Technoscience Journal for Community Development in Africa 1:1 (2020) 81–89
ISSN 2659-0573

Research Article

Comparative analysis of selected data mining algorithms for intrusion detection system

Aliyu Ishola Nasiru*, Musbau Dogo Abdulrahaman, Ameh Adams

Department of Information and Communication Science, University of Ilorin, Ilorin, Nigeria

Abstract: Due to the growth of information technology, there is a tremendous increase in demand for network connectivity by individuals and organisations, thereby making network security more worrisome than ever. An intrusion detection system is one of the security solutions employed to protect network-based information systems from unauthorised access or information misuse. Data mining and machine learning techniques are important fields of study that have been applied to the area of intrusion detection domain with different classification algorithms due to their ability to learn from a very large amount of data. However, identification of an appropriate and efficient technique and algorithms for building an intrusion detection system remains an increasing challenge. This study focuses on evaluating the performance of five well-known classification algorithms: Decision Tree, Naive Bayes, C4.5, K – Nearest Neighbour, and ID3, on a popular intrusion detection benchmark dataset (KDD). The result of the evaluation shows that K-nearest neighbour outperformed other algorithms with high accuracy of 99.97%, a low error rate of 0.03, and fast building time of 0.07 seconds which demonstrate its appropriateness for efficient intrusion detection system.

Keywords: Data mining; machine learning; intrusion detection system; comparative analysis; KDD

1. Introduction

Owing to the growth in computing and widespread adoption of the network-based information systems in many sectors such as finance, government, health, education, commerce, and social network by individuals and organisations, maintaining security and privacy on the cyberspace is a serious challenge. This is due to the disappearance of network boundaries between connected devices and increasing sophistication of cyber-attacks and tools (Al-jarrah et al., 2018). Network intrusion- unauthorised access to the information system to compromise its confidentiality, integrity, and availability (CIA)- remains the most challenging problem facing information system users (Chourasiya et al., 2018). One of the most prominent solutions to information security threats is the development and deployment of Intrusion Detection System (IDS) (Colom, Gil, Mora& Volckaert, 2018).

IDS is the hardware, software, or their combinations deployed on systems/networks to monitor any unauthorised access or activities. IDS is used in combination with firewall and other security mechanisms such as authentication and encryption to provide better and effective information security (Abdulrahaman & Alhassan, 2018; Chourasiya et al., 2018). Generally, IDS is categorised into two classes: the network-based and host-based. The host-based IDS analyses system call logs generated by an individual computer system, whereas network-based IDS analyses network packets collected from the network. IDS can also be categorised as anomaly-based or signature-based in term of the methods used for data analysis. The signature-based IDS uses a set of predefined rules manually set by security experts for detecting any suspicious activities on the network or system. However, this method is not effective enough when it comes to detecting unknown, dynamic,

* Corresponding author:
Email: aliyu.in@unilorin.edu.ng

KWASU PRESS AND PUBLISHING

and sophisticated attacks. Unlike anomaly-based IDS that uses intelligent techniques to detect deviation of the normal behaviour of network users. This method is more effective in detecting both known and unknown attacks needed for the modern network attacks that are dynamic and sophisticated (Abdulrahaman and Alhassan, 2018).

In recent time, data mining techniques have been employed for anomaly-based IDS improvement characterised by a high false alarm, low detection rate, and high running time (Idhammad et al., 2018). Data mining technique is equipped with powerful algorithms that can help in improving the performance of IDS, but identification of appropriate techniques and algorithms remain a serious challenge in the areas of intrusion detection and data mining (Adebowale et al., 2013; Hajimirzaei & Navimipour, 2019). The purpose of this study is to compare the performance of some selected data mining algorithms in the intrusion detection domain to highlight their strengths and weaknesses.

The organisation of the remaining parts of this work is as follows: Section 2 describes the related works while section 3 describes the methodology used for this study. Section 4 highlights the results of the study as well as the implications of the results. The conclusion and future research directions are in Section 5.

2. Related works

Intrusion detection system remains a critical component of secure information systems and many security experts are carrying out studies to improve its performance. This section describes some selected relevant works that have used data mining algorithms in the intrusion detection domain.

