A Comparison of Three Machine Learning Algorithms in the Classification of Network Intrusion

Amir Zulhilmi[1], Salama A. Mostafa[1](✉), Bashar Ahmed Khalaf[2], Aida Mustapha[1], and Siti Solehah Tenah[3]

[1] Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia, 86400 Johor, Malaysia
amirzulhilmi1998@gmail.com, {salama,aidam}@uthm.edu.my
[2] College of Basic Education, University of Diyala, 32001 Diyala, Iraq
basharalzubaidy60@gmail.com
[3] Research Management Centre, University Tun Hussein Onn Malaysia, 86400 Johor, Malaysia
solehah@uthm.edu.my

Abstract. Intrusion Detection Systems (IDS) effort to detect intrusion and misuse attack computer systems by assembling and examining data of computer networks. The IDS is usually examining huge traffic data based on Machine Learning (ML) algorithms to identify harmful changes or attacks, however, which algorithm can manifest the best performance is an issue to be investigated. ML-IDS requires to decrease false alarm and increase true alarm rates. In this work, three tree-based ML algorithms which are Decision Tree (DT), Decision Jungle (DJ), and Decision Forest (DF) have been tested and evaluated in an IDS model. The main objective of this work is to compare the performance of the three algorithms based on accuracy, precision and recall evaluation criteria. The Knowledge Discovery in Databases (KDD) methodology and Kaggle intrusion detection dataset are used in the testing. The results show that the DF achieves the highest overall accuracy of 99.83%, the DJ achieves the second highest overall accuracy of 99.74% and the DT achieves the lowest overall accuracy of 95.59%. The obtained results can serve as a benchmark in the evaluation of advanced IDS.

Keywords: Intrusion Detection Systems (IDS) · Decision Tree (DT) · Decision Jungle (DJ) · Decision Forest (DF)

1 Introduction

Intrusion is a serious issue in the security and a prime issue of the security break. It is because a solitary example of interruption can take or erase the information from computer machines and system framework in almost no time. An interruption can make additional harm to the framework and related equipment. Besides, the interruption can cause tremendous loses of the monetarily and bargain the information technology basic foundation, in this way prompting data inadequacy in cyberwar [1]. In this manner,

M. Anbar et al. (Eds.): ACeS 2020, CCIS 1347, pp. 313–324, 2021.
https://doi.org/10.1007/978-981-33-6835-4_21

an interruption recognition framework is imperative to stay away from interruption. Subsequently, an Intrusion Detection System (IDS) is proposed to organize traffic that is utilized for dubious activities. A few IDS are equipped for making a move when bizarre traffic or vindictive action is recognized, including blocking traffic sent from a dubious IP address while abnormality discovery and revealing is the essential capacity. Even though IDS screen arranges for potential vindictive action that has been recognized, they are additionally inclined to bogus cautions (bogus positive). Throughout the most recent decade, there has been expanding altogether the measure of the system assault. These assaults have been enormously serious and complex in nature [2]. There are numerous programmer tests and assault computer machines. To make a guard of these different digital assaults and computer machines infection, there are bunches of computer security procedure that have been concentrated in the most recent decade. As models incorporate considered cryptography, firewalls and interruption identification framework and so on [3].

As of late, an alternate kind of Machine Learning (ML) methods and techniques have been proposed to improve the presentation of interruption recognition frameworks of IDS [4, 5]. The ML methods are a part of computerized reasoning base on exact information like sensor information or database. These methods are notable on account of their capacity in detecting anomalies based on pattern analysis and finding solutions [6]. Some of the ML methods that have been looking at in IDS tasks are SVM [4], Random Forest (RF) [5], software agent [6] and Decision Jungle (DJ) [7]. The ML has a wide scope of uses including web indexes, clinical analysis, text and penmanship acknowledgement, picture screening, load determining, showcasing and deals determination [8–11].

There are many existing components for an intrusion detection system. The significant issue for the difficult articulation is the security and precision of the framework [12]. An interruption discovery framework was made to improve the issue of exactness and the proficiency of the framework each regular characterization approach three calculations are utilized [6]. This exploration is made to know with the calculation is the best to decrease sorts of assault. These standards can decide interruption attributes than to actualize in the firewall strategy administers as anticipation. The mix of IDS and firewall supposed the IPS, with the goal that other than recognizing the presence of interruption additionally can execute by doing preclude from securing interruption as avoidance [1]. The target of this proposition is to introduce a KDD dataset procedure that diminishes IDS cautions and evaluates its danger [13]. To accomplish the point of this work, the accompanying goals will be considered: to apply the data gain proportion calculation to separate the best highlights of IDS alarms to survey the cautions, construct a conglomeration IDS ready technique dependent on three choices tree-based calculations that decrease the measure of bogus positive cautions and diminish the alarms excess and assess the exactness and accuracy of the three calculations utilizing a chose standard dataset [14–17].

