Intrusion Detection Model Using Naive Bayes and Deep Learning Technique

Mohammed Tabash, Mohamed Abd Allah, and Bella Tawfik
Faculty of Computers and Informatics, Suez Canal University, Egypt

Abstract: *The increase of security threats and hacking the computer networks are one of the most dangerous issues should treat in these days. Intrusion Detection Systems (IDSs), are the most appropriate methods to prevent and detect the attacks of networks and computer systems. This study presents several techniques to discover network anomalies using data mining tasks, Machine learning technology and dependence of artificial intelligence techniques. In this research, the smart hybrid model was developed to explore any penetrations inside the network. The model divides into two basic stages. The first stage includes the Genetic Algorithm (GA) in selecting the characteristics with depends on a process of extracting, Discretize And dimensionality reduction through Proportional K-Interval Discretization (PKID) and Fisher Linear Discriminant Analysis (FLDA) on respectively. At the end of the first stage combining Naïve Bayes classifier (NB) and Decision Table (DT) using NSL-KDD data set divided into two separate groups for training and testing. The second stage completely depends on the first stage outputs (predicted class) and reclassified with multilayer perceptrons using Deep Learning4J (DL) and the use of algorithm Stochastic Gradient Descent (SGD). In order to improve the performance in terms of the accuracy in classification of penetrations, raising the average of discovering and reducing the false alarms. The comparison of the proposed model and conventional models show the superiority of the proposed model and the previous conventional hybrid models. The result of the proposed model is 99.9325 of classification accuracy, the rate of detection is 99.9738 and 0.00093 of false alarms.*

Keywords: *Classification, intrusion detection, deep learning, NSL-KDD, genetic algorithm, naïve bayes.*

Received December 30, 2017; accepted April 17, 2018
https://doi.org/10.34028/iajit/17/2/9

1. Introduction

The complexity, importance and information resources of distributed computer systems have evolved very rapidly. On the basis of this fact, computers and their networks have become the target of Computer Crimes, which has grown more and more in recent years. Intrusion Detection Systems (IDS) detection systems have become one of the hottest fields in computer security research. and intrusion detection technique, the IDS is used as a countermeasure to maintain data integrity [14]. and the continued operation of the system during the intrusion. Intrusion Detection allows monitoring and analyzing user activity and system, checking system configurations and vulnerabilities, evaluating the system integrity and data files, statistical analysis of model activity based on known attack matching, analysis of anomalies, and running the audit system [24].

1.1. Intrusion Detection Systems

The intrusion can be defined as a set of events and threats that threaten the confidentiality and integrity of information or availability of network resources such as user accounts, file systems, and system kernel [12]. IDS is a software system that monitors network activity and system work by detecting attacks and malicious activities that are exposed to networks and then sending reports to the system security administrator as shown in Figure 1. The intrusion detection includes identifying a series of malicious events that threaten the network systems and its contents of information and their impact on decision sources, the traditional methods of intrusion detection are based on the extensive knowledge of known attack signatures, IDSs techniques rely on intrusion detection into one of two categories (misuse detection or anomaly detection) which we will refer to in details later [3, 19].

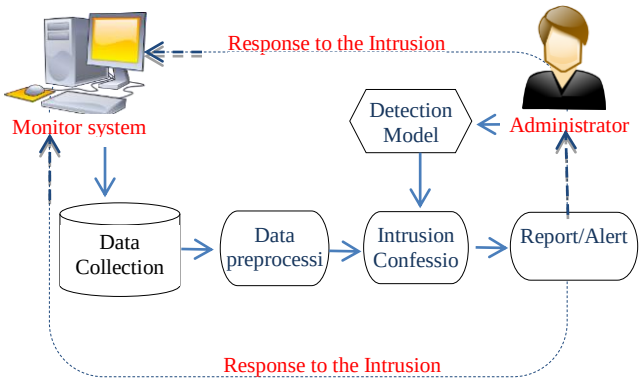

Figure 1. Overview of intrusion detection system.

1.2. Intrusion Detection Systems Techniques

1.2.1. Misuse Detection

This method is used to detect signatures that attack patterns which predetermined by domain specialists. The intrusion prevention system is based on signature and the network traffic control is to match the target with these signatures, and if a match is found the IDS will report an abnormality and will take action. In short, this section relies on detecting attacks when it deviates from normal data flow behavior [15].

1.2.2. Anomaly Detection

This method builds patterns of behavior of natural networks, called (profiles), These profiles are used to detect patterns that deviate greatly from these configurations, and such as these deviations may represent actual intrusions or new behaviors that need to be added to these profiles. The main advantage of anomaly detection is its ability to detect unusual, unobserved breaches, In short, the process of comparing patterns of data with known attack patterns to detect attacks with known behavior [15].

2. Related Work

Most of the Data Mining (DM) techniques, Machine Learning (ML) technology, and artificial intelligence is used in the development of IDSs, many former researchers in this field are focused on the use of classification algorithms and combining technologies to improve the intrusion detection operation. The latest research sets of ML techniques, artificial intelligence techniques, and DM tasks will be used to improve the performance of the IDS (high Accuracy (ACC), the Detection Rate (DR) and reducing False Alarm Rate (FAR)), and here we offer relevant works for each type depending on the approach we are using.

