Comparative Analysis of Intrusion Detection Models using Big Data Analytics and Machine Learning Techniques

Muyideen Ayodeji Alaketu
Department of Computer Science, Afe Babalola University, Nigeria
alaketumuyideen@gmail.com

Abiodun Oguntimilehin
Department of Computer Science, Afe Babalola University, Nigeria
ebenabiodun2@yahoo.com

Kehinde Adebola Olatunji
Department of Computer Science, Afe Babalola University, Nigeria
olatunjika@abuad.edu.ng

Oluwatoyin Bunmi Abiola
Department of Computer Science, Afe Babalola University, Nigeria
abiolaob@abuad.edu.ng

Bukola Badeji-Ajisafe
Department of Computer Science, Afe Babalola University, Nigeria
babukola@abuad.edu.ng

Christiana Olanike Akinduyite
Department of Computer Science, Afe Babalola University, Nigeria
akinduyiteco@abuad.edu.ng

Stephen Eyitayo Obamiyi
Department of Computer Science, Afe Babalola University, Nigeria
obamiyise@abuad,edu.ng

Gbemisola Olutosin Babalola
Department of Computer Science, Afe Babalola University, Nigeria
gbemibabz@abuad.edu.ng

Toyin Okebule
Department of Computer Science, Afe Babalola University, Nigeria
okebulet@abuad.edu.ng

Abstract: *Traditional cyber security measures are becoming less effective, leading to rise in modern attacks. However, the ability to analyze and use massive volume of data (big data) to train anomaly based systems that can learn from experience, classify attacks and make decisions can improve prediction of attacks before they actually occur. In this study, to ensure availability, integrity, and confidentiality of information systems, predictive models for intrusion detection that use Big Data and Machine Learning (ML) algorithms were proposed. The proposed approach used a big dataset (CIC-Bell-IDS2017) to independently train three ML classifiers before and after feature selection. Big data analytics tool was also employed for feature scaling and selection in order to normalize data and select the most relevant set of features. Performance evaluation and comparative analysis were done and the results showed there were improvements in the models' prediction accuracies.*

Keywords: *Cyber intrusion, intrusion detection system, machine learning, deep learning, ensemble learning, classifications, big data analytics.*

Received December 13, 2022; accepted June 19, 2023
https://doi.org/10.34028/iajit/21/2/14

1. Introduction

In this modern age where the use of information and computer technology is extremely important and constantly increasing, the significance of securing information systems, databases, and networks cannot be overemphasized, particularly as it contributes to our daily activities, security of personal information, improvement of national economy and facilitating successful business operations through digitalization, integration, and automation of processes [25]. Countries, businesses, organizations, and critical infrastructures now depend on information technology for their daily operations [46], leading to the rapid growth of information technologies, and also exposing information systems to different types of attacks and intrusions [52]. The effect of cyber-attack on information systems has increased globally by 900% over the last four years [23]. In 2020, Cybersecurity Ventures predict the cost of defending the cyberspace will increase by at least 15% annually, amounting to almost 11 trillion USD in 2025 [30]. Globally, Countries keep experiencing series of cyber warfare, such as the popular Stuxnet attack on Iranian uranium nuclear program in 2010 [9], Red October attack which targeted diplomatic, governmental and scientific agencies in 2012 [46] and more recently, the exfiltration attack and data theft of the North Atlantic Treaty Organization (NATO) confidential documents from the Portuguese Department of Defense System in September 2020, and the Iranian APT Log4Shell (Log4j) vulnerability attack against US federal agency's network in December 2021 [30].

Cyber-threat or cyber-intruder are human resources who seek to access networks and control systems without authorization and also steal information through various communication channels and protocols. They are capable of harming information systems, computers, and networks and at the same time obtaining illegal access to them. Cyber-attacks or cyber-intrusions on the other hand are intra-computer attacks that compromise the availability, integrity, and the confidentiality of networks, systems, and associated data [49]. They are mostly classified into different categories from different perspectives based on their architectural design and their effect on systems [11, 43] Malwares, Misuse of resources, Denial-of-Service (DoS), Web access compromise, User/Root access compromise, Advanced Persistent Threat (APT), Social engineering attacks, Phishing attack, and Zero-day attack [12, 16]. These security breaches and intrusions continue to increase because of the ability of modern attacks to keep evading traditional cybersecurity procedures and blacklist approaches [7], using highly sophisticated techniques and exploiting various new protocols mainly from the field of Internet of Things (IoT) to create zero-day attacks [37, 52]. However, ML technique primarily concerned with the design and development of algorithms that allow systems to improve by learning automatically from large-scale observation of data without being explicitly programmed [32, 50] can help provide more proactive predictive systems in and real-time [3, 42]. ML are generally divided into four categories based on the learning method and the data they learn from. Namely, reinforcement learning, unsupervised learning, supervised learning, and semi-supervised learning, [12, 20], which can all be used to develop predictive models for Intrusion Detection Systems (IDS) [5].

IDS are security systems (hardware or software) that monitor networks and/or hosts in order to identify and analyze computer systems or networks events, and automatically detect security issues, unauthorized access, attacks, or intrusions using data mining techniques, blacklist, and statistical approach to classify and detect anomalies [24, 36]. Generally IDS are grouped either by their approach (Network based or Host based), or technique (Anomaly-based, Signature-based, or Hybrid based IDS) [8, 19, 49]. The evolution of information technology over the last decade has led to the generation of huge amount of data from various sources and at high speed [36]. This data are referred to as big data, which can be employed for anomaly-based detection through Big Data Analytics (BDA) and data mining [28, 34]. The study of [28] defines big data in terms of five V's, representing, Volume: the size of data; Velocity: data streaming at unprecedented speeds; Variety: the different formats that data comes in, e.g. structured, unstructured, and semi-structured; Value: the value of new data added; Variability: constant change of data meaning, which was improved by [36] who described big data in terms of seven V's, adding Veracity: the trustworthiness of data; and Visualization: easy accessibility or readability of data. However, the definitions of big data are still counting till today.

On the other hand, BDA is the ability to analyze extensive data using existing BDA tools and platforms that allow a colossal amount of data that were formerly of no notable worth to be put into utmost use [12, 28], by using every bit of the behaviour and patterns to gain helpful insight [4]. This BDA platform can be employed for distributed processing of enormous data, advanced analytics, feature selection, data normalization, and preprocessing of big datasets faster and more efficiently [1], with different ML libraries and classifiers for analytics, mining, and classification task [14, 15].

