

Comparison and Analysis of Data Mining Techniques for Intrusion Detection

Ibrahim Adamu, Asabe Sandra Ahmadu, Usa John Guli and Ezekiel Peter

Department of Computer Science, Faculty of Physical Science Modibbo Adama University, Yola, Nigeria.

Corresponding author: mangas4real@gmail.com

Abstract

This research investigates the use of data mining techniques for intrusion detection. Decision Trees, Artificial Neural Network, Naïve Bayes and Support Vector Machine are the strategies examined. The dataset used was collected from different online repository where they are made available as an open-source data. The sample size used were 910, 10077, 10679 instances of KDDCUP'99, NSL-KDD-train and CICIDS2017 respectively, selected using random sampling technique. The Weka workbench tool was used for preprocessing and creation of the intrusion detection system. The accuracy, speed, and scalability of the techniques, among other criteria, are taken into account and contrasted. The study also examines the best method for detecting intrusions in dynamic networks and various applications. The performance of each technique in terms of accuracy, precision, recall, and F1-score is also examined in this research. The results revealed that the Decision Tree performs better with accuracy, precision, recall and F-measure of 99% than the other classifiers Support Vector Machine, ANN, NB and KNN in most of the tests on the three different datasets. Both Decision Tree and ANN classifier showed superior performance in detecting attacks. In conclusion, this paper reveals that artificial neural networks are the most accurate data mining technique for intrusion detection. However, in terms of implementation, Decision Tree classifier take a very short time implemented compared to ANN that takes a very long time to be implemented. Finally, this research recommended that employing these optimizing techniques to develop an intrusion detection model have a better accuracy rate.

Keywords: Artificial neural network, Computer network, Data mining, Intrusion detection, Decision tree, Support vector machine.

Introduction

Global organizations have changed their business models because of the Internet's expansion and potential. Every day, more and more people connect to the Internet in order to benefit from the convenience of conducting business online. Through a variety of platforms, the Internet, which has a wealth of information, has had an impact on trading, education and learning, governance, communication, and socialization (Ngai, Moon, Lam, Chin & Tao, 2015). However, due to the proliferation of high-speed Internet access, cyber-attacks against computer networks are now common affecting almost every Internet connected device on a daily basis. Now, a lot of research has been done on intrusion

detection systems based on data mining techniques as an effective tool. Data mining is one of the technologies applied to intrusion detection to invent a new pattern from the massive network data as well as to reduce the strain of the manual compilations of the intrusion and normal behaviour patterns (Patel, Thakkar & Ganatra, 2012).

Intruders use different techniques like password cracking, sniffing unencrypted traffic to exploit the system vulnerabilities to gain access to different resources in a network. Various network security techniques have been developed to provide security for LANs and the internet, this includes, firewalls, antivirus and internet security has been deployed to prevent malicious activities on the internet. However, there are still many security risks and vulnerabilities (Bendovschi, 2015). Intrusion detection systems are the primary and one of the most important components in network security system. It collects sensitive information of network flow and warns of possible attacks. Network administrator can reinforce their network security and use of data log for forecasting network activities (Keegan, Ji, Chaudhary, Concolato, Yu & Jeong, 2016) . Sadly, intrusion detection technology is far from flawless.

An attempt to change, damage, or destroy computer systems or networks, or the data or programs on them, can be characterized as a cyber attack (Herr, 2014). The cybercriminal's device subsequently disconnects the client's computer and spoofs both its internal data and its number of sequence (Jacobson & Idziorek, 2016). Detection of intrusion accurately requires understanding the method to successfully attack a system. Generally, an attack can be classified into five phases. They are reconnaissance, exploitation, reinforcement, consolidation, and pillage (Vinayakumar, Alazab, Soman, Poornachandran, Al-Nemrat & Venkatraman, 2019) Primarily, an IDS system aims at detecting all these types of attacks to prevent the computers and networks from malicious activities (Sharafaldin, Lashkari & Ghorbani, 2018).

Statement of the Problem

Prior studies had shown that network security techniques have been brought up to provide security for computer networks and the internet society. Setting up of firewalls, antivirus as a form of internet security has been used over the years. However, there are still cases of leakages, theft and intrusion. Hodo, Bellekens, Hamilton, Tachtatzis and Atkinson, (2017) and Mohan, Jain, Suyal, and Kumar, (2020) reported that the modes by which attackers infiltrates the network have become dynamic, therefore there is need to have a more sophisticated means of identifying treats and warn of possible attacks so as to cope with the ever-changing trend of the attackers. Research by Ashraf, Ahmad and Ashraf (2018) compared three data mining algorithms for high detection rate in intrusion detection system on one dataset.

Objectives of the Study

The objectives are to:

- Extract vital input features *in building an intrusion detection system* from intrusion dataset.
- Simulate the selected data mining classification techniques on KDD99, NSL_KDD and CICD2017 dataset
- Compare the result of the selected data mining classification techniques for intrusion detection.
- Evaluate algorithm by generating recall, precision, and f-measure and root mean square error metrics.

This study employed five data mining techniques in order to compare how they perform when applied in high detection rate of intrusion of unauthorized access, policy violators and malicious users using three different standard datasets.

