COMPARATIVE ANALYSIS OF FIREWALL RULE SET USING CLASSIFICATION ALGORITHMS

Mohd Fazzly Rassis bin Md Kasim[1]
Mohamad Fadli bin Zolkipli, Ph.D[2]

Abstract

This study focuses on comparative analysis of firewall rule set using classification algorithms based on the fundamental concept of data mining to evaluate the accuracy and performance of several classification algorithms. Rule sets grow to large numbers written by different network administrators. This condition will cause increase the rule set policy and complexity poses problem among other inconsistencies in the firewall configuration. This led to firewall poses overload and used high process performance. The Knowledge Discovery in Database (KDD) is adopted as research methodology to illustrate how this study was conducted. In this study, classification algorithms namely JRIP, J48, Naïve Bayes, Random tree and Random forest were used for the classification of dataset. Waikato Environment for Analysis Knowledge (WEKA) was used in comparing these algorithms. Two firewall dataset were used, KUIPSAS 1098 dataset and PSDC 1024 dataset as training and testing data on different classification algorithms. The experiment used dataset that have been formatted into ARFF 10 folds cross validation and the results were compared for accuracy. Based on the comparative analysis, it can be concluded that using two different datasets from different sources indicated that the Random Tree algorithm shows the best performance in terms of accuracy which are 99.70% for PSDC and 99.80% for KUIPSAS.

Keywords: Firewall rule set, Data Mining Algorithm, Machine Learning

Abstrak

Artikel ini bertujuan membuat analisis perbandingan dataset firewall menggunakan algoritma klasifikasi berdasarkan konsep asas perlombongan data dan menilai ketepatan serta prestasi algoritma. Peningkatan Rule set firewall yang dikonfigurasi oleh pentadbir rangkaian menyebabkan prestasi pemprosesan yang tinggi terhadap peranti firewall perlu dipertingkatkan. Metodologi The Knowledge Discovery in Database (KDD) digunakan untuk menggambarkan bagaimana kajian ini telah dijalankan. Dalam kajian ini, algoritma pengklasifikasian iaitu JRIP, J48, Naïve Bayes, Random tree dan Random forest digunakan untuk klasifikasi dataset firewall. Perbandingan algoritma-algoritma ini dijalankan dengan menggunakan perisian Waikato Environment for Analysis Knowledge (WEKA). Dua dataset firewall iaitu KUIPSAS sebanyak 1098 dataset dan 1024 dataset dari PSDC dipilih sebagai dataset untuk latihan dan pengujian dengan menggunakan algoritma klasifikasi yang berbeza. Data-data ini telah diformatkan ke dalam bentuk ARFF 10 folds cross validation dan digunakan untuk eksperimen ini di mana hasilnya dibandingkan dari segi ketepatan. Keputusan analisis perbandingan menunjukkan bahawa dengan menggunakan dua dataset firewall yang berbeza dari sumber yang berbeza menunjukkan algoritma Random Tree mempunyai prestasi terbaik dari segi ketepatan iaitu 99,70% untuk PSDC dan 99,80% untuk KUIPSAS.

Kata Kunci: Rule set firewall, Algoritma Perlombongan Data, Pembelajaran Mesin

[1] Penulis merupakan Pensyarah Fakulti Teknologi Maklumat dan Komunikasi, Kolej Universiti Islam Pahang Sultan Ahmad Shah (KUIPSAS)

[2] Penulis merupakan Pensyarah kanan Fakulti Sistem Komputer dan Kejuruteraan Perisian, Universiti Malaysia Pahang (UMP)

INTRODUCTION

Firewall is a first line device of network defence for the malicious attack and unauthorized traffic (Liu, 2009). In the beginning, firewall were simple packet filter by using small set rules to determine traffic would be allowed into the network (Cherian & Chatterjee, 2016). Over the year, the network has increased in the number and type of network consists applications, web-based services, communication tools and more.

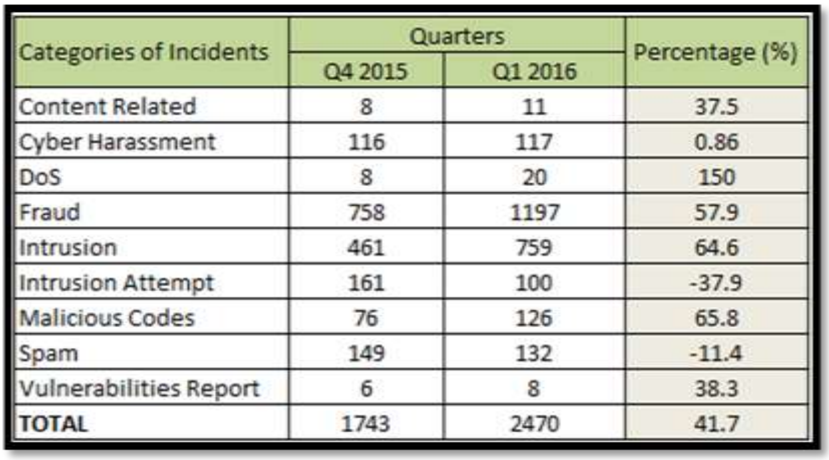
The firewall is comprised of software and hardware used to develop security policies that controlling the flow of traffic transmitted over two or more networks (Sheth & Thakker, 2014). It works to filter and manage packets in a computer network environment. There is a firewall definition mentioned by others shown in Table 1.

Table 1 Firewall definition

Reference	Definition
(Kadam & Bhusari, 2014)	The most important of firewall are is settled packet filtering at the entry point into a network which includes the highest secure access inbound and outbound from the network.
(Trabelsi *et al.*, 2013)	Additionally, the firewall contains security policies to check outbound and inbound network traffic.
(Cuppens-boulahi *et al.*, 2013)	In another reviewer, the firewall has been used to examine every incoming and outgoing data and deployed in field institution and business for securing a private network.

Nowadays, the attacks on the Internet are accumulating. Figure 1 as a result by MyCERT Malaysia indicates comparison incidents between Q4 2015 and Q1 2016 (MyCERT, 2016). It shows the highest percentage of an increase in malicious attempt code which is 65.8%. Meanwhile, the lowest percentage of a decreased incident in intrusion attempt is -37.9%.