Different techniques and methods have been proposed, developed, and evaluated to safeguard internet users against attacks. There are many research studies in IDS including the work of Li, et al. [17] which proposed an interruption recognition framework dependent on Online Sequence Extreme Learning Machine (OS-ELM) is built up, which is accustomed to identifying the assault in AMI and completing the near investigation with different calculations. Reproduction results show that contrasted and other interruption location techniques, interruption discovery strategy dependent on OS-ELM are increasingly predominant in identification speed and precision. Shakya and Kaphle [18], work propose another learning approach towards building up a novel interruption discovery framework (IDS) by backpropagation neural systems (BPN) and self-arranging map (SOM) and analyse the exhibition between them. The principle capacity of Intrusion Detection System is to shield the assets from dangers. It dissects and predicts the practices of clients, and afterwards, these practices will be viewed as an assault or typical conduct. The proposed strategy can fundamentally decrease the preparation time required.

This research is conducted by focusing on the intrusion detection system classification using the popular ML methods which are Decision Tree (DT), Decision Jungle (DJ), and Decision Forest (DF). The characteristics of Kaggle intrusion detection dataset are multivariate, medium sizes (126000 raws and 42 columns) and have some missing values. Among the most important factors to be considered are identifying the categories of illegal activities that lead to intrusions. The ML methods are selected to overcome intrusion problems using the same dataset. This work is segmented into five sections starting with Sect. 1 that represents the Introduction. The literature review has been discussed in Sect. 2. Next, the research methodology is illustrated in Sect. 3. Section 4 shows the testing results. Whereas, Sect. 5 concludes the work and proposes future research.

2 Methods and Materials

This research will use Knowledge Discovery in Database (KDD). KDD is the process of discovering useful knowledge from a collection of data [12]. The experiments were carried out using the Azure Machine Learning tool with 10-fold validation method for training and testing [19]. This method is being used because data is obtained from a dataset. KDD methodology involves seven steps of (1) data cleaning to removal noisy and irrelevant data (2) data integration to combine heterogeneous data of multiple sources (3) data selection to retrieve relevant data from the data collection (4) data transformation to prepare the data in the appropriate form (5) data mining to extract potentially useful patterns (6) pattern evaluation to identify related patterns based on given measures and (7) knowledge representation to represent and visualize results. Figure 1 shows the KDD methodology.

2.1 Testing Dataset

The data that have been used in the research is introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection taken from the Kaggle website [20]. This dataset has 42 attributes and 126000 instances. This data was selected by using Placement.

2.2 Machine Learning Methods

There are three methods that been used in this research which are Decision Tree (DT), Decision Jungle (DJ), and Random Forest (RF) have been discussed in detail. Decision

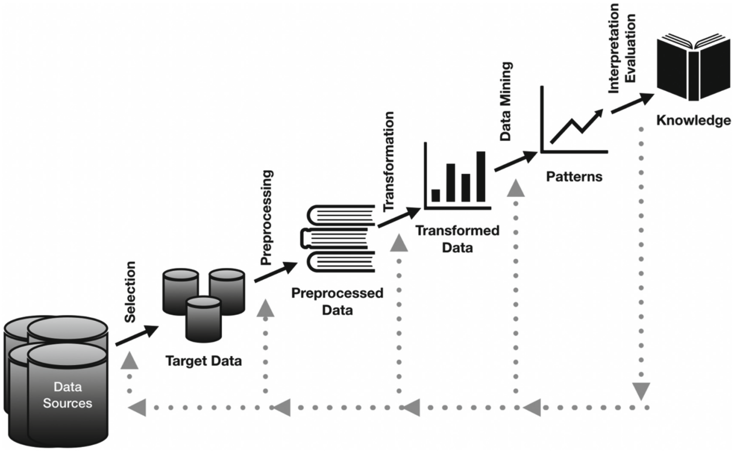


Fig. 1. Knowledge Discovery in Database (KDD) [12]

Tree (DT) is one of the most powerful and simple data mining method that has been employed in IDS. The decision tree is a kind of a tree that consists of branch nodes representing a choice among a number of alternatives, and each leaf nodes representing a class of data [1]. The architecture of the DT is illustrated in Fig. 2 in which TI, T2, T3, and T4 are branch nodes that assign a class number to an input pattern by filtering the pattern down through the tests in the tree. Subsequently, any input patterns can be categorized to class 1, 2, or 3 when the input pattern reaches the leaf nodes [3]. Therefore, the DT is valuable to categorize the data from large datasets.

