Received: February 4, 2021. Revised: March 9, 2021. 294

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Decision Tree for Multiclass Classification of Firewall Access

Hayder Naser Khraibet AL-Behadili[1]*

[1]*Computer Science Department, Shatt Alarab University College, Iraq*
* Corresponding author's Email: hayderkhraibet@sa-uc.edu.iq

Abstract: Internet usage is increasing rapidly worldwide, allowing numerous connected computer objects or devices to run and communicate with mass digital information. As Internet usage becomes pervasive, attacks against them are also rising aiming to penetrate the target network and remain undiscovered. Therefore, analyzing the behavior of Internet traffic manually is not possible due to its complexity and the large number of user activity. Incoming and outgoing Internet traffic are controlled using a firewall through an automated Internet security system using a predefined set of rules. Machine learning algorithms are used for Repeated Stemanalysis of the activities on firewall devices and to control traffic on the basis of the results. However, the output models (i.e., classification models) lack explanatory power insight into the relative influence of the main factors in the classification and thus have low accuracy. In this study, a decision tree classification algorithm with a tree-structured model is used for firewall activity analysis, which produces high classification accuracy. Empirical results on firewall access with different actions were compared against six benchmark classification algorithms, namely, SVM, OneR, ANN, Multi class classifier, PSO and ZeroR, in five popular evaluation metrics. The experimental results show that the performance of the proposed classifier in all evaluation metrics is higher than the state-of-the-art classification algorithms, such as SVM, ANN, Multi class classifier, PSO, and the most related classification algorithms that provide comprehensible models (i.e., OneR and ZeroR). The proposed classifier offers interpretation ability by presenting the classification model into a tree representation, which is a further advantage.

Keywords: Data mining, Firewall, Knowledge discovery, Machine learning, Network security.

1. Introduction

Network systems are the software and hardware elements used for network transactions, such as sending and receiving emails, making reservations, reading news, storing or sharing personal documents, shopping, and education [1]. In computing, network security is crucial to protect the critical information of the entire network and users [2]. A firewall is a network system that controls and monitors the incoming and outgoing network traffic in accordance with security rules and regulations [3, 4]. A firewall acts as the gate designed to detect unauthorized access from an untrusted network [5].

Machine learning algorithms have become powerful tools for finding patterns in data [6, 7]. Machine learning algorithms have been applied in various real world problems such as classification [8, 9], clustering [10, 11], medical diagnosis [12], and anomaly detection [13]. Firewall systems produce clever patterns that traditional systems cannot efficiently generate, and thus firewalls based on machine learning may outperform the classical network systems in preventing unauthorized Internet access [14].

Firewall log activities require analysis for further protection and damage assessment. The analysis can determine exactly what is allowed, dropped, and denied, but is neither easy nor straightforward due to work with large raw data (log files) collected from the Internet Access Management in different periods of time. Combining and collecting data from various sources in different periods of time, firewall log files become high volume and the systems used are incapable of handling extensive data. Thus, analysis of firewall log activities requires a powerful artificial

intelligence tool. To classify the log files, Breier and Branišová [15] proposed different anomaly detection methods on the basis of data mining techniques, including K-nearest neighbors (KNN), Decision Table, HyperPipes, and Naive Bayes. Abnormality in firewall rules can be automatically discovered by using a large amount of firewall log instances with machine learning algorithms. The experiment results show that KNN has the best performance. Fatih and Mustafa [14] used support vector machine (SVM) classification algorithm to classify firewall activities into three activation functions (polynomial, Radial Basis Function, Linear and Sigmoid), but the proposed technique obtained minimal improvement. Polpinij and Namee [16] used Generalized Sequential Pattern (GSP) algorithm to study the user behaviors in networks or the Internet using firewall event logs. The data were collected from an organization in Thailand that contains log events from September 20, 2015 to September 30, 2015 and October 15, 2016 to November 10, 2016. The classification result was promising, but several errors remained. Jakub and Branisova [17] proposed an approach for anomaly detection in log records data using Apache Hadoop framework together with Java implementation for classification. New types of breaches can be obtained with less than 10% error rates. Ussath in [18] used artificial neural network (ANN) to identify suspicious actions on the basis of user log-on and log-off activities behavior. Three different datasets on contextual user activities were utilized to evaluate the performance of this classification, and the proposed model achieved an accuracy of up to 98%. Allagi and Rachh [19] presented K-means and Self-organizing feature map method to detect anomalies through analysis of network log data. The proposed method was tested across a log dataset and the classification result obtained 97.2% accuracy.

