Enhancing Classification of Network Intrusion Attacks using Feature Reduction

Hajar Elkassabi[1], Mohammed Ashour[2], Fayez Zaki[3]

Abstract— In recent years Intrusion detection systems (IDS) considered an important approach to secure the network. The importance of IDS is due to the increasing of unauthorized access and policy violations. Machine learning approaches have been used in recent years in the field of network intrusion detection. These approaches can classify anomalous and normal patterns. Most of the databases used in the intrusion detection systems contain duplicates and irrelevant records. To improve detection systems and learning rate feature selection or feature reduction has been used in most approaches. In this paper NSL-KDD and UNSW-NB15 datasets have been used to evaluate the performance. Correlation and information gain have been used as feature selection method. Comparative study shows that the detection accuracy by UNSW-NB15 dataset is better than NSL-KDD dataset. WEKA tool has been used as simulation tool.

Index Terms— Intrusion Detection Systems (IDS), Feature Selection, Correlation, Information Gain, Weka, NSLK-DD, UNSW-NB15

1 INTRODUCTION

NETWORK security is one of the most important characteristic in network communication, no matter how small or big your business is. As long as there are networks there will be a need for network security. As networks grow and devices get added then there will be more and more demand for network security. While there is no network that is immune to attacks, a stable and efficient network security system is essential to protecting client data. Advantages of network security:

- Network Security can save our computers from hacking.
- Network security simplifies the process of protecting information shared on the network.
- It helps protecting the personal information of people on the network.

The good security system helps preventing users from falling victim to hackers. Network security consists of:

- Protection: You should configure your systems and networks as correctly as possible
- Detection: You must be able to identify when the configuration has changed or when some network traffic indicates a problem
- Reaction: After identifying problems quickly, you must respond to them and return to a safe state as rapidly as possible.

To secure the information within an organization the CIA triad (Confidentiality, integrity, and availability) has been designed to focus on policies for information security [1]. Organization's security team depends on firewall, intrusion Detection and prevention system (IDPS) to secure their assets. Intrusion detection is the operation of detecting whether the system is under attack or not. This process is done by monitoring the data flow and insure that there are no doubtful activity or malicious attacks and give an alert to the system against suspected data. Intrusions is the process of violating network security and threatens the confidentiality, integrity and availability of system network. IDS can be divided according to its deployment in two categories, Network based intrusion detection system (NIDS): is used for monitoring, analysis and detection of the different network attacks. Host based intrusion Detection system (HIDS): is used for monitoring, analysis and detection of the single host for any unusual activity [2].

Intrusion detection system classified in two mainly approaches: anomaly based detection and rule based or signature based detection. In first technique, IDS looks for data outside of the Ordinary and treated them as attack, Anomaly based detection is established on confirmed statistical behavior methods. The deviation from normal flow is detected as an anomaly. Helping detection of unknown attacks considered one of The advantage of this technique, also an attack can be accurately detected by this mechanism with Low false positives and negatives alarms .The disadvantage of this mechanism is, due to changes occurring in the network on the regular basis, The profile of normal traffic need to be updated. In other hand the signature based detection also called (Misuse based detection) is used for searching among a list of signatures or patterns of an intrusion to detect malicious data. This type of detection working as well as there are an regular updates to its database. when an attack happened the signatures of these attack are generated. The signature of known attacks helps in detecting the future attacks .Analysis and detecting the known attacks in accurate and efficient way which generate low false alarm is considered an advantage of these technique. The problem with signature based is the zero day attacks cannot be detected [2].

Machine learning is considered an effective approach to generate rules to be used with intrusion detection in computer network [4]. Not all features are required for the detection process. Having all the features will only add extra burden. Selecting only relevant feature will help in improving the

1,2,3 Department of Communication and Electronics Faculty of Engineering, Mansoura University, Mansoura, Egypt Emails: eng.hajar.elkasaby@gmail.com, mohmoh2@yahoo.com, fwzaki2017@gmail.com

efficiency and reduce the learning time. after this process the relevant feature is used for further processing [3]. There are two main types of feature selection techniques: wrapper and filter methods. Filter-based feature selection methods utilize statistical measures to score the relationship or reliance between input factors that can be filtered to pick the most important features. While Wrapper feature selection methods creates numerous models with various subsets of input features, the features that have the best performance have been selected according to a performance metrics. Our study is conducted with NSL-KDD dataset which considered an enhanced version of KDD99 dataset [7] and then compared with a new dataset UNSW-NB15 [17]. To measure the performance WEKA tool have been used.

In this paper correlation based feature selection and information gain has been used as a feature selection method. Related work is described in the next section. Data set Visualization is showed in section 3. Section 4 describes preprocessing and feature selection techniques. Experiments and results is discussed in section 5. Finally, conclusion and future work disrobed in section 6.