Mukherjee and Sharma [21] an approach was proposed Feature Vitality Based Reduction Method (FVBRM) based on selecting 24 out of 41 features in the NSL-KDD dataset and comparing them through 3 parameters of Feature Selection which are Information Gain (IG), Gain Ratio (GR), and Correlation-based Feature Selection (CFS) where the results showed, by using common NB classifier on discretized values. The FVBRM method achieved 97.78% overall classifier's ACC. Kanagalakshmi and Naveenantony [17] propose a model based on HNB classifier with discretization was created extensively as it focuses on intrusion detection problems, this based on hidden NB model classification of multiple class which improves significantly in terms of ACC and DR of attacks and in particular Denial-of-Service (Dos). Farid *et al*. [9] Produced two models based on the application of two separate algorithms in a hybrid way Both NB classifiers and Decision Table (DT) to improve classification ACC and to classify multi-layered problems.Two separate models were proposed, the first proposal is a hybrid between DT classifier with an NB classifier to remove repetitive, noisy and disturbing situations at the same time from the training dataset University of California, Irvine (UCI) before DT induction. In the second model and the proposed performance. Is Identified a subset of attributes to produce the naive assumption of conditional independence of the class. After testing the experiments and using sensitivity analysis of classification accuracy, 10-fold validation, on 10 standard data sets from UCI. Azad and Jha [4] produced a hybrid model which was proposed using a DT and NB to identify and discover the most possible breakthroughs for computer networks, This hybrid model is trained and tested with the NSL-KDD dataset, using different parameters such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and kappa statics. The result shows ACC 97.1399%. Canbay and Sagiroglu [6] proposed a hybrid model that combines the K-Nearest Neighbors (K-NN) and the Genetic Algorithm (GA) algorithm of the proposed model, to achieve accuracy and detection rate for attacks. KDD-CUP99 was used as a dataset for the proposed system, For the large volume of data, five different subgroups were randomly selected from the original dataset and the researchers used 10-fold validation, Where the results indicated the superiority and preference of the approach proposed on the classic K-NN. Tahir *et al*. [28] produced a model based on the improvement and the development of the performance of the detection of infiltration based on anomalies by dealing with the technology of aggregation to achieve accuracy and high detection rate, the researchers used the K-Means algorithm with the NB classifier with the discretization technique and applied it to the ISCX 2012 data set, and achieved satisfactory results with DR 99.3%, ACC 99.5% and FAR 1.2%. Aljawarneh *et al*. [1] proposed the implementation of several models and the comparison of two sets of work on a single basis, and then use the hybrid approach and collect those classifications together, Which can be used to estimate the penetration threshold based on the ideal characteristics of the data available for 80% training and 20% test, In this study, the researchers analyzed the data using the voting algorithm by increasing the proportion of information that combines probability distributions through these features in order to determine the most important and the best ones, The combined algorithm was composed of Meta Pagging, J48, Random Tree, AdaBoost M1, REPTree, NB and Decision Stump. The results were obtained through the accuracy of proposed approach 99.81% for bilateral groups and 98.56% for NSL-KDD data collection. Tang *et al*. [29] proposed the use of the DL algorithm for intrusion detection and storming the network and the evaluation of the proposed model for IDS network algorithm, As this approach still has some advantages

and large potential for further development, This approach has been applied to detect abnormalities based on the flow in the environment Software Defined Networking (SDN), where he trained this model with NSL-KDD and made using a data set 6 features where facilitates the process treatment where it is taken from base the 41 existing feature of NSL-KDD dataset. Niyaz *et al*. [22] proposed a DL approach based on the implementation of a Network Intrusion Detection System (NISD) to be flexible and effective, using Self-Taught Learning (STL) and applied to the NSL-KDD dataset to assess the abnormality and accuracy of detection and to compare it with previous works in terms of DR, ACC, and f-measure values. Soft-max regression and auto-encoder were applied to NISD. The comparison with what was previously implemented for the intrusion detection mechanism was positive and very good for this approach. Niyaz *et al*. [23] proposed a model based on DL in the SDN environment, eliminates the need for third-party devices and achieving different objectives and reducing the advantage of a wide range of features. The system was evaluated based on performance measures where the results DR 95.65 % and the normal traffic is classified at a very high resolution 99.82 % and a marked decrease in FAR. Dhanabal and Shantharajah [8] The NSL-KDD dataset has been tested to be as data to simulate and evaluate actual performance of the IDS. The CFS method was used to reduce the detection frequency and to reduce the dimensions of communication observed in the dataset, as well increase classification ACC. Modi and Jain [20] A survey of the use of classification techniques for the intrusion detection model was conducted using KDDCUP99 dataset via the WEKA tool. A large number of different methods and techniques have been scanned. This questionnaire included the 20 most important algorithms of the computer learning technology. It displays the data collected in its details and its contents which consists of 4,900,000 communication processors And 41 feature attacks are divided into 4 classes. Ghazali *et al*. [11] proposed a model for the detection of abnormal communication processes in the computerized network. This research is based on the testing of five different classification techniques with the NSL-KDD dataset. The classification techniques were used and were as follows and SimpleCart, PART, BFTree, NB and Ridor. The system was evaluated based on performance measures were selected PART as the best results. Achieve DR 95.5%, ACC 96.7% and the FAR 4.7%. Putchala [25] Proposed a light-weight architecture for an IDS in the Internet of Things (IoT) network. Based on TCP/IP layer architecture and the attack types at each layer, he suggested placing IDS classifiers at each layer. The Deep learning algorithms have been applied to classify the data at each IDS classifier, have used the full KDD 99'cup 21% data set of the experiments. The ACC and FAR were of All-Layer IDS are 98.91% and 0.76% respectively. Kim *et al*. [18] proposed a model with the DL approach, by applying Long Short Term Memory (LSTM) architecture on Artificial Neural Networks (RNN) and trained the IDS using KDD Cup '99 dataset. LSTM-RNN recorded 96.93% ACC with a DR of 98.88% and 10.04 the FAR. Gao *et al*. [10] produced by work that was trained on the KDD data set. The authors proved that deep learning of the Deep Belief Networks (DBN) can be used successfully as an effective identifier. They concluded a layer-by-greedy learning algorithm when used to pre-train and refine the DBN and gives high accuracy. The results showed that DBN recorded the best ACC of 93.49%, which is a TP value of 92.33 and FP by 0.76%. Alom *et al*. [2] proposed a model of the Deep Belief Networks (DBNs) capabilities to detect intrusion through a series of experiments. Trained DBN with NSL-KDD dataset to identify the unknown attack on it. They concluded by proposing DBN as a good IDS based on an ACC of 97.5% achieved in the experiment. Hadi [12] proposed an approach by applying Random forest algorithm through the information gain method which was used to select significant features. After Information gain is used the most 13 significant subset features from the original 41 features from the NSL-KDD standard data were employed to examine the performance of the proposed model. The result of the proposed model is the 99.33% ACC, 0.001 TP, 0.001, FP and 0.993 Precision. Rathore *et al*. [26] produced a model using Hadoop single node using MapReduce programming with various machine learning approaches. Intrusion datasets DARPA is used for evaluation and testing. The system generates better results by taking the proposed features with an overall ACC of more than 99 % TP and less than 0.001 % FP. Jayakumar *et al*. [16] proposed an approach work on intrusion detection, which is done with the help of the supervised learning Neural Network (NN). The feature selection is done with the help of IG algorithm and genetic algorithm. The Multi Layer Perceptron (MLP) supervised NN is used to train the relevant features alone. The results in detecting intrusions with higher ACC, especially for R2L, U2R and DoS attacks. Sujendran and Arunachalam [27] produced a structure based on the Neuro-fuzzy method to generate fuzzy rules and Wiener filter is used to filter out the attack as a noise signal using fuzzy rule generation.The experiment was evaluated on live network data collected, the proposed system achieves 98.46% of ACC and 0.08 % of the FAR.