2. Related Works

Xin *et al*. [51] presented a key literature survey and a brief explanation of different Deep Learning (DL) and ML classifiers commonly used for detecting intrusion in networks. The study reviewed seven classification algorithms. Namely, the Naive Bayes Classifier (NBC), convolutional neural network, Support Vector Machine (SVM), long short-term memory classifier, K-Nearest Neighbor, deep boltzmann machine, and Decision Tree (DT) classifier, and concluded that no particular classifier/algorithm could be ultimately selected over another, because each classifier has its merit and demerit, and its performance can also be influenced by the set of features and the training data size. In the study of Patel *et al*. [29], machine learning approach was used to develop an advanced predictive model for smart grid control system. SVM algorithm was employed for training, using binary classification (normal and attack) and 70:30 percentage split text with the highest portion for training. The Gaussian kernel and confusion matrix techniques were used for performance analysis and at the end of the experiment when compared with traditional IDPSs, the result showed an increase in the performance of the proposed system with higher detection accuracy and lower False Alarm Rate (FAR). The study of Wang and Jones [49] described the importance of data science particularly BDA in modern predictive models and systems, how the mining and analysis of huge data/big dataset are applied alongside ML classifiers to correlate information sources into knowledge that can be leverage against attacks, and finally achieve effective and efficient detection system. The study use BDA to analyse huge network data in other to gather valuable information and pattern, ML classifiers for data classification and clustering, and feature selection to improve classification accuracy. The result showed improvement in the performance of the model compared to older systems.

Sabnani [35] used two supervised learning algorithms- the Voting Frequency Intervals (VFI) based algorithm, and the naive bayes tree algorithm to create a predictive model for networks, employing feature selection and 10-fold cross validation on a fraction of the KDD 99 dataset. The algorithms were implemented in Weka environment for both the full and the reduced set of data. The system was evaluated using classification accuracy, precision, kappa statistic, f-measure and recall. The results of both algorithms were also compared, showing that Naive Bayes Tree produced higher detection accuracy of over 99.7% for the two set of data. Rai *et al*. [31] used the hybrid DT classifier to create a predictive model for IDS, employing information gain and segmentation technique to improve the model performance through feature selection and values segmentation. The author finally selected sixteen (16) features from the NSL-KDD dataset, however the experimental result only showed 79.52% prediction accuracy with low FAR. Krishnan and Raajan [22] developed a predictive model for network intrusion detection, the study used the Recurrent Neural Network (RNN) classification algorithm on the full Cup99 dataset for four-class classification -DoS, Probe, Root to location, and User to Root. The model performance was calculated using confusion matrix. The results were analysed showing that DoS, Probe, U2R, and R2L have overall accuracies of 97.4%, 96.6%, 86.5%, and 29.73% respectively. Vishwakarma *et al*. [47] propose a comparison between two intrusion detection models using both binary classification and multiclass classification on the fraction of Cup 99 dataset (NSL-KDD dataset). The study employed correlation based feature selection and the DT classification algorithm for modelling. However, in order to have a better result, the performance of the model was evaluate prior and after feature selection. The results obtained were studied and compared, showing that the overall accuracy for multi-class classification was beneath that of binary classification that generated an overall accuracy of 83.7% for the full set and 90.3% for the reduced, with a FAR of 2.5% and 9.7% respectively.

Mabayoje *et al*. [24] proposed a DT based IDS architecture that used Gain Ratio technique for feature selection and the KDD Cup99 dataset for training. The study employed 10-fold cross-validation technique for both full and reduced dataset, using multi-class classification method. For the full dataset the experiment generate a prediction accuracy of 100% for DoS attacks, 99.49% for probe attacks, 98% for Remote to Local attacks, and 75% for User to Root attacks. While for the reduced dataset the prediction accuracy for DoS attacks was 100%, probe attacks was 99.49%, Remote to Local attacks was 98%, and User to Root attacks was 75%. Using the subset of KDD Cup99 dataset, Vishwakarma *et al*. [47] developed a predictive model for cyber intrusion. In the study hybrid K-Nearest Neighbour classification algorithm known as Ant Colony Optimization (ACO) algorithm was used for attack classification. However, the model performance were evaluated using just two evaluation matrix -FAR and classification accuracy. At the end of the experiment, result showed a low FAR and a classification accuracy of 94.2%. Shakil and Farid [39] provided a detailed explanation of how feature selection can help improve model competence by significantly reducing the input features in the training set, and how the selected number of the features can affect the performance of an IDS models. The approach created three input feature subsets using the Correlation Based feature selection and the SVM classifier on the NSL-KDD dataset with the aim of knowing the number of feature that produce the best model. The experimental results showed that selecting 36 (thirty-six) features was as efficient as selecting 41 (forty-one) features, with both achieving 99% classification accuracy, while just 3 (three) features achieved 91% classification accuracy.

Kotpalliwar and Wajgi [21] presented an approach that used ten percent of the KDD Cup99 dataset and the mixed dataset, using SVM to train each of the datasets separately. In the study the model performance was evaluated using only classification accuracy. The 10 percent KDD Cup99 dataset produced a classification accuracy of 99.9%, while the Mixed datasets produced a classification accuracy of 89.85%. Sharifi *et al*. [40] developed two K-Nearest Neighbours (KNN) based intrusion detection models on the NSL-KDD dataset. The Principal Component Analysis technique was employed for features selection, picking only 10 features from the entire dataset. Furthermore, to evaluate and compare the models, the study implemented two case scenarios where the test data were excluded entirely from the training set, and where some test data were added to the training set, using only classification accuracy for the performance measure. At the end of the experiment both scenario produced an overall accuracy of 90%. Matin and Rahardjo [26] proposed an architectural design for predicting malware intrusion using honeypot for package data gathering (trapping), and ML classifier for data classification. The study proposed to use the SVM and the DT classification algorithms individually on the honeypot data to create a predictive model that distinguished between malware and good-ware, using 90:10 percentage split which will help achieve overall validation and also help produce high classification accuracy. Relang and Patil [33] proposed a predictive model for network intrusion, using two hybrid DT algorithms namely the C4.5 DT and the C4.5 DTWP (with pruning) on the KDD Cup 99 and NSL-KDD datasets. The study employed Information Gain technique for feature selection considering only the discrete features in the classification process. Precision and false alert rate were used to evaluate the performance of the models, showing that C4.5 DTWP has the best performance with 98.45% precision rate and 1.55% false alert rate compared to C4.5 DT that generated 89% precision rate and 6.9% false alert rate.

