INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2022.16.102 Volume 16, 2022

Use of Naive Bayesian Filtering in the Intrusion Detection System (IDS)

Zakaria Suliman Zubi[1], Abdul Wahab Mohamed Ibrahim[2]

[1]Department of Computer Science, Faculty of Science, Sirte University, Sirte, Libya
[2]Department of Electricity and Computer Engineering, School of Applied Science and Engineering, the Libyan Academy, Tripoli, Libya

Received: August 28, 2021. Revised: February 12, 2022. Accepted: February 27, 2022. Published: March 11, 2022.

Abstract: **- Recently, the main critical part of all organizational data systems is the security since it is threatened by several network attacks which in turn influences on the world financial system. Thus, the most used system in dealing with networks problems is the Intrusion Detection System (IDS). It is used to monitor the system performance and send alerts when there is anomalous activity existence in which the administrator of the system should respond to these alerts rapidly. In this paper, we proposed a statistical Naïve Bayesian method which will be used in the Intrusion Detection Systems (IDS) systems in different scenarios such as analyzing the HTTP service based traffic and identify the HTTP normal connections and attacks. On the other hand, a comparative study between them based on the performance parameters will be analyzed in order to determine the most effective and efficient statistical method in detecting various types of attacks.**

Key-Words: **- Circuits and Systems, Filtering, Systems Theory, Signal Processing, Intrusion Detection Systems (IDSs), Multivariate Statistical Analysis (MSA), Naive Bayesian Filtering.**

I. INTRODUCTION

Several researchers working on Circuits and Systems, Filtering, and Systems Theory has described the design of Intrusion Detection Systems (IDSs) which offers details and descriptions of the main characteristics of these systems that are applicable and relevant in the detection of attacks. Dacier and Alessandri in (1999) explored that the system design depends on the experiences which are resulted from the improvement and use of IDS and also on the analysis of several types of attacks [1].

Some of the IDS characteristics are the information which is used in the analysis, the level of the interpretation and verification of protocols and the used methods in the discovering of activities that may signify attacks. These systems range from simple to complex ones and differ in their characteristics [2].

Dacier and Alessandri (1999) described a model of IDS that consists of sensors and detectors. Porras et al described the used sensors which are signifying e-box of the Common Intrusion Detection Framework Architecture (CIDFA) and also recovering information from a data source. This recovered information goes to the detector. The proposed researchers explored that this detector is considered as a combination of both the a-box and d-box [1], [10].

The IDS characteristics can be represented by using two relatively simple parameters. The first parameter indicates a general characteristic of the system, like the capability to concern accepted expression matching on information. This parameter has not the ability to identify the scope where this characteristic is obtainable. In other words, this parameter cannot discover the type of information expression matching that should be concerned. The second parameter has the ability to discover the IDS scope which determines the validity of the system characteristics [2].

Alessandri (2004) explored that the IDS scope is an iterative method that consists of three main high level scopes, which are: Networking, user and host. Both networking and host are divided into several low level scopes, like application layer and process. User scope is the human that uses the IDS [2].

Several works about the analysis of IDSs are published in order to detect attacks. These works classify the IDSs, ID and attacks. One of these works is the MAFTIA proposal that uses several concepts, models and terminologies that are derived from reliability fields [3].

The main aim of the MAFTIA proposal is to combine between the developed concepts of ID society and the concepts of reliability society. These two societies have diverse roots. Anderson explored that the ID society was born in the seventies; while Spafford (1988) investigated that this society had impulsion with the internet worm incidence [4], [5].

In the eighties, Anderson described the main concepts. The main concept is the dependable computing. Avizienis et al (2000) explored that the first appearance of dependable computing was in the 1830s. In the 1940, the first electronic computers generation allowed the improvement of advanced reliability techniques due to their undependable components [6],[7]

II. THE STATEMENT OF THE PROBLEM

The purpose of this paper is to introduce and explain the use of Multivariate Statistical Analysis (MSA) method in the Intrusion Detection System (IDS) using Naive Bayesian Filtering. The following objectives must be met in order to achieve this problem:

(1) Understand the basic concepts of Intrusion Detection System (IDS).
(2) Theoretically analyze the Multivariate Statistical Analysis (MSA) method which is the Multi-way Principal Component Analysis (MPCA).
(3) Apply the Multivariate Statistical Analysis (MSA) in the Intrusion Detection System.
(4) Theoretically analyze several previous methods, such as Bayesian, data mining and statistical process control methods.
(5) Show the importance of using the (MSA) method.
(6) Evaluate the performance of (MSA) method.
(7) Compare between the MSA method and the previous methods.
(8) Realize the most common difficulties of the (MSA) method.

III. METHODS AND MODEL

IDSs are advanced security tools that can be used to detect various types of attacks in networks. On the other hand, the main problem of these systems is their low accuracy. The current IDSs are not precise enough to offer reliable detection where this problem resulted in a high number of generated false alarms: False Positives (FP) and False Negatives (FN). This large number of false alarms makes the process of filtering out false attacks without missing real ones a real challenge. Furthermore, it makes security administrators unable to respond correctly for risks.

This proposal aims to construct advanced IDS with enhanced efficiency to decrease the number of generated false alarms: false positives and false negatives alarms enhance the network security and improve the detection rate of various types of attacks. In this paper, two IDSs are presented and compared: Naive Bayesian classifier using HTTP services and maximum likelihood estimation (MLE) in Naive Bayesian method using HTTP services also. In both methods, both the training and testing stages are prepared through a categorization process where the HTTP traffic only is chosen. The resultant database includes 38 features. All the HTTP traffic records have identical protocol and service. Furthermore, the attacked HTTP traffic is categorized depending on the attack type. The Naïve Bayesian classifier based IDS utilizes the Naïve Bayesian equation for classifying network traffics depending on their features. The resultant database includes 41 features. For both systems, the NSL-KDD database is used to measure the system performance. This database includes 41 features of the network connection. In this work, the MATLAB program is used to apply both systems with the use of this database. All the used features should have numeric values in order to be implemented in the classifier

IV. DATASET

The DARPA evaluation database will be used in this paper to train and test the intrusion detection system. In this dataset, all network traffic, including the whole payload of packets were recorded in a tcpdump format and offered for evaluation. Tcpdump is a packet analyzer software function, which used to monitor and log Transmission Control Protocol/ Internet Protocol (TCP/IP) traffic passing among a network and the computer where it runs.