2 RELATED WORKS

Over the last few decades, researchers carried out studies using NSL-KDD dataset [14,15]. This involves implementation of several approaches of data mining and machine learning algorithms in intrusion detection. These studies concentrated on training and testing several machine learning algorithms as shown in Table (1).

Sabhani and Serpen [16] utilized the decision trees algorithm and get good accuracy, but the technique didn't do well with R2L and U2R attacks as they contain new attack types.

Dhanabal and Shantharajah [7] applied J48, SVM and Naïve Bayes algorithms for classification. Dataset was categorized based on four clasess of attacks . It was noticed that when CFS was used for feature selection , J48 classifier has a good accuracy rate . Application of correlation feature selection increases the accuracy and reduces the
detection time.

Shrivastava, Sondhi and Ahirwar [18] proposed a conceptual IDS framework model which improves the classification performance based on machine learning algorithmes . The proposed model was tested on the basis of Accuracy, False Alarm rate,Detection rate, Error rate.

Deshmukh, Ghorpade and Padiya [19] focusd on increasing the accuracy by using classifiers such as NBTree ,Naïve Bayes and HiddenNaive Bayes. several preprocessing steps have implemented on NSL-KDD dataset as Discretization and Feature selection. The proposed system shows that between algorithms used in the study NBTree algorithm performs well in Accuracy and Error rate.

The performance of NSL-KDD dataset was evaluated by Ingre and Yadav [20] using Artificial Neural Networks. Results applied based on several performance measures such as false positive rate, accuracy and detection rate and better accuracy was found. The proposed model achieved higher detection rate compared with existing models. It was found that for five class classification, the proposed model has good ability to detect the attack in NSL-KDD dataset.
Mostafa and Slay published a new dataset (UNSW-NB15) In 2015 which includes various features to the ones in the NSL-KDD dataset [17]

TABLE 1 OVERVIEW OF PREVIOUS MACHINE LEARNING TECHNIQUES FOR INTRUSION DETECTION

Ref.	Algorithms	Dataset	Simulation tool	Year
[3]	J48 PCA	NSL-KDD	WEKA	2012
[4]	Random Forest J48 SVM CART Navie Bayes	NSL-KDD	WEKA	2013
[5]	J48	NSL-KDD	WEKA	2014
[6]	LSSVM-IDS	KDDCUP99 NSL-KDD KYOTO2006+	Microsoft Word	2014
[7]	J48 SVM Naïve Bayes	NSL-KDD	WEKA	2015
[8]	Naïve Bayes	NSL-KDD	WEKA	2015
[9]	J48 Naïve Bayes	KDDCUP99 Kyoto2006+	-	2017
[10]	SVM-CART	KDDCUP99	-	2017
[11]	J48 Random Forest PART	NSL-KDD	WEKA	2018
[12]	RIPPER PART C4.5	NSL-KDD	WEKA	2018
[13]	SVM ANN	NSL-KDD	WEKA	2019

3 DATA SET VISUALIZATION

(Network Security Laboratory Knowledge Discovery and Data Mining) NSL-KDD which downloaded from (https://www.unb.ca/cic/datasets/nsl.html) is a reduced version of the original KDD dataset. It consists of the same features as KDD. There are 41 features and one class attribute in each record. Each connection is labeled as either as an attack type or as normal. The Total number of attacks presented in NSL-KDD are 39 attacks, each one of them is grouped into four main categories:

1. DOS: denial-of-service, which means preventing legitimate users from accessing a service, e.g. syn flooding.
2. R2L: Remote-to-Local, which means accessing the victim machine by intruding into a remote machine, it also means unauthorized access from a remote machine, e.g. guessing password.
3. U2R: User-to-Root is an attack category, in which a normal account has been used to login in to a victim system and tries to obtain root privilege. it also means unauthorized access to administrator (root) privileges, e.g. buffer overflow
4. Probing: checking and scanning vulnerability on the victim machine to gain information about it, e.g., port scanning.

The NSL-KDD data set consists of (training and testing data). It is important to note that the test data includes specific attack types which not exist in the training data, that's because it isn't from the same probability distribution as the training data. This makes the task more realistic.

TABLE 2

DISTRIBUTIONS OF ATTACKS AND NORMAL IN NSL-KDD

KDD dataset	Total records	Dos	U2R	PROBE	R2L	Normal
KDD train	125973	45972	52	11656	995	67343
		36.46 %	0.04 %	9.25 %	0.79 %	53.46 %
KDD test	22544	7458	200	2421	2754	9711
		33.08 %	0.89 %	10.74 %	12.22 %	43.07 %

As shown in figure (1& 2), The NSL-KDD dataset available in three versions:

a. KDDTrain+ with a total number of 125974 instances.
b. KDDTrain+_20Percent which consists of 20% of the training data with 25192 records.
c. KDDTest+ with a total number of 22544 records.