Saxena and Richariya [38], presented Information Gain and Binary Particle Swarm Optimization (BPSO) were employed to select 18 features from the KDD cup99 dataset. In the model, classification accuracy was used as evaluation matrix for multi-class classification. The result showed the classification accuracy for DoS, Probe, R2L and U2R were 99.4%, 99.3% and, 98.7%

and 98.5% respectively. Moore *et al.* [27] presented an approach that combined classification and feature reduction for cyber network threat detection, using the Artificial Neural Network (ANN) classification algorithm on the Department of Defense Cyber-Defense Exercises (CDX) network traffic data. After non-salient features was removed using signal-to-noise ratio, feature extraction was used to extract 248 features, which was later reduced to 18 features. Various features were considered in analysing the data, and the result showed that 18 features was efficient enough to train the model by generating 97.29% accuracy and low FAR of 2.71%, while the entire 248 features generated 82.56% accuracy higher FAR. Umara *et al*. [45] proposed an architectural framework that extracted threat actor's pattern and profile from cyber threat intelligence report in other to understand their attributes. Specifically, they collected over three hundred unstructured CTI reports from different sources, which was cleaned and analysed to create a CTI report dataset used with the public CTA dataset to individually train the model using the Naive Bayes, the Deep Neural Network (DNN) and the Random Forest (RF), classification algorithms with and without feature selection. The result of the experiment showed that the DNN achieve the best overall performance in both cases, followed by the RF and the NBC, and that feature selection did not improve the overall accuracy of any of the models. Furthermore, it was also observed that the CTI report dataset obtained product higher precision than the publicly available CTA profile database.

This section describes the techniques, classifiers, dataset and evaluation metrics used in the reviewed literature. However, the review shows that most of the earlier studies attempted to build an intrusion detection model using datasets which cannot be classified as a big data, and without employing BDA tools. It was also noticed from the review that rather than utilizing distinct, newer, bigger or more recent intrusion datasets, the majority of the reviewed studies employed the use of the same dataset, mostly KDD Cup99 and its subset.

In this study, the proposed model used a more modern big dataset (CIC-Bell-IDS2017), BDA tool (PySpark) and several ML classifiers for the creation of intrusion detection models. PySpark-an analytics tool for big data was used on CIC-Bell-IDS2017 big dataset to handle several complications that come with analysing the data, such as sheer volume, velocity, variety, scalability, and data complexity, which were not mentioned in the reviewed studies.

This work shows the efficacy of a big data tool in increasing the performances of the models generated from CIC-Bell-IDS2017 dataset. It exposes readers to the fact that performances of DNN and other modern predictive techniques can be improved using BDA tool (PySpark). Our approach on the dataset (CIC-Bell-IDS2017) has not been used by anyone, and can serve as a future reference or benchmark.

3. Methodology

This part of the study explain the technique, dataset and the proposed system architecture used in stages. It also describes the implementation of the ML classifier (RF, SVM, and DNN) used to classify normal and attack traffic in networks. The Proposed System Architecture is shown in Figure 1.

- *Stage* 1: data collection and description of dataset the proposed method used a big dataset from the Canadian Institute of Cybersecurity (CIC) and Bell Canada (BC) Cyber Threat Intelligence (CTI) otherwise referred to as the CIC- Bell-IDS2017 dataset, which is an IDS dataset that aims to address the shortcomings of previous intrusion detection datasets such as lack of traffic diversity, traffic volume, failure to cover a broad range of known attacks, failure to reflect current trends, as well as lack of feature set and metadata, among other things. CIC-IDS2017 was published in 2017 and can be found at https://www.unb.ca/cic/datasets/ids-2017.html. The dataset entails 79 features, and millions of benign (goodware) and malicious samples of common and modern intrusion attacks, namely Dos, DDos, Brute force, Infiltration, SSH-Patator, FTP-Patator, Heart-bleed, Port Scan, Cross-site Scripting (XSS), and SQL Injection. The distribution of attacks in the dataset is described in Table 1 of this study.

The key characteristics of the CIC-Bell-IDS2017 dataset include size and scope of data, data labelling, traffic volume, number of traffic features considered, and variety of attacks type with millions of individual flow records making it suitable for predicting a wide range of network intrusion scenarios and for assessing the effectiveness of IDSs against different attack vectors.

CIC-Bell-IDS2017 dataset provides a rich set of features and attributes extracted from real-life network traffic data and generated from a real network environment. These features capture information related to network traffic including flow-based and host-based features such as network traffic patterns, communication protocols, source and destination IP addresses, port numbers, packet sizes, packet length, flags, flow duration, IAT and lot more.

However, while the dataset offers many advantages, the potential limitation, include handling null and infinite values, dealing with data/class imbalance (where benign traffic vastly outweighs malicious traffic), and selecting relevant features for analysis.

Table 1. Distribution of selected attacks.

S/N	ATTACK TYPE	COUNT
0	BENIGN (GOOD WARE)	149113
1	DENIAL OF SERVICE (DoS HULK)	22,380
2	DENIAL OF SERVICE (GOLDEN EYE)	10,293
3	DISTRIBUTED DENIAL OF SERVICE (DDoS)	20,317
4	PORT SCAN	11,002
5	FTP-PATATOR	7,938
6	SSH-PATATOR	5,897
7	BRUTE FORCE (WEB ATTACK)	1,470
8	CROSS SITE SCRIPTING (WEB ATTACK)	652
9	INFILTRATION	36
10	SQL INJECTION (WEB ATTACK)	21
11	HEARTBLEED	11

- *Stage* 2: data cleaning large, redundant, and unprocessed data make up the majority of big dataset, therefore posing significant obstacles to knowledge discovery and data modeling. The act of cleaning data is the remover and modification of incorrect, incomplete, irrelevant, infinite, and wrongly formatted data before analysis. The cleaning operations done in this study involved dropping unwanted column, removing infinite values, replacing whitespace, calculating statistical distribution, data balancing (to avoid over fitting), and label encoding for binary classification. All attacks were grouped as Attack and encoded as “1”, and good-wares were grouped as Benign and encoded as “0”.
- *Stage* 3: development of the predictive models for cleaned Dataset At this stage, three ML classifiers, namely: SVM, which makes prediction creating boundaries between classes; DNN, which is based on human brain biological neural; and RF which is based on creating multiple Decision Trees were separately used on the cleaned dataset to generate three different predictive models using 80:20 split test (where 80% was used to train and 20% to test the model). The models were then evaluated using confusion matrix to determine the accuracy, precision, recall, and F-1 score of each model. Furthermore, model development time was evaluated to determine its usability in real-time.
- *Stag*e 4: data Normalization and Feature Selection At this stage, considering the size of the dataset, BDA tool (PySpark) was used for data normalization and feature selection. Data normalization also referred to as feature scaling was used to correct the compromises that come with handling features with similar and drastically different scale, which is a common attribute of big data.

Correlation-based feature selection that uses correlation analysis technique to check the level of similarity between input features and the relationship each has with the output feature and then select the subset of features that have a high correlation with the target variable (the variable you want to predict) and low correlation among themselves (the feature set) was employed for selecting twenty-four (24) most relevant features out of the 79 (seventy-nine) features in the dataset.