The evaluation, data were stored in the form of Windows NT audit data, sniffed network traffic, file

system snapshot and Solaris BSM audit data to recognize attacks which performed against a certain testing network through the data collection stage. This network includes a group of real and simulated machines, which used to artificially generate the background traffic while intrusions were performed against the real machines. Particularly, the DARPA 1999 dataset includes one, two and three weeks for the training data and weeks four and five for the testing one. For the training data, both the week one and three include normal network traffics, while week two includes labeled attacks.

The DARPA 1999 dataset includes 190 instances with 57 attacks that consist of 37 Probes, 53 R2L attacks, 37 U2R/Data attacks and 63 DoS attacks [8].

A Bayesian based IDS is developed and evaluated in this work with the use of the MATLAB programming tool. The system is evaluated using the NSL-KDD database that consists of 41 features as illustrated in the table 1. Practically, three sets of features are selected from the database to evaluate the system:

F1=[5,10,23,24,29,33,38,40] [11]
F2 =[5,10,24,29,33,34,38,40] [12]
F3=[2,5,8,23,30,34,35,38] [11]

These features are chosen according to previous section that focuses on that features number.

A. NSL-KDD Dataset

NSL-KDD is a data set suggested to solve some of the inherent problems of the KDD'99 data set which are mentioned in [15]. Although, this new version of the KDD data set still suffers from some of the problems discussed by McHugh and may not be a perfect representative of existing real networks, because of the lack of public data sets for network-based IDSs, we believe it still can be applied as an effective benchmark data set to help researchers compare different intrusion detection methods. Furthermore, the number of records in the NSL-KDD train and test sets are reasonable. This advantage makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research work will be consistent and comparable.

Features #	Name	Features #	Name
1	Duration	22	is_guest_login
2	protocol_type	23	Count
3	Service	24	srv_count
4	Flag	25	serror_rate
5	src_bytes	26	srv_serror_rate
6	dst_bytes	27	rerror_rate
7	Land	28	srv_error_rate
8	wrong_fragment	29	same_srv_rate
9	Urgent	30	diff_srv_rate
10	Hot	31	srv_diff_host_rate
11	num_failed_logins	32	dst_host_count
12	logged_in	33	dst_host_srv_count
13	num_compromised	34	dst_host_same_srv_rate
14	root_shell	35	dst_host_diff_srv_rate
15	su_attempted	36	dst_host_same_src_port_rate
16	num_root	37	dst_host_srv_diff_host_rate
17	num_file_creations	38	dst_host_serror_rate
18	num_shells	39	dst_host_srv_serror_rate
19	num_access_files	40	dst_host_rerror_rate
20	num_outbound_cmds	41	dst_host_srv_error_rate
21	is_hot_login		

Table 1. NSL-KDD database features

B. Developed Bayesian based on IDS

An IDS is developed in this work with a features selection property and with the application of the Bayesian network classifier, which an anomaly dependent classifier. It depends on the differences in probabilities of attacks occurrence in normal traffics for various feature values. The system performance is measured using the three sets of features as illustrated above in table 1. The main stages of the system are training, testing and running.

The Bayesian classifier is trained using a pre-classified training dataset. It then adjusts the probabilities of each feature. After completing the training stage, the Bayesian computes the probabilities of all records and classify them into normal or attack ones. Thus, the training engine of the Bayesian classifier can be represented as follows:

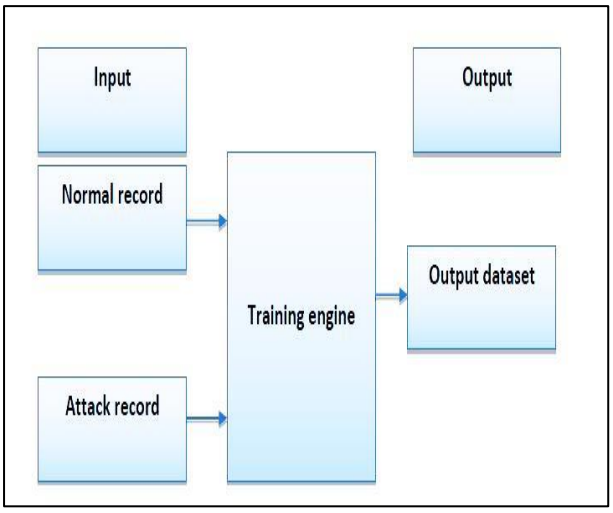

Figure 1. The training engine of the Bayesian classifier

For the training diagram, each input record is considered as a label that describes the connection type. These labels are used to train the system based on initially computing the number of good and bad records in the training dataset. After that, two has tables are generated to include each attribute frequency for normal and bad records. At the end, a third has table is generates, where it includes attributes from both types of records, which scored using the expression below:

$$Score\ of\ attribute = \frac{\frac{B}{number\ of\ bad\ records}}{\frac{B}{number\ of\ bad\ records} + \frac{G}{number\ of\ good\ records}}$$

Where: *B* represents the attribute frequency in the bad hash table file, while *G* represents the attribute frequency in the good hash table file.

After completing the training stage, the system is tested using the presented sets of features from the KDD dataset. The probability of where a record is a normal or bad one is then computed using the following expression:

$$Probability\ of\ record = \frac{\prod_{i=0}^{n} score(i)}{\prod_{i=0}^{n} score(i) + \prod_{i=0}^{n} (1 - score(i))}$$

Where: *n* represents the number of attributes required to test the record, while the score(*i*) represents the attribute score. If the computed probability of a record is bigger than a specific threshold, the record is classified as an attack one.