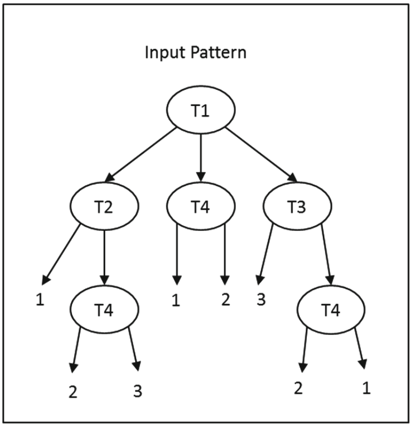


Fig. 2. Decision Tree Architecture [1, 3]

Decision Jungle (DJ) algorithm is a troupe learning strategy for grouping. The calculation works by building different choice trees and afterwards deciding on the most mainstream yield class. The trees that have high expectation certainty have a more noteworthy load in an official conclusion of the group. Furthermore, Choice Jungles are an expansion of Decision Forests [13]. Both create and afterwards total choice trees, yet with Decision Jungles there is the extra alternative of permitting branches to consolidate, bringing about a much-diminished memory impression. Choice Jungles are profoundly adaptable, non-parametric and non-straight, which means they are additionally exceptionally clamoring lenient. A choice wilderness comprises of a group of choice coordinated non-cyclic diagrams (DAGs) [1]. Choice wildernesses are non-parametric models, which can speak to non-direct choice limits. They perform incorporated component determination and characterization and are flexible within the sight of boisterous highlights.

Decision Forest (DF) algorithm is a gathering learning strategy for arrangement. The calculation works by building numerous choice trees and afterwards deciding on the most famous yield class as shown in Fig. 3. The trees that have high expectation certainty have a more noteworthy load in an ultimate conclusion of the outfit. DF is outfit classifiers, which are utilized for characterization and relapse investigation on the interruption discovery information. DF works by making different choice trees in the preparation stage and yield class marks those have the lion's share vote [13]. The DF accomplishes high grouping exactness and can deal with exceptions and clamor in the information. DF is utilized in this work since it is less defenseless to over-fitting and it has recently demonstrated great characterization results.

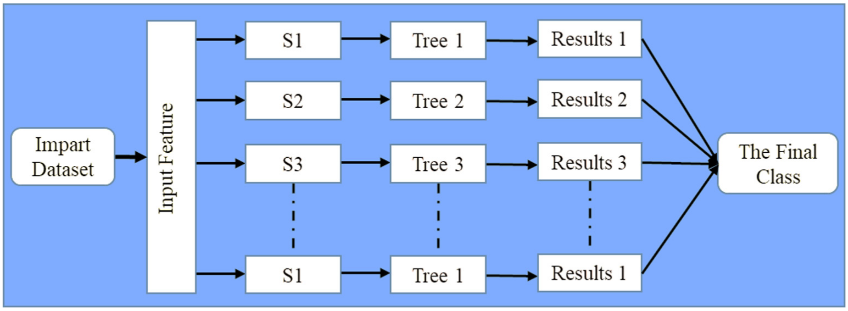


Fig. 3. The architecture of DF for IDS [13]

Figure 3 shows the execution of the irregular timberland grouping model in the information characterization in the proposed framework. A pre-prepared example of n tests is taken care of to the choice backwoods classifier. DF makes n various trees by utilizing a few element subsets. Each tree delivers a grouping result, and the consequence of the order model relies upon the greater part casting a ballot [14]. The example is allocated to the class that gets the most noteworthy democratic scores. The recently achieved characterization results demonstrate that DF is sensibly reasonable in the order of such information on the grounds that now and again, it has acquired preferable outcomes over have different classifiers. Different focal points of the RF incorporate its higher precision than Adaboost and less odds of overfitting.