Duque and Omar, (2015) proposed a data mining based model on the K-means algorithm to identify important clusters for intrusion detection. A benchmark IDS dataset NSL-KDD with a total number of 25,192 instances was used to generate four different clusters: 11, 22, 44, and 88 respectively. The evaluation results revealed that the best performance was obtained when the number of clusters corresponds to the number of data types which is 22. In their work, Al-jarrah et al., (2018) developed a multilayer clustering-based intrusion detection model for IDS with the help of K- means algorithm and evaluated on two separate benchmark intrusion detection datasets, Kyoto 2006+ and the popular NSL-KDD. The results of the evaluation show that the model performs well on both datasets using performance parameters such as detection rate, Mathew's correlation coefficient, and false alarm rate.

Ambusaidi et al. (2016) proposed an IDS model based on Least Square Support Vector Machine based IDS (LSSVM-IDS) to obtain high-efficiency rate, low false-positive rate and false negative rate using three popular benchmark datasets including KDD CUP 99, NSL-KD, and Kyoto 2006+ datasets. Before feeding the datasets into the algorithm, two feature selection techniques based on Mutual Information (MIFS) and Flexible Mutual Information Feature Selection (FMIFS) were performed to improve the predictive power of the algorithm. The evaluation results produced an improved model with better accuracy and lower computational cost when compared to other feature selection methods in the literature such as the Linear Correlation Coefficient. The performance metrics used are accuracy, detection rate, false-positive rate, F-measure, precision and recall respectively and it was revealed that IDS with Flexible Mutual Information Feature Selection (FMIFS) outperformed that of the Mutual Information Feature Selection (MIFS).

Similarly, Gupta and Kulariya (2016) proposed an intrusion detection framework for an efficient cyber security IDS using five different classification algorithms including Random Forest (RF), Logistic Regression (LR), Support Vector Machines (SVM), Gradient Boosted Decision Trees (GBDT), and Naive Bayes. The IDS was built with two different datasets: DARPA's KDD'99 and NSL-KDD pre-processed using Apache Spark and it MLlib library (a sophisticated tool for big data processing) and enhanced with correlation-based feature selection (CFS) and Chi-squared feature selection. It was observed that the two feature selections improved the performance of the algorithms significantly especially, RF and GB trees in terms of accuracy and time for training and predictions. However, despite the improvement in the performance of some of the used algorithms as a result of the feature selection, the removal of the highly correlated features in the dataset resulted into a negative implication which reduces Accuracy of Naïve Bayes, LG, and SVM.

In their work Verma and Ranga, (2018) built an IDS based on Coburg Network Intrusion Detection Dataset (CIDDS-001) dataset with k-nearest neighbour (KNN) classification algorithm and k-means clustering algorithms that classified and clustered network traffic into several categories such as a normal, attack, victim, suspicious and unknown. Based on the performance evaluation with popular metrics such as precision, F-measure, and false-positive rate, the simulation in WEKA shows a relatively improve performance. Gautam & Om (2016) developed a host-based IDS using Generalised Regression Neural Network and Multilayer perceptron Neural Network. However, this paper did not provide a concise developmental framework used for the study.

Mehibs and Hashim (2018a) proposed IDS model based on the construction of back propagation artificial neural network in the cloud-computing environment. The testing and training dataset was generated from KDD CUP 99. The experimental results show that a multilayer perceptron is a commendable approach to intrusion attack detection in the cloud with high detection rate and low false alarm rate. Also, Mehibs and Hashim (2018b) put forth a network IDS based on Fuzzy C-means Algorithm for Cloud Computing Environment. The algorithm was evaluated on KDD 99 dataset with four cyber-attack types. The proposed model performed well on the two-benchmark dataset used for the evaluation.

Jabbar et al. (2017) proposed a novel IDS based on the ensemble approach to enhance the performance of classification algorithms. The ensemble classifier was built with two well-known algorithms, Random Forest (RF) and Average One-Dependence Estimator (AODE). The performance of the built model was evaluated on Kyoto 2006+ dataset and produced not too encouraging results with 90.51% as its accuracy, detection rate of 92.38%, and false alarm rate of 0.14 respectively. Similarly, Abdulrahaman and Alhassan (2018) presented an ensemble learning-based IDS model, forming the classifier with the combination of Multilayer Perceptron Neural Network (MLPNN) and Sequential Minimal Optimisation (SMO) algorithms using Kyoto 2006+ intrusion detection dataset. The experimental results produced an improved performance of IDS with the accuracy of "95.02%", a detection rate of "96.92%", false alarm rate of "0.01" and Hubert Index of "90". Table 1 summarises some of the reviewed articles on the intrusion detection system and data mining or machine learning techniques.