V. CONDUCTED EXPERIMENTS

Conducted experiments to evaluate the system differ mainly in the deployed training data for building the dataset. This in turn effects on the test accuracy. In addition, both the number of features and the threshold value deployed in the training stage results in a large difference in results. Thus, the conducted experiments have different results due to the difference in training data, used features and threshold.

The used training data includes normal and bad records, where attacks records are known and distributed between four attacks types; R2L, U2R, Probe and DoS. Normal records are used in all experiments based on adding then to the attack record.

The selected features in each one of the presented three sets are selected in a way where key features are selected, such as basic features content, traffic and hot-based features. The basic features are the features from 1 to 9, which are derived from packet header with no examining for payload. The content features are the features from 10 to 22, in which the domain knowledge is deployed for evaluating the packets' payload. The traffic features or as called the time dependent traffic features is the features from 23 to 31, which capture properties established above two seconds. The last type represents features that use an estimated historical window across a specific number of connections rather than time to evaluate attacks that span for more than 2 seconds.

Features are selected based on analyzing the score map used in the training stage to determine the highest value, which can result in a score bigger that the defined threshold value. This in turn results in an increase in the detection rate.

The threshold represents the used level to distinguish among normal and attack records. It was adjusted among experiments to enhance the detection rate. In order to filter the training data, a control chart is deployed to limit the training data between an upper control values; mean+3sigma and a lower control value; mean-3sigma. Any value outside this range is considered as an attack value. This in turn ensures applying the testing data on the same range.

The results obtained after carrying out all experiments are analyzed depending on the number of both normal and attacks records based on using the True Negative (TN), True Positive (TP), False

Positive (FP) and False Negative (FN), where these expressions are described in the next sections. With the use of these expressions, the detection rate and correctness rate can be computed as follows:

$$Detection\ rate = \frac{TP}{TP + FN}$$

$$Correctness\ rate = \frac{TP + TN}{TP + FN + TN + FP}$$

VI. BAYESIAN CLASSIFIER

Bayesian networks are directed graphical probabilistic models that have related probability distribution functions and used for multivariate analysis. These probability distribution functions represent the strength of relations among problem variables in the graph. The Bayesian classifier is applied in the training stage to classify data into normal and abnormal ones. The main stages of the Bayesian based IDS are summarized in the below figure.

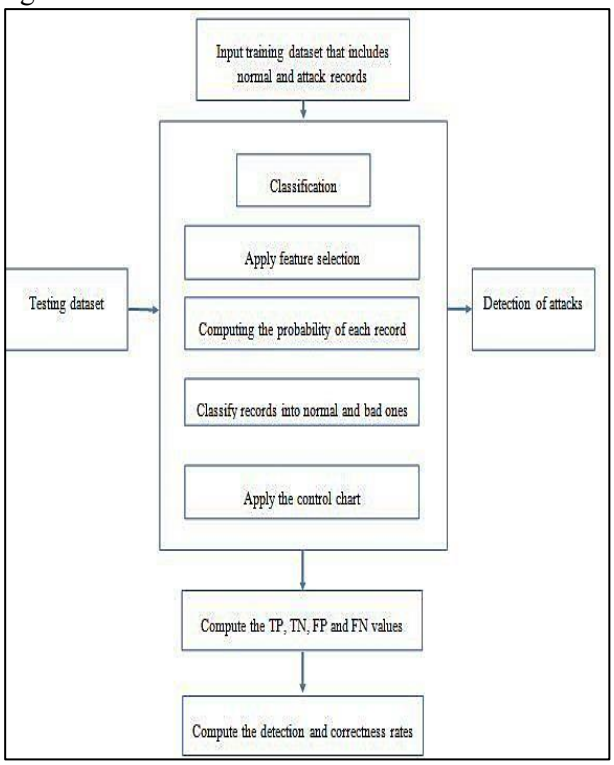

Figure 2 System Stages

Firstly, data from NSL-KDD are used, where the dimension and feature selection is reduced. The data are divided into two sets; training and testing sets, where both the Bayesian classifier is applied on the training data as well as on the testing data to measure the IDS performance. The Control Chart is used to filter data into normal and attack ones where the FP, FN, TP, TN, DR, and CR metrics are computed. The performance of classifier is compared to determine the optimal classifier that offer the lowest FP and FN.

This research work illustrates our proposed methodology that provides a security solution based on Bayesian classifier. For reaching a better level of evaluation on unknown attacks, in our methodology, the detection of suspicious traffic using the clustering strategy will tested the integration of the control chart filtered. The following attractive points are interesting in the proposed method is point out as follows:

1. As a first step there is a process of classifying the network traffic using Bayesian.

2. Second step by applying, clustering based detection as a stage and prevention of intrusion on real time traffic as another stage instead of control chart filter.

VII. ASSESSMENTS

The performance of the IDS will be evaluated based on computing three evaluation metrics for each one and comparing the obtained results; sensitivity, specificity and positive predictive value. These metrics depend initially on the following parameters for each system as follow:

True Positive (TP)	False Positive (FP)
False Negative (FN)	True Negative (TN)

Table 2 shows the parameters on which metrics depend

- True Positive (TP): Attacked records which are correctly detected.
- True Negative (TN): Normal records which are correctly detected
- False Positive (FP): Normal records which are incorrectly detected as attacked records

- False Negative (FN): Attacked records which are incorrectly detected as normal records

After defining those parameters, the three evaluation metrics can be calculated using the following formulas [9]:-

$$\textbf{Sensitivity} = \frac{TP}{TP + FN}\%$$

$$\textbf{specificity} = \frac{TN}{FP + TN}\%$$

$$\textbf{Positive predictive value} = \frac{TP}{TP + FP}\%$$

VIII. EXPERIMENTS AND RESULTS

In this work, it is expected to compute overall accuracy of the network with and without Naive Bayesian Filtering, in each method the accuracy, true negative, true positive, false positive and false negative.