TABLE 3

LIST OF ATTACKS(NAMES & TYPES) PRESENTED IN NSL-KDD

Attack Class	Attack Name
Dos	Neptune, Processtable, Teardrop, Back , Smurf, Apache2, Land, Pod, Mailbomb ,Udpstorm.(10)
R2L	Named, Warezmaster, Imap, Warezclient, Guess_Password, Snmpguess, Phf, Sendmail, Spy, Ftp_write, Xsnoop, Multihop, Snmpgetattack, Xlock, Worm. (15)
U2R	Rootkit, Buffer_overflow, Ps, Perl, Xterm, Loadmodule, Sqlattack, Httptunnel. (8)
PROBE	Portsweep, Saint, Ipsweep, Satan, Nmap, Mscan. (6)

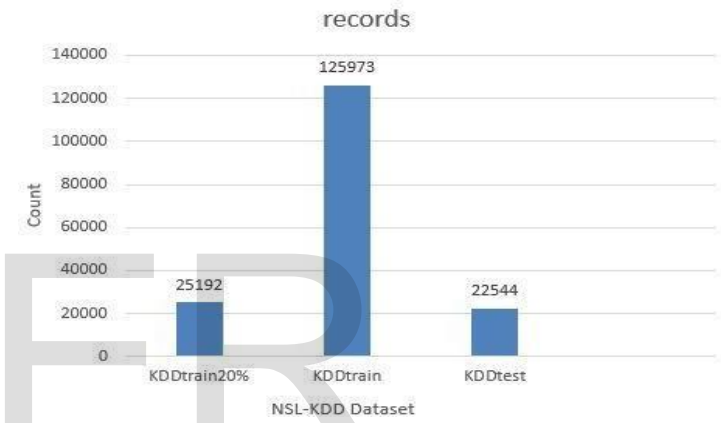

Fig. 1. NSL-KDD dataset versions (train & 20%train & test)

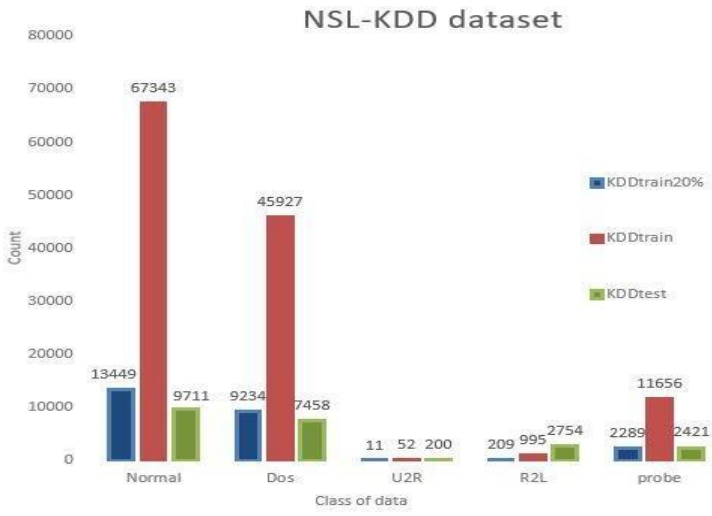

Fig. 2. Statistics of NSL-KDD total records

Despite the fact that NSL-KDD dataset had a few issues, it is an extremely successful dataset that can be utilized for research purposes [18], [21]. In addition, it is hard to acquire certifiable security datasets considering the idea of the security area and keeping in mind that there are other datasets, for example, UNSW-NB15 dataset [17], the NSL-KDD dataset is considered as probably the best one for IDS researches.

4. Preprocessing Techniques

Before machine learning algorithms can be applied to the data, it needs to be converted into a format that is suitable for data analysis by the chosen machine learning algorithm.

4.1 Discretization

Researchers demonstrated that discretization significantly enhance the overall classification performance as well saving storage space considering that the discretized data require fewer space [22]. Several classifiers using discrete data so discretization considers a critical step before classification. Discretization is the way toward quantizing Continuous attributes by gathering those values into various discrete intervals [23].

4.2 Feature selection methods

The high dimensionality of the dataset produced challenges in analyzing the data, hence dimension reduction or feature selection approaches are utilized for data analysis.

4.2.1 Information gain

Information gain is utilized as a measure for estimating the value of an attribute depending on the idea of entropy, the higher the entropy the more the data content. Entropy can be seen as a measure of uncertainty of the system [24].
Equation (1) define The Entropy of a discrete component X

$$\boldsymbol{H(X) = -\sum_{x \in X} p(x) \log_2 (p(x))} \quad \textbf{(1)}$$

Information gain between two attributes is defined as

$$\boldsymbol{IG(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)} \quad \textbf{(2)}$$

In this paper, the information gain is evaluated between individual features and the class. Therefore, we ranked features by their relevance to the class. The higher the gain, the more relevant the feature for determining the class labels.