The above-mentioned classification algorithms have drawbacks. For instance, the SVM is highly sensitive to discrete and noise data, requires validation, must be practiced to determine a suitable kernel function, and its kernel function is sensitive to the number of attributes [20]. In addition, ANN classifier requires numerous data cases for learning, is time consuming, and has difficulty in determining the number of necessary layers and neurons [21]. KNN requires wide computational value and will be affected by the large number of neighbors, compared with unlabelled instances. Irrelevant or anomaly attributes may be present in the data, thus, the KNN assigns an equal weight to each attribute [22]. Consequently, these drawbacks can produce very poor classification performance.

Furthermore, conventional algorithms (i.e., SVM, ANN, KNN) used to detect unauthorized access produce incomprehensible, complex, and highly difficult to understand classification models. These drawbacks prevent usage of these classifiers to construct rules and regulations for further protection. Meanwhile, recent research directions have proposed a new objective, which is the understandability of the classification model to detect unauthorized access by the firewall.

Therefore, a Decision Tree (DT) algorithm is proposed as a classification algorithm to analyze and construct a novel model to detect the type of firewall access by focusing on log files. A total of 65,532 logs activities are taken from the firewall with control connection activities to the Internet on the basis of the rules. Therefore, each firewall action (i.e., allow, drop, deny, and reset-both) is used for classification. Through the logs, firewall activity can provide all the necessary requirements to classify actions.

The remainder of this paper is structured as follows. Section 2 explains how this DT algorithm is applied for access authentication. Section 3 describes the research method and techniques. Section 4 presents the experimental results of the paper. Section 6 discusses the conclusions and future directions.

2. The proposed method

In this section, the classification model is presented on the basis of generalizing powerful decision trees. DT algorithm is considered one of the most popular machine learning algorithms for transparency (interpretable) [23]. DT uses a separate-and-conquer method to construct the classification model. Quinlan [24] described the algorithm as constructing a tree upside down from *D* instances in the data, with roots at the top. The *D* instances of firewall log activities are the "leaves" associated with the most frequent class in *D*. The DT generates a feature list and attributes for each feature.

Example: Feature List: Destination Port, NAT Destination Port, Bytes Received, NAT Source Port, Source Port, Bytes Sent, Bytes, Packets, Elapsed Time, pkts_received, and pkts_sent. Attributes for Elapsed Time are Elapsed Time>0 and Elapsed Time <=0.

Decision tree uses gain ratio as an information-based measurement that considers various numbers (and different probabilities) of test outcomes of attributes for each feature. Therefore, the maximum gain ratio is found among all features and assigned to the root node. Let *C* indicate the number of classes (i.e., allow, drop, deny, and reset-both) and *p(D, j)*

