminate any features with 0 min and 0 max based on the statistics provided by RapidMiner Studio that will not contribute to the training process. The filter example function in RapidMiner Studio is used to remove rows containing 'Infinity' and 'NaN' or Not a Number, which will disrupt the data during the fitting and training processes. This preliminary phase compresses the dataset to X columns and Y attributes. Table II depicts a list of the zero data characteristics, while Fig. 3 depicts the RapidMiner Studio procedures.

TABLE II. LIST OF ZERO ATTRIBUTES

No	Name
1	Fwd Avg Bytes/Bulk
2	Fwd Avg Packets/ Bulk
3	Fwd Avg Bulk Rate
4	Bwd Avg Bytes/ Bulk
5	Bwd Avg Packets/ Bulk
6	Bwd Avg Bulk Rate

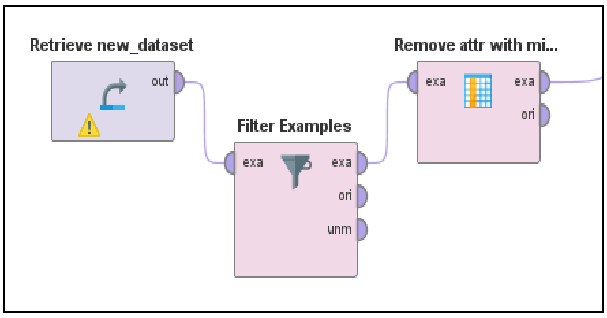


Fig. 3. RapidMiner procedures

B. Sequential Forward Selection

Forward selection is an iterative technique. Sequential Forward Selection (SFS) is one of the wrapper feature choices that began with an empty collection of features. After each iteration, a new feature is added and its effect on performance is evaluated. The procedure is repeated until the addition of a new variable or feature no longer enhances the performance of the model [14]. Python is utilised with 'sklearn' for training the model, 'pandas' for importing the data from a.csv file and separating it into dependent and independent variables, and 'mlxtend' for the feature selection procedure. This method employs the Random Forest algorithm in conjunction with the sequential forward selector. Table III shows features chosen for the training process.

TABLE III. LIST OF SELECTED FEATURES

No	Name
1	Packet Length Mean
2	Destination Port
3	Flow Packet/s
4	Flow Duration
5	Init Win bytes forward

C. Finding K Value

To find the optimal k for the algorithm, k-NN is trained multiple times ranging from 1 to 10 and the error is measured and shown in the graph in Fig. 4. The implementation of the method using Python with ‘sklearn’ to train the model and find the metrics, and ‘Pandas’ to import and split the data, and ‘matplotlib’ for plotting the result of error against the value of K. From Fig. 4 it can be observed that K=1 results in minimal error compared to the others. So, 1 is chosen for K in this experiment.

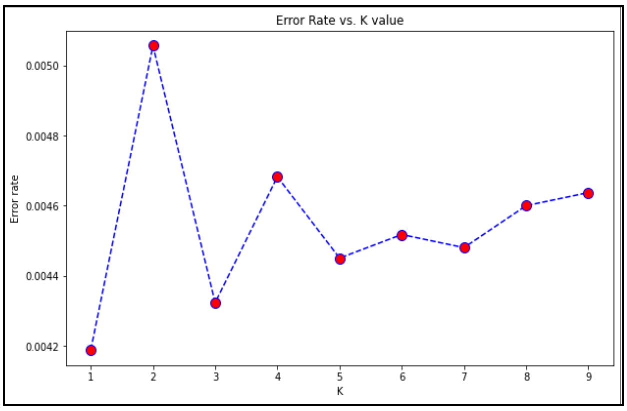


Fig. 4. Error rate vs K-value

V. RESULT ANALYSIS AND DISCUSSION

The performance of the classifiers is evaluated based on the confusion matrix. The confusion matrix is the most frequent and well-known metric applied to determine how exact and accurate a machine learning model is. The confusion matrix is appropriate in our research as the classification result contains equal or more than two types of classes. The comparison of confusion matrix on k-NN and Decision Tree in Table IV are made up of two dimensions of real and expected values. Some of the terms and data connected to the confusion matrix are true negatives (TN), true positives (TP), false negatives (FN) and false positives (FP). The Table V projects the result of the experiment in terms of accuracy, precision, recall and f1-score for each class for both of the classifiers.