The importance of this intrusion detection system is that it enable the automation of the detection process, improve accuracy, and provide real-time protection against a wide range of cyber threats, to safeguard their critical data and systems from malicious activities.

Literature Review

According to Adeleke (2020) anomaly-based detection methods can be very effective at detecting previously unknown threats. Network packets are monitored by anomaly-based IDS, which contrast any unusual behaviour with a predetermined baseline if it detects it. Compared to signature-based IDS, it is more responsive to emerging threats, but is more challenging to implement (Carlin, Hammoudeh & Aldabbas, 2015). Numerous studies indicate that classification techniques and clustering are by far the most widely used data mining techniques (Wahono, 2015). The classification technique is used to detect email spamming and String comparator is used to detect deceptive information in criminal record (Parvathi & Rautaray, 2014). Due to its versatility and validity, data mining is recognized as a solution to handling data analysis, and it is currently widely utilized for network security objectives. In order to find network assaults and anomalies, data mining technique is essential. These techniques help in selecting and refining useful and relevant information from large data sets (Dewa & Maglaras, 2016).

A Bayesian network is a type of model that represents the probabilistic connections between various variables. This method is typically combined with statistical techniques for intrusion detection. The results of utilizing Bayesian networks are similar to those obtained from threshold-based systems, but a significant increase in processing effort is necessary, which is a big drawback (Patel *et al.*, 2012). Both anomaly intrusion detection and misuse intrusion detection have used neural networks. The neural network would accept data from the network stream and analyze the information to look for instances of misuse in misuse intrusion detection (Patel *et al.*, 2012). According to Cherian and Bindu (2017) decision trees are popularly employed for prediction and classification purpose. Compared to other models, decision trees signify rules. The rules can be presented for humans understanding or directly employed in language for accessing database such as SQL (Structured Query

Language) so that records can be retrieved base on the rules. Patel *et al.*, (2012) performed a review to compares various data mining techniques used to outline the benefits and drawbacks of each technique, such as decision trees, artificial neural networks, SVM, and the k-nearest neighbor algorithm, intrusion detection. The result revealed the ability of computer systems to detect intrusion is provided and current research challenges are pointed out in the field of intrusion detection system.

Mohan *et al.*, (2020) used data mining classification techniques for intrusion detection in order to build a safe network. Considered algorithms are Random tree, Naïve Bayes, J48 and Random Forest. The study reported that random forest classification technique gives a comparatively better accuracy rate than other algorithms. According to Thakkar and Lohiya (2020), intrusion detection is a classification challenge where different machine learning and data mining approaches are used to divide network data into legitimate and malicious traffic. Additionally, as network attack types have evolved over time, it is necessary to update the datasets used to assess IDS. Hindy, Brosset, Bayne, Seeam, Tachtatzis, tkinson *et al.* (2020) exposes the research gaps and shortcomings of current datasets, their impact on building Network Intrusion Detection Systems (NIDS) and the growing number of sophisticated threats. The detection accuracy of existing machine learning IDS techniques is hampered by the fact that current datasets clearly lack attack representation, real-network threats, and a significant number of deprecated threats.

The Proposed Approach

This section gives the description of materials and methods; data set preparation methods, hybrid feature selection method, the different data mining algorithms applied and the results obtained in this research. WEKA data mining tool was used to apply the data mining algorithms. This particular dataset was chosen since it will not leak confidential information, compromise security, or harm network systems.

Table 1: shows the population of the study and the sample sizes used

Dataset	Sum of weight	Total Attributes	Instances used	Attributes used
KDD CUP 1999	494020	42	910	13
NSL-KDD-train	125973	49	10077	9
CICIDS2017	170366	79	10679	67

<https://www.openml.org/data/download/53996/KDDCup99.arff>

<http://205.174.165.80/CICDataset/NSL-KDD/Dataset/NSL-KDD.zip>

<http://205.174.165.80/CICDataset/CIC-IDS-2017/Dataset/GeneratedLabelledFlows.zip>

Dataset Preparation

The KDD Cup dataset consists of 494020 numbers of instances and 42 umber of attributes. There are no missing values in the dataset and there are duplicate data and may be

classified as either normal or attacking. There are four types of attacks in a KDD cup dataset.

- Denial of Service: In this form of attack, the device memory is occupied and overflowing to the point where it is unable to answer when the request is made. Turning off the gadget is the best defense against this type of assault.
- Attack via probing: In a probing attack, an attacker tries to learn as much as possible about a target system or network in order to find any potential weaknesses. Attackers investigate the network using scanning techniques to find open ports, active services, and other details that can be used in later attacks.
- Remote to Local (R2L) Attack: In an R2L attack, the attacker attempts to acquire unauthorized access to a remote machine on the network and uses vulnerabilities to gain local user privileges on that machine. The goal is to infiltrate the local network and compromise a remote system.
- User to Root Attack: In this form of attack, a cybercriminal with specific access to a device tries to get control of the router by taking advantage of system flaws.