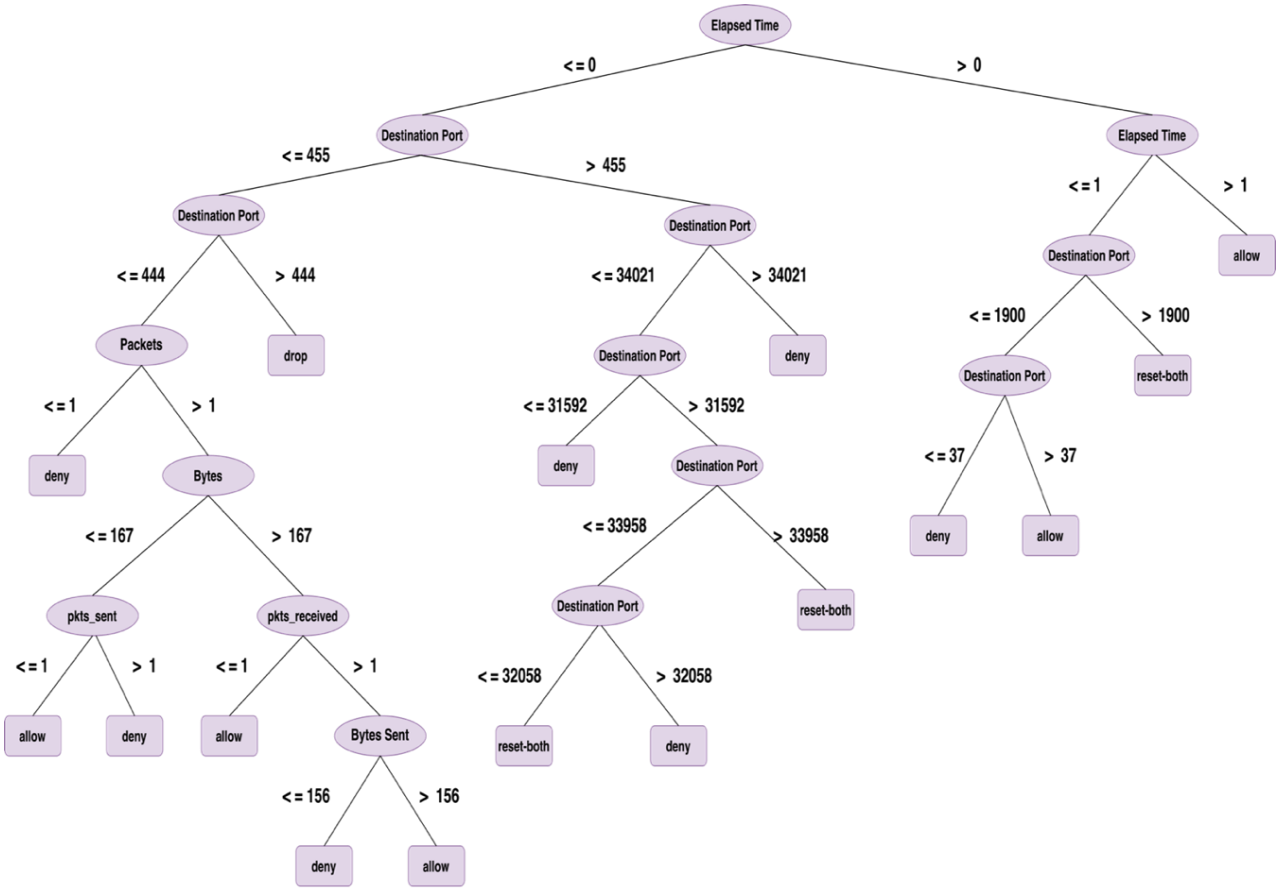

Figure. 1 Learning a decision tree for firewall access

denote the ratio of D instances in the data that are associated with the *jth* class. The uncertainty measures the class to which D instances belong and is calculated as follows:

$$Info\ (D) = -\sum_{j=1}^{C} p(D,j)\ log_2\ (p(D,j)) \qquad (1)$$

where:

- *Info (D)* is the current set of instances for which entropy is computed.
- *j* is the set of classes in the data, {allow, drop, deny, and reset-both}.
- the ratio of elements in class *j* to the number of element in the instances list *D*.

The discrete attribute has one constant value *A*=" " whereas the continuous attribute has two outcomes *A*<=*t* and *A*>*t* (i.e., *Elapsed Time*>*0*, and *Elapsed Time* <=*0*). DT determines the best criterion where the class changes using info gain. The identical information gained by test T with k consequence is computed as follows:

$$Info\ Gain\ (D,T) = Info(D) - \sum_{i=1}^{k} \frac{|Di|}{|D|}\ Info(Di) \qquad (2)$$

where:

- the entropy set for *Info (D)*.
- *Di* is the subsets created from the splitting set of *D* by the attribute *Di*.
- *i* is the value of the feature.
- the entropy set for *Info (Di)*.

The info gain is highly affected by the number of consequence and its maximum when each subset D has only one single instance. Moreover, the i information achieves decreased instances by dividing a group of instances according to the known subset *D*. The split information is measured as follows:

$$Split\ (D,T) = -\sum_{i=1}^{k} \frac{|Di|}{|D|}\ log_2\left(\frac{|Di|}{|D|}\right) \qquad (3)$$

where

- *Di* is the subsets created from splitting set of *D* by the attribute *Di*.
- *i* is the value of the feature.
- *T* is the specific value for specific attribute.

The split information increases with the number of consequences of the test. In addition, a criterion tests the period of the split information. The gain ratio for every split is tested to determine and select its maximum. In several cases, every data split *D* has the

same distribution among classes. Thus, all splits have zero gain, which DT algorithm employs as an extra stopping criterion.

Finally, the recursive partition strategy is applied to handle noisy data when attribute values are incorrectly assigned and instances are incorrectly classified. The DT algorithm prunes the initial tree by determining the parts with less predictive accuracy and changes a "leaf." Fig. 1 above shows the learning decision tree for the firewall access.

3. Research method

This section presents the research methodology by analyzing the data collected from firewall log activities using data mining techniques suitable for processing such data. Data mining incorporates collection, extraction, evaluation, and insights of the most important information [25, 26]. Data mining finds rules to the different access with network security system, and thus allows the firewall to make proactive, knowledge-driven decisions. Data mining (see Fig. 2) involves the following phases:

3.1 Application domain identification

Study of the log files in the firewall devices is extremely important for monitoring the Internet traffic on the basis the results. In this study, data were collected from the firewall device Palo Alto 5020 at Firat University. The firewall log activities comprise 65,532 instances collected as a result of approximately 30 seconds of access. This data has 11 main features and four main classes. Tables 1 and 2 display the data and class descriptions, respectively.