TABLE IV. COMPARISON OF CONFUSION MATRIX ON K-NN AND DECISION TREE

Algorithm		True Benign	True FTP	True SSH	Precision
Decision Tree	Pred Benign	129425	25	17	99.97%
	Pred FTP	3	2389	0	99.87%
	Pred SSH	16	0	1819	99.13%
	Recall	99.99%	98.96%	99.07%	
k-NN	Pred Benign	129467	19	68	99.93%
	Pred FTP	57	2317	2	97.52%
	Pred SSH	39	1	1724	97.73%
	Recall	99.93%	99.14%	96.10%	

TABLE V. PERFORMANCE COMPARISON ON DECISION TREE AND K-NN

Algorithm	Class	Accuracy	Precision(%)	Recall(%)	F1-score(%)
Decision Tree	BENIGN	0.99954	99.97	99.99	99.98
	FTP		99.87	98.96	99.13
	SSH		99.13	99.07	99.10
k-NN	BENIGN	0.99861	98.4	99.93	99.93
	FTP		97.52	99.14	97.94
	SSH		97.73	96.10	96.91

Based on the Table V shows that Decision Tree has the highest precision in each class compared to the kNN. The decision tree also has a high recall rate except for the ftp class which is k-NN higher 0.18% which mean k-NN return a more relevant result in classifying SSH attack. For overall performance, it can be seen that the Decision Tree outperformed the Decision Tree in terms of F1 score which means that the Decision Tree has better performance in classifying attacks compared to the k-NN. Thus, it can be concluded that the Decision Tree has better performance in classifying the attacks compared to the k-NN.

VI. SUGGESTION FOR IMPROVEMENT AND FUTURE WORKS

For improvement of this research, improving the dataset processing, removing the huge class differences, and improving the machine learning parameters further. The machine learning algorithm, especially Decision Tree can also be improved by using the AdaBoost to further improve the result of the classification. Integration of the Intrusion Detection System (IDS) with this experiment will also improve the effectiveness of the IDS and increase protection against brute force attacks.

VII. CONCLUSION

This research has successfully achieved the study aims which is to examine the performance of the k-NN and Decision Tree algorithms by contrasting their precision, recall, and F1 score. The dataset was undergo the process of data cleaning and feature selection to further reduce the attributes and to make sure that only attributes are optimally selected for the training process. Besides, this research also demonstrated the optimal parameters for the machine learning algorithms to achieve the best results. This had done with kNN where the K is found by using error rate vs K. The lowest error rate for K will be chosen to be used in the training process. In this experiment, 1 has been chosen for K as it has the lowest error rate between the range of 1 to 10. The best performance algorithm which is Decision Tree can be used in detecting FTP and SSH brute force attack by using precision, recall and F1-score from this research. In the nutshell, the precision, recall and F1 score must be calculated based on the confusion matrix to determine the detection of FTP and SSH attacks.

ACKNOWLEDGMENT

The authors wish to thank all the experts involved for evaluation contribution of this research and the Ministry of Higher Education and University Teknologi Malaysia for funding this work under vot number (Q.J130000.3851.19J91).

REFERENCES

[1] Srivastava, P. and Khan, R. (2018) A review paper on cloud computing. *International Journals of Advanced Research in Computer Science and Software Engineering, 8,* 17-20. https://doi.org/10.23956/ijarcsse.v8i6.711.

[2] Barona, R. & Anita, Mary. (2017). A survey on data breach challenges in cloud computing security: Issues and threats. 1-8. https://doi.org/10.1109/ICCPCT.2017.8074287.

[3] Bonguet, Adrien & Bellaiche, Martine. (2017). A survey of denial-of-service and distributed denial of service attacks and defenses in cloud computing. *Future Internet, 9*, 43. https://doi.org/10.3390/fi9030043.

[4] Rajaguru, Harikumar & Chakravarthy, Sannasi. (2019). Analysis of decision tree and K-nearest neighbor algorithm in the classification of breast cancer. *Asian Pacific Journal of Cancer Prevention: APJCP. 20*, 3777-3781. https://doi.org/10.31557/APJCP.2019.20.12.3777.

[5] Harrison, O. (2019, July 14). *Machine learning basics with the K-nearest neighbors algorithm*. Medium. Retrieved March 20, 2023, from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

[6] Ikeda, S. (2020, April 13). 250 million microsoft customer service records exposed; exactly how bad was it? *CPO Magazine.* Retrieved March 20, 2023, from https://www.cpomagazine.com/cyber-security/250-million-microsoft-customer-service-records-exposed-exactly-how-bad-was-it/.

[7] Shung, K. P. (2020, April 10). Accuracy, precision, recall or F1? Medium. Retrieved March 20, 2023, from https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9.

[8] Najafabadi, M. M., Khoshgoftaar, T. M., Calvert, C. L., & Kemp, C. (2015). Detection of SSH brute force attacks using aggregated netflow data. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA),* 283-288. https://doi.org/10.1109/ICMLA.2015.20.