Therefore, despite enormous efforts put forth by the various security experts to improve the performance of the intrusion detection system, identification of appropriate techniques and algorithms remain a serious challenge in the intrusion detection and data mining areas. Hence, a comparative analysis of some selected data mining algorithms in the intrusion detection domain is needed and the reviewed pieces of literature shall form the basis of the research in the quest to providing a reasonable solution to the intrusion detection problems.

3. Materials and methods

This section describes some of the important components of the method used for the comparative analysis of some classifiers for intrusion detection systems.

3.1. KDD Cup 99 Dataset

KDD'99 dataset is one of the most popular benchmark datasets for the evaluation of algorithms by researchers in intrusion detection system domain. It was first used at the international knowledge discovery data mining competition and was selected from DARPA 98 network traffic dataset in 1999 by collecting single TCP dump into TCP connection (Mehibs & Hashim, 2018b). It consists of 41 attributes that can be categorised as basic features, content features, and traffic features. The dataset features consist of values in different formats such as numeric, binary, and real number. It also has an additional class attribute that depicts whether an instance of connection is normal or malicious (Karatas & Sahingoz, 2018). KDD dataset is made up of normal traffic data and four known categories of attack as shown in Table 2.

Denial of Service (DOS) attack: This is the kind of attack where an attacker attempts to overwhelm the information system resource in order to prevent legitimate users from accessing the system.

Remote-to-Local (R2L) attack: The attacker sends packets to the victim's system through a network and gets unauthorised local access to the machine through the exploited vulnerability.

User-to-Root (U2R) attack: In this type of attack, the hacker gets access to a normal user account and attempts to exploit the system's vulnerability in order to gain illegitimate super-user privileges.

Probe attack: The attacker scans the network for sensitive information about the user of the system for stealing sensitive data for an attack.

3.2. Selected classification algorithms

Data mining and intrusion detection domains use algorithms to learn patterns in complex data to make decisions or predictions. This research work sought to perform an empirical comparative analysis of five popular classification algorithms used for building data mining or machine learning-based intrusion detection model. There exist several classification algorithms in the literature that may be explored. Due to the limited experimentation time, only five popular classifiers were comparatively analysed, namely Naïve Bayes, Decision Tree, ID3, C4.5, and K – nearest neighbour (KNN).

Naïve Bayes Algorithm (NB): This probabilistic classification algorithm is based on Bayes theorem. It counts the frequency and combination of values in a given dataset (Patil & Sherekar, 2013). Naïve Bayes works on the assumption that the effect of a feature on a given class is independent of the values of the other attributes. This refers to as class conditional

Table 1: Summary of the reviewed papers

No	Author	Year	Methodology	Strengths	Weaknesses
1	Duque and Omar	2015	Generates four clusters with K-Means using NSL-KDD with 25,192 instances	The clustering algorithm used and generated new clusters for intrusion detection	The efficiency of the algorithm is low and produced a high false-negative rate (4.03%). No classifier used.
2	Al-Jarrah et al.	2018	Developed a semi-supervised multilayer Clustering-based IDS using K-means Algorithm. Evaluated on NSL_KDD and Kyoto 2006+ IDS datasets.	It can handle both labelled and unlabelled data for intrusion detection. Performance metrics used are accuracy, detection rate, Mathew's correlation coefficient, and false alarm rate.	It has high testing time compared to the benchmarked models such as bagging and Tri-training
3	Ambusaidi et al.	2016	Lest Square Support Vector Machine based IDS with NSL-KDD and Kyoto 2006+ as benchmark datasets. Feature Selection with (1) Mutual Information and (2) Flexible Mutual Information. Evaluation parameters are: accuracy, detection rate, false-positive rate, F-measure, precision and recall with Least Square SVM (LS-SVM)	The feature selection introduced improved the performance of least square SVM classifier on the two-benchmarked dataset.	The impact of unbalanced sample distribution for IDS was not considered. Though the model shows an encouraging performance, the result obtained is not optimal.
4	Gupta and Kulariya	2016	IDS model based on five classifiers: Logistic Regression, Support Vector Machines, Random Forest, Gradient Boosted Decision Trees, and Naive Bayes using DARPA KDD' 99 and NSL KDD datasets using Apache Spark using Chi-squared feature selection.	Feature selection enhanced the performance of the classifiers.	The sample size for training and testing of the model was not specified.
5	Verma and Ranga	2018	CIDDS-001 dataset, used k-means for clustering and KNN for classification with 1, 2,3,4, and 5 neighbours in open Stack server.	The result shows the optimal performance of the model with an average accuracy of 100%	There is evidence of bias in the random selection of the samples which could have resulted in over-fitting
6	Mehibs and Hashim	2018a	ANN based on back propagation using KDD CUP 99. Evaluated with Accuracy, Detection Rate, and False Alarm Rate in three different experiments.	Achieved a better performance on the dataset.	The data sample used for training and testing is very small.
7	Jabbar, Aluvalu, Satyanarayana, & Reddy	2017	An ensemble of RF and AODE for Kyoto 2006+ classification. Achieved 90.51% accuracy	The ensemble method improved the performance of base classifiers.	The proposed method has a low detection rate and accuracy. The feature selected was done manually with no clear reason for the ones selected.