This work presents the development of Bayesian based Intrusion Detection System (IDS) using the MATLAB program. The Network Security Laboratory-Knowledge Discovery and Databases (NSL-KDD) database, which includes 41 features, is used to measure and evaluate the system performance. The main stages of the system are training, testing and running.

The presented NSL-KDD database includes a large number of unimportant features, which must be eliminated without affecting the system performance. Thus, the NSL-KDD database is mainly classified into various classes of features based on the types of attacks. The implemented system in this work is applied on the following sets of features:

F1=[5,10,23,24,29,33,38,40] [11]
F2 =[5,10,24,29,33,34,38,40] [12]
F3=[2,5,8,23,30,34,35,38][11]

The performance of the system is assessed based on measuring the following metrics:

- True Positive (TP): represents the correctly detected attacked records.
- True Negative (TN): represents correctly detected normal records.
- False Positive (FP): represents incorrectly detected normal records as attacked records.
- False Negative (FN): represents incorrectly detected attacked records as normal records.
- Detection Rate (DR): represents the number of intrusion instances detected by the system (True Positive) divided by the total number of intrusion instances presented in the testing set.
- Correctness Rate (CR): represents the summation of correctly detected attacked and normal records divided by the summation for the four types of records above.

A. Results of Applying Bayesian Based IDS on F1

In this subsection, the achieved results of applying the Bayesian based IDS on the first set of features; F1=[5,10,23,24,29,33,38,40] [11] are presented. The below figure shows the obtained Bayesian probabilities of the training dataset. Where x-axis represents the size of data and y-axis represent the probability to each data.

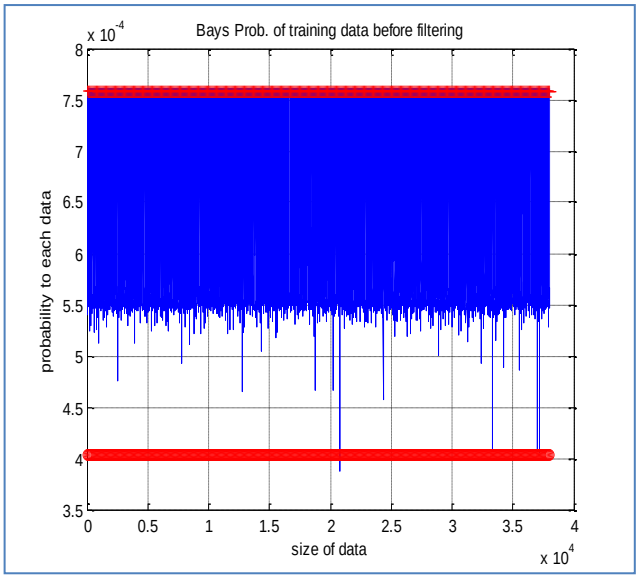

Figure 3. Bayesian probabilities of the training data for F1

As shown above, the probabilities of the training data exceed the range from Mean-3sigma to Mean+3sigma, which represented in red lines. Therefore, the control chart is applied to improve these results based on filtering the data and

removing those outside which specified range with considering them as attacks. The resultant probabilities of the training data after filtering are illustrated below:

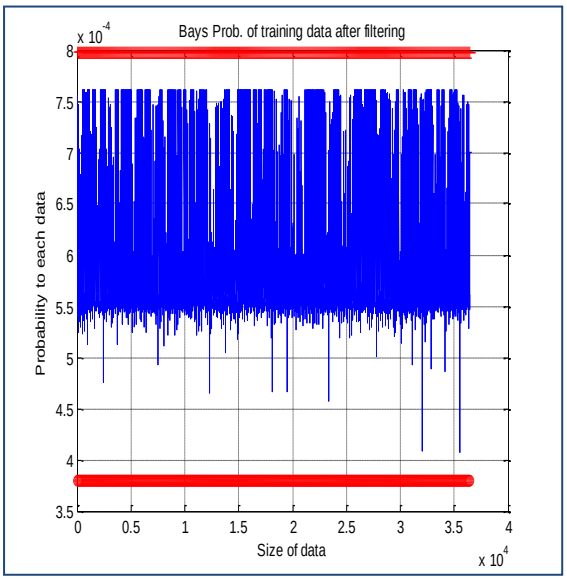

Figure 4. Bayesian probabilities of the training data for F1 after filtering

The following figure shows the obtained probabilities of the testing data.

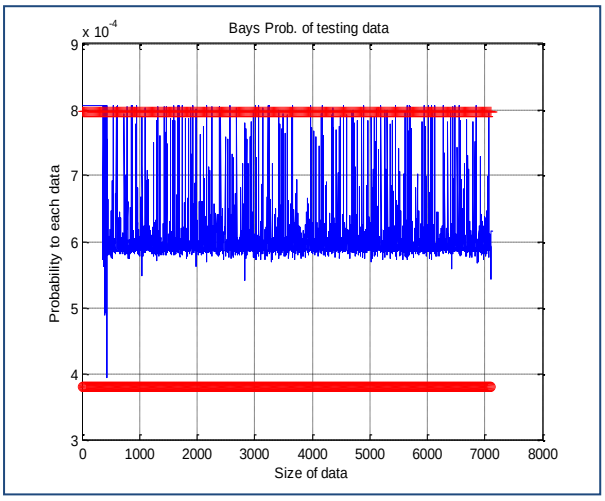

Figure 5. Bayesian probabilities of the testing data for F1

The next table and figure mentioned below demonstrate the obtained results of the presented Bayesian based IDS for F1 where maximum detection rate reach 86.56%.