The comparison of the proposed model and relevant previous models show the superiority of the proposed approach vs the relevant previous models, through available evaluation tools, As shown in Table 1.

Table 1. Comparisons of other works.

Authors(s)/Year	Data Set	Evaluation Standard (Performance)		
		ACC	DR	FAR
Our Model	NSL-KDD	99.9325	99.974	0.00093
Hadi [12]	NSL-KDD	99.33	Tp 0.993	Fp 0.001
Aljawarneh *et al.* [1]	NSL-KDD	99.81	-	-
Putchala [25]	KDD Cup '99	98.91	-	0.76
Tahir *et al.* [28]	ISCX 2012	99.5	99.3	1.2
Niyaz *et al.*[23]	NSL-KDD (DDoS)	99.82	95.65	-
Kim *et al.* [18]	KDD Cup '99	96.93	98.88	10.04
Ghazali *et al.* [11]	NSL-KDD	96.7	95.4	4.7
Rathore *et al.* [26]	DARPA	99	-	Fp 0.001
Gao et *al.* [10]	KDD Cup '99	93.49	Tp 0.923	Fp 0.76
Alom *et al.* [2]	NSL-KDD	97.5	-	-
Sujendran and Arunachalam [27]	live network data	98.46	-	0.08
Azad and Jha [3]	NSL-KDD	97.14	-	-
Mukherjee and Sharma [21]	NSL-KDD	97.78	-	-

3. Data Sets

We used the NSL-KDD dataset and concluded that it was the best after comparing KDDCUP99 [13] as an effective reference dataset to assist researchers with intrusion detection techniques, General data sets were used with intrusion detection systems based on many connection logs with the network, and we note that they are limited and few in this area in terms of accuracy, effectiveness, and non-repetition [30].

The dataset used in the training and testing process NSL-KDD is the most widely used in the construction of intrusion detection systems since 2000 [7]. To ensure that the quality of the proposed model is controlled, this data will be used after it enters into the preprocessing stage of the system's data configuration unit, and we will present this in detail later in section 4.

3.1. Data Collection NSL-KDD

The NSL-KDD dataset is based on an actual database of data extracted which includes a wide variety of intrusions simulated in a military network environment. Data was monitored through Internet traffic for seven weeks and simulated attacks were classified into four categories: DOS Attack, R2L Attack, U2R Attack and Probing Attack [5, 14, 31].

3.2. Data Description

NSL-KDD dataset resolved the data problems in KDD-CUP99 mentioned previously [7]. The NSL-KDD reductive copy of the data KDD original and is composed of the same data KDD-CUP99 containing in each contact record to 41 feature with an extra feature which is to determine that is this contact normal or type of attacks, there are 38 digital feature and 3 symbolic features as shown in Tables 2 and 3 [14, 24].

Table 2. Number of Records in detail [14, 24].

NSl-KDD		KDDTrain	KDDTest	KDDTrain+_20Per	KDDTest-21
Normal		67343	9711	13449	2152
Abnormal	DoS	45927	7458	9234	4342
	Probe	11656	2421	2289	2402
	U2R	52	200	206	2421
	R2L	995	2754	12	533
Total		**125973**	**22544**	**25190**	**11850**

Table 3. Type of features [5, 14].