4.2.2 Correlation based feature selection (CFS)

Correlation is considered a popular and effective technique for choosing the most related features in any dataset, it describes strength of association between features. CFS depends on the presumption that features are conditionally independent given the class. It is based on the following hypothesis [24]: A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. The following equation described the evaluation function

$$\boldsymbol{M_s = \frac{K\,\overline{r_{fc}}}{\sqrt{K + K(K-1)\overline{r_{ff}}}}} \quad \textbf{(3)}$$

Where S is feature subset containing K features, $\boldsymbol{\overline{r_{fc}}}$ is the mean feature-class correlation, and $\boldsymbol{\overline{r_{ff}}}$ is the average feature-feature correlation.

5. Experiments and Results

5.1 Performance Metrics

The following performance metrics have been used in our work

- True Positive (TP): The record is correctly detected as an attack.
- True Negative (TN): Correctly detected as a normal instance.
- False Positive (FP): When a classifier detected normal instance as an attack.
- False Negative (FN): When a detector identifies an attack as a normal instance.
- One of the evaluation metrics which is consists of a combination of recall and precision or detection rate called F-measure. It is obtained from the following equations

$$\boldsymbol{precision = \frac{TP}{TP+FP}} \quad \textbf{(4)}$$

$$\boldsymbol{Recall = TP + FN} \quad \textbf{(5)}$$

$$\boldsymbol{F - measure = \frac{2 \times precision \times recall}{recall + precision}} \quad \textbf{(6)}$$

5.2 Results

5.2.1 Sampling Process

To understand the data set, We extracted only 100 records randomly from train and test data where 53% are normal and 47% are distributed between different types of attacks.

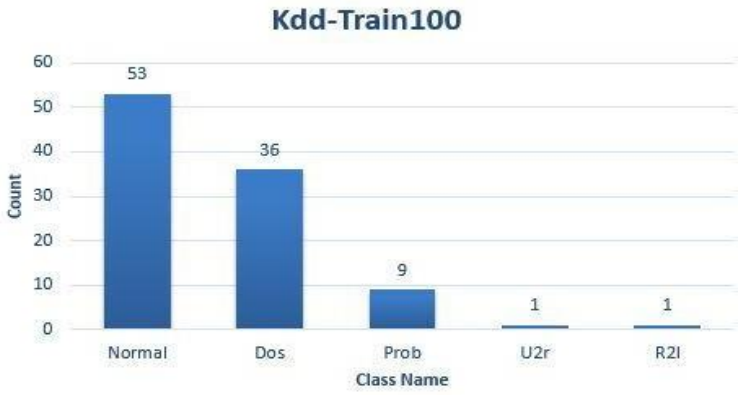

Fig. 3. NSL-KDD train (Normal&Attack)

As shown that figure (3,4,5) illustrate histograms for the sample which has taken from the original KDD dataset. These histograms explain the distribution of normal and attack records.

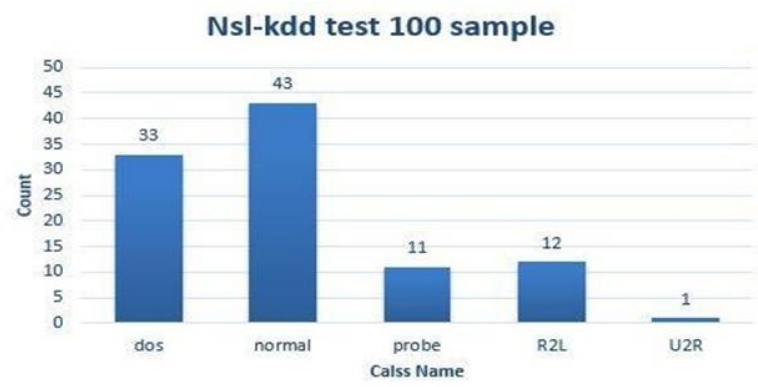

Fig. 4. NSL-KDD test (normal&attack)

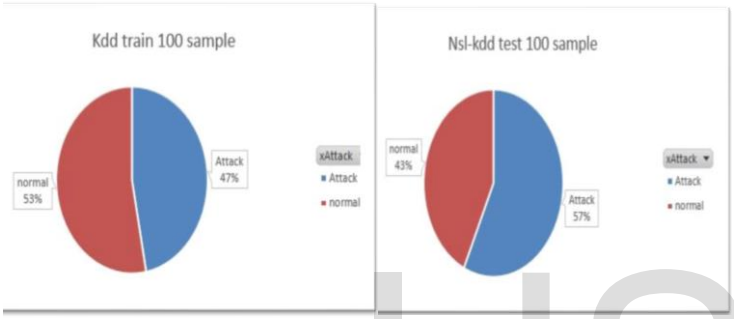

Fig. 5. Distribution of KDD (Train & test)

5.2.2 Correlation coefficient

We calculated correlation for the sample we taken from the dataset with R package. we converted protocol column to numeric [icmp=1, tcp=2, udp=3], Xattack column to [normal=0, Dos=1, U2R=3, probe=4] and replace all 42 features with[x1,x2,x3,…….x42]. After removing null values and zeros columns we have correlation shown in fig (8).
From figure (8) we can notice that the width and the intensity of the circles represent the strength of correlation, thus easily see which features are positively and negatively correlated with each other.