In this study the author calculate the correlation between each feature in the dataset and the target variable. For each feature, the study obtain a correlation value that represents its relationship with the target variable. Features with higher correlation values to the target variable were considered to be more relevant as they are more likely to contain information that is useful for making predictions, while features that are highly correlated with each other are dropped as they are redundant and provide similar information which can lead to overfitting and increase the model complexity. Pearson correlation coefficient presented in Equation (1) was used to calculate the correlation between two variable x and y.

$$r = \frac{\Sigma[(x-\bar{x})(y-\bar{y})]}{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2} \tag{1}$$

Where r is the correlation coefficient

x and y is the value of x-variable and y-variable in a sample

$\bar{x}$ is mean of x,

$\bar{y}$ is the mean of y.

NOTE:

a. if r is between 0.6 and 1 then a Positive correlation exist
b. if r is between -0.6 and -1 then a Negative correlation exist
c. if r is 0 then no correlation whatsoever (Neutral)
d. if r is closer to 0 than 1 (<=0.5) then weak correlation exit

Correlation-based feature selection typically uses a search algorithm such as forward selection, backward elimination, or a heuristic search) to search for the optimal feature subset, and then select the best feature subset.

- *Stage* 5: Development of the Predictive Models for Reduced Dataset At this stage of this work, after big data analytic tool has been used for feature selection and data normalization, SVM, DNN, and RF (ensemble learning technique), were again used separately to train the reduced dataset, in order to generate three different predictive models, using 80:20 split test. The new models were also evaluated using standard metrics. A comparative analysis between the two sets of models (full and reduced) was then carried out.

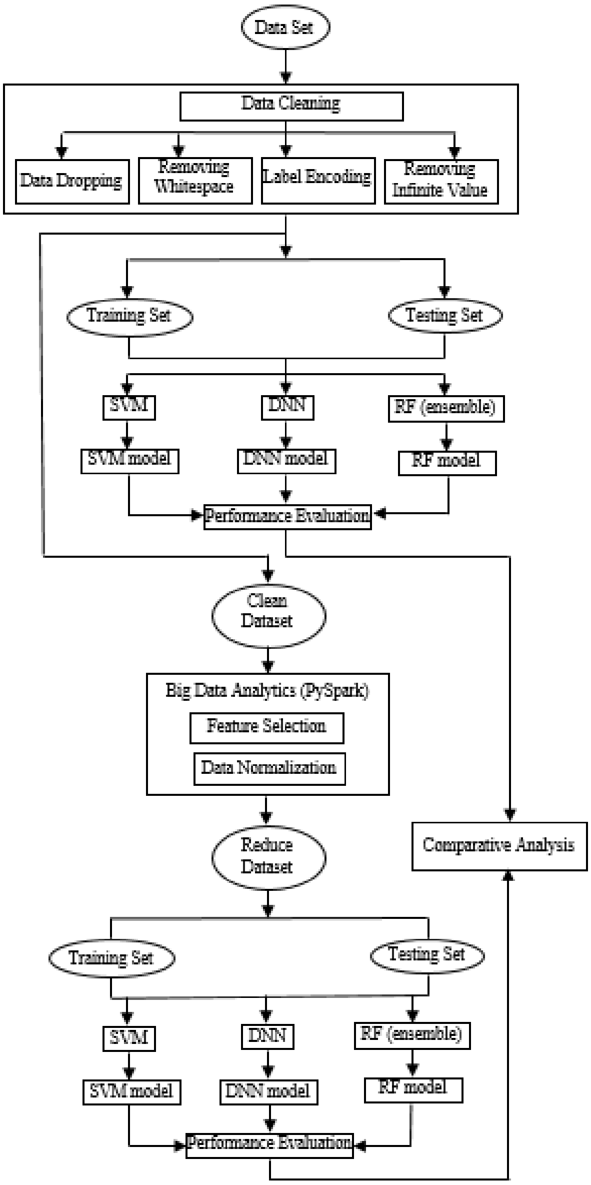


Figure 1. Proposed system architecture.

3.1. System Performance Evaluation Metrics

Confusion matrix which uses true detection and false detection rate to calculate IDS model performance-accuracy, precision, recall, and F-1 score [35, 52] was employed for performance evaluation. Furthermore [22] identified that accuracy alone can be bias in respect to the size of the dataset, therefore using precision and recall which are not can help produce a more suitable evaluation process. According to [26] Table 2 represent a simple confusion matrix for binary classification.

Table 2. Confusion matrix for a binary class problem.

PARAMETER	PREDICTION	
(REAL LIFE)	ATTACK	BENIGN
ATTACK	True Positive	False Positive
BENIGN	False Negative	True Negative

Where True Positive (TP) is the correctly predicted Attack.

True Negative (TN) is the correctly predicted good-ware (Benign).

False Negative (FN) are Attack that failed to be identified, or predicted as good-ware (Benign).

False Positive (FP) are good-ware (Benign) that failed to be identified, or predicted as Attack.

a) Accuracy: This is the most basic matrix of measuring ML model performance. It determines the percentage of correctly classified instances over the total number of instances [19], by gives the ratio of TPs and TNs to the total number of instances [35]. The formula for calculating accuracy is given in Equation (2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

b) Precision: this is the percentage of correctly predicted Attacks to all samples predicted as Attack [52], it is the percentage ratio of the number of TPs records divided by the sum of TPs and FPs (FP) classified records [22]. Calculate as presented in Equation (3).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

c) Recall: this is the percentage of all samples correctly classified as Attacks to all samples that are truly Attacks [52]. Calculated as presented in Equation (4).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

d) F Score: this is defined as the weighted harmonic mean of recall and precision [3]. It represents a balance between both, and helps to address any classification problems [19, 35]. Calculated as presented in equation (5).

$$F\text{-}measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

e) Model Development Time: model development time is also another performance measure that is used for evaluating ML performance, as it has considerable effect in real-time environment [24]. Model development time shows the time taken to train a particular algorithm with a set of training data, and to build a model in respect to the dataset. This is calculated to help show the usability of the model in real-time.

3.2. Description of Machine Learning Algorithms and Data Analytics Tool Explored

The description of the ML algorithms and BDA tool used are presented in this section.

3.2.1. Support Vector Machine

The SVM is a ML classification algorithm that uses non-random criteria to divide samples into classes, with the intension of increasing the hyperplane (the distance between each group of class), and therefore solving the problem of over-fitting [29]. The SVM is seen as the most reliable and accurate supervised ML classification algorithms by several authors [51]. It is generally based on the idea of decision boundaries and risk minimization principle which separate different instances into greater than two classes for multi-class classification task and two classes for binary classification task [2], as shown in Figure 2. It works based on the formula presented in Equation (6)

$$MINIMIZE a_{0,\dots,}a_m\colon \sum_{j=1}^{n} MAX\ \{0,\ 1 - (\sum_{i=1}^{m} a_i\, x_{ij} + a_o)\ y_j\} + \lambda \sum_{i=1}^{m}(a_i) \qquad (6)$$

Where *m* is attributes number**,** *n* is data points number**,**
x_{ij} is the i^{th} attribute of j^{th} data point,
y_j is the boundary, which is *1* at one data point, and -1 at the other data point.

