Feature Selection in UNSW-NB15 and KDDCUP'99 datasets

JANARTHANAN, Tharmini and ZARGARI, Shahrzad <http://orcid.org/0000-0001-6511-7646>

Feature Selection in UNSW-NB15 and KDDCUP'99 datasets

Tharmini Janarthanan
Department of Computing, Sheffield Hallam University
Sheffield
United Kingdom
Tharmini_1@hotmail.com

Shahrzad Zargari
Department of Computing, Sheffield Hallam University
Sheffield
United Kingdom
S.Zargari@shu.ac.uk

***Abstract*— Machine learning and data mining techniques have been widely used in order to improve network intrusion detection in recent years. These techniques make it possible to automate anomaly detection in network traffics. One of the major problems that researchers are facing is the lack of published data available for research purposes. The KDD'99 dataset was used by researchers for over a decade even though this dataset was suffering from some reported shortcomings and it was criticized by few researchers. In 2009, Tavallaee M. et al. proposed a new dataset (NSL-KDD) extracted from the KDD'99 dataset in order to improve the dataset where it can be used for carrying out research in anomaly detection. The UNSW-NB15 dataset is the latest published dataset which was created in 2015 for research purposes in intrusion detection. This research is analysing the features included in the UNSW-NB15 dataset by employing machine learning techniques and exploring significant features (curse of high dimensionality) by which intrusion detection can be improved in network systems. Therefore, the existing irrelevant and redundant features are omitted from the dataset resulting not only faster training and testing process but also less resource consumption while maintaining high detection rates. A subset of features is proposed in this study and the findings are compared with the previous work in relation to features selection in the KDD'99 dataset.**

Keywords—anomaly detection; feature selection; data mining; machine learning; KDDCUP'99, UNSWNB15, IDS

I. INTRODUCTION

The complexity of modern days' network and the launch of sophisticated attacks on critical infrastructures by hackers bring challenges in the field of cybersecurity. According to Derek Manky, a Fortinet global security strategist, *"Every minute, we are seeing about half a million attack attempts that are happening in cyber space."* [8] The Network Intrusion Detection Systems (NIDS) have been used to monitor inbound and outbound network traffic and identify attacks on the network.

Commonly known NIDSs are signature based and Anomaly based. In signature based NIDS, a database of existing known attack signatures is compared with the current system activities in order to alert the network administrators [9]. On the contrary, anomaly based NIDS, deals with detecting of unknown attacks in network traffics [10].

Anomaly detection in intrusion detection systems could be automated by using data mining and machine learning techniques. This topic has attracted the attention of many researchers over the last decade, particularly after the publication of KDD'99[1] dataset [1,4,3,5,6] being one of the most widely used datasets in this field. Many reported some inherent problems in the KDD'99 dataset such as including a huge number of redundant records, missing values and being outdated since it does not reflect the current network threat environment [1,2]. In 2015, Moustafa and Slay [7,12] created a dataset called UNSW-NB15 dataset for research purposes which has a hybrid of the real modern normal and the contemporary synthesized attack activities of the network traffics [7]. In both datasets (i.e. KDD'99 and UNSW-NB15 datasets), more than 40 features have been considered which may not be significant in anomaly detection and they will increase the resources consumption in data mining. In general, the more features to be included in the data mining, the more difficult the problem is to solve and in case of machine learning algorithms, increasing the number of features significantly increases the training time required to learn the intrusion task [13]. Thus, feature selection is beneficial to both the training and classification processes where it reduces effectively the amount of data required to process, the dimensionality of the problem and memory and CPU usage. It is important to mention that KDD'99 and UNSW-NB15 datasets share only a few common features and the rest of the features are different which makes it harder to compare them.

This research examines the features included in UNSW-NB15 dataset to identify the significant features and reduce the number of features in the UNSW-NB15 dataset. Therefore, a subset of significant features in detecting intrusion can be proposed by using machine learning techniques. These features then can be used in the design of Intrusion Detection Systems (IDS), working towards automating anomaly detection with less overhead.

The remainder of this paper is organized as follows: Section II provides a summary of the previous work in intrusion detection. Section III describes the structure of UNSW-NB15 dataset. Section IV illustrates the methodology

[1] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

of this study, followed by Section V which is the discussion and findings. At the end, conclusions are drawn in Section VI.