8	Abdulrahaman and Alhassan	2018	An ensemble of MLPNN and SMO for intrusion detection using Kyoto 2006+ IDS dataset.	Achieved an improved performance over the base classifiers in term of accuracy, detection rate, false alarm rate, and Hubert index	There was no technical reason to justify the choice of the few features used for the classification. The result is not optimal.

Table 2: Attack categories in KDD cup 99 dataset (Mehibs & Hashim, 2018b)

S/N	Attack Categories	Attack Types
1	DoS	smurf, neptune, pod , back, land , teardrop
2	Probe	satan, ipsweep, portsweep, nmap
3	R2L	warezclient, guess_passwd,warezmaster, ftp_write, multihop, phf, spy, imap
4	U2R	buffer_overflow,rootkit, loadmodule, perl

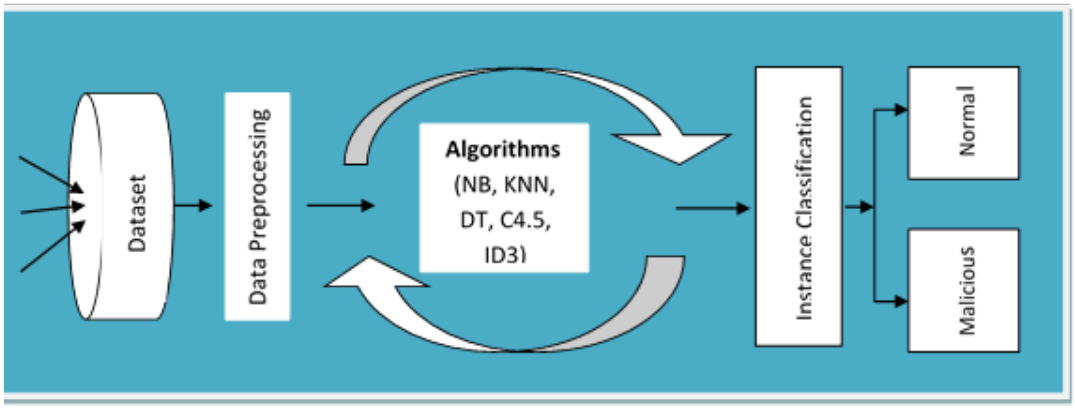

Figure 1: Framework for comparing data mining algorithms for intrusion detection system

independence. The term "Naïve" is as a result of the fact that the algorithm makes computation relatively simple as it represents dependencies among the subsets of attributes. Naïve Bayes simplifies computations with high accuracy and speed. The effect of the set of vectors $A=\{a_1, a_2, a_3, \ldots, a_n\}$, on a given class $C=\{C_1, C_2, C_3, \ldots, C_k\}$ in a training sample is said to be independent of other attributes values and can be represented mathematically as follows:

$$P(C \mid A) = \frac{P(A \mid C) * P(C)}{P(A)} \quad (1)$$

Where: P(C|A) is the posterior probability of target class given an attribute; P(C) refers to the class prior probability; P(A|C) means the probability of attribute given class. It is a likelihood; P(A) means the prior probability of attribute (Naïve Bayes, nd).