Type	Features
Nominal	Protocol_type(2), Service(3), Flag(4)
Binary	Land(7), logged_in(12), is_host_login(21), is_guest_login(22)
Numeric	Duration(1), src_bytes(5),dst_bytes(6), wrong_fragment(8),urgent(9), hot(10), num_failed_logins(11), num_compromised(13), root_shell(14), su_attempted(15), num_root(16), num_file_creations(17), num_shells(18), num_access_files(19), num_outbound_cmds(20), count(23), srv_count(24), serror_rate(25), srv_serror_rate(26), rerror_rate(27), srv_rerror_rate(28), same_srv_rate(29), diff_srv_rate(30), srv_diff_host_rate(31), dst_host_count(32), dst_host_srv_count(33), dst_host_same_srv_rate(34), dst_host_diff_srv_rate(35), dst_host_same_src_port_rate(36), dst_host_srv_diff_host_rate(37), dst_host_serror_rate(38), dst_host_srv_serror_rate(39), dst_host_rerror_rate(40), dst_host_srv_rerror_rate(41)

4. Research Proposal and Methodology

The methodology steps of the model which have been used were described, the main steps to build the model were followed and explained in details in this section, also, we provide a full explanation includes the methodology used in this research, and explain the steps of the proposed model in detail as presented in Figure 2. This section is split into four main parts, the first part presents the main steps of the methodology used for the use of the proposed effective model in this research. The second part contains and discusses the pre-processing of a data set, the use of methods and algorithms for the selection and extraction of features and the process of discretization. The third part is the most important section in this chapter, as it includes the process of building and completion of the proposed model and contains the basic experiences and an explanation of the parameters of each algorithm used. The fourth part is to evaluate the proposed model and the extent of success in achieving the desired goals.

4.1. Methodology Steps

- *Data Collection:* in our research, the NSL-KDD dataset will be used and we believe it is the best after comparison with KDDCUP99 as an effective benchmark dataset to help researchers in intrusion detection methods [14, 31], which includes a huge collection of intrusions simulated in a military network environment.
- *Data Preprocessing*: apply a number of preprocessing steps to deal with missing, noisy, inconsistent data and datasets cleaning.

- *Feature Selection*: this technique is used to select important data features and to clarify the relationship between them. It helps to simplify models and to reduce the time of implementation in training and testing for obtaining distinct results.
- *Feature Extraction*: This step is used when the data size is huge and difficult to process and it used also to convert the available original data into simple data and take advantage of the original selected data.
- *Discretization*: it is an essential pre-processing step for machine learning algorithms that can handle only discrete data. However, discretization can also be useful for machine learning algorithms that directly handle continuous variables. Our results indicate that the improvement in classification performance.
- *Implement Stage*: this phase is based on a set of algorithms and classifications: Naïve Bayes (NB), Decision Table (DT), DeepLearning4J and the use of optimization algorithm SGD and logistic regression.
- *Evaluation Stage*: in the methodological evaluation, we adopted a model for the use of misclassification rate, accuracy, F-measure, detection rate, false alarm, and the spent time in the classification process.
- *Comparison Stage*: the comparison process takes place on more than a stage:
- Compare the actual performance with each algorithm independently.
- Comparing the techniques of selection and extraction of the features on a data set.
- Comparison of actual performance with models based on the principle of combining algorithms.
- Performance comparison between the proposed model and the published work which is related to the intrusion detection system.

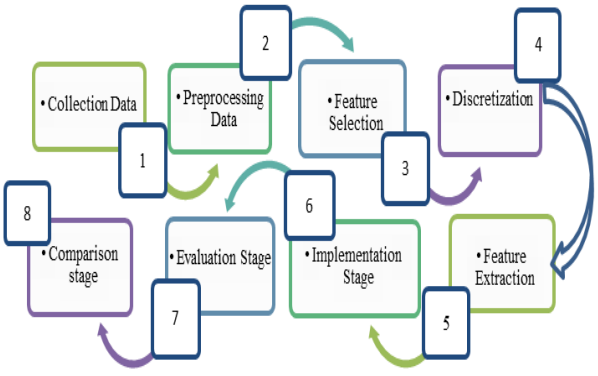

Figure 2. Methodological steps to build the proposed model.

4.2. Evaluation and Discussion

One of the most important steps in a scientific research of any work is the stage of evaluating the performance of the proposed final model so that some of the requirements are calculated through the results extracted. The quality and the performance of this model is judged, through the ability to make accurate and robust predictions. This evaluation is performed across a set of implementing accounts through the confusion matrix.

- *Confusion Matrix*: This matrix is one of the best methods that evaluated IDS. It depends on several measurements to determine the performance of the model where each column in this matrix represents the expected class while each row represents the actual class. The performance of the classifier is evaluated by calculating the number of the expected records correctly and the number of records classified incorrectly. Table 4 [19] shows the four basic elements that determine the content of the matrix will be presented as follows:

Table 4. Confusion matrix [19].

Actual		**Predicate Class**	
		Positive	**Negative**
Class	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- True Positive Rate |TP|= TP / (TP + FN) [4].
- False Positive Rate |FP|= FP / (FP + TN) [4]
- True Negative Rate |TN|= TN / (TN + FP) [4].
- False Negative Rate |FN|= FN / (FN + TP) [1].
- Accuracy |ACC|= (TP + TN) / (TP + TN + FP +FN) [1].
- Detection Rate |DR|= (TP*A + TN*N) / (N + A) [11].
- False Alarm Rate |FAR|= (FP) / (FP + TN) [11].

4.3. The Proposed Model

Figure 3 shows a comprehensive overview of the steps of the proposed model and the adoption of the stages mentioned in previously in the development of our effective model of intrusion detection to obtain the highest accuracy of the classification of the types of attacks in the dataset NSL-KDD and the highest rate of detection of expected attacks and so after compared with previous works. On the other hand, this research focused on the model implementation time, which are avoided by the searchers in the majority of intrusion research.In addition, it decreases the rate of false alarm, by using non-repetitive methods of machine learning techniques. In order to achieve the desired goals in the least time and effort possible, we have followed the following steps:

- *Phase* 1: collect a data set by selecting a high quality of an appropriate data that has several attributes as mentioned above (3.1). The NSL-KDD is a positive, standard, modified and Effective Global Dataset for the connection records of the KDD dataset.