3.2 Data Pre-processing

Considered one of the important steps in data mining, data pre-processing involves cleansing, editing, reduction, and wrangling raw data into an understandable format. In this study, the data were pre-processed by Fatih and Mustafa [14] and saved with an Attribute-Relation File (i.e., arff format), which describes a list of firewall log activities.

3.3 Data mining

The first step is to identify the suitable algorithm that is widely used in data mining. In addition, the problem is not only to decide which algorithm to use, but also to determine which classification model is needed and which data types and heuristic criteria are suitable for such data. Therefore, this research presents a novel DT classification model (C4.5) for a multi-class firewall access dataset. The DT represents a very comprehensible model (i.e., tree-structured) that comprises a set of nodes and edges [23]. Equally important, the C4.5 algorithm uses two heuristic functions: information gain and the gain ratio that such gain using the information provided by test outcomes [27]. C4.5 can deal with numeric and nominal data and with a multi-dimensional dataset

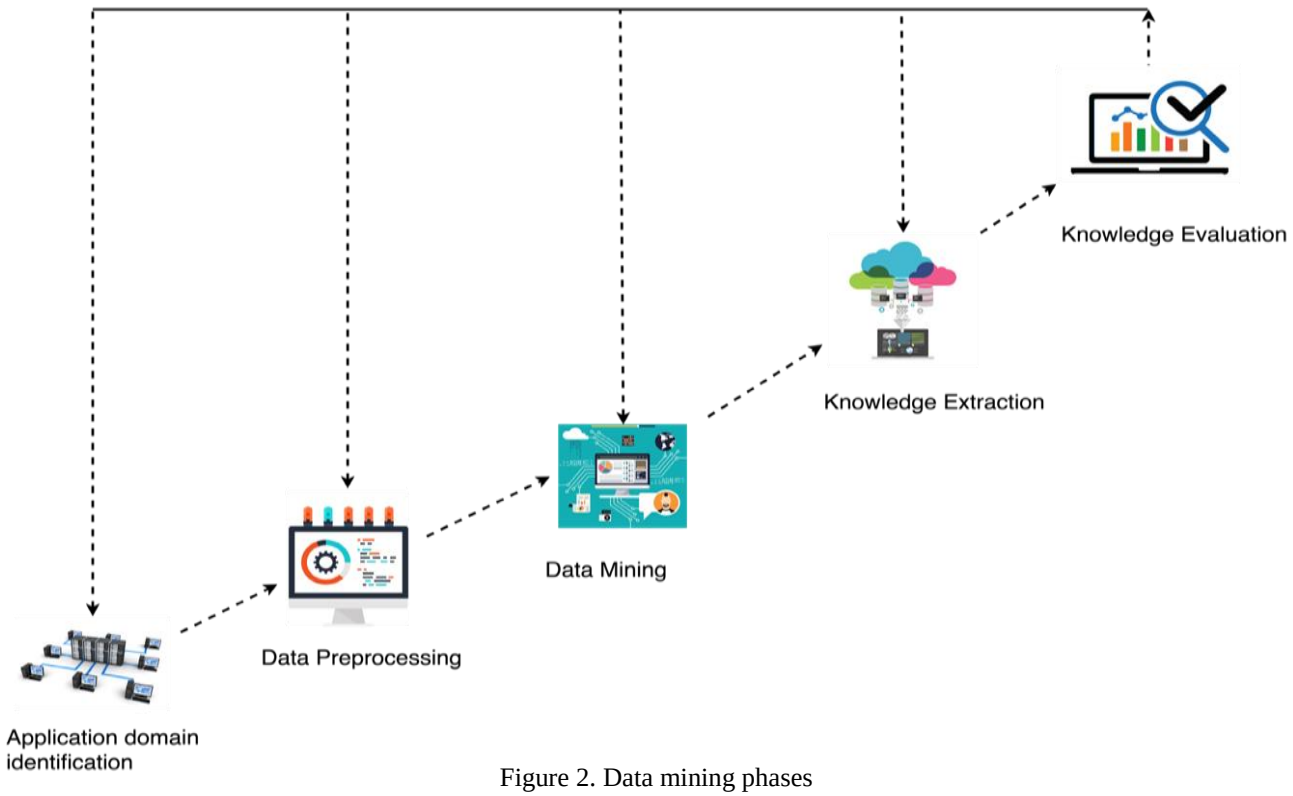

Figure 2. Data mining phases

Table 1. Descriptions of firewall log activities dataset

NO	Feature name	Feature type	Description
1	Source Port	Continuous attribute	Represents the source port of the client
2	Destination Port	Continuous attribute	Represents the destination port of the client
3	NAT Source Port	Continuous attribute	Represents the translation source port of the network address
4	NAT Destination Port	Continuous attribute	Represents the translation destination port of the network address
5	Bytes	Continuous attribute	Represents the number of bytes
6	Bytes Sent	Continuous attribute	Represents the number of bytes sent
7	Bytes Received	Continuous attribute	Represents the number of bytes received
8	Packets	Continuous attribute	Represents the packets number
9	Elapsed Time	Continuous attribute	Time elapsed for flow
10	pkts_sent	Continuous attribute	Represents the number of packets sent
11	pkts_received	Continuous attribute	Represents the number of packets received
12	Action	Discrete attribute	Firewall actions

Table 2. Descriptions of firewall log activities classes

NO	Action name	Description	Instance Number
1	Allow	Authorize internet traffic	37640
2	Drop	Use the default deny action procedure by blocking traffic for the application that is being denied	12851
3	Deny	Use the drops traffic procedure by silently resetting TCP and does not send the host/application	14987
4	Reset-both	Reset TCP and send information to client and server devices	54

class [28]. Finally, C4.5 algorithm is considered one of the top 10 algorithms used in data mining [29].

3.4 Knowledge extraction