Decision Tree Algorithm: This is a classification algorithm that learns inductively to construct a predictive model from a labelled dataset. Its decision is based on a hierarchical structure where each data item is defined by the attribute value in the dataset. A decision tree divides classification problems into several sub-problems and then creates a decision tree that can be used for classification purposes (Aljawarneh et al., 2017).To classify a particular data item, it starts at the root node and traverses down until a terminal node is reached for a decision to be made. One of the advantages of using a decision tree for IDS is its ability

to construct a predictive model that determines whether an instance of the traffic data is malicious or benign. It can also handle high dimensional network traffic well and respond to the dynamic nature of huge traffic data (Adebowale et al., 2013). A decision tree is known for a certain constraint which is ensuring that the output attribute is always categorical and usually produces a complex and unstable tree

ID3 Algorithm: This is a simple decision tree classifier developed by Ross Quinlanin 1983. ID3 constructs decision tree using greedy search method in a top-down manner for testing each attribute on every node. To select the most relevant features for classification purpose in a given set, a metric information gain is introduced and the depth of the tree must be minimised. Thus, some functions that will provide the most balanced splitting are needed. This algorithm is useful in IDS as it provides simple classification techniques for predicting the class of traffic data in a quicker manner.

C4.5 Algorithm: This is a classification algorithm proposed by Ross Quinlan in 1993 as an improvement over the earlier ID3 algorithm. C4.5 is highly sensitive to attributes with a large number of values. This limitation needs to be overcome in order to use it in robust applications such as internet search agent. C4.5 algorithm allows for the measurement of a gain ratio defined as follows:

$$GainRation = \ (P,T) = \frac{Gain(P,T)}{SplitInfo(P,T)} \qquad (2)$$

$$\text{Where } SplitInfo(P,T) = -\sum_{j=1}^{n} P(\frac{j}{p}) * \log(p(P^{'}(\frac{j}{P}) \qquad (3)$$

P' (j/P) is the proportion of elements present at the position p, taking the value of j^{th} test. It should be noted that unlike the entropy, the examples inside different classes in the foregoing definition do not dependent on the distribution.

K-Nearest Neighbour: This is a predictive learning algorithm that classifies objects with respect to the closeness of the feature space of the training samples. It is usually referred to as lazy learners as its function is usually approximated locally which makes computations defer until classification. KNN is one of the simplest algorithms. It classifies object base on majority voting of its neighbours. That is, it assigns a class to an object based on it closeness or k-nearest to its neighbour (Adebowale et al., 2013). KNN uses similarity-based search strategy to determine a local hypothesis function. The test instances are compared to the stored instances and then assigned the same class as the k most similar stored instances. KNN is used in IDS due to its simple implementation features and with the way, it adapts to a new environment quickly and effectively. Its major disadvantage is its high storage requirement which sometimes susceptible to misclassification of instances in high dimensional data instances.

3.3. Experimental setup

This section briefly explained the procedure and some tools used in performing a comparative analysis of the selected classification algorithms for intrusion detection systems.

Figure 1 describes the processes involved while performing the comparative analysis of the five selected data mining algorithms, namely, naïve bayes, k-nearest neighbour, decision tree, C4.5, and ID3 algorithm. The analysis was performed using data mining and machine learning techniques including the following phases:

1. Dataset selection (KDD Dataset)
2. Loading of the dataset
3. Data pre-processing
4. Algorithm selection, training and evaluation
5. Network traffic Classification by individual algorithm

The KDD intrusion detection benchmark dataset was selected and loaded into the WEKA data-mining tool. A total number of 25,587 instances were randomly selected as training and evaluation instances. The data sample was pre-processed through data transformation by converting some symbolic based features to numeric and class to nominal depending on the nature or classification requirements of each of the algorithms. The pre-processed data were fed into the individual algorithms and used for training iteratively. For the purpose of training and testing of the model, a standard 10-fold cross-validation method was selected. The individual algorithm classified instance of network traffic as an intrusion (malicious) or benign. These processes are depicted in Figure 1.

3.4. Performance evaluation

To measure the performance of the evaluated algorithms on the KDD dataset, some standard performance parameters in data mining and machine learning techniques were used and described as follows:

Confusion matrix: describes the parameters to measure the performance of the individual algorithms in term of correctly and incorrectly classified instances of network traffic. The components of a typical confusion matrix include True Positive, False Positive, False Negative, and True Negative. These parameters are described in Table 3.