This can be accomplished using a variety of techniques, including phishing assaults, sniffer, also referred to as packet controlling, or social engineering. Attacks of this kind occur when a hacker who does not have physical access to the target computer is able to deliver packets from the computer to the network system and is then able to take advantage of security flaws in the system to get access to the target machine. Attacks of this kind attempt to gain data via computer networks with the clear intention of evading security safeguards on the system.

The NSL-KDD dataset has a total weight of 125973 and total attributes of 49. The four subcategories of this dataset are KDDTest+, KDDTest-21, KDDTrain+, and KDDTrain-21. While KDDTest+ and KDDTrain+ are complete datasets with all of their components, KDDTest-21 and KDDTrain-21 only make up 20% of each dataset, respectively. Each of the NSL-KDD datasets contains 43 attributes. The first 41 attributes in every dataset are referred to as the traffic input, while the remaining two are referred to as label and score. This dataset includes the contemporary activities of benign and attacks which depicts the real-time network traffic. The main interest is given towards collecting the real-time background traffic during creating this dataset. Using B-profile system, benign background traffic was collected. This benign traffic contains the characteristics of 25 users based on the HTTP, HTTPS, FTP, SSH, and email protocols. The network traffic for five days was collected and dumped with normal activity traffic on one day, and attacks injected on other days. Brute Force FTP, Brute Force SSH, "DoS," Heart bleed, Web Attack, Infiltration, Botnet, and "DoS" were among the different attacks injected.

Selected Data Mining Techniques

Table 2: shows the data mining techniques and corresponding names in Weka.

S/N	Algorithm	Name in WEKA
1	Multilevel Perceptron Artificial Neural Network	Multilevel Perceptron
2	Naïve Bayes	Naïve Bayes
3	Decision Tree	DTree
4	K-Nearest Neighbour	KNN
5	Support Vector Machine	SMO

Dataset Preprocessing

The Weka workbench version 3.6.2 was used for preprocessing and creation of the intrusion detection system. Dataset preprocessing was performed to remove useless attributes from the dataset and to select valuable attributes. The filters applied to preprocess in each dataset include randomization, remove duplicate, binarization and discretization where necessary. Specific features are selected from a set of features, which have a high discrimination capability between class labels. It also reduced the irrelevant features by eliminating them using two main methods; wrapper and Filter. Feature selection is used to select sets with minimum length while ensuring maximum classification accuracy attributes of most importance and increase performance speed by reducing the irrelevant features through elimination. At the first stage, the dataset is chosen and pre-processing techniques are used to clean data. After applying these filters, 13 attributes are left from 42 attributes of KDD CUP 1999, 9 out of the 49 Attributes of NSL-KDD-train dataset, and 10679 instances of the CICID2017 dataset with 67 attributes were used to reduce time complexity for the model building.

Simulation of the Intrusion Detection System

The simulation tool that was used in this study is Waikato Environment for Knowledge Analysis (WEKA). It is a set of machine learning techniques for handling actual data mining issues. The algorithms can either be applied directly to a dataset or called from other Java code. Among the benefit of WEKA are its free availability (open-source), portability, comprehensive collection of data pre-processing and modelling techniques and lastly it is ease of use due to its graphical user interfaces. The simulation process was implemented on the preprocessed KDD CUP 1999, CICIDS2017, and NSL-KDD datasets.

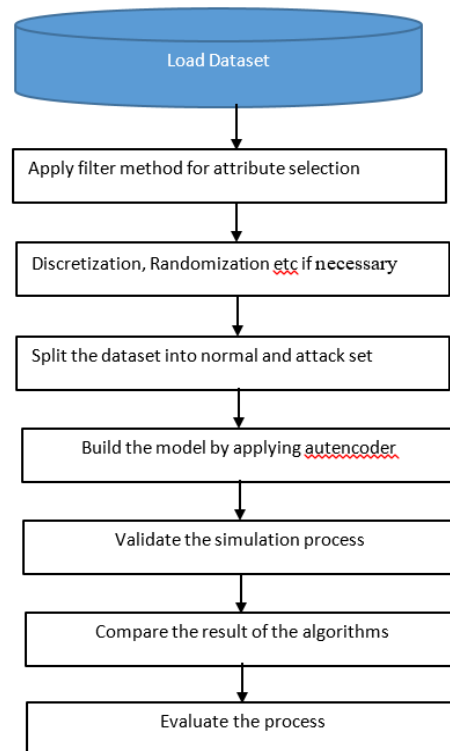


Figure 1 shows the Implementation flow Diagram of the IDS.

Performance Evaluation Metrics

The next step was comparing the five classification algorithms according to the six statistical criteria (accuracy rate, RMSE, precision, recall, and F-measure). To measure the performance of each classifier following metrics are required:

Recall: This represents the percentage of intrusions that are correctly detected. It is calculated by:

$$\text{Recall, } R = \left(\frac{TP}{TP + FN} \right) * 100 \quad (1)$$

Where FN stands for false negative and TP for genuine positive.

Overall Accuracy (OA) or Precision: This represents the percentage that is designed by the total number of intrusions that are properly determined allocated by the total amount of considerations.

$$OA = \left(\frac{TP + TN}{TP + TN + FP + FN} \right) * 100 \quad (2)$$

Where:

TP: is True Positive. When a circumstance is appropriately identified as an intrusion

TN: is True Negative that occurs when an occasion is correctly determined as normal

FN: is False Negative that arises when an event is considered as normal, but actually, it is intrusive

FP: is False Positive that arises when an event is considered as intrusive, but actually, it is normal.