TN	FP	TP	FN	DR	CR
99.3240%	0.6760%	86.5639%	13.4361%	86.5639%	92.9439%

Table 3. Obtained results of applying the presented Bayesian based IDS on F1

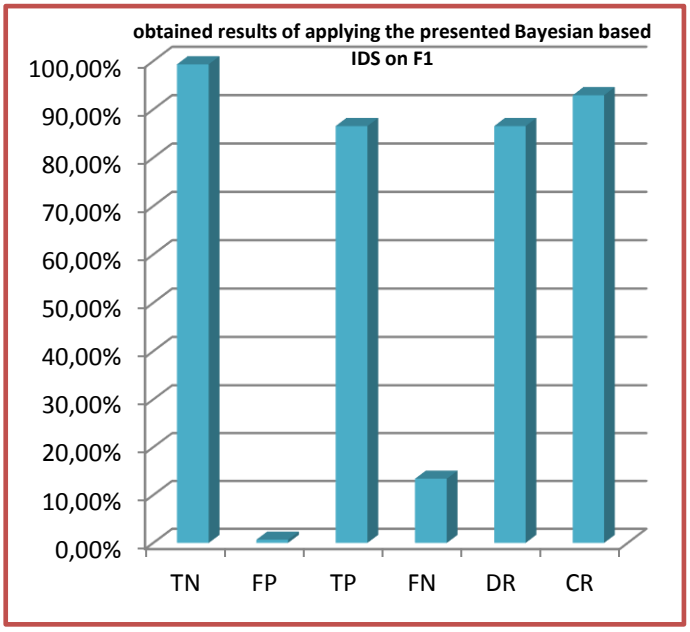

Figure 6. Obtained results of applying the presented Bayesian based IDS on F1

As shown in the table 3 and figure 6 , there are 13.4361% False Negative (FN) and 0.6760% False Positive (FP) alarms. Therefore, records which cause those alarms should be removed from the database and then the test must be repeated to achieve highest True Negative (TN) and True Positive (TP) rates. On the other hand, the system has 86.5639% and 92.9439% detection and correctness rates, respectively.

B. Results of Applying Bayesian Based IDS on F2

The obtained results of applying the Bayesian based IDS on the second set of features; F2=[5,10,24,29,33,34,38,40] [12] for all records using HTTP service are illustrated in this subsection. In the below figure, we displays the obtained Bayesian probabilities of the training data before filtering.

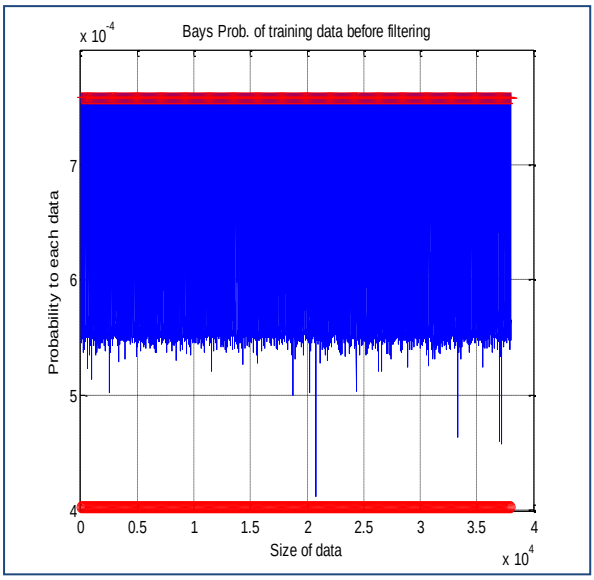

Figure 7. Bayesian probabilities of the training data for F2

As viewed in the above figure, the achieved probabilities of the training data exceed the limits. Thus, those data are filtered, where the achieved probabilities of the training data after filtering are illustrated below.

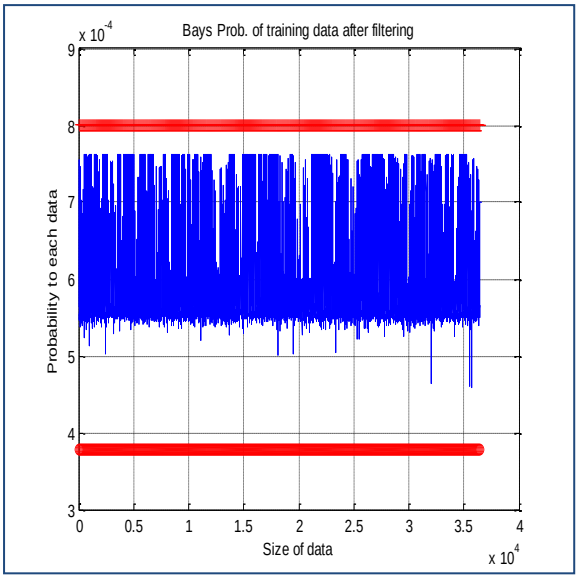

Figure 8. Bayesian probabilities of the training data after filtering for F2

The below figure indicates the expected obtained results in case of probabilities of the testing data.

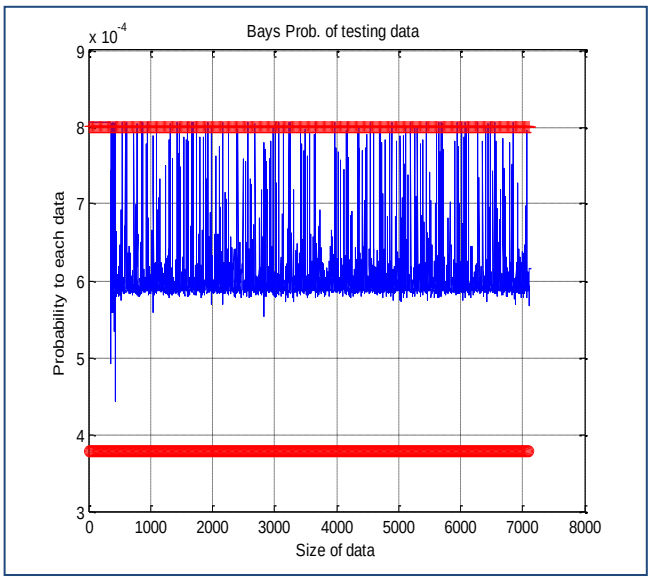

Figure 9. Bayesian probabilities of the testing data for F2

Table 4 and figure 10 demonstrate the obtained results of applying the presented Bayesian based IDS on F2 where maximum detection rate reach 88.16%.