Table 2 Comparison of number of incidents between Q4 2015 and Q1 2016



Categories of Incidents	Quarters		Percentage (%)
	Q4 2015	Q1 2016	
Content Related	8	11	37.5
Cyber Harassment	116	117	0.86
DoS	8	20	150
Fraud	758	1197	57.9
Intrusion	461	759	64.6
Intrusion Attempt	161	100	-37.9
Malicious Codes	76	126	65.8
Spam	149	132	-11.4
Vulnerabilities Report	6	8	38.3
TOTAL	1743	2470	41.7

Network security consists of protection of policies on computer network for entire infrastructure (Mohan, 2015). It is clear the firewall can protect critical computer files and information to help prevent viruses, theft, spyware, malware, and more. When the internet has become widely used, everyone can easily gain access and threats has increased. The data of any organization can be easily accessed by intruders. The rest of the topics will be explained in the next section.

RELATED WORK

This section will give a brief description of the previous work related to comparative analysis of various algorithms. Some of these algorithms will be used in this study. According to (Ucar & Ozhan, 2017) Naïve Bayes, kNN, Decision Table and HyperPipiesis is used to build a model in which to detect anomaly in firewall rule repository. This research combines the precision and recall for performance evaluations. The experiment showed that kNN is the best performance.

Dash (2013) stated that Lazy, Meta, Rules and Tree classifier algorithm is used for the classification of a few dataset in format ARFF 10 fold cross validation. Analysis for these algorithms are performed by using WEKA tool and help in the correctly classifier.

Masud et al (2014) stated that the data mining can be used for packet filtering. DDF Data Driven Firewall proposes to predict class for incoming packet either accept or deny. The advantage for this algorithm in data mining offers a much faster and better accuracy.

In another study, Verbruggen (2014) used Random Tree and Random Forest to classify based on malware intrusion detection. They ran test data for both classifications to check performance of several algorithms on WEKA tool.

According to Urvashi & Jain (2015) a data mining technique is used to classify different detection methodology for intrusion attack. They used dataset in network security research KDDCUP 99 its contain various component. They have compared the performance various classification algorithms used in WEKA such as Bayes, Lazy, Meta, Rules, Misc and Tree.

FIREWALL RULESET

In order to understand how the firewall operates it need to be understood the relationship between the access control rules and the packet they govern. This subsection provides a formal explanation of the terms.

According Abdul Aziz *et al* (2012) the rule base is a set of rules that control what is allowed or denied through a firewall. The decision to deny or allow contained in the packet. Internet firewalls are usually set to five and consist of protocol, destination address, destination port, source address, source port and action (Mustafa *et al*., 2013). The source is IP address which initiates the packet while the destination is the IP address of receiver packet. From Sheth & Thakker (2014) addressed when a packet received, the rule base scan from start to the end. Next action is associated the first match is taken if the packet header is matched all that rule.

The protocol field specifies a protocol as documented in the IP Packet header protocol field, Internet Protocol (RFC 791). For an internet firewall, this will be either TCP, UDP or ICMP (Abedin *et al*., 2010). The TCP and UDP protocol use port numbers in the range 0-65535. The Internet Assigned Number Authority (IANA) recommends global unique names and number use for TCP and UDP. Adopted Abedin *et al* (2010) the port number are divided into three groups mention in Table 2 and Table 3 outlines the action field values.

Table 3 Port number assign by IANA

Group	Port number
Well known port	0-1023
Register port	1024-49151
Private port	49152-65535

Table 4 The action field values and its effect

Action	Effect
Deny	Forward the packet
Allow	Drop the packet

Firewall configuration is a difficult task (Mustafa *et al*., 2013). This is because set rules evolve into thousands of rules and the trend of network traffic continues to change. From Table 4 shows an example of firewall rule set filter. The way it works is that this rule identifies clearly any incoming packet with any source address from 192.168.50.8 source port 80 and destination address is 216.58.199.206 destination port 80. Next, these ruleset action are allowed to access.

Table 5 The actual sample packet filter firewall Rule set

Pro	Source Address	Source Port	Dest Address	Dest port	Action
TCP	192.168.50.8	80	216.58.199.206	80	Allow
TCP	192.168.51.2	80	179.60.194.35	80	Deny
TCP	192.168.51.3	80	179.60.194.35	80	Deny
TCP	192.168.51.4	80	179.60.194.35	80	Allow
TCP	192.168.51.5	80	179.60.194.35	80	Allow
UDP	192.168.51.6	80	179.60.194.35	80	Allow
UDP	192.168.51.7	80	179.60.194.35	80	Deny

From the sample rule set above, troubleshooting or editing a thousand rules is not easy. Each packet will be checked for sequence in order to see whether the match is allowed or pushed into the network (Mustafa *et al.*, 2013). This study proposes to classifier rules set using algorithm to find out the suitable algorithm for packet filtering firewall performance. The next section will discuss more depth about machine learning.

MACHINE LEARNING

Generally, Machine Learning has become more popular and it use has more common. Machine Learning is subset of algorithm develop in Artificial Intelligent and these algorithm use different features to learn a set of rule to identify different classes (Liao *et al.*, 2012). This is a study to learn new skills and knowledge regarding machine learning. Therefore, many algorithms are used in data mining technique to solve real life problem.

According to Sharma & Niranjan (2012) Machine Learning has various application as mentioned. There are many advantages of data mining such as those that assist in finance, banking , retail, marketing, medical science, image screening, search engine and more. The input of machine learning is a simple dataset or example is derived from features also known as discriminator and data set is an example matrix compared to discriminator. Lastly, the processes are the knowledge that the machine learned.

There are three major types of Machine Learning in context of rule set classification that is Supervised (classification), Unsupervised (clustering) and Reinforcement (Praveena & Jaiganesh, 2017). In this study will focus on supervised learning. To understand of a brief introduction machine learning and state of the research work are presented here, figure 2 shows Machine Learning type.

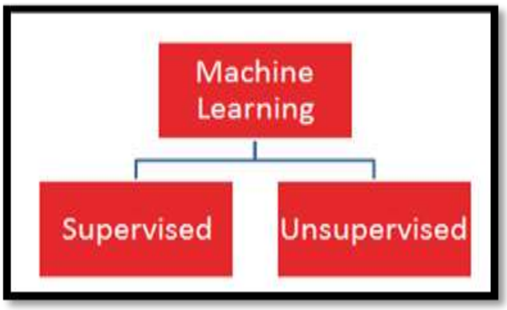


Figure 2 Machine Learning Type adopted (Praveena & Jaiganesh, 2017)

Supervised machine learning is a mission to get the meaning of label data that has set training example. As far as supervised learning each instance is a mainstay that contains an input object and enforceable output values. Supervised learning algorithm initially perform analysis tasks from practise data construct a function in order top map new example. The supervised methods may be used in many areas of the application including testing, market prediction, finance and so on. To apply Machine Learning mechanism for this study WEKA classifier algorithms is chosen as the data mining tasks.