An effective intrusion detection has high Recall and Precision with low false measures.

Precision (P): the fraction of correctly predicted positive observations among the total predicted positive observations.

$$P = \frac{TP}{TP + FP} \quad (3)$$

In which FP stands for false positive and TP for true positive.

F-measure: It is possible to interpret the Precision and Recall requirements jointly rather than separately. To accomplish this, we consider the F-Measure values generated by the harmonic mean of the Precision and Recall columns, as the harmonic mean provides the average of two separate factors produced per unit. Therefore, F gives both the classification's level of accuracy and its level of robustness (the amount of data loss it experiences):

$$F - measure = \frac{2 * P * R}{P + R} \quad (4)$$

Where R is the recall and P is the precision.

ROC area: The various classification algorithms' ability to anticipate outcomes is determined by the ROC field curve. One of the crucial evaluation factors used to choose the best classification method is the area under the ROC curve. The categorization was correctly carried out when the area under the curve is close to 1.

RMSE: the root mean squared error deviation is obtained by determining the square root of the mean squared error (MSE). The RMSE is typically used to quantify the discrepancy between a model's or estimator's estimated values and actual values. To put it another way, the RMSE displays the standard deviation of the variance between the estimated and observed values. The RMSE value should be as low as possible.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (5)$$

Results and Analysis

Simulation Results and Analysis

The sample classification results produced by Weka tool are as presented in figure 2

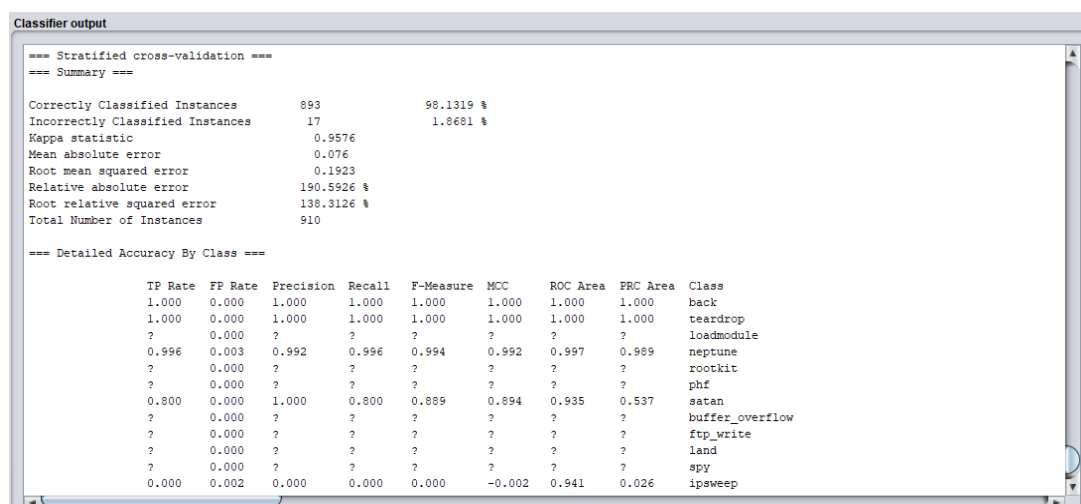


Figure 2: Visualization of the result of intrusion according to SVM data mining technique on KDD CUP '99 dataset.

Comparison of the Simulation Results of the Selected Data Mining Techniques

In this study, WEKA data mining tool was used to apply the data mining algorithms. The snapshot of experimental results from implementation of five selected classification algorithms, SVM, Naive Bayes, ANN DTree and KNN Classifiers on each of the three datasets; KDD CUP'99, NSL-KDD-train, and CICID2017 Dataset are analyzed and compared as shown in table 3 to 5.

Table 3: presents the result of Performance of SVM, Naive Bayes, ANN DTree and KNN Classifiers on KDD CUP'99 Dataset

Algorithm used	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy
SVM	893	17	98.1319
NAÏVE BAYES	895	15	98.3516
ANN	900	10	98.9011
DTREE	890	20	97.8022
KNN	897	13	98.5714