The DT, DJ and DF consist of several steps for the training and testing phases as shown in Fig. 4. The first step includes importing the dataset, then obtaining the labels. Subsequently, the labels will be checked one by one based on the original dataset features. Furthermore, in the step of traffic analysis, a setting function is employed to analyze and monitor the incoming traffics and set the threshold. Subsequently, the DT, RF, and DJ will analyse the features of the incoming traffics, then, the IDS will forward it to the decision function to determine whether the incoming traffics are attack traffics or not. In case of the incoming traffics have anomalies, the IDS saves the IP address which sends the attack traffic for a permanent block. Whereas in case of the incoming traffics do not have anomalies this means that the traffics identified as normal traffic and pass it to the webserver.

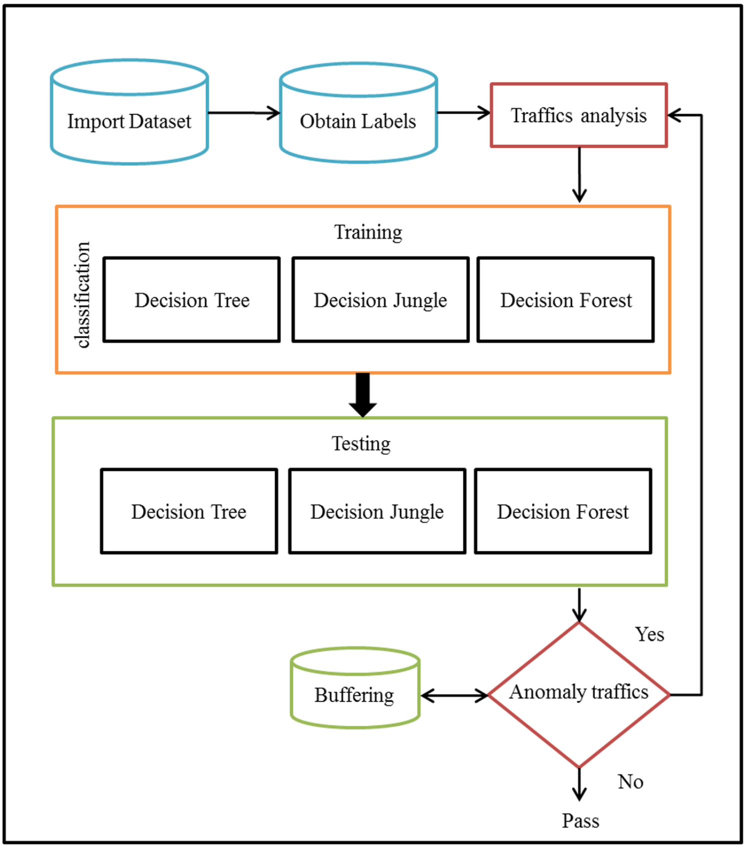


Fig. 4. The architecture of the ML-IDS

2.3 Evaluation Metrics

The evaluation metric includes the following:

- **Micro-average method:** In Micro-average method, you sum up the individual true positives, false positives, and false negatives of the system for different sets and apply them to get the statistics [3, 21].

$$\text{Micro - average of precision } = \frac{TP_1 + TP_2}{TP_1 + TP_2 + FP_1 + FP_2} \tag{1}$$

and,

$$\text{Micro - average of recall } = \frac{TP_1 + TP_2}{TP_1 + TP_2 + FN_1 + FN_2} \tag{2}$$

- **Macro-average Method:** The method is straight forward. Just take the average of the precision and recall of the system on different sets [22, 23].

$$\text{Macro - average precision} = \frac{P_1 + P_2}{2 * 3} \tag{3}$$

and,

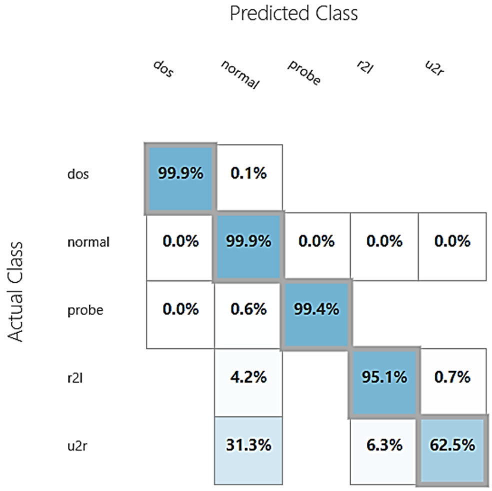
$$\text{Macro - average recall} = \frac{R_1 + R_2}{2} \tag{4}$$