In order to ensure that machine learning process accomplishment to improve performance of the classification algorithms. The following algorithms will be used and will be discussed for the next section.

CLASSIFICATION ALGORITHM

Classification algorithm also known as classifier is used to classify the packet filter rule as Allow or Deny. According to Urvashi & Jain (2015) WEKA has a vast collection of algorithms for solving problem. For this study, the algorithms used are J48, Random Tree, Random Forest, Naïve Bayes and JRIP is discussed below. These algorithms are discussed below.

JRIP is one of basic and most popular algorithm. This algorithm is optimized version algorithm proposed by William W Cohen. It optimizes the rules set using discretionary length (Choudhury & Bhowal, 2015). Classes are examined in growing size and an initial set of rules for the class is generated using incremental reduced error JRIP (RIPPER) proceeds by treating all the examples of a particular decision in the training data as class, and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered. Proposed rule leaner and cumulative error pruning method to reduce mistakes. Finally, this algorithm will try to add every possible rule until it becomes accurate.

Decision Tree Algorithm is to find out the way the attributes vector behaves for a number of instances. This algorithm generates the rules for the prediction of the target variable. J48 is an extension of ID3 introduce by Ross Quinlan Is an upgraded of C4.5 in WEKA with some additional function to solve inefficient on ID3 (Barnaghi, Sahzabi, & Bakar, 2012)**.** Therefore, this technique is a time and space consuming. Initially, it builds a tree using split and conquer algorithms and then uses heuristic criteria. The main purpose works on the supervised learning, classification to produce decision tree. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

Naïve Bayes method is also used one of classification solution in data mining. Naive Bayes is an extension of the Bayes theorem as it considers the independence of attributes (Nwulu, 2017). Classification base of extracting text from a document where a relationship between the words accumulates into the concepts. Naïve Bayes algorithm is based on the rule of conditional probability and takes discrete data as input. Naïve Bayes is easy to construct without any need for complicated parameter estimation. This algorithm may be sue for the large datasets. It's robust, easy to understand, and often not surprising though it may not be the best classifier in any particular application. The data in Naïve Bayes are symbol as n size feature vector, X = (x_1, x_2,…x_n). Figure 3 shows general outlook of system.

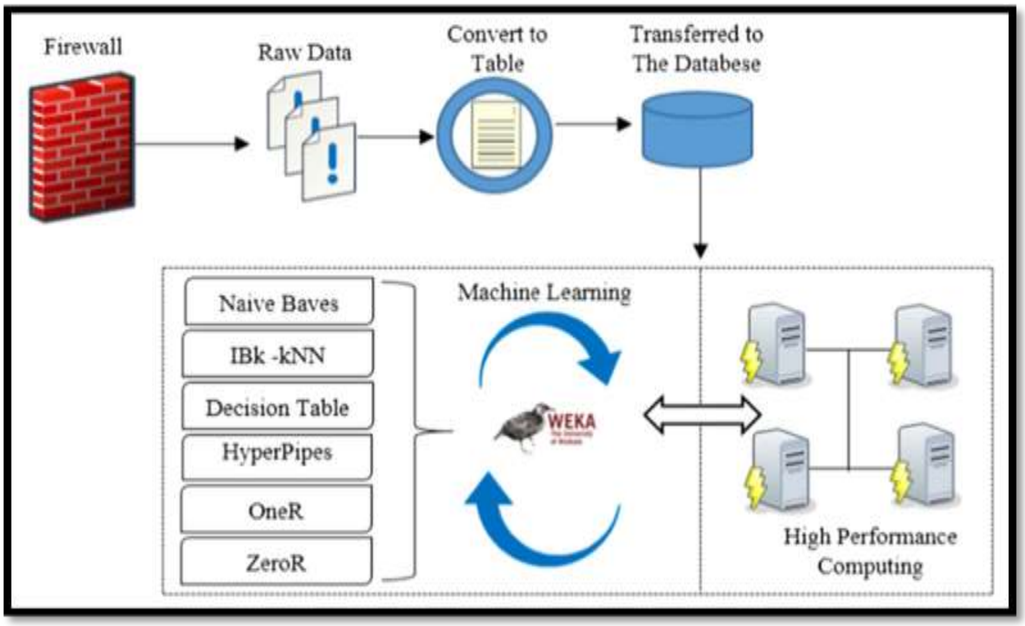


Figure 3 The general outlook of system taken from (Ucar & Ozhan, 2017)

According to Praveena & Jaiganesh (2017) introduced the Random Tree is defined as a prediction modeling technique from the machine learning field and the statistics that builds a simple tree such as a structure to model pattern. Random tree are example of classification algorithm. Classifier algorithm has solved various problems such as diagnosing patients with heart problem, card credit theft detection and so on. From the traffic network classification algorithm can characterize network data such as scanning, malicious, packet filter performance using information like protocol, source IP/destination IP, port and number of byte. Figure 4 shows example random tree for known malicious port 8787.

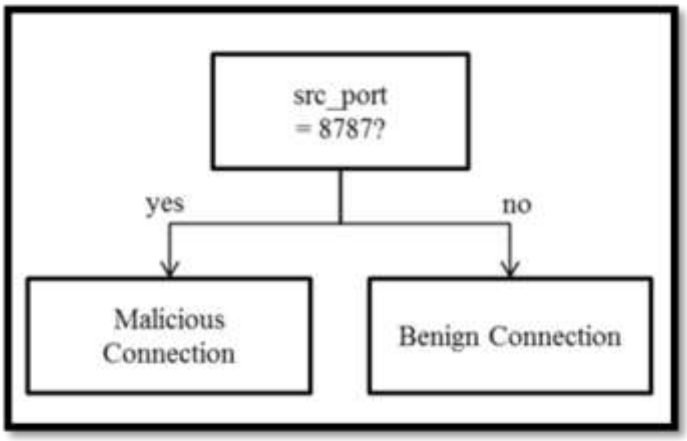


Figure 4 Random tree example

Random forests is an idea of the general technique of random decision forests that are an ensemble learning technique for classification, regression and other tasks, that control by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Adopted from (Choudhury & Bhowal, 2015) random forest uses an ensemble method to get the better predictive performance. It produces output in the individual tree and based on the decision random tree algorithm. Therefore, a random forest is a highly accurate algorithm and can handle multiple variables. The following chapter will go into detail regarding research methodology.

METHOD

The general methodology of Knowledge Discovery in Databases (KDD) adopted from (Guruvayur, 2017), (Liao *et al*., 2012) are used as the research methodology to make the overall methodology complete. Figure 5 illustrates the KDD process that explains the iterative and interactive procedure.