This step is responsible for extracting knowledge from raw data source. The resulting knowledge must be in an interpretable and readable manner and must produce information that facilitates inferencing. This study extracted the valuable information from firewall log data and is intended to be a guide for organizations seeking the most appropriate rules for the Internet via firewall access.

3.5 Knowledge evaluation

This section presents the evaluation for the classification performance, which is usually assessed using performance metrics from information retrieval. Such common metrics include classification accuracy, classification error, Kappa Statistic, F-measure, and Mean absolute error [30 –33], presented in Eqs. (4)–(8), respectively.

$$Classification\ accuracy = \frac{TP + TN}{TP+FN+FP+TN}, \quad (4)$$

$$Classification\ error = \frac{FP+ FN}{TP+FN+FP+TN}, \quad (5)$$

$$Kappa\ Statistic = \frac{Accuracy-randomAccuracy}{1-randomAccuracy},$$

$$RandomAccuracy = \frac{(TN+FP)(TN+FN)+(FN+TP)(FP+TP)}{Total\ .\ Total}, \quad (6)$$

$$F\text{-}Measure = \frac{2(Precision\ .\ Recall)}{(Precision+ Recall)}, \quad (7)$$

$$Mean\ absolute\ error = \frac{\sum_{i=1}^{n}\left|y_{est,i} - y_i\right|}{n}, \quad (8)$$

where the confusion matrices that used to describe the classification results are *FN*, *FP*, *TN*, and *TP*.

- *TP* and *TN* are the instances correctly classified for data classes.
- *FP* and *FN* are the instances falsely classified the data classes.

Table 3. Performance of the classification algorithms

Classifier	Classification accuracy	Classification Error
DT	**99.839**	**0.160**
SVM	97.474	2.525
OneR	97.326	2.673
ANN	98.712	1.287
Multi class classifier	99.064	0.935
PSO	97.021	2.979
ZeroR	57.437	42.562

4. Results and discussion

DT algorithm is widely used in many real world datasets, corresponding to multiclass and binary application domains. Training, testing, and validation sets are determined by using ten runs where the data are randomly split to ignore any instances with unusually bad or good testing or training sets.

To compare the performance of DT, a benchmarking study is implemented that includes state-of-the-art classification algorithms, namely, SVM, ANN, and Multi class classifier [34, 35], and the most related classification algorithms that provide comprehensible models (i.e., OneR and ZeroR) [36, 37]. The SVM, ANN, and Multi class classifier versions, PSO, are performed by the author in Matlab®. Meanwhile, the popular OneR and ZeroR classifier is implemented by the Weka workbench.

Furthermore, the SVM is a supervised learning model in machine learning used for regression, classification. This model is considered one of the most robust classification algorithms, being based on statistical learning frameworks. Meanwhile, the ANN is a computing system inspired by human brain processes and analyzes the information. ANN has a self-learning ability to enable the understanding of a broad data to produce better results. Multi class classifier is a popular classifier used for multi-class datasets.