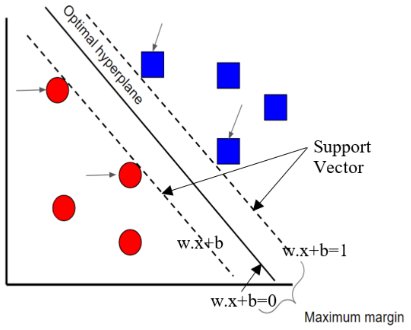


Figure 2. Diagrammatic and mathematical representation of SVM [48].

3.2.2. Deep Neural Network (DNN)

DNN is a DL algorithm, a significant subset ML that simulate he ideas of the human brain biological neural network. DNN builds security models by accepting input and using different layers of interconnected neurons to process them [37]. It has currently gained wide spread recognition in ML models and for network-based anomaly detection, because they are scalable, they mostly outshined traditional ML methods, they allow usability of unstructured data and can also handle complex operations [51]. According to [44], DNN is more efficient in terms of performance than most traditional ML classifiers because it trains a model using complex algorithms and DNNs. Figure 3 shows how DNN can be represented mathematically.

DNN can be represented mathematically using matrix representation presented in Equation (7).

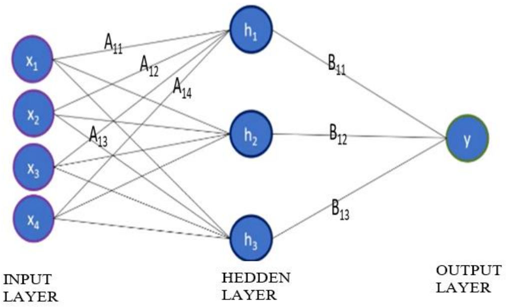


Figure 3. DNN with multiple variable and hidden layer [47].

1. From input node x, three set of linear equation were generated using class weight A.

$$\begin{aligned} h_1 &= A_{11\,x_1} + A_{12\,x_2} + A_{13\,x_3} + A_{14\,x_4} \\ h_2 &= A_{21\,x_1} + A_{22\,x_2} + A_{23\,x_3} + A_{24\,x_4} \\ h_3 &= A_{31\,x_1} + A_{32\,x_2} + A_{33\,x_3} + A_{34\,x_4} \end{aligned} \qquad (7)$$

2. This set of linear equation can be combined into a single matrix, represented as:

Matrix vector $[h_n]$ = matrix weight $[A_{nn}]$ multiply by input vector x_n.

$$[h_n] = [\ A_{nn}\ ] * [\ x_n\ ] \qquad (8)$$

3. Vector h, is equal to the matrix A, multiply by the input vector x**:** $h = Ax$, then h is given to a function of Ax, represented as:

$$h = \mathcal{F}_1(Ax) \qquad (9)$$

4. For output value y using the same approach, y is calculated as: $y = Bh$, then given to a function of Bh, represented as:

$$y = \mathcal{F}_2(Bh) \qquad (10)$$

5. Substituting Equation (9) into Equation (10), we have the output

$$y = \mathcal{F}_2(B\mathcal{F}_1(Ax)) \qquad (11)$$

Where y is output, h is the hidden node (learner), x is the input variable, A and B are the class weight, $\mathcal{F}_1$ and $\mathcal{F}_2$ are activation functions.

3.2.3. Random Forest

RF is a class of ensemble that create multiple decision trees in parallel during training time, and then select the class chosen by the most trees as the output for classification problems using bagging technique and random selection of features. It is an enhanced form of the DT that predicts future occurrences with multiple classifiers rather than a single classifier, to improve prediction accuracy and correctness [6]. RF performs well in variety of predictive modeling situations and it is a variation of bagging that chooses subsets of characteristics in each data sample at random. Using bagging, each DT in the ensemble forest is constructed using a sample with replacement from the training dataset [18]. This approach employs randomization for picking the optimum node to split during modeling, which is equal to the root of the number of features in the data set. This technique uses the class and probability to determine the Gini index (a function used to measure the impurity of data and uncertainty of event) of each branch on a node, determining which of the branches is more likely to occur [13]. Calculated as presented in Equation (12).

$$Gini = 1 - \sum_{i=1}^{N}(P_i)^2 \qquad (12)$$

Where P_i represents the relative frequency of the class, *N* is the number of classes in the data set, and *i* is the i^{th} class label in the data set.

3.2.4. Big Data Analytics Tool

This study used PySpark which is a standalone framework for Apache Spark. PySpark is a collaborative application programming interface (API) that was released to support Python and Spark for real-time and large-scale data processing [15]. Through PySpark the Apache Spark framework provides fast advanced analytics libraries that can help analyze and preprocess massive dataset with hundreds of thousands if not millions of records more effectively. These libraries can be used for several machine leaning task, such as data preprocessing, feature scaling, and feature selection which are all important stages in developing ML models, since data quality can affect performance [32]. The ability to analyze and process big data for knowledge discovery, the ability to access multiple libraries, the ability to develop more scalable analytic pipelines, and the compatibility with other libraries such as Sklearn and Pandas, are part of the benefits of utilizing PySpark.

4. Implementation and Results

4.1. Model Development

The experiment was done in Jupyter Notebook Conda environment, which is a python web-based interactive graphic user programing interface and development environment for data mining, machine leaning and advance analytics libraries and tools: PySpark, MLLib, SkLearn, Tensorflow, Pandas, Numpy, etc., within a single python programming interface.

4.2. Experimental Results

This section shows the performance of SVM, DNN and RF for the task of detecting intrusions on both full and reduced CIC-IDS2017 dataset, as shown in Table 3.

Table 3. Performance results and comparative study of the three classifiers for both full and reduced set.

CLASIFIERS		Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score	Model Development Time
SVM	Full set	59%	66%	66%	100%	79%	3751sec
	Reduced set	72%	75.7%	73%	100%	84%	2067sec
DNN	Full set	80.2%	65.1%	65%	100%	100%	227sec
	Reduced set	85.5%	85.7%	84%	97%	100%	200sec
RF	Full set	95%	90.6%	89%	97%	93%	93sec
	Reduced set	89%	90.9%	89%	98%	93%	43sec

4.3. Discussion of Results

From the experiment results, it shows that there were improvements in the performances of the three classifiers after BDA tool (PySpack) was used for feature scaling and selection in all metrics of evaluation: Accuracy, Precision, Recall, F1-score, and Model Development Time, except for DNN's recall that slightly reduced by 3% and RF training accuracy that reduced by 6%.