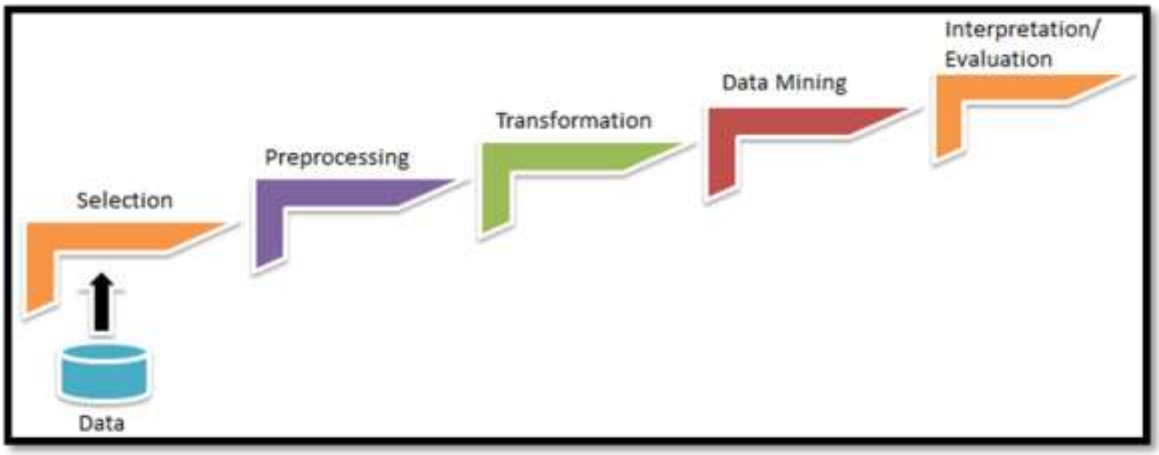


Figure 5 Modeling Steps KDD Adopted from (Guruvayur, 2017)

The KDD process in Figure 5 consists of five major steps which are selection, preprocessing, transformation, data mining and interpretation/evaluation. The following sections describe the respective steps of the development stage.

a) Selection

In this sub stage, selection refers to the process of obtaining data process, the data may come from different sources. Therefore, the data needs to be selected before it can be used as an input for the data mining process. This step is important in order to select only significant data that will be used for the entire study.

In this sub stage, selection refers to the process of obtaining data process, the data may come from different sources. Therefore, the data needs to be selected before it can be used as an input for the data mining process. This step is important in order to select only significant data that will be used for the entire study.

In this study, the real data are used as the target data for classification purpose. The data are collected from firewall policy rule set of organization education KUIPSAS and PSDC is used as target data. The raw data originally consists of eight attributes. This attributes can be viewed in Table 6.

Table 6 Raw Data Attributes for KUIPSAS and PSDC

Num.	Attributes Name for KUIPSAS	Num.	Attributes Name for PSDC
1	Protocol	1	Protocol
2	Sources Address	2	Sources Address
3	Source Port	3	Source Port
4	Destination Address	4	Destination Address
5	Destination Port	5	Destination Port
6	Action	6	Action
7	Packet size	7	Schedule
8	Packet Arrival time	8	Count

Generally, several features were selected to be classified using machine learning algorithm. The detail of selection process is explained in the following section. In the next section, pre-processing data will be described in detail.

b) Pre-processing

Every Machine learning often differs in design or structure of database requirement. Database is usually contains missing value and error. Therefore, it is needed to be processed before it can be used in machine learning. The raw data can be found in various formats. Once the target data has been determined, the raw data need to go through a data cleaning process. This process contributes the raw data into good data to fit into the classification. These processes include eliminating the unwanted attributed to ensure the classification process run smoothly and obtain a good result.

According Nath *et al* (2011) stated dimensionality reduction can improve efficiency of the classification. This process reduces the effective number of variable under consideration or to find invariant representation for the data. Dimensionality reduction is an essential data processing technique for a large scale data.

In this study, dimensionality reduction has been chosen as the feature reduction. Verbruggen (2014) mentioned that used six features consist of protocol, destination address, destination port, source address, source port and action. Based on this, after the features reduction process, only six attributes are selected from the original eight attributes. Figure 6 and 7 illustrate the attributes reduction process for KUIPSAS and PSDC.

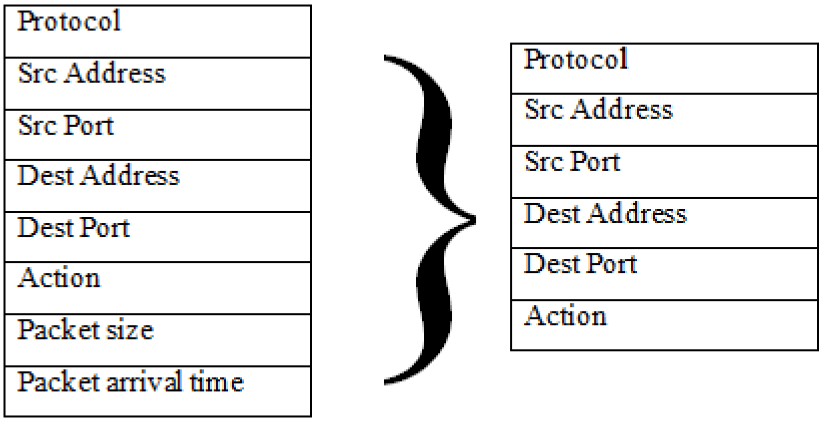


Figure 6 Attributes Reduction on Features Selection for KUIPSAS

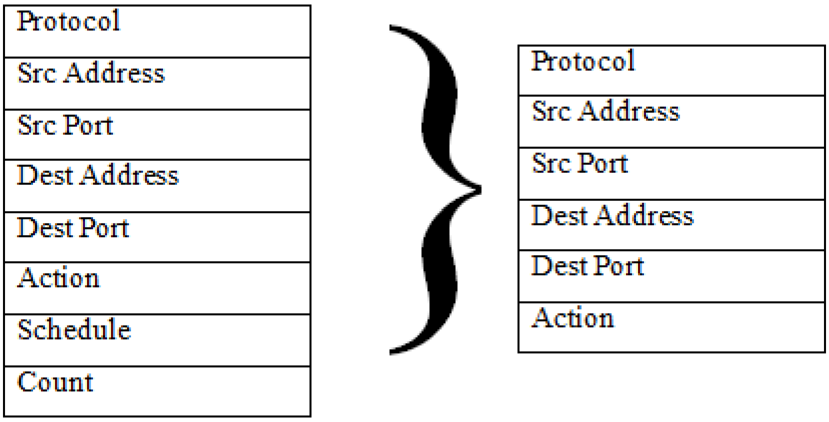


Figure 7 Attributes Reduction on Features Selection for PSDC

These six attributes are selected after completing review process. The output of this stage is clean rule set firewall from education organization KUIPSAS and PSDC. Each data a rule set in firewall in Table 5 and Table 6 explained the entire attributes has been used.