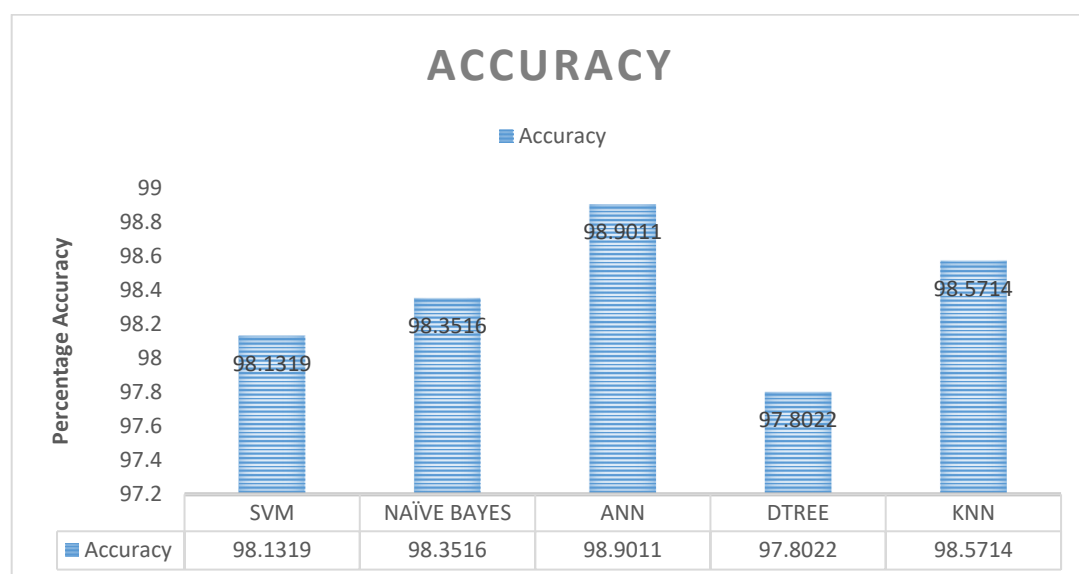


Figure 3: Chart showing the comparison of accuracy achieved by five classifiers on KDD CUP'99 datasets.

The results showed that the best classification accuracy is 98.90% that is achieved by applying the ANN data mining techniques on KDD CUP'99 datasets. The performance of each algorithm was determined based on correctly classified, incorrectly classified and accuracy. Table 3 shows that ANN has the best classification that classified 900 correctly out of 910 instances, followed by KNN with 897 correct classifications in that order. Decision Tree fall short among the techniques considered with 890 correctly classified. Correctly classified instance and incorrectly classified instances show the percentage of test instances that were correctly and incorrectly classified. Accuracy or sample accuracy are terms used to describe the proportion of instances that are correctly classified. The performances of SVM, Naive Bayes, DTree, ANN and KNN are compared. 10-fold cross-validation is used to avoid overfitting. From the chart in figure 3, the results depict that ANN achieved better results than the other algorithms on KDD CUP'99 dataset. Similar Experimental Result on NSL-KDD Dataset using the five selected data mining techniques is obtained as presented in table 4.

Table 4: presents the result of Performance of SVM, Naive Bayes, ANN DTree and KNN Classifiers on NSL-KDD Dataset

Algorithm used	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy
SVM	9210	867	91.3962
NAÏVE BAYES	8847	1230	87.7940
ANN	9914	163	98.3825
DTREE	9977	100	99.0076
KNN	9459	163	93.8672

The experiment was carried out on NSL-KDD dataset on the selected algorithm for comparison. The algorithms used are Artificial Neural Network (ANN), Decision Tree, K-Nearest Neighbour (KNN), Naïve Bayes (NB) and Support Vector Machine (SVM). The performance of each algorithm was determined based on accuracy. Table 4.3 shows that DTree algorithm has the highest accuracy as 99% follow by 98.38% for ANN, 97.79% for Naïve Bayes, 93.86% for KNN while SVM gives 91.99% which is the least

Figure 4 shows the graphical representation of how the selected data mining techniques performed on NSL-KDD dataset in terms of accuracy. From the chart, the results depict that DTree achieved better results than the other algorithms on NSL-KDD dataset.

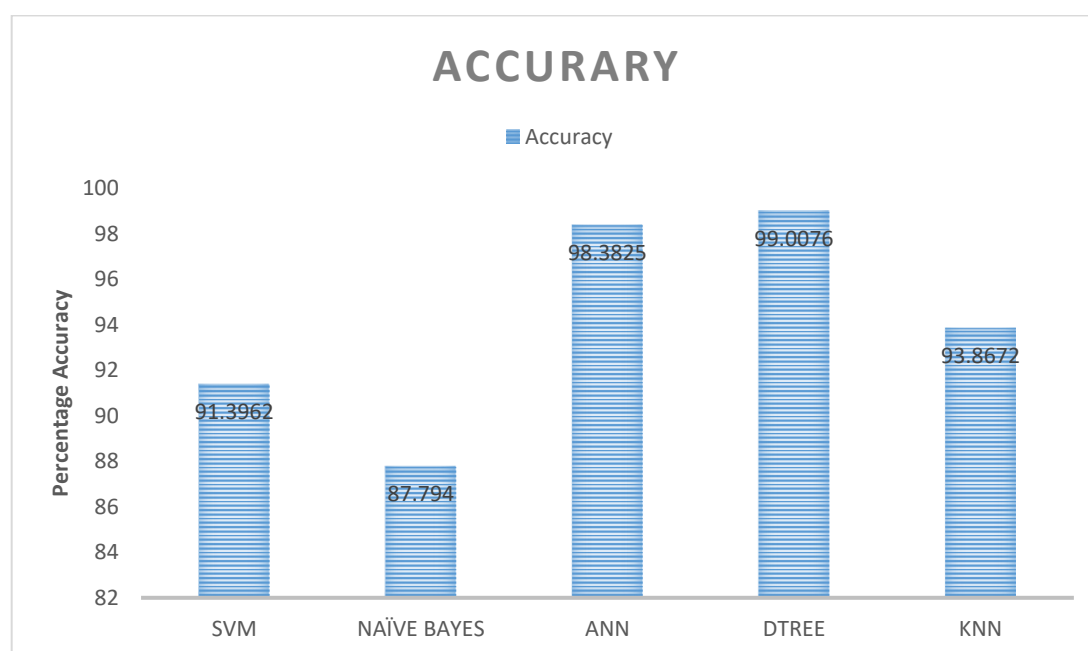


Figure 4: Chart showing the comparison of accuracy achieved by five classifiers on NSL-KDD datasets

The next experimental result is on CICIDS2017 Dataset as shown in table 5.