The accuracies of SVM and DNN were significantly improved after BDA was used, with SVM testing and training accuracies increasing from 66% to 75.7% and 59% to 72% respectively while DNN testing and training accuracies increase from 65.1% to 85.7% and 80.2% to 85.5% respectively. Furthermore, with the ensemble learning ability of the RF classifier, and the reduction in its training accuracy, the result still shows an increase in the testing accuracy of RF, which is a shred of evidence that even the much celebrated ensemble learning algorithm can be improved with big data analytics.

The best testing accuracy for both the full and the reduced dataset, was produced by the RF classifier, achieving 90.6% and 90.9% respectively. RF also produced the best precision and model development time in both cases, with precision of 89% for both set, and model development time of 93sec and 43sec respectively, which is important in using ML model in a real-time. SVM produced the best recall of 100% for both set of data, which is the same for DNN algorithm with 100% recall for the full dataset and 97% for the reduced dataset. DNN also produced 100% F1-Score for both full and reduced dataset. Even though the RF classifier produced the best accuracy, precision, and model development time for both sets of the data, however, considering the improvement done by using BDA (PySpark) for feature scaling and feature selection, DNN showed the best improvement with 21% increase in accuracy, followed by SVM with 10% increase in accuracy, and then RF with 0.3%.

5. Conclusions, Recommendations and Challenges

The security industry has been highly criticized recently based on the fact that traditional security solutions have failed over and over again to detect sophisticated and new attacks. However, Anomaly Intrusion Detection System (AIDS) that uses ML technique is an effective method in detection of these attacks, as shown in this work. The results of the three classification algorithms: RF, SVM and DNN for both full and reduced datasets were analyzed towards their suitability for detecting intrusions from a large dataset containing multiple modern attacks.

This study has shown how BDA can be used to enhance the prediction capability and improve the performance of ML models, even on ensemble model thereby leading to a better IDS.

In this study, predicting intrusions using ML and big data presents several challenges, obstacles, and complexities due to the dynamic nature of the problem and the big data involved. The challenges include dealing with class imbalance or imbalanced datasets (when the number of good-ware is significantly higher than the number of attacks), handling missing and infinite values, and selecting the best set of relevant features for analysis. Another major challenge came with the size of the big dataset used, which was impossible to model without BDA tool (PySpark) and a very powerful system's processor.

6. Acknowledgments

The authors are grateful to Afe Babalola University, Ado-Ekiti, Nigeria for supporting this research.

7. Conflict of Interest

The authors declare that there is no known competing financial or personal interests to declare.

8. Future Directions

Specific avenues for improvement or exploration include:

a. The use of different classification algorithms which might have positive influence on the dataset, and helps to produce models with better performances.
b. The feature selection technique used in this study was based on correlation analysis. However, it is possible to conduct future experiments where different a feature selection or reduction technique will be used, which will most likely select a different set of features.
c. The use of ensemble method to assemble the three algorithms, might generate a single better model.
d. The use of Apache Spark for distributed processing, to make use of the entire CIS-IDS2017 dataset and more dataset collectively for training, which might improve the performance of the model.

References

[1] Ahmad Z., Khan A., Shiang C., Johari A., and Farhan A., "Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32 no. 1, pp. 1-29, 2021. https://doi.org/10.1002/ett.4150

[2] Ali A. and Abdullah M., "A Parallel Grid Optimization of SVM Hyperparameter for Big Data Classification using Spark Radoop," *Karbala International Journal of Modern Science*, vol. 6, no. 1, pp. 7-18, 2020. DOI: 10.33640/2405-609X.1270

[3] Al-Shamery A. and Al-Shamery E., "Prediction of Iraqi Stock Exchange using Optimized Based-Neural Network," *Karbala International Journal of Modern Science*, vol. 7 no. 4, 2021. https://doi.org/10.33640/2405-609X.3159

[4] Angin P., Bhargava B., and Ranchal R., "Big Data Analytics for Cyber Security," *Journal of Security and Communication Networks*, vol. 2019, no. 4109836, pp. 1-2, 2019. Doi: 10.1155/2019/4109836

[5] Apruzzese G., Colajanni M., Ferretti L., Guido A., and Marchetti M., "On the Effectiveness of Machine and Deep Learning for Cyber Security," *in Proceedings of the 10th International Conference on Cyber Conflict*, *IEEE Access Journal*, Tallinn, pp. 371-389, 2018. DOI: 10.23919/CYCON.2018.8405026

[6] Balyan A., Ahuja S., Lilhore U., Sharma S., Manoharan P., and Algarni A., "A Hybrid Intrusion Detection Model Using Ega-PSO and Improved Random Forest Method," *MDPI Sensors Journal*, vol. 22 no. 16, pp. 59-86, 2022. https://doi.org/10.3390/s22165986

[7] Barriga J. and Yoo S., "Malware Detection and Evasion with Machine Learning Techniques: A Survey," *International Journal of Applied Engineering Research*, vol. 12, no. 18, pp. 7207-7214, 2017. https://api.semanticscholar.org/CorpusID:31201315

[8] Beigh B. and Peer M., "Intrusion Detection and Prevention System: Classification and Quick Review," *Journal of Science and Technology*, vol. 2, no. 7, pp. 661-675, 2012. https://api.semanticscholar.org/CorpusID:15921252

[9] Buchanan S., "Cyber-Attacks to Industrial Control Systems since Stuxnet: A Systematic Review," Ph.D Thesis, Capitol Technology University ProQuest Dissertations Publishing, 2022.

[10] *CCNA, Introduction to Cybersecurity*, Cisco Networking Academy (NetAcad), 2018.

[11] Dasgupta D., Akhtar Z., and Sen S., "Machine Learning in Cybersecurity: A Comprehensive Survey," *Journal of Defense Modeling and Simulation*, vol. 19, no. 1, pp. 1-50, 2020. https://doi.org/10.1177/1548512920951275 doi: 10.1109/CSNT.2015.185

[12] Duc T., Leiva R., Casari P., and Ostberg P., "Machine Learning Method for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey," *Association for Computing Machinery Computing Surveys Journal*, vol. 52, no. 5, pp. 1-39, 2019. https://doi.org/10.1145/3341145

[13] Fawagreh K., Gaber M., and Elyan E., "Random Forests: From Early Developments to Recent Advancements," *Systems Science and Control Engineering Journal*, vol. 2, pp. 602-609, 2014. DOI: 10.1080/21642583.2014.956265

[14] Hariri R., Fredericks E., and Bowers K., "Uncertainty in Big Data Analytics: A Survey Opportunities, and Challenges," *Journal of Big Data*, vol. 6, no. 1, pp. 44-46, 2019. https://doi.org/10.1186/s40537-019-0206-3

[15] Harshal K., Phadnis M., Chittar P., Zarkar K., and Bodhke B., "A Review of Data Analysis and Visualization Of Olympics Using PySpark and Dash-Plotly," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 4, no. 6, pp. 2093-2097, 2022.