[9] Odun-Ayo, I., Ananya, M., Agono, F., & Goddy-Worlu, R. (2018). Cloud computing architecture: a critical analysis. 2018 *18th International Conference on Computational Science and Applications (ICCSA),* 1-7. https://doi.org/10.1109/ICCSA.2018.8439638.

[10] Suryateja, P. S. (2018). Threats and vulnerabilities of cloud computing: A review. *International Journal of Computer Sciences and Engineering, 6*(3), 297-302.

[11] Park, Jeonghoon & Kim, Jinsu & Gupta, Brij B & Park, Namje. (2021). Network log-based SSH brute-force attack detection model. *Computers, Materials & Continua., 680,* 887-901. https://doi.org/10.32604/cmc.2021.015172.

[12] Radivilova, T., Kirichenko, L., Ageiev, D., & Bulakh, V. (2019, September). Classification methods of machine learning to detect DDoS attacks. *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)* (Vol. 1, pp. 207-210). IEEE.

[13] Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J. Open Source Softw., 3*, 638.

[14] Wankhede, S., & Kshirsagar, D. (2018, August). DoS attack detection using machine learning and neural network. *2018 Fourth International Conference on Computing*

Communication Control and Automation (ICCUBEA) (pp. 1-5). IEEE.

[15] Singh, A., & Chatterjee, K. (2017). Cloud security issues and challenges: *A survey. Journal of Network and Computer Applications*, *79*, 88-115.

[16] Cloud computing. Cambridge Dictionary. (n.d.). Retrieved March 20, 2023, from https://dictionary.cambridge.org/dictionary/english/cloud-computing.

[17] Wanjau, S. K., Wambugu, G. M., & Kamau, G. N. (2021). SSH-brute force attack detection model based on deep learning.

[18] Subramanian, N., & Jeyaraj, A. (2018). Recent security challenges in cloud computing. *Computers & Electrical Engineering*, *71*, 28-42.

[19] Zamfiroiu, A., Petre, I., & Boncea, R. (2019, September). Cloud computing vulnerabilities analysis. *Proceedings of the 2019 4th International Conference on Cloud Computing and Internet of Things* (pp. 48-53).

[20] Yao, C., Luo, X., & Zincir-Heywood, A. N. (2017, November). Data analytics for modeling and visualizing attack behaviors: a case study on SSH brute force attacks. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1-8). IEEE.

[21] Shmagin, D. (2019). Utilizing Machine Learning Classifiers to Identify SSH Brute Force Attacks.

[22] Sadasivam, G. K., Hota, C., & Anand, B. (2018). Detection of severe SSH attacks using honeypot servers and machine learning techniques. *Software Networking*, 2018(1), 79-100

[23] Hossain, M. D., Ochiai, H., Doudou, F., & Kadobayashi, Y. (2020, May). Ssh and ftp brute-force attacks detection in computer networks: Lstm and machine learning approaches. *2020 5th International Conference on Computer and Communication Systems (ICCCS)* (pp. 491-497). IEEE.

[24] Hynek, K., Beneš, T., Čejka, T., & Kubátová, H. (2020). Refined detection of ssh brute-force attackers using machine learning. *ICT Systems Security and Privacy Protection: 35th IFIP TC 11 International Conference, SEC 2020, Maribor, Slovenia, September 21–23, 2020, Proceedings*, 35(pp. 49-63). Springer International Publishing.

[25] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp, 1*, 108-116.

[26] Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: A review. *Big Data*, *7*(4), 221-248.

[27] He, Y., Ma, J., & Ye, X. (2017). A support vector machine classifier for the prediction of osteosarcoma metastasis with high accuracy. *International Journal of Molecular Medicine*, *40*(5), 1357-1364.

[28] Ooi, K. B., Lee, V. H., Tan, G. W. H., Hew, T. S., & Hew, J. J. (2018). Cloud computing in manufacturing: The next industrial revolution in Malaysia? *Expert Systems with Applications*, *93*, 376-394.

[29] Kulkarni, R. (2020, May 23). Summary of KNN algorithm when used for classification. Medium. Retrieved March 20, 2023, from https://medium.com/analytics-vidhya/summary-of-knn-algorithm-when-used-for-classification-4934a1040983.

[30] Das, A. (2020, July 18). Decision tree algorithm for multiclass problems using Python. Medium. Retrieved March 20, 2023, from https://towardsdatascience.com/decision-tree-algorithm-for-multiclass-problems-using-python-6b0ec1183bf5.