Table 7 Field of each record firewall Rule Set

Pro	Source Address	Source Port	Dest Address	Dest port	Action
TCP	192.168.50.8	80	216.58.199.206	80	Allow
TCP	192.168.51.2	80	179.60.194.35	80	Deny
TCP	192.168.51.3	80	179.60.194.35	80	Deny
TCP	192.168.51.4	80	179.60.194.35	80	Allow
TCP	192.168.51.5	80	179.60.194.35	80	Allow
UDP	192.168.51.6	80	179.60.194.35	80	Allow
UDP	192.168.51.7	80	179.60.194.35	80	Deny

Table 8 Define attributes

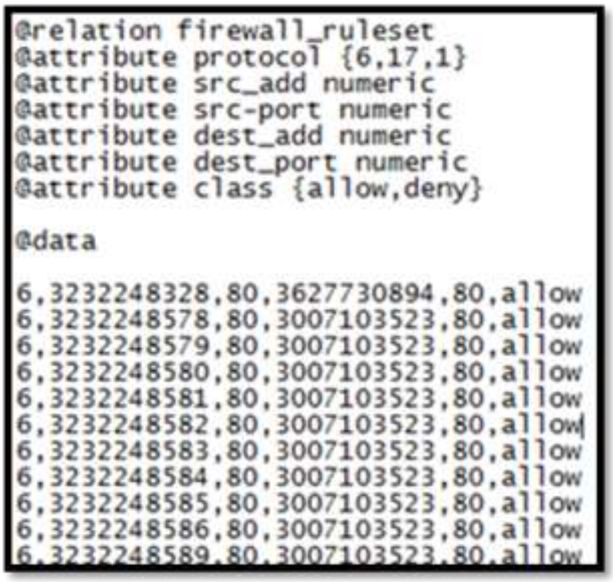
Instances	Explanation
Pro	The protocol applies at the session packet. (TCP, UDP, ICMP)
Source address	IP (internet Protocol) address of the device send the IP packet.
Source port	Port from 1024 to 65535.
Dest address	The IP address of the device to which the packet is being sent.
Dest port	Port from 1024 to 65535.
Action	Status match packer either allow or deny.

c) Transformation

Transformation is the process of making data more useful and provide more meaningful data format (Guruvayur, 2017). The aim for data transformation is quite similar to data prepossessing, this is to pledge a good input for the classification process. This process involves converting or transforming the data in to usable format. Data from previous process (Processing) may be modified to facilitate usage by selecting technique that require specific data format. Some attribute may be changed to value.

According to Masud *et al* (2014) the data information consists ICMP, TCP, UDP and IP address must convert before applying with the association rule technique. In this study, first attribute is the Protocol such as TCP, UDP and ICMP will represent to value 6,17,1. The protocol format reserved from the Internet Assigned Numbers Authority (IANA). Next attribute are IP address and its port for the source and destination, while format for attribute

IP address such as 117.121.250.149 will change to decimal number 1970928277. Finally the last attribute is the action either allow or deny the packet. Figure 8 below is the sample attribute and data format of WEKA.



```
@relation firewall_ruleset
@attribute protocol {6,17,1}
@attribute src_add numeric
@attribute src-port numeric
@attribute dest_add numeric
@attribute dest_port numeric
@attribute class {allow,deny}

@data

6,3232248328,80,3627730894,80,allow
6,3232248578,80,3007103523,80,allow
6,3232248579,80,3007103523,80,allow
6,3232248580,80,3007103523,80,allow
6,3232248581,80,3007103523,80,allow
6,3232248582,80,3007103523,80,allow
6,3232248583,80,3007103523,80,allow
6,3232248584,80,3007103523,80,allow
6,3232248585,80,3007103523,80,allow
6,3232248586,80,3007103523,80,allow
6,3232248589,80,3007103523,80,allow
```

Figure 8 Rule set format

d) Data mining

The last steps in development stage is refer to the process of applying suitable algorithm to transform data into desired result. According to Ucar & Ozhan (2017) WEKA officially known as Waikato Environment for Knowledge Analysis, is a computer program developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data collected from agricultural domains. WEKA supports many different standard data mining tasks such as preprocessing, classification, grouping, regression, visualization and selection of data features. WEKA has been used by several researchers in data mining domain.

For this study, WEKA classifier algorithms is chosen as the data mining stage. The classifier has been evaluated in Waikato Environment For knowledge Analysis (WEKA) has been used for this experiment. The major WEKA package is classifier, filter, cluster, association and attribute selection. We use 10-fold cross validation to test and evaluate the algorithm. In 10-fold cross validation process the data set is divided into 10 subsets. Performances are calculated across all 10 trials. So here is the scenario for 100 label data. WEKA will take 100 label data and it produces the same 10 sizes set. Each set divided into two groups, which is 90 label data is used for training and 10 label data is used for testing. It produces an algorithm from 90 label data and use them in 10 test data for set one. It does the same to set two to 10 and produces nine more classification algorithms. Finally, it averages the performances of the 10 classifier produced from this method. Figure 9 is example of 10 fold cross validation method.

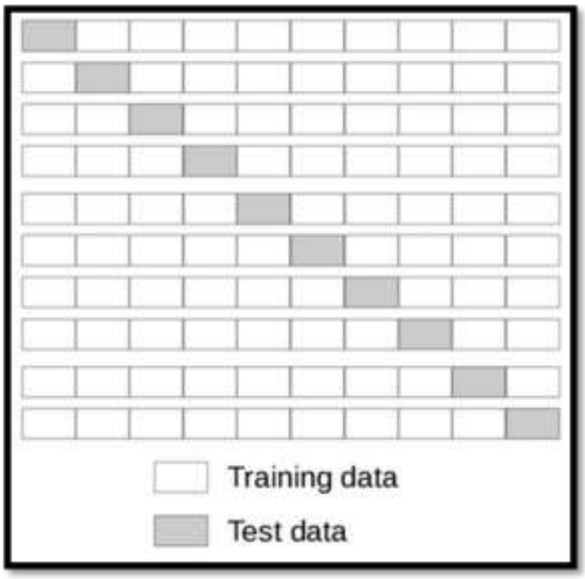


Figure 9 Example of 10 fold cross validation

Based study on related work, classification algorithm has been selected are J48, JRIP, Naïve Bayes, Random Tree and Random Forest used in this experiment. We use dataset from education organization KUIPSAS and PSDC for this training and testing.