The particle swarm optimization (PSO) is considered a swarm-based classification algorithm. It is designed to represent the simulation of bird flocking graphically. This algorithm classifies the data on the basis of the simulation of birds searching through their environment to evade predators or to

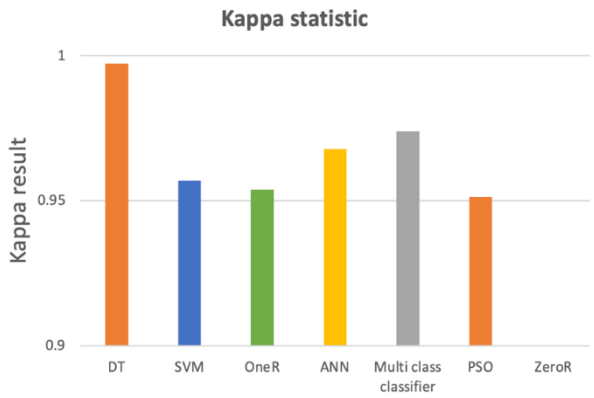

Figure. 3 Kappa statistic result

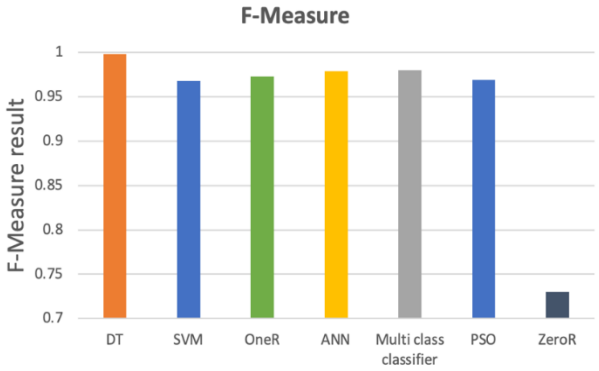

Figure. 4 F-Measure result

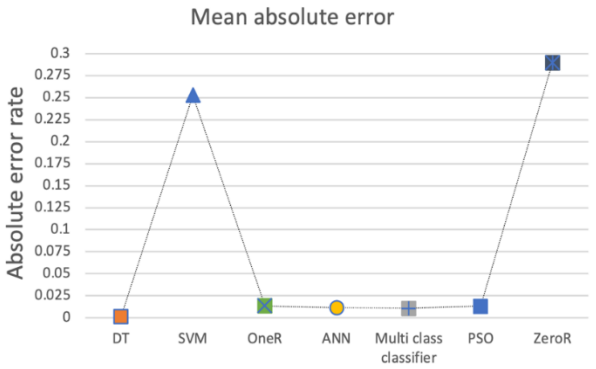

Figure. 5 Mean absolute result

seek food. Each particle in the swarm has its specific velocity and location which represent a solution in the search space. Thus, the birds will be guided by the best classification patterns found by all neighboring particles. Therefore, the PSO will converge on the best classification patterns for each data and produce the classification model.

Different from other complex machine learning models, OneR and ZeroR is simple and an accurate classification algorithm. The OneR or "One Rule" produces one classification rule for each predictor in the data, then selects the best one with high classification accuracy. Meanwhile, ZeroR is the simplest rule-based classification algorithm, which

relies on the target classes and eliminates all predictors. ZeroR classifier simply predicts the most frequent value of class on the basis of the frequency table. Thus, more specifically, it does not contain any rule that elaborates on the non-target (class) attributes.

The evaluation results of the DT algorithm for firewall access are determined using five benchmark evaluation metrics, namely, classification accuracy, classification error, Kappa statistic, F-Measure, and Mean absolute error. The first observation from our findings show the effectiveness of the proposed classifier compared with state-of-the-art classification algorithms, swarm-based algorithm, and the most related classification algorithms. The firewall comprises 65,532 log activities divided into training and testing using l0-fold cross validation methods. The results of the application on the multiclass classification problems are summarized in Table 3 and Figs. 3–5 that are used to determine the best classifier.

The classification results in Table 3 show that the DT is better than all other classifiers, namely, SVM, OneR, ANN, Multi class classifier, PSO, and ZeroR. The proposed classifier can produce high classification accuracy with 99.839 and 0.160 error rates. The Multi class classifier follows with the second best result.

Figs. 3–5 show the results of the Kappa statistic, F-Measure, and Mean absolute error, respectively. In Kappa statistic and F-Measure, the highest result indicates a good algorithm performance. Therefore, the proposed classifier dominates the other six classifiers in Kappa statistic metrics with a 0.9972 Kappa statistic result. The Multi class classifier follows the second best result with 0.9739 Kappa statistic. Fig. 4 shows that DT achieves a performance better than all other classification algorithms across firewall activity detections. The proposed classifier can produce high F-Measure accuracy with 0.998 result. Fig. 5 proves that the results obtained by the proposed classifier are better than the other classifiers when considering the Mean absolute error with 0.0014 absolute error rates. The lowest error indicates a good algorithm performance.