TN	FP	TP	FN	DR	CR
98.9100%	1.0900%	88.1643%	11.8357%	88.1643%	93.5371%

Table 4. Obtained results of applying the presented Bayesian based IDS on F2

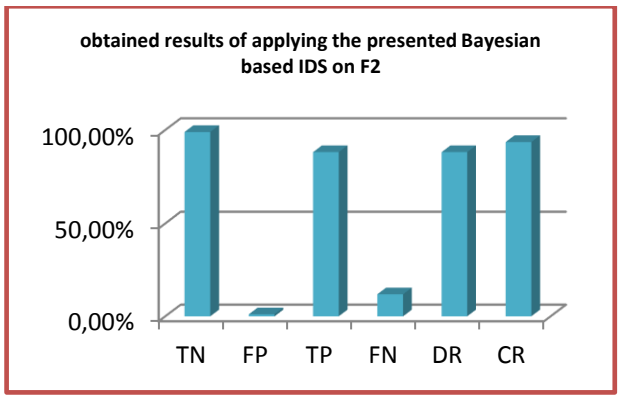

Figure 10, obtained results of applying the presented Bayesian based IDS on F2

As it is illustrated from the above table and figure for F2, there are 11.8357% False Negative (FN) and

1.09% False Positive (FP) alarms. Therefore, records which cause those alarms will be removed from the database and then the test must be repeated to achieve highest True Negative (TN) and True Positive (TP) rates. On the other hand, the system has 88.1643% and 93.5371% detection and correctness rates, respectively.

C. Results of Applying Bayesian Based IDS on F3

This subsection explores the achieved results of applying the Bayesian based IDS on the third set of features; F3=[2,5,8,23,30,34,35,38] [11] for all records using HTTP service. The incoming figure below viewed the obtained Bayesian probabilities of the training data for F3 before filtering.

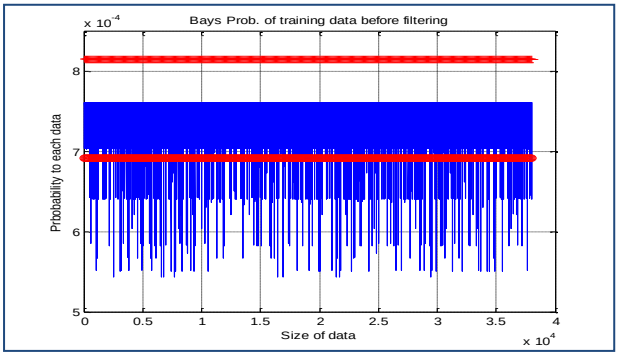

Figure 11. Bayesian probabilities of the training data for F3

As shown in the figure 11 for F3, the achieved probabilities of the training data exceed the specified limits. Thus, those data are filtered, where the achieved probabilities of the training data after filtering are displayed in the figure 12 as follows:

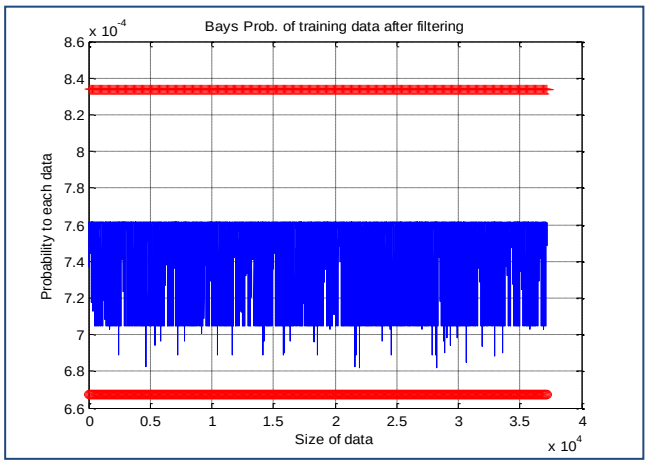

Figure 12. Bayesian probabilities of the training data after filtering for F3

The obtained probabilities results of the testing data are illustrated in the following figure:

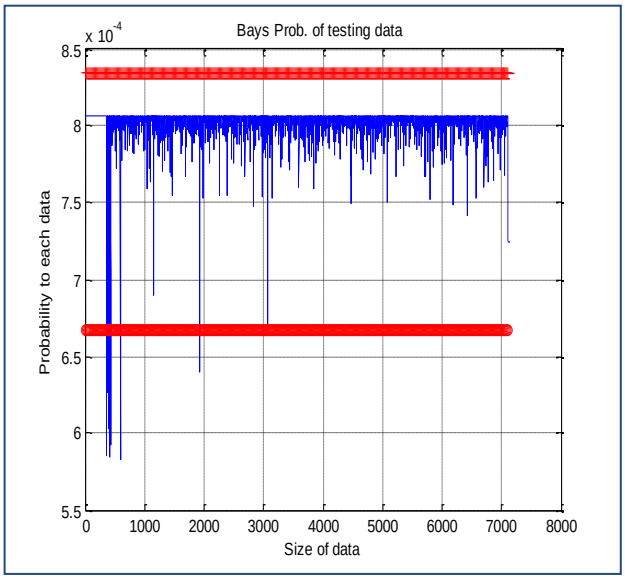

Figure 13. Bayesian probabilities of the testing data for F3

Table 5 and figure 14, demonstrate the obtained results of applying the presented Bayesian based IDS on F3 where maximum detection rate reach 75%.