- **Overall accuracy:** Overall Accuracy is essentially told us out of all of the reference sites what proportion were mapped correctly. The overall accuracy is usually expressed as a percent, with 100% accuracy being a perfect classification where all reference sites were classified correctly [19, 24].

$$\text{Overall Accuracy } = \frac{TP + TN}{P + N} \tag{5}$$

3 Results

The IDS prevents hackers from hacking the systems and makes networks secure from the threat of attack include DDoS, Benign, DoS GoldenEye, Heartbleed, DoS Hulk, DoS Slowhttp, DoS slowloris, SSH-Patator, FTP-Patator, Web Attack, Infiltration, Bot and PortScan [1, 24]. The DT, DJ and DF algorithms that are integrated into the IDS help to detect the threats that attack the computer or network systems. The outcome of this research decides the best ML algorithm from the three by comparing the results of them. Intrusion detection performance depends on accuracy as well as decreases false alarm and increases true alarm rates. The evaluation metrics of accuracy, precision and recall are calculated to measure the performance of the algorithms. The testing experiments were carried out on Windows 7 using the Azure ML tool and 10-fold cross-validation. Whereas, the hardware specifications of the implementation and testing are Intel (R) Core (TM) i7-5500U processor, 2.40 GHz, and 16 GB RAM. Subsequently, Fig. 5 gives information about the actual classes and predicted classes of the multiclass confusion matrix of the DJ test.



	dos	normal	probe	r2l	u2r
dos	99.9%	0.1%			
normal	0.0%	99.9%	0.0%	0.0%	0.0%
probe	0.0%	0.6%	99.4%		
r2l		4.2%		95.1%	0.7%
u2r		31.3%		6.3%	62.5%

Fig. 5. The confusion matrix of the DJ

Initially, a data cleaning and multiple testing are performed to ensure that the dataset and the algorithms are ready for the training, testing and evaluation phases. Meanwhile, 10-folds cross-validation is performed to obtain reliable results. Table 1 shows the results of the tests for all the three DT, DJ and DF algorithms in terms of accuracy, precision and recall with the range of the dataset splitting. From the table, we can see that all three algorithms have high performance.

The results show that the DF got a higher overall accuracy of 99.83%, the DJ got the medium overall accuracy of 99.74% and the DT got the lowest accuracy of 95.59%. Moreover, the DF has a higher recall compared to the DT and DJ. However, the DJ has

Table 1. The result of accuracy, precision and recall of DT

Test	Split	Accuracy		Precision		Recall	
		Overall	Average	Micro	Macro	Micro	Macro
DT							
1	90:10	0.95018	0.98007	0.95018	0.83176	0.95018	0.60897
2	80:20	0.95295	0.98118	0.95295	0.81787	0.95295	0.69845
3	70:30	0.95345	0.98138	0.95345	0.81452	0.95345	0.69748
4	60:40	0.95355	0.98142	0.95355	0.81399	0.95355	0.71109
5	50:50	0.95439	0.98176	0.95439	0.83153	0.95439	0.72911
6	40:60	0.99835	0.99934	0.99835	0.92122	0.99835	0.91366
7	30:70	0.95559	0.98224	0.95559	0.84226	0.95559	0.72135
8	20:80	0.95616	0.99935	0.95616	0.86956	0.95616	0.73319
9	10:90	0.95773	0.98309	0.95773	0.90660	0.95773	0.75162
10	66:34	0.95856	0.98343	0.95856	0.91925	0.95856	0.67565
DJ							
1	90:10	0.99523	0.99809	0.99523	0.96954	0.99523	0.81184
2	80:20	0.99659	0.99864	0.99659	0.92928	0.99659	0.86997
3	70:30	0.99643	0.99857	0.99643	0.95278	0.99643	0.84843
4	60:40	0.99661	0.99865	0.99661	0.94899	0.99661	0.85865
5	50:50	0.95439	0.99897	0.99743	0.93313	0.99743	0.87424
6	40:60	0.99724	0.99890	0.99724	0.96688	0.99724	0.86048
7	30:70	0.99701	0.99881	0.99701	0.92195	0.99701	0.84906
8	20:80	0.99717	0.99887	0.99717	0.95519	0.99717	0.84003
9	10:90	0.99694	0.99878	0.99694	0.99329	0.99694	0.85561
10	66:34	0.99754	0.99902	0.99754	0.99828	0.99754	0.98166
DF							
1	90:10	0.99637	0.99855	0.99637	0.97537	0.99637	0.83276
2	80:20	0.99734	0.99894	0.99734	0.90504	0.99734	0.87721
3	70:30	0.99753	0.99901	0.99753	0.87130	0.99753	0.84022
4	60:40	0.99780	0.99912	0.99780	0.89400	0.99780	0.86677
5	50:50	0.99829	0.99931	0.99829	0.89620	0.99829	0.87534
6	40:60	0.99835	0.99934	0.99724	0.92122	0.99835	0.91366
7	30:70	0.99844	0.99881	0.99844	0.91402	0.99844	0.89678
8	20:80	0.99839	0.99935	0.99839	0.92274	0.99839	0.87002
9	10:90	0.99853	0.99941	0.99853	0.94677	0.99853	0.87322
10	66:34	0.99913	0.99965	0.99913	0.99935	0.99913	0.99930