5.2.3 CFS & IG Results

NSL-KDD dataset is considered a well-known available dataset in the area of intrusion detection system. It is still widely used in evaluating the performance of intrusion detection. There are two forms of training sets in the NSL-KDD data set, the full training set and a 20% subset of the full training set. In their experiments, Dhanabal and Shantharajah [7] used 20% of the NSL KDD dataset that was handled with WEKA tool. Authors in [13] used 20% of the NSL-KDD dataset in their proposed model also with WEKA.

In the summary of this part, we discovered that most researchers have utilized the 20% training data set. Figure (6) shows distribution of label (class) column in Nsl-kddtrain20% dataset.

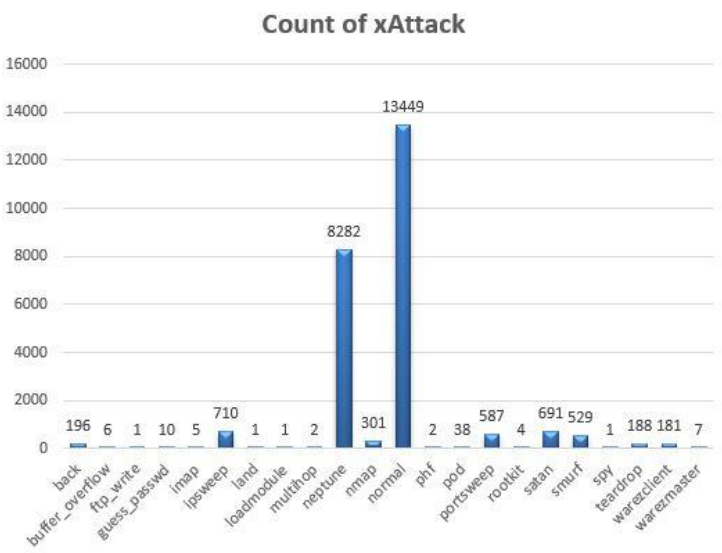

Fig. 6. Distribution of Xattack in Nsl-kddtrain20%

In this paper we applied a hybrid feature selection method based on IG and CFS to reduce the number of features using 20%Nsl-kdda dataset. the results are compared with Unsw-nb15 dataset which was published by Mostafa and Slay [17] in 2015 as containing up-to-date attacks and different features to the ones in the NSL-KDD dataset. For the simulation we used Waikato Environment for Knowledge Analysis (Weka). It is a data mining tool applied widely with machine learning approaches. It has various machine learning algorithms and tools for data preprocessing, Clustering, classification visualization and data analysis. Weka get the data file either in attribute-relation file format (arff) file format or comma separated value (csv). The experimental steps are

1. Import and preprocess the dataset.
2. Choose and run the classifier.
3. Compare the results.

The classification for Nsl-kddtrain20% dataset before preprocessing step By WEKA is 89.5 % as shown:

```
Correctly Classified Instances       22570               89.5919 %
Incorrectly Classified Instances      2622               10.4081 %
Kappa statistic                          0.7906
Mean absolute error                      0.1034
Root mean squared error                  0.3152
Relative absolute error                 20.7817 %
Root relative squared error             63.1897 %
Total Number of Instances            25192

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.912    0.123      0.895     0.912     0.903    0.791     0.968     0.969     normal
                 0.877    0.088      0.897     0.877     0.887    0.791     0.963     0.946     anomaly
Weighted Avg.    0.896    0.106      0.896     0.896     0.896    0.791     0.966     0.958

=== Confusion Matrix ===

     a     b   <-- classified as
 12272  1177 |     a = normal
  1445 10298 |     b = anomaly
```

Fig. 7. Classification before preprocessing

First, we deceritized the NSL-KDD dataset by the same method used in [28] as a preprocessing step. The Correlation feature selection algorithm (CFS) was used with greedy search technique as a first step in feature selection method, we selected five features {5,6,12,26,30} from CFS and added them to our final features set. We notice that CFS not guaranteed to

detect the feature dependence and select all relevant features, so we used IG as a second step in the feature selection method.

considerably low false positive rate. To confirm that our feature selection technique delivers a higher detection rate, we