TN	FP	TP	FN	DR	CR
93.9569%	6.0431%	75%	25%	75%	84.4748%

Table 5. Obtained results of applying the presented Bayesian based IDS on F3

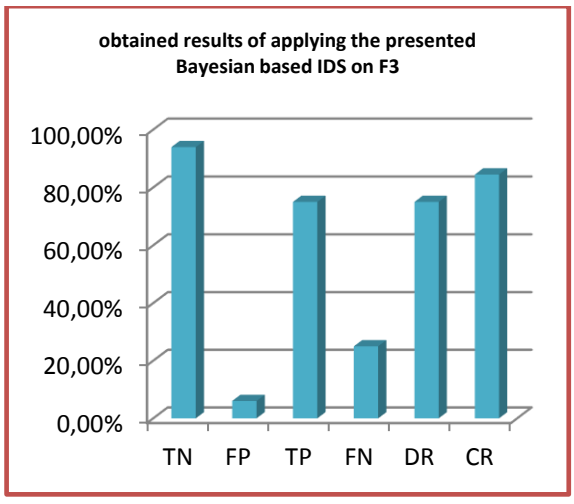

Figure 14 Obtained results of applying the presented Bayesian based IDS on F3

As illustrated above, there are 25% False Negative (FN) and 6.0431% False Positive (FP) alarms. Thus, records which cause those alarms should be removed from the database and then the test must be repeated to achieve highest True Negative (TN) and True Positive (TP) rates. Conversely, the system has 75% and 84.4748% detection and correctness rates, respectively.

IX. COMPARISONS BETWEEN THE PRESENTED SETS OF FEATURES

The following subsections show comparisons among the presented three sets of features.

A. Comparison based on FP and FN Percentages

The table 6 and figure 6, show comparisons among the three sets of features based on both the FP and FN percentages.

Set of features	FP	FN
F1:[5,10,23,24,29,33,38,40]	0.6760%	13.4361%
F2: [5,10,24,29,33,34,38,40]	1.09%	11.8357%
F3: [2,5,8,23,30,34,35,38]	6.0431%	25%

Table 6. Comparisons among the three sets of features based on both the FP and FN percentages

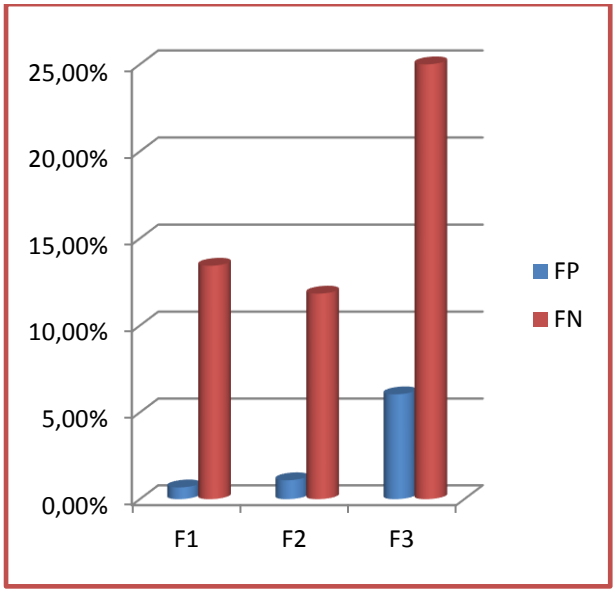

Figure 15. Comparisons among the three sets of features based on both the FP and FN percentages

Table 6 and figure 15 show that the presented Bayesian based IDS offers the lowest percentage of false alarms for F2 and the highest percentages for F3.

B. Comparison based on DR Percentages

A comparison among the three sets features based on DR percentages is shown in the table 7 and figure 16.

Set of features	DR
F1=[5,10,23,24,29,33,38,40]	86.5639%
F2: [5,10,24,29,33,34,38,40]	88.1643%
F3: [2,5,8,23,30,34,35,38]	75%

Table 7 Comparison among the three sets of features based on DR percentages

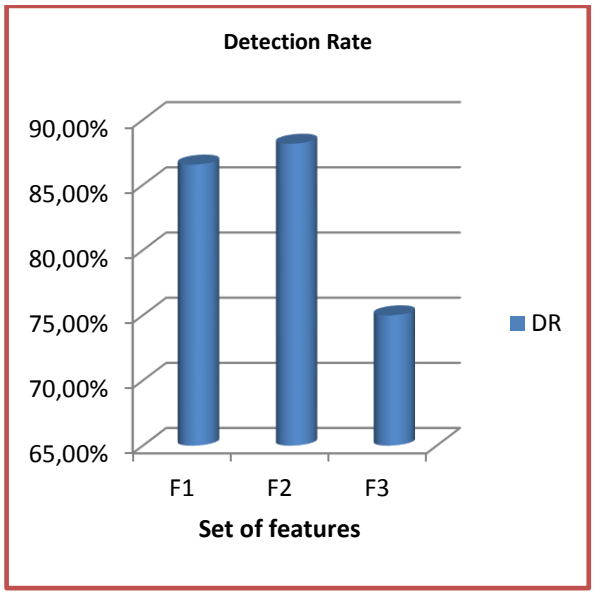

Figure 16 Comparison among the three sets of features based on DR percentages

Based on the results above, the Bayesian based IDS detects attacks in F2 in a higher detection rate than the other two sets of features.

C. Comparison based on CR Percentages

A comparison among the three sets of features based on CR percentages is illustrated in the table 8 and figure 17.

Set of features	CR

F1=[5,10,23,24,29,33,38,40]	92.9439%
F2: [5,10,24,29,33,34,38,40]	93.5371%
F3: [2,5,8,23,30,34,35,38]	84.4748%

Table 8. Comparison among the three sets of features based on CR percentages

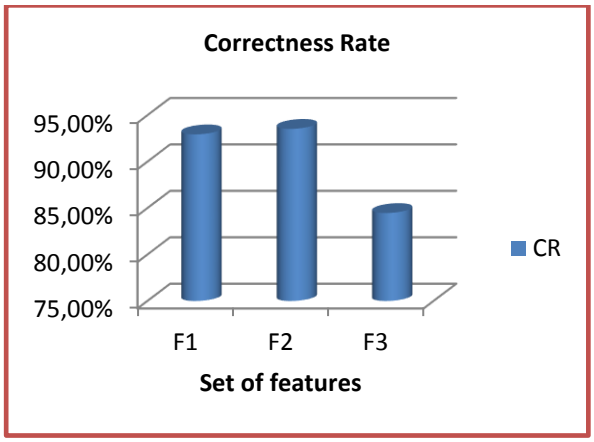

Figure 17. Comparison among the three sets of features based on CR percentages

Based on the results above, the Bayesian based IDS offers the highest correctness rate -for F2, while it offers the lowest rate for F3. Thus, it can be concluded that the Bayesian based IDS offers the best results and minimum number of false alarms for the second set of features.