Table 5: presents the result of Performance of SVM, Naive Bayes, ANN DTree and KNN Classifiers on CICIDS2017 Dataset

Algorithm used	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy
SVM	10543	136	98.7265
NAÏVE BAYES	10230	448	95.8049
ANN	10618	61	99.4288
DTREE	10625	54	99.4943
KNN	10596	83	99.2228

The experiment was carried out on CICIDS2017 dataset on the selected algorithm for comparison. The algorithms used are Artificial Neural Network (ANN), Decision Tree, K-Nearest Neighbour (KNN), Naïve Bayes (NB) and Support Vector Machine (SVM). The performance of each algorithm was determined based on accuracy. Table 5 shows that DTree algorithm has the highest accuracy as 99.49% follow by 99.42% for ANN, 99.22% for KNN SVM gives 98.73% while 95.80% for Naïve Bayes, which is the least.

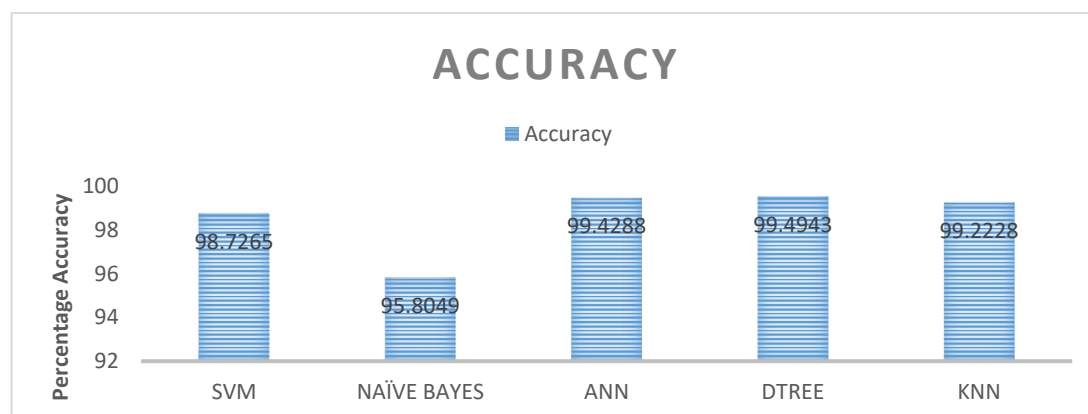


Figure 5: Chart showing the comparison of accuracy achieved by five classifiers on CICIDS2017 datasets

Figure 5 shows the graphical representation of how the selected data mining techniques performed on CICIDS2017 dataset in terms of accuracy. From the chart, the results depicts that DTree achieved better results than the other algorithms on CICIDS2017 dataset. The comparison of the result of the experiment based on the three datasets KDDCup99, CICIDS2017 and NSL-KDD dataset is presented in table 6.

Table 6: comparison of the Performance on three Datasets used base on Accuracy

Algorithm used	KDDCup99 dataset	CICIDS2017 dataset	NSL-KDD dataset
SVM	98.1319	98.7265	91.3962%
NAÏVE BAYES	98.3516%	95.8049	87.7940%
ANN	98.9011%	99.4288	98.3825%
DTREE	97.8022%	99.4943	99.0076%
KNN	98.5714%	99.2228	93.8672%