- *Phase* 2: we have identified two separate parts of the NSL-KDD dataset which is a training group and a testing group, the training group-contains (25190 communication processes) and the testing group contains (11850 communication processes). There are records in the testing group that are not in the training group to ensure the quality control of the results of the proposed model, where the above details are found in Table 2.
- *Phase* 3: this phase is considered very important before the operation of processing the data set, such as deleting recurring records, deleting values and records that do not have any usefulness, converting nominal values to numeric, In order to achieve non-bias of a certain class, to ensure the quality and effectiveness of the proposed model.
- *Phase* 4: this stage is based on the selection of attributes that will be used in the proposed model through the use of the evaluator (CfsSubsetEval) to evaluate the value of a subset of attributes by the individual predictive ability of each feature along with the degree of repetition between them, through using the Genetic Algorithm (GA) based on num population(20), ratio of set the probability of crossover (0.25) and set the probability of mutation occurring (0.033). The 14 attributes were chosen as follows: (2, 4, 5, 6, 11, 15, 21, 24, 26, 29, 30, 36, 37, 40).
- *Phase* 5: at this stage, we improve performance by applying discretization by using the Proportional K-Interval Discretization (PKID) method. By sorting the numeric attributes only using the equal frequency, where the number of boxes is equal to the quadratic root of the figure of values, not lost, performs an adjustment the interval size and interval number symmetrical to the number of training.
- *Phase* 6: it is based on reducing the dimensions of features. We have reduced the dimensions of all 41 features in the dataset by applying the FLDA method to reduce the cost of the account, through implementing fisher's linear discriminant analysis for dimensionality reduction and to convert the available original data into simple data and take advantage of the original selected data. Leading to improved processor performance and the memory of the computer.
- *Phase* 7: it is an important part of the building of the model and the actual stage towards the construction of the final model, a classification of the stored communication process in the dataset, which depends on the hybrid model technique (Voting) which combines the NB classifier and the DT classifier to rely on evaluation search feature subsets using (best first). Which raises the working level of the proposed hybrid model and that is through doing the exact match of each attribute value altogether, and thus removes the strong independence assumption caused by NB classifier.
- *Phase* 8: in this stage, we rely on the outputs of the eighth stage to complete the building of the final model by fully processing and depending the predicted class, resulting from step 8 (predicted class) and applying it through the Classification and regression with multilayer perceptrons using DeepLearning4J with the optimization algorithm SGD and the works with the output layer at a learning rate 0.01.
- *Phase* 9: calculate performance measurements, calculate the accuracy of the model rating, the rate of false alarms resulting from the final model, and obtain the actual detection rate of infiltrations and attacks using mathematical calculations via the confusion matrix.

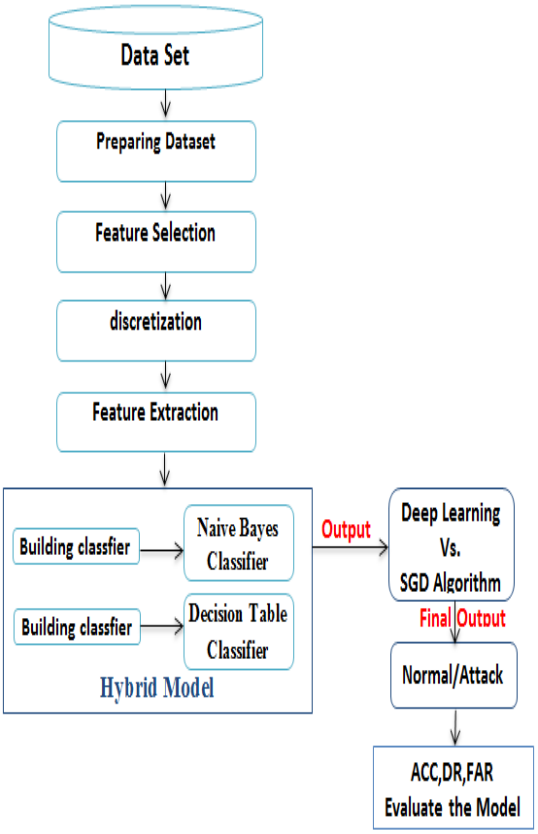

Figure 3. General view of proposed model.

5. Experimental Results

In this section, we review the final stage of our work and explain the experience in detail and to discuss the results that extracted and to analyze them scientifically and mention tools and the environment used in the work of experience. We also describe the events of the experiment and make effective scientific comparisons with the work of the same scientific direction in range the work of intrusion detection systems or data mining techniques or machine learning field.

5.1. Experimental Setup

In this section, we provide an Inclusive explanation of the method of preparation of the experiment through the tools used, description of the experimental environment, the list of the evaluation of the performance of the works and the proposed model, as well as the proposed final model.

5.1.1. Experimental Environments and Tools

The test procedures based on our steps are applied in the previously mentioned research work done on a computer using processor Intel (R) Core (TM) i5-CPU 2.20GHz and 8GB RAM and working on the Windows 7 64-bit operating system, used several software by method directly or help by doing each tool or program according to its tasks in the experiment and it is as follows:

- *WEKA Version* (3.8.1): Waikato Environment for Knowledge Analysis University of Waikato-New Zealand, a program that contains a set of algorithms to learn the machine that works on the extraction of data, where these algorithms are applied directly on the data set or to introduce of tools and methods of pre-processing.
- *Microsoft Excel* (2010): is a program that belongs to the family of electronic tables and specializes in complex and simple operations and calculations.