[16] Haseeb K., Jan Z., Alzahrani F., and Jeon G., "A Secure Mobile Wireless Sensor Networks Based Protocol for Smart Data Gathering with Cloud," *Computers and Electrical Engineering Journal*, vol. 97, pp. 1075-1084, 2022. https://doi.org/10.1016/j.compeleceng.2021.107584

[17] Ingre B., Yadav A., and Soni A., "Decision Tree Based Intrusion Detection System for NSL-KDD Dataset," *in Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems*, pp. 207-218, Ahmedabad, 2017. https://doi.org/10.1007/978-3-319-63645-0_23

[18] Jain, V., Machine Learning Khanna, Publishing House, 2018. https://doi.org/10.3390/s22165986

[19] Javaid A., Niyaz Q., Sun W., and Alam M., "A Deep Learning Approach for Network Intrusion Detection System," *in Proceedings of the 9th European Alliance for Innovation International Conference Endorsed Transactions on Security and Safety*, Braga, pp. 3-12, 2016.

[20] Khraisat A., Gondal I., Vamplew P., and Kamruzzaman J., "Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges," *Cybersecurity Journal*, vol. 2 no. 1, pp. 1-22, 2019. DOI: 10.1186/s42400-019-0038-7

[21] Kotpalliwar M. and Wajgi R., "Classification of Attacks Using Support Vector Machine on KDD Cup 99 IDS Database," *in Proceedings of the 5th International Conference on Communication Systems and Network Technologies*, Gwalior, pp. 987-990, 2015.

[22] Krishnan R. and Raajan N., "An Intellectual Intrusion Detection System Model for Attacks Classification Using RNN," *International Journal of Pharmacy and Technology*, vol. 8, no. 4, pp. 23157-23164, 2016.

[23] Lornov K., Applying Emerging Data Techniques and Advanced Analytics to Combat Cyber Threat, Master's Thesis, African University of Science and Technology Abuja, 2017.

[24] Mabayoje M., Abimbola A., Balogun A., and Opeyemi A., "Gain Ratio and Decision Tree Classifier for Intrusion Detection," *International Journal of Computer Applications*, vol. 126, no. 1, pp. 56-59, 2015. DOI: 10.5120/ijca2015905983

[25] Marjani M., Fariza N., Gani A., Karim A., Hashem I., and Siddiqa A., "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges," *IEEE Access Journal*, vol. 5, pp. 5247-526, 2017. doi:10.1109/ACCESS.2017.2689040

[26] Matin I. and Rahardjo B., "Malware Detection Using Honeypot and Machine Learning," *in Proceedings of the 7th International Conference on Cyber and IT Service Management*, Brisbane, pp. 1-4, 2019.

[27] Moore K., Bihl T., Bauer K., and Dube T., "Feature Extraction and Feature Selection for Classifying Cyber Traffic Threats," *Journal of Defense Modeling and Simulation-Applications, Methodology and Technology*, vol. 14, no. 3, pp. 217-231, 2017. https://doi.org/10.1177/1548512916664032

[28] Oguntimilehin A. and Ademola O., "A Review of Big Data Management, Benefits and Challenges," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 5 no. 6, pp. 433-438, 2014.

[29] Patel A., Alhussian H., Pedersen J., Bounabat B., and Júnior J., "A Nifty Collaborative Intrusion Detection and Prevention Architecture for Smart Grid Ecosystems," *Computers and Security Journal*, vol. 64, pp. 92-109, 2017. https://doi.org/10.1016/j.cose.2016.07.002

[30] Rabia A., Aftab H., Sharma P., and Kumar P., "Machine Learning-Based Soft Computing Regression Analysis Approach for Crime Data Prediction," *Karbala International Journal of Modern Science*, vol. 8 no. 1, pp. 1-19, 2022. https://doi.org/10.33640/2405-609X.3197

[31] Rai K., Devi M., and Guleria A., "Decision Tree Based Algorithm for Intrusion Detection," *International Journal of Advanced Networking and Applications*, vol. 7, no. 4, pp. 2828-2834, 2016.

[32] Ranganathan G., "Real Time Anomaly Detection Techniques Using Pyspark Frame Work," *Journal of Artificial Intelligence and Capsule Networks*, vol. 2, no. 1, pp. 20-30, 2020. DOI:10.36548/jaicn.2020.1.003

[33] Relang N. and Patil D., "Implementation of Network Intrusion Detection System Using Variant of Decision Tree Algorithm," *in Proceedings of the International Conference on Nascent Technologies in Engineering* Navi Mumbai, pp. 1-5, 2015. doi: 10.1109/ICNTE.2015.7029925.

[34] Rizvi S., Labrador G., Guyan M., and Savan J., "Advocating for Hybrid Intrusion Detection Prevention System and Framework Improvement," *Procedia Computer Science*, vol. 95, no. 1, pp. 369-374, 2016. https://doi.org/10.1016/j.procs.2016.09.347

[35] Sabnani S., Computer Security: A Machine Learning Approach: Master's Thesis, Department of Mathematics, Royal Holloway University, 2008.

[36] Saranya T., Sridevi S., Deisy C., Chung T., and Khan M., "Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review," *Procedia Computer Science Journal*, vol. 171, pp. 1251-1260, 2020. https://doi.org/10.1016/j.procs.2020.04.133

[37] Sarker I., Kayes A., Badsha S., Alqahtani H., and Watters P., "Cybersecurity Data Science: An Overview from Machine Learning Perspective," *Journal of Big Data*, vol. 7 no. 41, pp. 2-29, 2020. https://doi.org/10.1186/s40537-020-00318-5

[38] Saxena H. and Richariya V., "Intrusion Detection in KDD99 Dataset Using SVM-PSO and Feature Reduction with Information Gain," *International Journal of Computer Application*, vol. 98 no. 6, pp. 25-29, 2014. DOI: 10.5120/17188-7369

[39] Shakil P., and Farid D., "Feature Selection and Intrusion Classification in NSL-KDD Cup 99 Dataset Employing SVMs," *in Proceedings of the 8th International Conference on Software, Knowledge, Information Management and Applications*, Dhaka, pp. 1-6, 2014.

[40] Sharifi A., Kasmani S., and Pourebrahimi A., "Intrusion Detection Based on Joint of k-Means and KNN," *Journal of Convergence Information Technology*, vol. 10, no. 5, pp. 42-51, 2015.