The reason is that the DT can determine the best performance in all evaluation metrics: DT creates understandable classification model (i.e., tree model), DT can manage categorical and continuous data attributes. DT implements classification without requiring much computation time and can remarkably provide a presentation of which attributes are most important for classification the target class for the data.

5. Conclusion

The amount of data exchanges in the Internet is rapidly increasing. Incoming and outgoing network traffic from such large data are controlled by applying the DT classification algorithm to the Internet security system. To classify firewall logs access, a drastic experimental analysis was carried on real data collected from the firewall device. The performed results are compared to check the performance of the DT algorithm. The efficiency of the DT algorithm determines the best classification accuracy with the utmost competence, in comparison with six states of art machine learning algorithms, namely, SVM, ANN, Multi class classifier, PSO, OneR, and ZeroR. The performance evaluation is also conducted with another common metrics for evaluation, including classification error, Kappa Statistic, F-measure, and Mean absolute error. The outcomes show that DT is superior in all performance metrics with different types of firewall activities. In addition, the model allows decision makers to act early to develop the trust firewall system on the basis of discovered intelligent rules. The limitation of this study is the limited size of instances in the collected data. As a future work, this study aims to expand the instances by adding more data from different sources of Internet devices and environments to deeply analyze and detect unauthorized access from untrusted networks and increase the accuracy of classification.

Conflicts of Interest

The author declares no conflict of interest.

Author Contributions

For this research article all contributions (e.g., conceptualization, research methodology, programming, results, and validation) have been completed by the main author (“Hayder”).

Acknowledgments

The author would like to thank the Department of Computer Science, Shatt Alarab University College, for financially supporting this research.

References

[1] B. Khan, M. Mahmud, M. K. Khan, and K. S. Alghathbar, “Security analysis of firewall rule sets in computer networks”, In: *Proc. of Fourth International Conference on Emerging Security Information, Systems and Technologies*, pp. 51–56, 2010.

[2] N. Gupta, V. Naik, and S. Sengupta, "A firewall for Internet of Things", In: *Proc. of 2017 9th International Conference on Communication Systems and Networks*, No. January, pp. 411–412, 2017.

[3] A. Voronkov, L. H. Iwaya, L. A. Martucci, and S. Lindskog, "Systematic literature review on usability of firewall configuration", *ACM Comput. Surv.*, Vol. 50, No. 6, 2017.

[4] K. Yao, H. Li, W. Shang, and A. E. Hassan, "A study of the performance of general compressors on log files", *Empir. Softw. Eng.*, Vol. 25, No. 5, pp. 3043–3085, 2020.

[5] I. Phillips, "Assessing risk associated with firewall rules", *AT&T Glob. Netw. Serv. UK BV, assignee.*, Vol. 2, pp. 1–34, 2018.

[6] A. K. Tiwari, "Introduction to machine learning", *Ubiquitous Machine Learning and Its Applications*. 2017.

[7] Y. Tu, "Machine learning", In: *Proc. of EEG Signal Processing and Feature Extraction*, 2019.

[8] H. N. K. Al-behadili, K. R. Ku-Mahamud, and R. Sagban, "Hybrid Ant Colony Optimization and Iterated Local Search for Rules-Based Classification", *J. Theor. Appl. Inf. Technol.*, Vol. 98, No. 4, pp. 657–671, 2020.

[9] H. N. K. Al-Behadili, R. Sagban, and K. R. Ku-Mahamud, "Adaptive Parameter Control Strategy for Ant-Miner Classification Algorithm", *Indones. J. Electr. Eng. Informatics*, Vol. 8, No. 1, pp. 149–162, 2020.

[10] A. M. Jabbar, K. R. Ku-Mahamud, and R. Sagban, "An improved ACS algorithm for data clustering", *Indones. J. Electr. Eng. Comput. Sci.*, Vol. 17, No. 3, pp. 1506–1515, 2020.

[11] A. M. Jabbar, K. R. Ku-Mahamud, and R. Sagban, "Modified ACS Centroid Memory for Data Clustering", *J. Comput. Sci.*, Vol. 15, No. 10, pp. 1439–1449, 2019.

[12] J. Wahid and H. F. Al-Mazini, "Classification of Cervical Cancer Using Ant-Miner for Medical Expertise Knowledge Management", In: *Proc. of Knowledge Management International Conference (KMICe)*, 2018.

[13] H. Almazini and K. R. Ku-Mahamud, "Grey Wolf Optimization Parameter Control for Feature Selection in Anomaly Detection", *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 2, pp. 474–483, 2021.

[14] F. Ertam and M. Kaya, "Classification of firewall log files with multiclass support vector machine", In: *Proc. of 6th Int. Symp. Digit. Forensic Secur.* ISDFS 2018 - Proceeding, No. July, pp. 1–4, 2018.