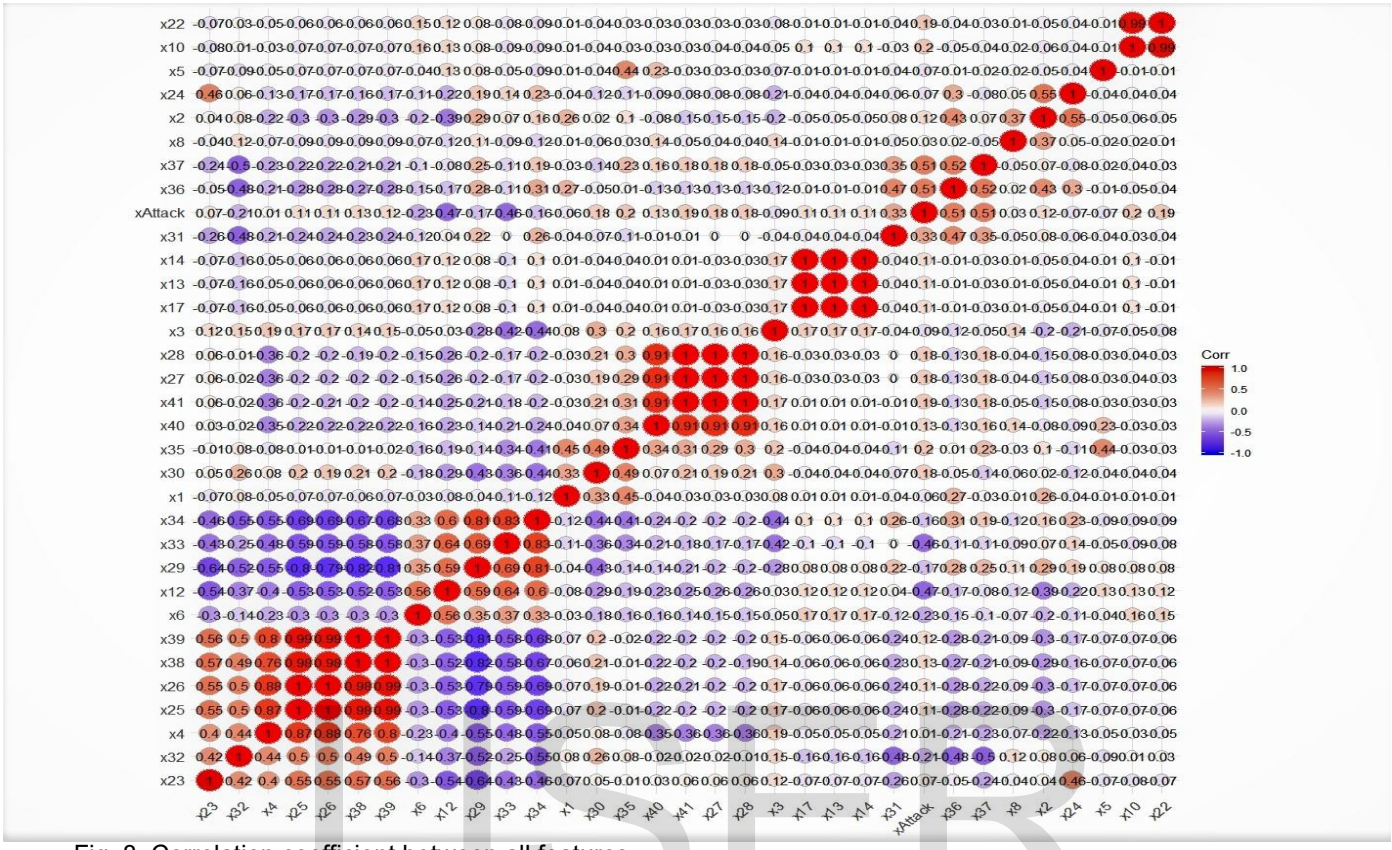

Fig. 8. Correlation coefficient between all features

According to a predetermined threshold, the features were ranked and the top ranked ones are selected {3,4,29,33,34,35,38,25,23,39}. The number of features in our final set from both steps was 15. We used naive Bayes classifier to train the reduced dataset using the adaptive boosting technique. The Adaptive Boosting algorithm was first proposed by Freund and Schapire [25]. Boosting consider a public technique applied to enhance the performance of any learning algorithm. Authors in [26][27] applied naïve bayes as a weak learner which improved with adaboost and carried out performed various experiments to show the different results obtained when using various feature selection methods.

Table 4 shows that the higher detection accuracy is obtained from our algorithm. We applied the same steps on a new dataset called UNSW-NB-15 which was created for intrusion detection research purposes in 2015 [17]. The UNSW-NB15 data set includes nine moderns attack types different to the ones in NSL-KDD data set as well as 49 features. This dataset consists of two datasets, Training dataset (#175, 341 records) and a Testing dataset (#82, 332 records) containing all attack types and normal traffic records.

TABLE (4) COMPARISON BETWEEN DIFFERENT FEATURE SELECTION ALGORITHMS USING 23 CLASSES.