5.1.2. Experimental Measurements

The evaluation criteria for performing the model experiments are different according to the needs of each researcher. In our experiment, we depend on the evaluation using the confusing matrices to obtain accurate results, from which we conclude the accuracy rates, the detection rates, the false alarms generated by our experiments, the value of the classification error and F measurement, depending on the results of the equations in section (4.3).

5.2. Results and Discussion

- *Stage* 1: the results that appeared in the first stage in Tables 5 and 6 and Figure 4, was one of the reasons for selecting the Naïve Bayes classifier, which we see that the accuracy of the classification of the ratio of (80.73) is low compared with the most known 12 classified in field the intrusion detection, However, it was the fastest among them by completing the classification in a fraction of a second, because it is based on the classification of the training group only once to store the statistics and also has the advantage of easy implementation and works well on the actual data group.

Table 5. Comparison between the resultsof classifiers in stage1.

NO	Classifier	Weighted Avg		Correctly classified Instance	Incorrectly classified Instance	Time second
		TP Rate	FP Rate			
1.	Naïve Bayes	0.807	0.158	80.731	19.269	Part .S
2.	DecisionTable	0.974	0.027	97.3518	2.6482	22
3.	JRip	0.982	0.020	98.1503	1.8497	20
4.	OneR	0.946	0.057	94.615	5.385	6
5.	J48	0.986	0.014	98.5983	1.4017	8
6.	RandomForest	0.987	0.014	98.7048	1.2952	13
7.	BayesNet	0.951	0.048	95.1295	4.8705	4
8.	DecisonStump	0.817	0.150	81.7335	18.2665	4
9.	ZeroR	0.569	0.569	56.9242	43.0758	1
10.	SMO	0.946	0.059	94.6017	5.3983	116
11.	NBTree	0.987	0.013	98.7048	1.2952	68
12.	DTNB	0.973	0.031	97.2543	2.7457	420

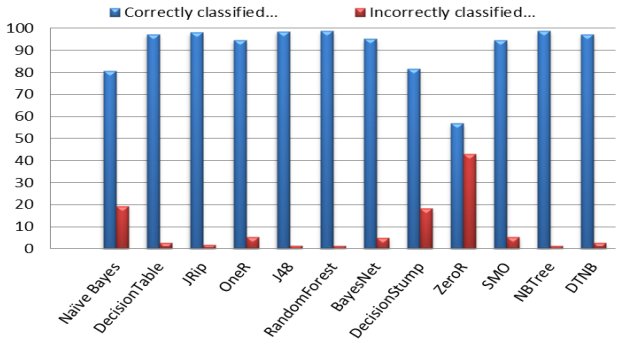

Figure 4. Comparison between the results of classifiers in Stage 1.

Table 6. Result naïve bayes classifier in stage 1.

Classifier	Class	Weighted Avg		Correctly classified instance	Incorrectly classified instance	Time/s
		FP Rate	FP Rate			
Naïve Bayes	Normal	0.950	0.301	80.731	19.269	1
	anomaly	0.699	0.050			

- *Stage* 2: the second stage is a phase based on taking several steps and procedures with parallel work to ensure that the implementation time is taken into account during the application of the model. This is done through the preparation of data procedures to obtain actual results free from Bias. Then the selection of the suitable feature for the process of optimization through using the evaluator (CfsSubsetEval), the GA and cross-linking between the characteristics of the relationship of features with each other. Then the discretization method where we discrete continuous values by using the PKID method, By sorting the numeric attributes only using the equal frequency which is implemented to improve the performance and efficiency of the model. Then the data extraction step is based on the technique of reducing the dimensions FLDA for dimensionality reduction and to convert the available original data into simple data and take advantage of the original data. The final step in the second stage where the work on the use of hybrid between NB classifier and D.TABEL classifier, This is due to the improved classification accuracy rate (96.721) and it was better at the time of implementation the comparison with the most known 12 classified in thefield the machine learning and view the development results in Table 7, 8, and

Figure 5, (note that each classifier was shared separately with NB classifier).

Table 7. Comparison between the resul of hybrid models IN STAGE 2.

No	Classifier+Naïve Bayes	Weighted Avg		Correctly classified instance	Incorrectly classified Instance
		TP Rate	FP Rate		
1.	DecisionTable	0.967	0.033	96.7215	3.2785
2.	JRip	0.959	0.041	95.8653	4.1347
3.	OneR	0.939	0.061	93.9089	6.0911
4.	J48	0.977	0.023	97.6753	2.3247
5.	RandomForest	0.966	0.034	96.6106	3.3894
6.	BayesNet	0.953	0.047	95.2664	4.7336
7.	DecisionStump	0.952	0.046	95.1599	4.8401
8.	ZeroR	0.952	0.048	95.2353	4.7647
9.	SMO	0.967	0.033	96.6505	3.3495
10.	NBTree	0.983	0.017	98.2521	1.7479
11.	DTNB	0.962	0.038	96.1847	3.8153
12.	REPTree	0.982	0.018	98.1811	98.1811

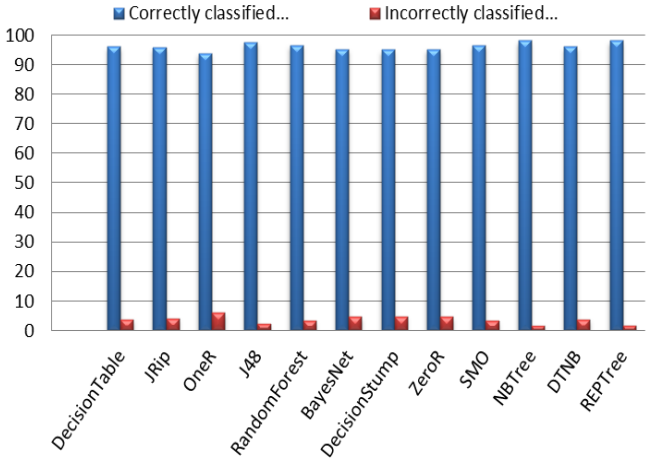

Figure 5. Comparison between the results of hybrid models in Stage 2.