[41] Siddiqi M., Mugheri A., and Oad K., "Advance Persistent Threat Defense Techniques: A Review," *Pakistan Journal of Computer and Information Systems*, vol. 1, no. 2, pp. 53-65, 2016. http://142.54.178.187:9060/xmlui/handle/123456789/826

[42] Suhaimi N. and Abas H., "A Systematic Literature Review on Supervised Machine Learning Algorithms," *Perintis eJournal*, vol. 10, no. 1, pp. 1-24, 2020.

[43] Tahir R., "Study on Malware and Malware Detection Techniques," *International Journal of Education and Management Engineering*, vol. 8, no. 2, pp. 20-30, 2018. DOI:10.5815/ijeme.2018.02.03

[44] Tang T., Mhamdi L., McLernon D., Zaidi S., and Ghogho M., "Deep Learning Approach for Network Intrusion Detection in Software Defined Networking," *in Proceedings of the International Conference on Wireless Networks and Mobile Communications,* Fez, pp. 258-263, 2016.

[45] Umara N., Anwar Z., Tehmina A., and Choo K., "A Machine Learning-Based Fintech Cyber Threat Attribution Framework Using High-Level Indicators of Compromise," *Future Generation Computer Systems Journal*, vol. 9, no. 6, pp. 227-242, 2019. https://doi.org/10.1016/j.future.2019.02.013

[46] Virvilis-Kollitiris N., Detecting Advanced Persistent Threats through Deception Techniques, Ph.D. Thesis, Athens University of Economics and Business, 2015. https://www.infosec.aueb.gr/Publications/Virvilis-Kollitiris%20Dissertation%20Text.pdf

[47] Vishwakarma S., Sharma V., and Tiwari A., "An Intrusion Detection System Using KNN-ACO Algorithm," *International Journal of Computer Application*, vol. 171, no. 10, pp. 18-23, 2017. DOI:10.5120/ijca2017914079

[48] Wang L. and Alexander C., "Big Data in Distributed Analytics, Cybersecurity, Cyberwarfare and Digital Forensics," *Journal of Science and Education Publishing,* vol. 1, no. 1, pp. 22-27, 2015. doi: 10.12691/dt-1-1-5

[49] Wang L. and Jones R., "Big Data Analytics for Network Intrusion Detection: A survey," *International Journal of Networks and Communications*, vol. 7, no. 1, pp. 24-31, 2017. doi:10.5923/j.ijnc.20170701.03

[50] Wolfgang E., *Introduction to Artificial Intelligence*, Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-58487-4

[51] Xin Y., Kong L., Liu Z., Chen Y., and Li1 Y., "Machine Learning and Deep Learning Methods for Cybersecurity," *IEEE Access Journal,* vol. 6, pp. 35365-35381, 2018. doi: 10.1109/ACCESS.2018.2836950.

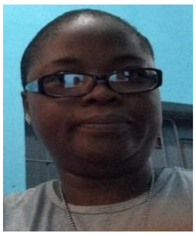
[52] Zeeshan A., Khan A., Shiang C., Johari A., and Ahmad F., "Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches," *Journal of Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, pp. 1-29, 2020. DOI:10.1002/ett.4150

Muyideen Ayodeji Alaketu obtained B.Sc Computer Science from the University of Jos, Jos-Plateau, Nigeria and M.Sc Computer Science form the Afe Babalola University, Ado-Ekiti, Nigeria. He is at present an Information Technology expert and his research interests include Cybersecurity, Data Science, and Machine Learning.

Abiodun Oguntimilehin (Ph.D) is an Associate Professor of Computer Science, Department of Computer Science, Afe Babalola University, Nigeria. He obtained Ph.D in Computer Science from the Federal University of Technology, Akure, Nigeria. He is a chartered member of Computer Professionals (Registration Council of Nigeria), Member, International Association of Engineers (IAENG) and Nigeria Computer Society. His research interests are Medical Informatics, Data Science and Machine Learning.

Kehinde Adebola Olatunji (Ph.D) is an Associate Professor of Computer Science, Department of Computer Science, Afe Babalola University, Nigeria. She obtained Ph.D in Computer Science from the Federal University of Technology, Akure, Nigeria. She is a chartered member of Computer Professionals (Registration Council of Nigeria). Her research interest is in Artificial Intelligence, Machine Learning, Biometrics and Data Science.

Oluwatoyin Bunmi Abiola (Ph.D) is an Associate Professor of Computer Science, Department of Computer Science, Afe Babalola University, Nigeria. She obtained a Ph.D. in Computer Science from the Federal University of Technology, Akure, Nigeria. She is a Chartered member of the Computer Professionals (Registration Council of Nigeria). Her research interests are Language Processing, Machine Learning and Deep Learning.



Bukola Badeji-Ajisafe (Ph.D) is of the Department of Computer Science, Afe Babalola University, Ado-Ekiti, Nigeria. Her research and academic experience is of modelling human immune system and its application to improve intrusion detection. Her research areas are related to Data Science, medical image analysis and immunology. She got awarded by African-German Network of Excellence in Science (AGNES) for her researches.

Christiana Olanike Akinduyite (Ph.D) is an academic and educational consultant. She currently lectures at the Department of Computer Science, Afe Babalola University, Nigeria. She obtained her PhD. in Computer Science from the Federal University of Technology, Akure, Nigeria. Her research interests are in the area of Security and Privacy, Cryptography and Machine Learning. She is a serial award winner of various prestigious awards and a registered member of Association for Computer Machinery (ACM), Organisation for Women in Science in the Developing World (OWSD) and For-Women-in-Science Sub-Saharan African.

Stephen Eyitayo Obamiyi (Ph.D) is a dynamic lecturer within the Department of Computer Science, Afe Babalola University, Nigeria where he actively contributes to shaping the next generation of computer scientists. With a profound academic background and a passion for cutting-edge technologies, Dr. Obamiyi's expertise encompasses a diverse range of fields, including Intelligent Systems, Blockchain Technology, and Cloud Computing.

Gbemisola Olutosin Babalola lectures at the Department of Computer Science, Afe Babalola University, Nigeria. She obtained Masters in Computer Science from the University of Ibadan, Nigeria. She is a member of Computer Professionals (Registration Council of Nigeria and Nigeria Computer Society. Her Research interest is in Internet of Things, Data Science and Machine Learning.

Toyin Okebule is currently a lecturer and a Ph.D student of Computer Science at the Department of Computer Science, Afe Babalola University, Nigeria. She is a chartered member of Computer Professionals (Registration Council of Nigeria), member, Nigeria Computer Society (NCS), National Association of Technologists in Engineering (NATE), and Society of Petroleum Engineers (SPE). Her research interest is in Machine Learning.