Method	*No. Of Features*	*F-measure*	*Selected Features*
CFS+Bestfirst	5	90.8%	5,6,12,26,30
CFS+Greedy	5	90.8%	5,6,12,26,30
IG (T=0.3)	15	91.8%	5,3,6,4,30,29,33,34,35,38,12,39,25,23,26
Gain ratio(T=0.2)	20	94%	12,26,25,30,4,39,38,6,29,5,37,34,3,33,35,8,23,31,32,41
Correlation	18	91.8%	12,30,4,39,25,38,26,6,5,39,35,34,37,32,33,36,3,31
CFS+IG	**15**	**95.2%**	**5,6,12,26,30,3,4,29,33,34,35,38,25,23,39**

In our work we extracted only 10000 records from UNSW-NB-15 dataset because WEKA can't handel large data. In table (5), same experiments are conducted using the new dataset, which shows that our feature selection technique delivers a higher detection rate comparing with NSL-KDD. Table (6) shows features in UNSW-NB-15 dataset. We applied CFS algorithm to the UNSW-NB15 as a first step in feature selection technique. Six features have been selected {1,8,9,11,28,37,44} and added them to our final set. The second step associated with ranking the features based on the measure of their information gain. The top ranked features are selected and added to our final set.

Based on table 4 and 5, it is clear that UNSW-NB-15 dataset has a higher detection rate that lies in the fact that it contains updated attacks than NSL-KDD dataset.

6. CONCLUSION

In this paper, we applied a feature selection technique used in [8] which based on correlation feature selection and information gain as preprocessing step. NSL-KDD & UNSW-NB-15 datasets have been used in this research. Because CFS is not guaranteed for selection features, we used information gain as a second step. The features have been selected based on a predetermined threshold and added to our final set. We applied the same steps on both datasets and compare results. We notice that UNSW-NB-15 dataset has detection rate higher than NSL-KDD dataset because this dataset is updated and newest than NSL-KDD dataset.
Our future work is to focus research in UNSW-NB-15 as anew dataset which contains up-to-date attacks. We also can apply deep learning instead of machine learning. Using NS3 or opnet we can try our system in life attack scenario.

TABLE 6
FEATURES IN UNSW-NB-15 DATASET

Attribute Number	Attribute Name	Attribute Number	Attribute Name
1	id	23	dtcpb
2	dur	24	dwin
3	proto	25	tcprtt
4	service	26	synack
5	state	27	ackdat
6	spkts	28	smean
7	dpkts	29	dmean
8	sbytes	30	trans_depth
9	dbytes	31	response_body_len
10	rate	32	ct_srv_src
11	sttl	33	ct_state_ttl
12	dttl	34	ct_dst_ltm
13	sload	35	ct_src_dport_ltm
14	dload	36	ct_dst_sport_ltm
15	sloss	37	ct_dst_src_ltm
16	dloss	38	is_ftp_login
17	sinpkt	39	ct_ftp_cmd
18	dinpkt	40	ct_flw_http_mthd
19	sjit	41	ct_src_ltm
20	djit	42	ct_srv_dst
21	swin	43	is_sm_ips_ports
22	stcpb	44	attack_cat
		45	label

TABLE 5 COMPARISON BETWEEN DIFFERENT FEATURE SELECTION ALGORITHMS IN UNSW-NB-15 DATASET

Method	*No.of Features*	*F-measure*	*Selected features*
CFS+Bestfirst	6	98.8%	8,9,11,28,37,43
CFS+Greedy	6	98.8%	8,9,11,28,37,43
IG (T=0.3)	5	91.3%	9,8,28,13,3
Gain ratio(T=0.3)	2	96.6%	43,41
Correlation	9	99.0%	11,4,12,23,24,22,5,21,33
CFS+IG	**9**	**99.8%**	**8,9,11,28,37,43,13,29,3**

REFERENCES

[1] Fenrich, Kim. "Securing your control system: the" CIA triad" is a widely used benchmark for evaluating information system security effectiveness." Power Engineering 112, no. 2 (2008): 44-49

[2] Liao, Hung-Jen, Chun-Hung Richard Lin, Ying-Chih Lin, and Kuang-Yuan Tung. "Intrusion detection system: A comprehensive review." Journal of Network and Computer Applications 36, no. 1 (2013): 16-24.

[3] Alazab, Ammar, Michael Hobbs, Jemal Abawajy, and Moutaz Alazab. "Using feature selection for intrusion detection system." In 2012 international symposium on communications and information technologies (ISCIT), pp. 296-301. IEEE, 2012

[4] Revathi, S., and A. Malathi. "A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection." International Journal of Engineering Research & Technology (IJERT) 2, no. 12 (2013): 1848-1853.

[5] Chae, Hee-su, and Sang Hyun Choi. "Feature Selection for efficient Intrusion Detection using Attribute Ratio." International Journal of Computers and Communications 8 (2014).

[6] Ambusaidi, Mohammed A., Xiangjian He, Priyadarsi Nanda, and Zhiyuan Tan. "Building an intrusion detection system using a filter-based feature selection algorithm." IEEE transactions on computers 65, no. 10 (2016): 2986-2998.