Table 8. Result of hybrid model classifier in stage 2.

Hybrid model (Naïve Bayes + Decision Table)					
Class	TP Rate	FP Rate	No Instance	Correctly classified instance	Incorrectly classified instance
Normal	0.968	0.034	21802	96.7215	3.2785
Anomaly	0.032	0.034	739		
Weight Avg	0.967	0.033	22541		

- *Stage* 3: the final stage in the construction of the proposed model based on the results of the second phase. Entirely, through the participation of the outputs (predicted class) of the second phase (hybrid phase) and taking it through the process of processing through classification and regression with multilayer perceptrons using DeepLearning4J. Which is used to improve the performance of the model by obtaining the highest accuracy and detection rate and low false alarm rate and obtain the excellent execution time. The result of the proposed model is (99.9325) of classification accuracy, the rate of detection is (99.9738) and (0.00093) of false alarms, as set in Table 9 and Figure 6 (note that the final results come is the best compared with previous works in the field of IDSs, as set in Table 1).

Table 9. Result of proposed model for IDS.

Output (Naïve Bayes+ Decision Table)+ Deep Learning					
Class	TP Rate	FP Rate	No Instance	Correctly classified Instance	Incorrectly classified Instance
Normal	1.00	0.001	11839	99.9325	0.0675
Anomaly	0.99	0.000	8		
Weight.Avg	0.99	0.000	11847		
Time taken to build the model				134 Seconds	
Accuracy(ACC)				99.9325	
Detection Rate (DR)				99.9738618	
False Alarm Rate (FAR)				0.000936658	

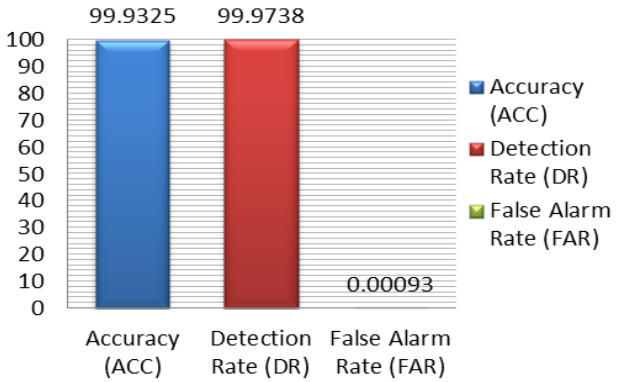

Figure 6. Result of proposed model for IDS.

6. Conclusions

Through the results of the experiments and conducting evaluations based on the proposed final model, we conclude that the NSL-KDD dataset is one of the best effective reference groups for use in the IDSs simulation process. The proposed hybrid model based on deep learning technology as the last stage, it improves the detection rate, accuracy and reduces false alarms through the analysis and evaluation procedures conducted on the NSL-KDD dataset. The most important steps to build an effective model preparing and processing the dataset before the processing stage, by selecting appropriate features, reducing dimensions and used the discretization technique to improve detection performance of the intrusion. As for future studies, the researchers recommend developing IDSs to work in a real environment (on-line). Used new techniques in the extraction, selection of features. Integrate the characteristics of modern algorithms of deep learning to deal with the huge volume of data transmitted over networks and they improve the performance of IDSs.

References

[1] Aljawarneh S., Aldwairi M., and Yassein M., "Anomaly-Based Intrusion Detection System Through Feature Selection Analysis and Building Hybrid Efficient Model," *Journal of*

Computational Science, vol. 25, pp. 152-160, 2018.

[2] Alom M., Bontupalli V., and Taha T., "Intrusion Detection Using Deep Belief Networks," *in Proceedings of National Aerospace and Electronics Conference*, Dayton, pp. 339-344, 2015.

[3] Azad C. and Jha V., "Data Mining in Intrusion Detection: A Comparative Study of Methods, Types and Data Sets," *International Journal of Information Technology and Computer Science*, pp. 75-90, 2013.

[4] Azad C. and Jha V., "Data Mining based Hybrid Intrusion Detection System," *Indian Journal of Science and Technology*, vol. 7, no. 6, pp. 781-789, 2014.

[5] Bhavsar Y. and Waghmare K., "Intrusion Detection System Using Data Mining Technique: Support Vector Machine," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 3, pp. 581-586, 2013.

[6] Canbay Y. and Sagiroglu S., "A Hybrid Method for Intrusion Detection," *in Proceedings of IEEE 14th International Conference on Machine Learning and Applications*, Miami, pp. 156-161, 2015.

[7] Chae H., Jo B., Choi S., and Park T., "Feature Selection for Intrusion Detection using NSL-KDD," *Recent Advances in Computer Science*, pp. 184-187, 2013.

[8] Dhanabal L. and Shantharajah S., "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, no. 6, pp. 446-452, 2015.

[9] Farid D., Zhang L., Rahman C., Hossain M., and Strachan R., "Hybrid Decision Tree And Naïve Bayes Classifiers For Multi-Class Classification Tasks," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1937-1946, 2014.

[10] Gao N., Gao L., Gao Q., and Wang H., "An Intrusion Detection Model Based on Deep Belief Networks," *in Proceedings of 2nd International Conference on Advanced Cloud and Big Data*, Huangshan, pp. 247-252, 2014.