Table 3: Confusion matrix

		Predicted Class	
		Normal traffic Instance	Malicious traffic Instance
Actual Class	Normal traffic Instance	TN	FP
	Malicious traffic Instance	FN	TP

True positive (TP): simply refers to the total number of malicious instances "correctly" labelled by the classifier

True N=negative (*TN*): The total number of normal instances "correctly" labelled by the classifier

False positive (FP): The total number of normal instances "incorrectly" labelled by the classifier as malicious

False negative (FN): refers to the total number of malicious instances "incorrectly" labelled by the classifier as normal

Accuracy: This is used to measure how accurate a model can detect whether an instance is normal or malicious. This is also known as the percentage of instances that have been correctly classified and can be expressed as follows:

$$\text{Accuracy (ACC)} = \frac{TN + TP}{TP + FP + FN + TN} \quad (4)$$

Error rate: this describes the number of misclassifications made by the classifier. This is also known as the percentage of incorrectly classified instances. This is expressed as follows:

$$\text{Error Rate} = \frac{FP + FN}{TP + FP + FN + TN} \quad (5)$$

Kappa statistics: refers to the measurement corrected with a chance. It measures the agreement between the classifications and the true classes. It is usually calculated by taking away the expected agreement from the agreement observed, divides by the maximum value of all possible agreements. The classifier is doing better than chance if a value greater than zero is achieved (Adebowale et al., 2013).

Other errors: Some of the performance metrics related to error rate are Mean Absolute Error, Relative Absolute Error, and Root Mean Square Error.

4. Results and discussion

The summary of the result for the comparative analysis of the selected algorithms on KDD dataset is given as follows.

To perform the comparison, the total number of 25,587 instances was randomly selected as training and evaluation instances. The dataset was pre-processed to suit the requirement of individual algorithms. For the purpose of achieving better classification accuracy, a 10-fold cross-validation method was used. Table 4 shows the comparison of different classification algorithm performance in terms of correctly and incorrectly classified instances as well as the time taken for the classification. As depicted in Table 4, K-nearest neighbour outperformed other algorithms with 25,579 correctly classified instances and 8 instances were misclassified at the lowest time of 0.07s. This is a reasonable result when compared with the like of Naïve Bayes that has 25,525 and 72 as correct and incorrectly classified instances respectively at the time of 0,43s; Decision Tree that with correctly classified as 25,575 and 12 as incorrectly classified instances within 5.85s; C4.5 has 25,569 and 18 misclassification respective at the time of 188.73; ID3 algorithm with correct classification as 25,515 and misclassified instances as 72 in 400.19s. This result shows that KNN is better when it comes to analysis and accurately detecting network intrusions followed by Decision Tree algorithm. In addition, when considering intrusion detection in real-time, KNN also has the smallest time and closely followed by Naïve Bayes and decision tree.

Table 4: Comparison of different classification algorithms

Algorithm	Correctly Classified Instance	Incorrectly Classified Instance	Classification Time (Seconds)
Naïve Bayes	25525	72	0.43
KNN	25579	8	0.07
Decision Tree	25575	12	5.85
C4.5	25569	18	188.73
ID3	25515	72	400.19

In addition, Table 5 shows the comparison of the selected algorithms in terms of correct instance classification rates, KA Statistics, error rates and other error related measurements including, Relative Absolute Error, Mean Absolute Error, and Root Mean Square Error. The results show that KNN performed better than other algorithms with the highest detection and accuracy of 99.9687% and the lowest error rate of 0.03 compared to other algorithms. KNN also shows its superiority with higher KA statistic value 0.9988%. Figures 2, 3 and 4 show the performance of individual algorithms in terms of accuracy, error rate and classification time respectively.