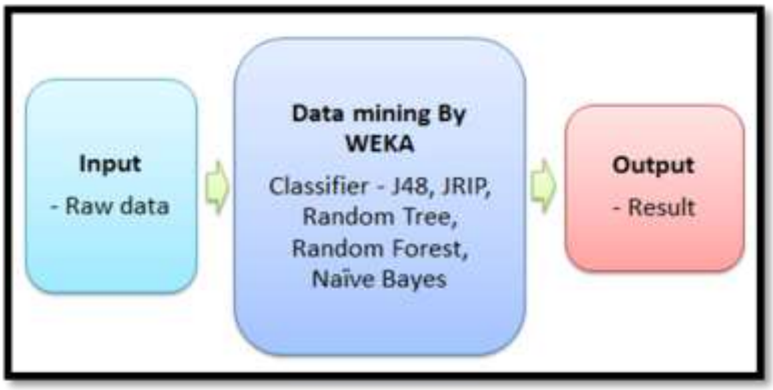

Input
- Raw data
Data mining By WEKA
Classifier - J48, JRIP, Random Tree, Random Forest, Naïve Bayes
Output
- Result

Figure 10 Working of WEKA (Sharma & Niranjan, 2012)

e) Evaluation measure

This stage conveys the result into understandable format to be used by user. All the data must be clearly interpreted and evaluated to ensure the resulting information is clear and express the result without any misunderstanding. This stage involves the evaluation to determine the accuracy.

The performance of classification algorithm is usually examined by evaluating the accuracy of classification algorithm. This situation implies the decision as either “ALLOW” or “DENY” rule set packet filtering for firewall. Generally, the output for this classification algorithm can be divided into two group of classes such as event and non-event. Event class represent the “ALLOW” and non-event class is for “DENY” the rule set. In this study, performance evaluation is adopted from (Garg & Khurana, 2014) as shown in Table 10.

Table 9 Class Prediction

	Actually an Event	**Actually non Event**
Event/deny	True positive (TP)	False Positive (FP)
Non Event/allow	False Negative (FN)	True Negative (TN)

The value of true positive indicates the number of event correctly predicted by the algorithm. False positive shows that the actual non-event is categorized as an event. If the non-event is correctly predicted, it will be categories as true positive. False negative refer to the group of the wrong predicted actual event as non-event. All the number is calculated by comparing the result of the algorithm versus the real result made by an expert. From this number (TP, FP, TN, FN), sensitivity, specificity and accuracy of the algorithm can be calculated.

The value of true positive indicates the number of event correctly predicted by the algorithm. False positive shows that the actual non-event is categorized as an event. If the non-event is correctly predicted, it will be categories as true positive. False negative refer to the group of the wrong predicted actual event as non-event. All the number is calculated by comparing the result of the algorithm versus the real result made by an expert. From this number (TP, FP, TN, FN), sensitivity, specificity and accuracy of the algorithm can be calculated.

PERFORMANCE MATRIC

The performance of classification algorithms is usually examined by evaluating the accuracy of the classification algorithm. Generally, output classification algorithm is divided into two group class which are event (allow) and non-event (deny). To measure the performance of the algorithm, the accuracy, sensitivity, specificity is used. These performances metric are readily usable for the evaluation. Sensitivity is a metric measurement of how classification test correctly identify the event. In this study, the event refers to “allow” rule set. A sensitivity of 1 or 100% means that the test recognises all “allow” rule set. Sensitivity formulation is:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

To indicate the specificity, a matric measurement of how well classification test correctly prediction non-event. A 1 or 100% specificity means that the test recognize to “deny” rule set. Specificity formulation is :

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Accuracy is the ratio of total correct prediction. Accuracy formulation is :

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

RESULT AND DISCUSSION

This study conducted two experiments on Waikato Environment for Knowledge (WEKA). Five different classification algorithms were selected such as JRIP, J48, Naïve Bayes, Random Tree and Random forest. In this study, real dataset from two source firewall rule set of KUIPSAS and PSDC have been used instead of standard dataset. Dataset contains five features and consists of two kind of action which are allow or deny. Table 11 shows the information about the dataset.

Table 10 Dataset Information

DATASET	NO.OF RULES	ALLOW	DENY
KUIPSAS	1098	583	575
PSDC	1024	417	607

This experiment has compared the performance of all algorithms based on the accuracy. Training and testing dataset used cross validation method 10-fold to test and evaluate the algorithm. Both the training and testing dataset are in ARFF file format. After completion of the experiments, comparative analysis will be made to evaluate the best result for classification algorithms.

Experiment 1: Classification using dataset KUIPSAS

Figure 10 shows the comparative analysis of various classifications according to accuracy, sensitivity and specificity for KUIPSAS and PSDC datasets. The evaluation algorithm process has shown a very interesting result. From the graph, we can observe all classification algorithms and the decision tree Random tree algorithms has performed with accuracy 99.80%. Second lower is Random forest with accuracy 99.70% and followed by JRIP, J48 and Naïve Bayes. The highest sensitivity for Random Tree is 99.70 % and lowest for Naïve Bayes is 95.60%. The specificity is maximum for Random Tree which is 100% and minimum for Naïve Bayes which is 85.30 %.