[15] E. Ucar and E. Ozhan, "The Analysis of Firewall Policy Through Machine Learning and Data Mining", *Wirel. Pers. Commun.*, Vol. 96, No. 2, pp. 2891–2909, 2017.

[16] J. Polpinij and K. Namee, "Internet usage patterns mining from firewall event logs", In: *Proc. of ACM Int. Conf. Proceeding Ser.*, pp. 93–97, 2019.

[17] J. Breier and B. Jana, "A Dynamic Rule Creation Based Anomaly Detection Method for Identifying Security Breaches in Log Records", *Wirel. Pers. Commun.*, Vol. 94, No. 3, pp. 497–511, 2017.

[18] M. Ussath, D. Jaeger, F. Cheng, and C. Meinel, "Identifying Suspicious User Behavior with Neural Networks", In: *Proc. of 4th IEEE Int. Conf. Cyber Secur. Cloud Comput. CSCloud 2017 3rd IEEE Int. Conf. Scalable Smart Cloud*, SSC 2017, pp. 255–263, 2017.

[19] S. Allagi and R. Rachh, "Analysis of Network log data using Machine Learning", In: *Proc. of 2019 IEEE 5th Int. Conf. Converg. Technol.* I2CT 2019, pp. 2019–2021, 2019.

[20] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", *Informatica*, Vol. 31, pp. 249–268, 2007.

[21] R. Kumar, R. K. Aggarwal, and J. D. Sharma, "Comparison of regression and artificial neural network models for estimation of global solar radiation", *Renew. Sustain. Energy Rev.*, Vol. 52, No. August, pp. 1294–1299, 2015.

[22] T. Phyu, "Survey of Classification Techniques in Data Mining", In: *Proc. of Int. MultiConference Eng. Comput. Sci.*, Vol. I, pp. 18–20, 2009.

[23] X. Hu, C. Rudin, and M. Seltzer, "Optimal sparse decision trees", In: *Proc. of 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019, Vol. 32, No. NeurIPS, pp. 1–9.

[24] J. R. Quinlan, "Improved Use of Continuous Attributes in C4.5", *J. Artificial Intell. Res.*, Vol. 4, No. 1, pp. 77–90, 1996.

[25] R. Sowmya and K. R. Suneetha, "Data Mining with Big Data", In: *Proc. of 2017 11th International Conference on Intelligent Systems and Control*, ISCO 2017, 2017.

[26] I. Witten, F. Eibe, H. Mark, and P. Christopher, "Data Mining: Practical Machine Learning Tools and Techniques", Vol. 40, No. 6. Elsevier, 2016.

[27] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree

ID3 and C4.5", *Int. J. Adv. Comput. Sci. Appl*., No. 2, pp. 13–19, 2014.
[28] R. Alsagheer, A. Alharan, and A. Al-Haboobi, "Popular Decision Tree Algorithms of Data Mining Techniques: A Review", *Int. J. Comput. Sci. Mob. Comput*., Vol. 6, No. 6, pp. 133–142, 2017.
[29] X. Wu, "Top 10 algorithms in data mining", *Knowl. Inf. Syst*., Vol. 14, No. 1, pp. 1–37, 2008.
[30] G. Canbek, T. Temizel, S. Sagiroglu, and N. Baykal, "Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights", In: *Proc. of 2nd International Conference on Computer Science and Engineering, UBMK 2017*, pp. 821–826, 2017.
[31] A. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology", *arXiv Prepr*. arXiv1809.03006, pp. 1–37, 2018.
[32] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Vol. 12. 2006.
[33] S. Saravi, R. Kalawsky, D. Joannou, M. R. Casado, G. Fu, and F. Meng, "Use of artificial intelligence to improve resilience and preparedness against adverse flood events", *Water (Switzerland)*, Vol. 11, No. 5, pp. 1–16, 2019.
[34] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms", *Expert Syst. Appl*., Vol. 82, pp. 128–150, 2017.
[35] F. Luo, W. Guo, Y. Yu, and G. Chen, "A multi-label classification algorithm based on kernel extreme learning machine", *Neurocomputing*, Vol. 260, pp. 313–320, 2017.
[36] K. Rajasekaran, P. Jayasheelan, S. Preethaa, "Predictive Analysis in Agriculture to Improve the Crop Productivity using ZeroR algorithm", *Int. J. Comput. Sci. Eng. Commun*., Vol. 4, No. 2, pp. 1397–1401, 2016.
[37] A. Gupta, A. Mohammad, A. Syed, and M. N., "A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA", *Int. J. Adv. Comput. Sci. Appl*., Vol. 7, No. 7, pp. 374–381, 2016.