a higher precision compared to the DT and DF. Ultimately, the DF outperforms the DT and DJ as Fig. 6 shows.

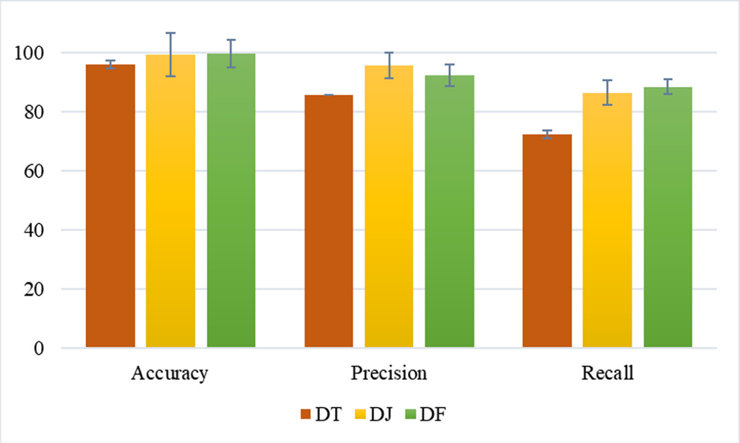


Fig. 6. The overall accuracy, precision and recall of the algorithms

4 Conclusion

This research about the technique that can give the best performance to detect an intrusion in the IDS. It presents an analysis for the detection of intrusion using ML-based classification algorithms for IDS. The algorithms are Decision Tree (DT), Decision Jungle (DJ), and Random Forest (RF). The performance assessment in the IDS models is made based on accuracy precision and recall measurements. The implementation of the models is performed by Azure ML tool. The test results show that the DF has a higher overall accuracy of 99.83%, DJ got the medium overall accuracy of 99.74% and the lowest score is made by the DT with an accuracy of 95.59%. In future research, we plan to explore more attributes along with other data mining classification tasks and platforms.

Acknowledgement. This paper is supported by Universiti Tun Hussein Onn Malaysia.

References

1. Khalaf, B.A., Mostafa, S.A., Mustapha, A., Mohammed, M.A., Abduallah, W.M.: Comprehensive review of artificial intelligence and statistical approaches in distributed denial of service attack and defense methods. IEEE Access **7**, 51691–51713 (2019)
2. Jubair, M.A., et al.: Bat optimized link state routing protocol for energy-aware mobile ad-hoc networks. Symmetry **11**(11), 1409 (2019)
3. Richariya, V., Singh, U.P., Mishra, R.: Distributed approach of intrusion detection system: survey. Int. J. Adv. Comput. Res. **2**(4), 358 (2012)