X. EVALUATIONS

In this section, different methods of IDS are compared with our proposed work. The table 9, illustrates some of these methods.

IDS index	Intrusion Detection Systems	FP rate (%)	FN rate (%)	DR rate (%)
IDSI	A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers[14]	2.98	1.72	95.3
IDSII	Significant of Features Selection for Detecting Network Intrusions[12]	1.1	0.175	97.45
IDSIII	IDS based on Naive Bayes classifier[13]	4.61	24	86
IDSIV	HTTP service based IDS[13]	2.54	2.73	97.5
IDSV	classified HTTP based IDS[13]	1	0.25	99.03

Table 9. Different methods of IDS

As declared from the above table, different methods were applied in IDS, different features and techniques were used in [12] , [13] [14], according to the above table and compared with the proposed results, different points can be achieved for evaluation:

- DR best result for the proposed system was 88.1643% for F2, this result is more better results than the result in IDSIII.
- FN best result for the proposed system was 11.8357% for F2, this result is a more promising result than the results presented in IDSIII.
- FP best result for the proposed system was 0.6760% for F1, this result shows that our proposed system improved better results than all scenarios in the above table.

XI. CONCLUSIONS

In this paper we presents the results of applying the developed Bayesian based IDS on three sets of features from the NSL-KDD database using the MATLAB program. The three sets were:

F1=[5,10,23,24,29,33,38,40];
F2=[5,10,24,29,33,34,38,40];
F3=[2,5,8,23,30,34,35,38].

Where each one of them is divided into training and testing sets. For the three cases, a control chart is applied on the Bayesian probabilities of the training data to keep the data in the range from Mean-3sigma to Mean+3sigma based on removing the out-bound data. A comparison was performed among the three cases using six evaluation metrics such as: TP, TN, FP, FN, DR and CR to determine the best set of features. Results demonstrate that the system has 13.4361% FN and 86.5639% DR percentages when applied on the first set, 11.8357% FN and 88.1643% DR percentages when applied on the second set and 25% FN and 75% DR percentages when applied on the third set. Thus, the system offers the minimum number of false alarms and the highest efficiency and detection rate for the second set of features.

This research work proposed the development of advanced Intrusion Detection Systems (IDSs) to decrease the number of generated false alarms: false positives and false negatives improve the network

security and enhance the detection rate of several types of attacks. the Naive Bayesian method were applied on the constructed IDSs in different scenarios using the MATLAB program, where then a comparative study among them and previous methods was conducted based on analysing the performance parameters and determining the most efficient statistical method in detecting various types of attacks. The NSL-KDD database was used to measure the performance of the implemented systems, where it is composed of 41 features of the network connection.

The performance of the IDS is measured through experiments using NSL-KDD data set. The results by Naïve Bayes classifier show that the best detection rate of the IDS is about 88.1643%, using HTTP service, the false-positive rate is about 0.6760%, and the false-negative rate is about 11.8357%.

REFERENCES:

[1] DACIER, M and ALESSANDRI, D. VulDa: *A Vulnerability Database*. 1999.
[2] ALESSANDRI, D. *Attack-Class-Based Analysis of Intrusion Detection Systems*. 2014.
[3] MAFTIA *Consortium. Architecture and revised model of MAFTIA. R*. In: STROUD 2013.
[4] ANDERSON, J.P. *Computer Security Technology Planning Study*. 1972.
[5] SPAFFORD, E.H. *The Internet Worm Program: An Analysis*. 1988.
[6] ANDERSON, J.P. *Computer Security Threat Monitoring & Surveilanc*e. 1980.
[7] AVIZIENIS, A, LAPRIE, J.C and RANDELL, B. *Fundamental Concepts of Dependability*. 2015.
[8] THOMAS, C, SHARMA, V and BALAKRISHNAN, N. *Usefulness of DARPA Dataset for Intrusion Detection System Evaluation*. 2014.
[9] ZHU, W, ZENG, N and WANG, N. *Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations*. 2012.
[10] PPORRAS, P, SCHNACENBERG, D, CHEN, S.S, STILLMAN, M and WU, F. *The Common Intrusion Detection Framework Architecture*,1998
[11] Bhoria, P. and Garg, K., "*Determining feature set of DOS attacks*", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, No. 5, pp. 875-878, 2013.
[12] Al-Sharafat, W. and Naoum, R., "*Significant of Features Selection for Detecting Network Intrusions*", Institute of Electrical and Electronics Engineers, 2019.
[13] Mohamed M. Abd-Eldayem,'"*A proposed HTTP service based IDS*" Egyptian Informatics Journal (2014) 15, 13–2 , pp. 13 –240.
[14] Adel Nadjaran Toosi, Mohsen Kahani," *A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers*", 2017.
[15] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "*A Detailed Analysis of the KDD CUP 99 Data Set*," Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2019.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Zakaria Suliman Zubi, carried out the idea and the optimization as well as the statistics of the article.

Abdul Wahab Mohamed Ibrahim, has implemented the algorithm with statistical used of Naïve Bayesian method in the IDS as well as the code.

Sources of funding for research presented in a scientific article or scientific article itself

The research work was supported by Department of Computer Science, Faculty of Science, Sirte University, Sirte, Libya.

It was also supported partly from the Department of Electricity and Computer Engineering, School of Applied Science and Engineering, the Libyan Academy, Tripoli, Libya

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)