Table 5: Classification algorithms comparison results

Parameters	Naïve Bayes (%)	KNN (%)	Decision Tree (%)	C4.5 (%)	ID3 (%)
Correctly Classified Rate (Accuracy)	99.7186	**99.9687**	99.9531	99.9297	99.7186
Incorrectly Classified Rate (Error Rate)	0.28	**0.03**	0.05	0.07	0.28
KA Statistics	0.9892	0.9988	0.9982	0.9973	0.9891
Mean Absolute Error	0.0002	0.0001	0.0001	0.0001	0.0004
Root Mean Square Error	0.0157	0.0052	0.0064	0.0076	0.0147
Relative Absolute Error	1.0902	0.4328	0.2944	0.2625	1.8386
Root Relative Squared Error	14.7197	4.4904	6.0077	7.1254	13.8589

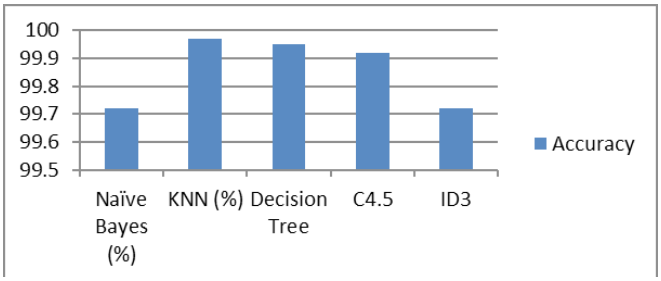

Figure 2: Detection accuracy of the selected data mining algorithms

Figure 2 compares the performance of the analysed algorithms showing the classification accuracy, it shows that KNN achieved the highest accuracy on KDD dataset with 99.9687% and closely follow by a Decision tree with 99.9531%. The least accuracy result was produced by Naïve Bayes and ID3 with the same classification accuracy of 99.7186 respectively. This result indicates that KNN is much appropriate for intrusion detection with less misclassification error.

Similarly, the misclassification rate of each of the classifiers analysed is plotted in Figure 3, the results show KNN to have the least error rate of 0.03, decision tree with 0.05, C4.5 with 0.07, and Naïve Bayes and ID3 to have 0.28 respectively. This also shows a clear superiority of KNN over other compared algorithms when it comes to intrusion detection using machine-learning technique.

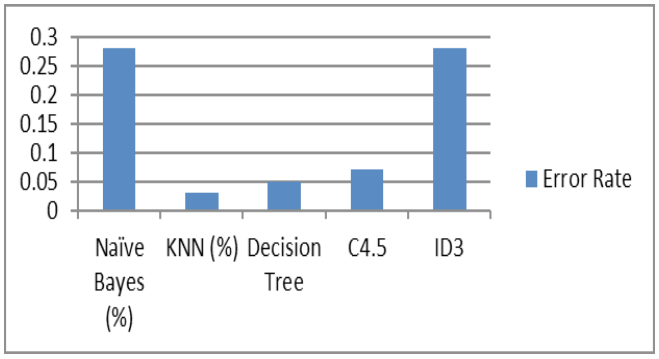

Figure 3: Misclassification Rate of the selected data mining algorithms

Figure 4 describes the time taken each of the algorithms to classify instances of network traffic as either benign or intrusion. The figure indicates that KNN outperformed others with 0.07 seconds and closely follow by Naïve Bayes with 0.43. Decision tree also performs better compares to C4.5 and ID3 with 188.73 seconds and 400.19 seconds respectively.

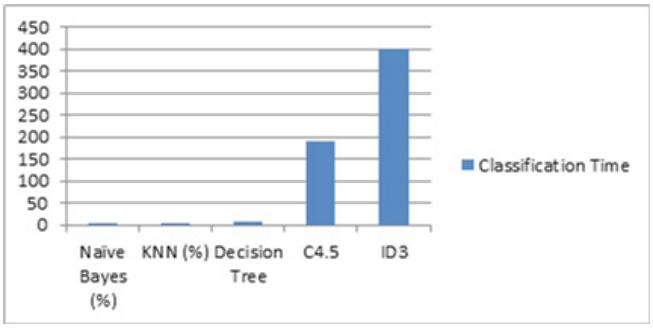

Figure 4: Classification Time of the selected data mining algorithms

5. Conclusion

Intrusion is a serious problem affecting the security and privacy of the information on the cyberspace. Despite the ability of data mining and machine learning algorithms to learn from a huge amount of network traffic data, it is very difficult to identify appropriate techniques and algorithms in the area of intrusion detection. In this study, comparative analysis of some data mining and machine learning classifiers has been performed using some important metrics in the fields. The results of our comparative analysis show that K-Nearest Neighbour performed better in terms of higher detection rate, low error rate, faster building and detection time, and higher accuracy when compared with the performance of other algorithms. This shows the superiority of KNN algorithm and proves its appropriateness to the intrusion detection system. For future works, a combination of more than one algorithm popularly known as ensemble technique may be considered in order to enhance the performance of base classifiers. Development of intrusion detection model for specific network environments such as Local Area Network (LAN), Cloud computing environment, and Internet of Thing (IoT) could also be considered.