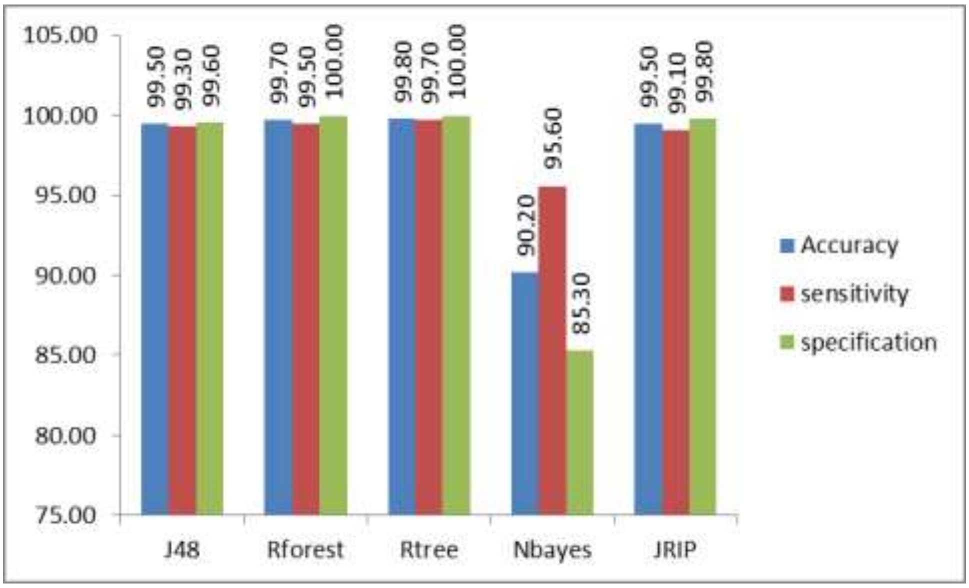


Figure 10 Performance graph for all classification algorithms for KUIPSAS

Table 12 shows the ranking of classification. The highest time is taken by Random Tree with the consuming 0.01 seconds while lowest time is taken by Random Forest with 0.64 seconds. According to the above analysis, Random Tree is higher based on accuracy than other classification because it requires a lot of examples tree for processing its random concept. It is considered to be in accordance with the characteristic and the ability to accumulate and store a large amount and facilitate it.

Table 11 Ranking of classification

Classifier	Accuracy	Sensitivity	Specificity	Time (s)	RANK
Randomtree	99.80	99.70	100	0.01	1
Rforest	99.70	99.50	100	0.64	2
J48	99.50	99.30	99.60	0.09	3
JRIP	99.50	99.10	99.80	0.50	4
Nbayes	90.20	95.60	85.30	0.05	5

Experiment 2: Classification using dataset PSDC

Naïve Bayes is lower in term of accuracy and sensitivity compared to the other classification which their performance are comparatively low as shown in Figure 11. Naïve Bayes is based on Bayesian theorem is highly scalable it performs for dataset like medical data. From the graph, it can be seen, the decision tree Random Tree algorithm has performed with highest accuracy 99.70% while the lowest accuracy is Naïve Bayes 72.40%. The highest sensitivity are Random Tree and Random forest with 99.80% and the lowest is Naïve Bayes is 60.40%. The specificity is maximum for Random Tree which is 99.70% and minimum for Naïve Bayes which is 92.90 %

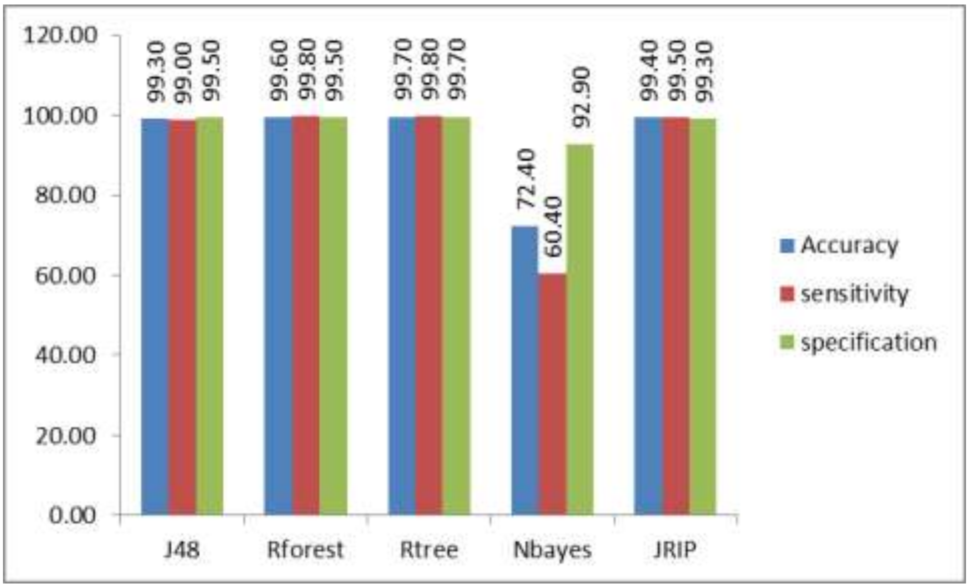


Figure 11 Performance graph for all classification for PSDC

Table 12 shows the ranking of classification. The highest time is taken similarly by Random Tree, J48 and Naïve Bayes with time consuming 0.02 seconds while lowest consuming taken by Random Forest with 0.56 seconds. According to the analysis, Random Tree performance is higher based on accuracy. (Guruvayur, 2017) Mentioned decision tree has better function with numeric information. Therefore, it can be found that reasonable in this experiments decision tree classifier have the highest performance.

Table 12 Ranking of classification

Classifier	Accuracy	Sensitivity	Specificity	Time (s)	RANK
RandomTree	99.70	99.80	99.70	0.01	1
Randomforest	99.60	99.80	99.50	0.17	2
JRIP	99.40	99.50	99.30	0.04	3
J48	99.30	99.00	99.50	0.01	4
Nbayes	72.40	60.40	92.90	0.01	5

Result Observation

Graph in Figure 12 shows the Random Tree is the best classification algorithm based on accuracy for dataset firewall KUIPSAS and PSDC. The accuracy of dataset KUIPSAS is the highest compared to PSDC. The accuracy of KUIPSAS is 99.80% while accuracy for PSDC is 99.70%. Based on comparative analysis, it can be concluded the Random Tree algorithm is the best performance in term of accuracy by using two different dataset from different source.