[11] Ghazali A., Nuaimy W., Al-Atabi A., and Jamaludin I., "*Comparison of Classification Models For Nsl-Kdd Dataset for Network Anomaly Detection*," Academic Journal of Science, vol. 4, no. 1, pp. 199-206, 2015.

[12] Hadi A., "Performance Analysis of Big Data Intrusion Detection System over Random Forest Algorithm," *International Journal of Applied Engineering Research*, vol. 13, no. 2, pp. 1520-1527, 2018.

[13] Hettich S. and Bay S., KDD cup 99 task description, http://kdd. ics. uci. edu/databases/kddcup99/task. html , Last Visited, 1999.

[14] Nsl-kdd Data Set for Network-Based Intrusion Detection Systems, Available on: http://nsl. cs. unb. ca/KDD/NSLKDD.html, Last Visited, 2009.

[15] Liao H., Lin C., Lin Y., and Tung K., "Intrusion Detection System: A Comprehensive Review," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 16-24, 2013.

[16] Kaliappan J., Revathi T., and Karpagam S., "Intrusion Detection using Artificial Neural Networks with Best Set of Features," *The International Arab Journal of Information Technology*, vol. 12, no. 6A, pp. 728-734, 2015.

[17] Kanagalakshmi R. and Naveenantony V., "Network Intrusion Detection Using Hidden Naive Bayes Multiclass Classifier Model," *International Journal of Science, Technology and Management*, vol. 3, no. 12, pp. 76-84, 2014.

[18] Kim J., Kim J., Thu H., and Kim H., "Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection," *in Proceedings of International Conference on Platform Technology and Service*, Jeju, pp. 1-5, 2016.

[19] Liao H., Lin C., Lin Y., and Tung K., "Intrusion Detection System: A Comprehensive Review," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 16-24, 2013.

[20] Modi U. and Jain A., "A Survey of IDS Classification Using KDD CUP 99 Dataset in WEKA," *International Journal of Scientific and Engineering Research*, vol. 6, no. 11, pp. 947-954, 2015.

[21] Mukherjeea S. and Sharmaa N., "Intrusion Detection using Naive Bayes Classifier with Feature Reduction," *Procedia Technology*, vol. 4, pp. 119-128, 2012.

[22] Niyaz Q., Sun W., Javaid A., and Alam M., "A Deep Learning Approach for Network Intrusion Detection System," *in Proceedings of the 9th International Conference on Bio-inspired Information and Communications Technologies*, pp. 21-26, 2016.

[23] Niyaz Q., Sun W., and Javaid A., "A Deep Learning Based DDoS Detection in System Software-Defined Networking," *EAI Endorsed Transactions*, vol. 4, no. 12, pp. 1-12, 2016.

[24] Noureldien N. and Yousif I., "Accuracy of Machine Learning Algorithms in Detecting Dos Attacks Types," *Science and Technology*, vol. 6, no. 4, pp. 89-92, 2016.

[25] Putchala M., "Deep Learning Approach for Intrusion Detection System (IDS) in the Internet of Things (IoT) Network using Gated Recurrent Neural Networks (GRU)," PhD Dissertation, Wright State University, 2017.

[26] Rathore M., Ahmad A., and Paul A., "Real Time Intrusion Detection System for Ultra-High-Speed Big Data Environments," *The Journal of Supercomputing*, vol. 72, no. 9, pp. 3489-3510, 2016.

[27] Sujendran R. and Arunachalam M., "Design and Development of Suginer Filter for Intrusion Detection Using Real Time Network Data," *The International Arab Journal of Information Technology*, vol. 15, no. 4, pp. 633-638, 2015.

[28] Tahir H., Said A., Osman N., Zakaria N., Sabri P., and Katuk N., "Oving K-Means Clustering Using Discretization Technique in Network Intrusion Detection System," *in Proceedings of 3rd International Conference on Computer and Information Sciences*, Kuala Lumpur, pp. 248-252, 2016.

[29] 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 International Conference on Wireless Networks and Mobile Communications*, Fez, pp. 258-263, 2016.

[30] Tavallaee M., Bagheri E., Lu W., and Ghorbani A., "A Detailed Analysis of The KDD CUP 99 Data Set," *in Proceedings of IEEE Symposium on Computational Intelligence for Security and Defense Applications*, Ottawa, pp. 1-6, 2009.

[31] Wutyi K. and Thwin M., "Heuristic Rules for Attack Detection Charged by NSL KDD Dataset," *in Proceedings of Genetic and Evolutionary Computing, Yangon*, vol. 1, pp. 137-153, 2015.

Mohammed Tabash is a holds a BSc degree in Computer Science from Al-Quds Open University (2002), studying Master of Information Systems at the faculty of computers and informatics Suez Canal University (2014). His research interests: data mining, machine learning, network security and information systems.

Mohamed Abd Allah is a lecturer at the Department of information systems and decision support Faculty of Computer Science & informatics Suez Canal University. He received his First degree in Computer Science and Operation Research, Faculty of Science, Master degree in Expert systems, Faculty of Science Cairo university. And his PhD degree in computer science, Faculty of Science, Zagazig University. His research interests: Machine learning, data mining, intelligent Bioinformatics, metaheuristic optimization, and predictive models.

Bella Tawfik received his B.Sc. in Electrical engineering from Military Technical Collage, Cairo, Egypt in 1986. He received his M. Sc. in Computer Engineering from the Military Technical Collage, Cairo in 1991. He received his Ph.D. in Electrical Engineering from Colorado State University in August 1999. He got his Post Doctor in Computer Engineering from Colorado State University in October 2006. He is currently assistance professor in Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt. His current research interests are Networks, Modeling, simulation, and Image Processing.