References

Abdulrahaman, M. D., & Alhassan, J. K. (2018). Ensemble learning approach for the enhancement of performance of intrusion detection system. In 2nd *International Conference on Information and Communication Technology and its Applications (ICTA)*, 190-196.

Adebowale, A., Idowu, S. A., & Amarachi, A. (2013). Comparative study of selected data mining algorithms used for intrusion detection. *International Journal of Soft Computing and Engineering (IJSCE)*, *3*(3), 237-241.

Al-Jarrah, O. Y., Al-Hammdi, Y., Yoo, P. D., Muhaidat, S., & Al-Qutayri, M. (2018). Semi-supervised multi-layered clustering model for intrusion detection. *Digital Communications and Networks*, *4*(4), 277-286. https://doi.org/10.1016/j.dcan.2017.09.009

Aljawarneh, S., Aldwairi, M., & Yasin, M. B. (2017). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *Journal of Computational Science*, *25*, 152-160. https://doi.org/10.1016/j.jocs.2017.03.006

Ambusaidi, M. A., He, X., Nanda, P., & Tan, Z. (2016). Building an intrusion detection system using a filter-based feature selection algorithm. *IEEE transactions on computers*, *65*(10), 2986-2998. https://doi.org/10.1109/TC.2016.2519914

Chourasiya, R., Patel, V., & Shrivastava, A. (2018). Classification of cyber attack using machine learning technique at microsoft azure cloud. *International Research Journal of Engineering & Applied Sciences*, *6*(1), 4–8.

Colom, J. F., Gil, D., Mora, H., Volckaert, B., & Jimeno, A. M. (2018). Scheduling framework for distributed intrusion detection systems over heterogeneous network architectures. *Journal of Network and Computer Applications*, *108*, 76-86. https://doi.org/10.1016/j.jnca.2018.02.004

Duque, S., & bin Omar, M. N. (2015). Using data mining algorithms for developing a model for intrusion detection system (IDS). *Procedia Computer Science*, *61*, 46-51. https://doi.org/10.1016/j.procs.2015.09.145

Gautam, S. K., & Om, H. (2016). Computational neural network regression model for Host based Intrusion Detection System. *Perspectives in Science*, *8*, 93-95. https://doi.org/10.1016/j.pisc.2016.04.005

Gupta, G. P., & Kulariya, M. (2016). A framework for fast and efficient cyber security network intrusion detection using apache spark. *Procedia Computer Science*, *93*, 824-831. https://doi.org/10.1016/j.procs.2016.07.238

Hajimirzaei, B., & Navimipour, N. J. (2019). Intrusion detection for cloud computing using neural networks and artificial bee colony optimization algorithm. *ICT Express*, *5*(1), 56-59.

Idhammad, M., Afdel, K., & Belouch, M. (2018). Distributed intrusion detection system for cloud environments based on data mining techniques. *Procedia Computer Science*, *127*, 35-41.

Jabbar, M. A., & Aluvalu, R. (2017). RFAODE: A novel ensemble intrusion detection system. *Procedia computer science*, *115*, 226-234. https://doi.org/10.1016/j.procs. 2017.09.129

Karatas, G., & Sahingoz, O. K. (2018, March). Neural network based intrusion detection systems with different training functions. In *2018 6th International Symposium on Digital Forensic and Security (ISDFS)* (pp. 1-6). IEEE.

Mehibs, S. M., & Hashim, S. H. (2018a). Proposed network intrusion detection system in cloud environment based on back propagation neural network. *Journal of University of Babylon*, *26*(1), 2-40.

Mehibs, S. M., & Hashim, S. H. (2018b). Proposed network intrusion detection system based on fuzzy c mean algorithm in cloud computing environment. *Journal of University of Babylon*, *26*(2), 27-35.

Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, *6*(2), 256-261.

Verma, A., & Ranga, V. (2018). Statistical analysis of CIDDS-001 dataset for network intrusion detection systems using distance-based machine learning. *Procedia Computer Science*, *125*, 709-716. https://doi.org/10.1016/j.procs.2017.12.091