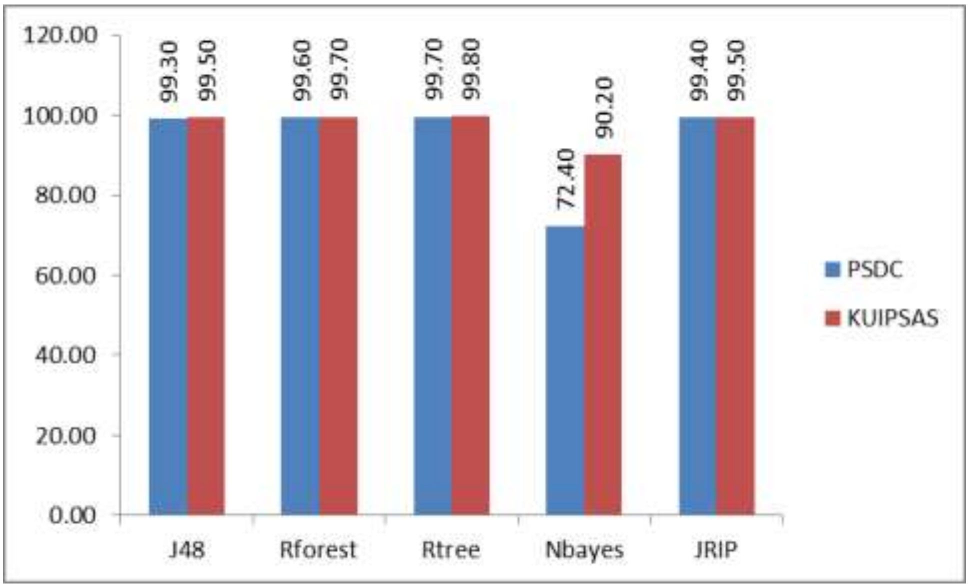


Figure 12 Comparison Between dataset KUIPSAS and PSDC based on Accuracy

CONCLUSION

The experimental result and performance evaluation of classification algorithms model has been performed. Original dataset from KUIPSAS and PSDC are used for this testing and training to compare the result from five different classification algorithms. From the performance evaluation of the classification algorithms by WEKA tool. The result has produced on both dataset for Random Tree as the higher with accuracy 99.80% for KUIPSAS and 99.70% for PSDC. Finally, this research provides a basic comparative analysis of different classification algorithms for firewall ruleset and build a useful model that can be used in a real environment firewall. This research is the initial step and would be a benchmark for other researcher in data mining field can be considered as preliminary study.

ACKNOWLEDGEMENT

Assalamualaikum Wrt., Wbt. In the name of Allah, the Most Gracious, The Most Merciful. All praise to Allah for giving me health, strength, courage and patience those make possible the completion of this research. The author would like to thank UMP and KUIPSAS in providing the facilities for this study. We are also grateful and would like to acknowledge and thank the Faculty of Computer Systems and Software Engineering, UMP and Faculty of Information Technology and Communication, KUIPSAS that has supported this research. Any correspondence should be emailed to fazzly@kuipsas.edu.my.

REFERENCES

Abdul Aziz, M. Z., Ibrahim, M. Y., Omar, A. M., Ab Rahman, R., Md Zan, M. M., & Yusof, M. I. (2012). Performance analysis of application layer firewall. *2012 IEEE Symposium on Wireless Technology and Applications (ISWTA)*, *3*(600), 182–186.

Abedin, M., Nessa, S., Khan, L., Shaer, E. Al, & Awad, M. (2010). Analysis of firewall policy rules using traffic mining techniques. *International Journal of Internet Protocol Technology*, *5*(1/2), 3.

Barnaghi, P. M., Sahzabi, V. A., & Bakar, A. A. (2012). A Comparative Study for Various Methods of Classification. *International Conference on Information and Computer Networks*, *27*(Icicn), 62–66.

Cherian, M. M., & Chatterjee, M. (2016). Firewall Optimization with Traffic Awareness Using Binary Decision Diagram, *8*(1), 9–14.

Choudhury, S., & Bhowal, A. (2015). Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network Intrusion Detection, (May), 89–95.

Cuppens-boulahia, N., Garcia-alfaro, J., Martinez, S., & Cabot, J. (2013). Management of stateful firewall misconfiguration, *9*. https://doi.org/10.1016/j.cose.2013.01.004

Dash, R. kumari. (2013). Selection Of The Best Classifier From Different Datasets Using WEKA. *International Journal of Engineering Research & Technology (IJERT)*, *2*(2), 1–5.

Garg, T., & Khurana, S. S. (2014). Comparison of classification techniques for intrusion detection dataset using WEKA. *International Conference on Recent Advances and Innovations in Engineering, ICRAIE 2014*.

Guruvayur, S. R. (2017). a Detailed Study on Machine Learning Techniques for Data, 1187–1192.

Kadam, P. R., & Bhusari, V. K. (2014). REVIEW ON REDUNDANCY REMOVAL OF RULES FOR OPTIMIZING FIREWALL, 397–401.

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications - A decade review from 2000 to 2011. *Expert Systems with Applications*, *39*(12), 11303–11311.

Liu, A. X. (2009). Firewall policy verification and troubleshooting. *Computer Networks*, *53*(16), 2800–2809.

Liu, A. X., Chen, F., & Member, S. (2011). of Firewall Policies in Virtual Private Networks, *22*(5), 887–895.

Liu, A. X., & Gouda, M. G. (2010). Complete Redundancy Removal for Packet Classifiers in TCAMs, *21*(4), 424–437.

Masud, M. M., Mustafa, U., & Trabelsi, Z. (2014). A data driven firewall for faster packet filtering. *4th International Conference on Communications and Networking, ComNet 2014 - Proceedings*.

Mustafa, U., Ain, A., & Wood, T. (2013). Firewall Performance Optimization Using Data Mining Techniques. *World*, 934–940.

MyCERT. (2016). MyCERT 1st Quarter 2016 Summary Report. Retrieved June 6, 2018,from https://www.mycert.org.my/en/services/advisories/mycert/2016/main/ de tail/1190/index.html

Nath, B., Bhattacharyya, D. K., & Ghosh, A. (2011). Dimensionality Reduction for Association Rule Mining. *IJIIP: International Journal of Intelligent Information Processing*, *2*(1), 9–21.

Nwulu, N. I. (2017). Evaluation of Machine Learning Classification Algorithms & Missing Data Imputation Techniques, 1–5.

Pinjan, T. S., & Samvatsar, P. M. (2014). Study of Efficient Firewall Packet Filtering and, *2*(Xi), 539–544.

Praveena, M., & Jaiganesh, V. (2017). A Literature Review on Supervised Machine Learning Algorithms and Boosting Process. *International Journal of Computer Applications*, *169*(8), 975–8887.

Sharma, N., & Niranjan, S. (2012). OPTIMIZATION OF WORD SENSE DISAMBIGUATION USING CLUSTERING IN WEKA, *3*(August), 1598–1604.

Sheth, A. C., & Thakker, B. R. A. (2014). Performance Optimization of Network Firewalls by Rulebase Reordering based on Traffic Conditions, *2*, 1–11.

Trabelsi, Z., Zhang, L., Zeidan, S., & Ghoudi, K. (2013). Dynamic traffic awareness statistical model for firewall performance enhancement. *Computers & Security*, *39*, 160–172.

Ucar, E., & Ozhan, E. (2017). The Analysis of Firewall Policy Through Machine Learning and Data Mining. *Wireless Personal Communications*, *96*(2), 2891–2909.

Urvashi, M., & Jain, M. A. (2015). A survey of IDS classification using KDD CUP 99 dataset in WEKA. *International Journal of Scientific & Engineering Research*, *6*(11), 947–954.

Verbruggen, R. (2014). Creating firewall rules with machine learning techniques, 1–53. Retrieved from http://www.ru.nl/oii/onderwijs/afstuderen/vm/reports/