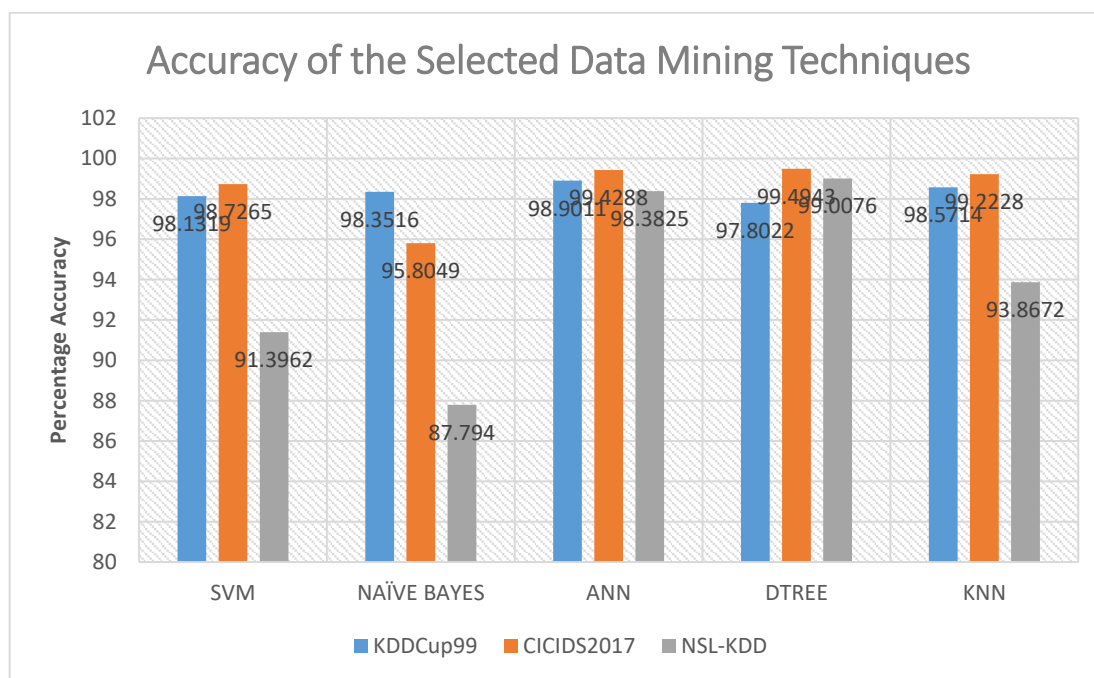


Figure 6: Presents the performance accuracy of all selected data mining techniques

From table 6 and figure 6, it can be seen that the accuracies of each of the data mining techniques shows that the accuracy of ANN for each of three datasets shows the best possible result. This result reflected that ANN gives the best result in terms of accuracy than other classifiers used. Decision tree has better results next to ANN. Whereas KNN Algorithm yielded good result on KDDCup99 and CICID2017 dataset. All the classifiers did very well in all the metrics. In all the experiments on the three dataset, Naïve Bayes algorithm is the least performed in all the metrics used for our classification. However, the evaluations revealed that the five algorithms are performing very well that give promising results from the individual dataset from the experiment carried out. Decision Tree classifier outperformed all other classifiers in two out of the three datasets tested for this experiment.

Evaluation of the Classification Data Mining Algorithm based on Recall, Precision, and F-Measure and Root Mean Square Error Metrics

The different types of evaluation metrics employed to evaluate the performance of the selected data mining classifiers are the commonly used evaluation metrics used with respect to intrusion detection. They are recall, precision, and f-measure and root mean square error metrics. The outcomes obtained after implementing the SVM, Naive Bayes, ANN, DTree and KNN technique as shown Tables 7, 8 and 9 regarding Precision, Recall RMSE and F-Measure respectively.

Table 7: Presents the evaluation of the Result of the simulation on NSL-KDD dataset

Data mining techniques	Precision	Recall	f-measure	RMSE
SVM	89%	95%	92%	0.3140
NAÏVE BAYES	86%	92%	89%	0.1447
KNN	98%	98%	98%	0.0531
DTREE	98%	99%	99%	0.0435
ANN	90%	98%	99%	0.0544

Table 7 shows classification result that is achieved by five classifiers in terms of performance measures including F-Measure, precision RMSE and recall. From this table, it can be observed clearly that in terms of Precision metrics, DTree and KNN perform better than the other techniques with precision of 98% each. Naïve Bayes is the least compared to the other techniques with precision of 86%. The DTree and ANN achieved maximum F-measure of 99% each followed by KNN with F-measure of 98% and the least performed is Naïve Bayes compared to the other techniques. The DTree had the minimum RMSE of 0.0435, followed by KNN and ANN with RMSE of 0.531 and 0.544 respectively. The DTree achieved the highest recall of 99%, while the KNN and ANN each achieved a recall of 98%. The SVM achieved 95% recall while Naïve Bayes achieved the least recall of 92%

Table 8: Presents the evaluation of the Result of the simulation on CICIDS2017 dataset

Data mining techniques	Precision	Recall	f-measure	RMSE
SVM	98%	99%	99%	0.2928
NAÏVE BAYES	99%	96%	97%	0.3348
ANN	99%	99%	99%	0.1231
DTREE	99%	99%	99%	0.0949
KNN	99%	98%	94%	0.2203

It can be observed form table 8 clearly that in terms of precision metrics, DTree, ANN, Naïve Bayes and KNN performs better than the other techniques with precision of 99% each. SVM is the least compared to the other techniques with precision of 98%. The DTree, SVM and ANN achieved maximum F-measure of 99% each followed by Naïve Bayes with F-measure of 97% and the least performed is KNN compared to the other techniques. The ANN had the minimum RMSE of 0.1231, followed by KNN and SVM with RMSE of 0.2203 and 0.2928 respectively. The DTree, SVM and ANN achieved the highest recall of 99%, while the Naïve Bayes achieved the least recall of 96%.