[7] Dhanabal, L., and S. P. Shantharajah. "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms." International Journal of Advanced Research in Computer and Communication Engineering 4, no. 6 (2015): 446-452.

[8] Wahba, Yasmen, Ehab ElSalamouny, and Ghada ElTaweel. "Improving the performance of multi-class intrusion detection systems using feature reduction." arXiv preprint arXiv:1507.06692 (2015).

[9] Kumar, D. Ashok, and S. Venugopalan. "The effect of normalization on intrusion detection classifiers (Naïve Bayes and J48)." Int. J. Future Revolut. Comput. Sci. Commun. Eng 3 (2017): 60-64.

[10] Puri, Aastha, and Nidhi Sharma. "A novel technique for intrusion detection system for network security using hybrid svm-cart." International Journal of Engineering Development and Research 5, no. 2 (2017): 155-161.

[11] [Abdullah, Manal, A. Alshannaq, A. Balamash, and Soad Almabdy. "Enhanced intrusion detection system using feature selection method and ensemble learning algorithms." International Journal of Computer Science and Information Security (IJCSIS) 16, no. 2 (2018).

[12] Gnanaprasanambikai, L., and Nagarajan Munusamy. "Data Pre-Processing and Classification for Traffic Anomaly Intrusion Detection Using NSLKDD Dataset." Cybernetics and Information Technologies 18, no. 3 (2018): 111-119.

[13] Taher, Kazi Abu, Billal Mohammed Yasin Jisan, and Md Mahbubur Rahman. "Network intrusion detection using supervised machine learning technique with feature selection." In 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 643-646. IEEE, 2019.

[14] Tavallaee M., Stakhanova N., and Ghorbani A., 2010, 'Toward Credible Evaluation of Anomaly-Based Intrusion-Detection Methods', IEEE Transactions on Systems, MAN, and Cybernetics, pp. 516-524.

[15] Tavallaee M., Bagheri E., Lu W., and Ghorbani A.,2009, "A Detailed Analysis of the KDD CUP 99 Data Set", IEEE Symposium on Computational Intelligence in Security and Defence Applications (CISDA), 2009.

[16] Sabhnani M., and Serpen G., "Application of machine learning algorithms to KDD intrusion detection dataset within misuse detection context", International Conference on Machine Learning, Models, Technologies and Applications, pp. 209-215, 2003,

[17] Moustafa N. and Slay J., 2015, "Unsw-nb15: A comprehensive data set for network intrusion detection," in MilCIS-IEEE Stream, Military Communications and Information Systems Conference. Canberra, Australia, IEEE publication, 2015.

[18] [17] A. Shrivastava, J. Sondhi, and S. Ahirwar, "Cyber attack detection and classification based on machine learning technique using nsl kdd dataset," Int. Reserach J. Eng. Appl. Sci., vol. 5, no. 2, 2017.

[19] D. H. Deshmukh, T. Ghorpade, and P. Padiya, "Improving classification using preprocessing and machine learning algorithms on NSL-KDD dataset," in Proceedings - 2015 IEEE International Conference on Communication, Information and Computing Technology, ICCICT 2015, 2015.

[20] B. Ingre and A. Yadav, "Performance analysis of NSL-KDD dataset using ANN," Int. Conf. Signal Process. Commun. Eng. Syst. - Proc. SPACES 2015, Assoc. with IEEE, pp. 92–96, 2015.

[21] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," J. Comput. Sci., vol. 25, pp. 152– 160, 2016.

[22] Y. Bhavsar and K. Waghmare, "Improving Performance of Support Vector Machine for Intrusion Detection using Discretization," vol. 2, no. 12, pp. 2990–2994, 2013.

[23] H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: An enabling technique," Data Min. Knowl. Discov., vol. 6, no. 4, pp. 393–423, 2002.

[24] M. a Hall, "Correlation-based Feature Selection for Machine Learning," Methodology, vol. 21i195–i20, no. April, pp. 1–5, 1999.

[25] Y. Freund and R. R. E. Schapire, "Experiments with a New Boosting Algorithm," Int. Conf. Mach. Learn., pp. 148–156, 1996.

[26] W. Li and Q. Li, "Using naive Bayes with AdaBoost to enhance network anomaly intrusion detection," Proc. -3rd Int. Conf. Intell. Networks Intell. Syst. ICINIS 2010, pp. 486–489, 2010.

[27] D. Farid, "Adaptive Intrusion Detection based on Boosting and Naive Bayesian Classifier," Int. ..., vol. 24, no. 3, pp. 12–19, 2011

[28] U. M. Fayyad and K. B. Irani, "Multi-Interval Discretization of Continuos-Valued Attributes for Classification Learning," Proceedings of the International Joint Conference on Uncertainty in AI. pp. 1022–1027, 1993.