4. Aburomman, A.A., Reaz, M.B.I.: A novel SVM-kNN-PSO ensemble method for intrusion detection system. Appl. Soft Comput. **38**, 360–372 (2016)
5. Farnaaz, N., Jabbar, M.A.: Random forest modeling for network intrusion detection system. Procedia Comput. Sci. **89**(1), 213–217 (2016)
6. Khalaf, B.A., Mostafa, S.A., Mustapha, A., Abdullah, N.: An adaptive model for detection and prevention of DDoS and flash crowd flooding attacks. In: 2018 International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR), pp. 1–6. IEEE, August 2018
7. Elmasry, W., Akbulut, A., Zaim, A.H.: Empirical study on multiclass classification-based network intrusion detection. Comput. Intell. **35**(4), 919–954 (2019)
8. Ishak, A.M., Mustapha, A., Idrus, S.Z.S., Abd Wahab, M.H., Mostafa, S.A.: Correlation impact by random forest towards prediction of phishing website. In: IOP Conference Series: Materials Science and Engineering, vol. 917, no. 1, p. 012043. IOP Publishing (2020)
9. Razali, N., Mostafa, S.A., Mustapha, A., Abd Wahab, M.H., Ibrahim, N.A.: Risk factors of cervical cancer using classification in data mining. In: Journal of Physics: Conference Series, vol. 1529, no. 2, p. 022102. IOP Publishing, April 2020
10. Rajagopal, S., Hareesha, K.S., Kundapur, P.P.: Performance analysis of binary and multi-class models using azure machine learning. International Journal of Electrical & Computer Engineering (2088-8708), 10 (2020)
11. Razali, N., Mustapha, A., Abd Wahab, M.H., Mostafa, S.A., Rostam, S.K.: A data mining approach to prediction of liver diseases. In: Journal of Physics: Conference Series, vol. 1529, no. 3, p. 032002. IOP Publishing, April 2020
12. Dhanabal, L., Shantharajah, S.P.: A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. Int. J. Adv. Res. Comput. Commun. Eng. **4**(6), 446–452 (2015)
13. Shamim, A., Balakrishnan, V., Kazmi, M., Sattar, Z.: Intelligent data mining in autonomous heterogeneous distributed and dynamic data sources. In: 2nd International Conference on Innovations in Engineering and Technology (ICCET'2014), pp. 19–20, Sept 2014
14. Gao, X., Shan, C., Hu, C., Niu, Z., Liu, Z.: An adaptive ensemble machine learning model for intrusion detection. IEEE Access **7**, 82512–82521 (2019)
15. Ghosh, P., Mitra, R.: Proposed GA-BFSS and logistic regression based intrusion detection system. In: Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT), pp. 1–6. IEEE, February 2015
16. Stibor, T., Timmis, J., Eckert, C.: A comparative study of real-valued negative selection to statistical anomaly detection techniques. In: Jacob, C., Pilat, M.L., Bentley, P.J., Timmis, J.I. (eds.) ICARIS 2005. LNCS, vol. 3627, pp. 262–275. Springer, Heidelberg (2005). https://doi.org/10.1007/11536444_20
17. Li, Y., Qiu, R., Jing, S.: Intrusion detection system using Online Sequence Extreme Learning Machine (OS-ELM) in advanced metering infrastructure of smart grid. PLoS ONE **13**(2), e0192216 (2018)
18. Shakya, S., Kaphle, B.R.: Intrusion detection system using back propagation algorithm and compare its performance with self organizing map. J. Adv. Coll. Eng. Manag. **1**, 127 (2016)
19. Microsoft Azure Machine Learning Studio. https://studio.azureml.net/. Accessed on June 2016
20. Introducing Kaggle Simulations. https://www.kaggle.com/what0919/intrusion-detection. Accessed on 2019
21. Micro Average vs Macro average Performance in a Multiclass classification setting, Data Science (2018). https://datascience.stackexchange.com/questions/15989/micro-average-vs-macro-average-performance-in-a-multiclass-classification-settin
22. Khalaf, B.A., et al.: A simulation study of syn flood attack in cloud computing environment. AUS J. **1–10**, 2019 (2019)

23. Al-Ta'i, Z.T.M., Abass, J.M., Abd Al-Hameed, O.Y.: Image steganography between Firefly and PSO Algorithms. Int. J. Comput. Sci. Inform. Secur. **15**(2), 9 (2017)
24. Babatunde, O.S., Ahmad, A.R., Mostafa, S.A., khalaf, B.A., Fadel, A.H., Shamala, P.: A smart network intrusion detection system based on network data analyzer and support vector machine. In: International Journal of Emerging Trends in Engineering Research, vol. 8, no. 1, pp. 213–220 (2020)
25. Fadel, H., Hameed, R.S., Hasoon, J.N., Mostafa, S.A.: A Light-weight ESalsa20 Ciphering based on 1D Logistic and Chebyshev Chaotic Maps. Solid State Technol. **63**(1), 1078–1093 (2020)