Table 9: Presents the evaluation of the Result of the simulation on KDD CUP'99 dataset

Data mining techniques	Precision	Recall	f-measure	RMSE
SVM	99%	99%	98%	0.1923
NAÏVE BAYES	98%	99%	99%	0.0369
ANN	99%	99%	99%	0.0287
DTREE	98%	99%	99%	0.0422
KNN	99%	99%	98%	0.0016

It can be observed from table 9 clearly that in terms of precision metrics, all the techniques performed well with precision almost the same precision, recall, and f-measure. However, the KNN had the minimum RMSE of 0.0016, followed by ANN with RMSE of 0.0287. The Naïve Bayes and DTree had RSME of 0.0369 and 0.0422 respectively while SVM had the highest RSME of 0.1923.

Conclusion

The aim of this study is to compare some selected data mining classification techniques that are used to detect intrusions inside three selected datasets. WEKA data mining tool was used to apply the data mining algorithms. The results obtained are analyzed and compared. From the analysis, the results revealed that the Decision Tree algorithm performs better with accuracy, precision, recall and F-measure of 99% than the other classifier of Support Vector Machine, ANN, NB and KNN in most of the tests on the three different datasets. Compared with other datasets that also applied the selected algorithm, both DT and ANN classifier showed superior performance in detecting attacks. Nevertheless, the results of the accuracy of almost all the algorithms that are used for detection were relatively good. However, in terms of implementation, Decision Tree classifier takes a very short time to be implemented unlike Multilevel Perceptron ANN that takes a very long time to be implemented. It is recommended that employing these optimizing techniques to develop an intrusion detection model have a better accuracy rate. This study has contributed to knowledge by revealing which features are most important to analyze and to detect any intrusions in each of the dataset used, and provided the knowledge of weaknesses and strengths for each technique used.

Recommendations

- Financial institutions handle large volumes of valuable data, making them attractive targets for cybercriminals. The approach can help them identify and respond to security breaches promptly.
- Security analysts, network administrators, and IT professionals can benefit from incorporating data mining techniques into their cyber security arsenal to analyze large amounts of data more effectively.

- Government organizations, such as defense departments, and intelligence agencies, deal with highly sensitive and critical information. This system can safeguard security and infrastructure.
- Educational Institutions often manage sensitive data related to students, staff, and research projects. Implementing data mining techniques for intrusion detection can protect their IT infrastructure from potential cyber threats
- Companies that provide online services, such as cloud storage, web hosting, rely on secure networks to protect their customers' data. They need this in monitoring and defending against potential attacks.
- Data mining techniques can offer them cost-effective ways to bolster their security posture to small and medium-sized enterprises.

References

- Adeleke, O. (2020, March). Intrusion detection: issues, problems and solutions. In *2020 3rd International Conference on Information and Computer Technologies (ICICT)* (pp. 397-402). IEEE.
- Bendovschi, A. (2015). Cyber-attacks–trends, patterns and security countermeasures. *Procedia Economics and Finance*, 28, 24-31.
- Carlin, A., Hammoudeh, M., & Aldabbas, O. (2015). Defence for distributed denial of service attacks in cloud computing. *Procedia computer science*, 73, 490-497.
- Cherian, V., & Bindu, M. S. (2017). Heart disease prediction using Naive Bayes algorithm and Laplace Smoothing technique. *Int. J. Comput. Sci. Trends Technol*, 5(2), 68-73.
- Dewa, Z., & Maglaras, L. A. (2016). Data mining and intrusion detection systems. *International Journal of Advanced Computer Science and Applications*, 7(1).
- Herr, T. (2014). PrEP: A framework for malware & cyber weapons. *Journal of Information Warfare*, 13(1), 87-106.
- Hindy, H., Brosset, D., Bayne, E., Seeam, A. K., Tachtatzis, C., Atkinson, R., & Bellekens, X. (2020). A taxonomy of network threats and the effect of current datasets on intrusion detection systems. *IEEE Access*, 8, 104650-104675.
- Jacobson, D., & Idziorek, J. (2016). *Computer security literacy: staying safe in a digital world*. CRC Press.
- Keegan, N., Ji, S. Y., Chaudhary, A., Concolato, C., Yu, B., & Jeong, D. H. (2016). A survey of cloud-based network intrusion detection analysis. *Human-centric Computing and Information Sciences*, 6(1), 1-16.
- Mohan, L., Jain, S., Suyal, P., & Kumar, A. (2020, September). Data mining Classification Techniques for Intrusion Detection System. In *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 351-355). IEEE.
- Ngai, E. W., Moon, K. L. K., Lam, S. S., Chin, E. S., & Tao, S. S. (2015). Social media models, technologies, and applications: an academic review and case study. *Industrial Management & Data Systems*.
- Patel, R., Thakkar, A., & Ganatra, A. (2012). A survey and comparative analysis of data mining techniques for network intrusion detection systems. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(1), 265-260.
- Parvathi, I., & Rautaray, S. (2014). Survey on data mining techniques for the diagnosis of diseases in medical domain. *International Journal of Computer Science and Information Technologies*, 5(1), 838-846.
- Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1, 108-116.

- Thakkar, A., & Lohiya, R. (2020). A review of the advancement in intrusion detection datasets. *Procedia Computer Science*, 167, 636-645.
- Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *Ieee Access*, 7, 41525-41550.
- Wahono, R. S. (2015). A systematic literature review of software defect prediction. *Journal of software engineering*, 1(1), 1-16.