II. RELATED WORK

Numerous research has been carried out using the KDD'99 dataset over the last few decades [4,6]. This includes implementation of various techniques and data mining algorithms in intrusion detection. An example of this is the application of decision trees in the KDD'99 competition by the winner, Pfanhringer [14]. Sabhani and Serpen [15] applied the decision trees approach and obtained good accuracy but the approach did not perform well with U2R and R2L attacks as they are minor classes and include a large proportion of new attack types. These classes obtained higher detection rates with an Artificial Neural Network and k means clustering. Work carried out by Gharibian and Ghorbani [3] showed that the decision trees and Random Forests are very sensitive to the selection of different subsets compared to the probabilistic techniques such as Naïve Bayesian and Gaussian which were more robust and produced higher detection rates on the minor classes.

Kumar et al. [16] applied Binary Particle Swarm Optimization (BPSO) and Random Forests (RF) classifier algorithms to classify the PROBE attacks. They found out that the increase in the number of trees in the forest reduces the false positive rate when determining the attacks. The collaborative filtering technique and random forest algorithm have been successful in finding patterns suitable for prediction in large volumes of data. Bajaj and Arora [17] examined the contribution of all the features (#41) in NSL-KDD dataset and found that J48, Naïve Bayes, NB tree, SVM and simple cart methods were applied for binary classification. Three out of 41 features [***urgent*** (#9), ***num_outbound_cmds*** (#20) and ***is_hot_login*** (#21)] using the NSL-KDD training dataset had no significant role in the detection of attacks. Five out of 41 features [***su_attempted*** (#15), ***num_file_creations*** (#17), ***num_access_files*** (#19), ***dst_host_count*** (#32) and ***dst_host_rerror_rate*** (#40)] had little significant role in detection of attacks. Pervez et al. [18] also proposed an approach consisting of merging feature selection and classification for multiple class NSL-KDD intrusion detection dataset by using Support Vector Machine (SVM). The proposed method achieved 91% classification accuracy using only three input features and 99% classification accuracy using 36 input features, while 41 input features achieved 99% classification accuracy. It is important to mention that some of the researchers have been working on KDD'99 dataset samples rather than the complete training dataset due to the size of this dataset [10].

Ingre et al. [19] evaluated the performance of NSL-KDD dataset using ANN. Their work was based on the findings of Bajaj and Arora [17]. Further, they found that features [***land*** (#7), ***wrong_fragment*** (#8), ***num_failed_login*** (#11) and ***root_shell*** (#14)] have all zero values in the dataset. Thus, they reduced the number features is NSL-KDD training and testing datasets to 29 features. For five class classification, the system had good capability to find the attack for the particular class in NSL-KDD dataset. In 2015, Moustafa and Slay [7] criticised that the KDD'99 and NSL-KDD datasets did not represent the modern attacks in an intrusion detection system and introduced a comprehensive network-based dataset known as the UNSW-NB15. This dataset included different features from KDD'99 dataset and only shared a few common features [12]. Further, they examined the characteristics of the UNSW-NB15 and KDD'99 dataset. The UNSW-NB15 was replicated to the KDD'99 dataset to measure efficiency and an Association Rule Mining (ARM) approach was used in feature selection from both the datasets however, it is not clear to how different features were compared in this work since the considered features were different. The results obtained showed that the original KDD'99 features are less efficient than the replicated UNSW-NB15 attributes though the accuracy of the KDDCUP'99 dataset had a higher accuracy than the UNSW-NB15 dataset. The False Alarm Rate (FAR) of the KDD'99 dataset is lower than the UNSW-NB15 dataset [12]. At the time of writing only the above work was found on UNSW-NB15 dataset [7,12].

Aghdam Hosseinzadeh and Kabiri [20] applied ant colony optimisation method on the KDD'99 dataset. A set of 5 best features [***urgent*** (9), ***num_failed_logins*** (#11), ***count*** (#23), ***rerror_rate*** (#27) and ***dst_host_srv_diff_host_rate*** (#37)] were selected under the category of Normal, a set of 4 best features [***durations*** (#1), ***flag*** (#4), ***root_shell*** (#14) and ***dst_host_srv_diff_host_rate*** (#37)] were selected under the category of DoS, a set of 4 best features [***service*** (#3), ***dst_bytes*** (#6), ***count*** (#23), ***serror*** (#25)] was selected under the category of U2R, a set of 3 best features [***count*** (#23), ***srv_count*** (#24), ***diff_srv_rate*** (#30)] under R2L and a set of 8 best features [***protocol_type*** (#2), ***flag*** (#4), ***hot*** (#10), ***logged_in*** (#12), ***num_compromised*** (#13), ***num_access_files*** (#19), ***diff_srv_rate*** (#30***), dst_host_diff_srv_rate*** (#35)] under the category of Probe. They found that the proposed method reduced the number of features by approximately 88% and the detection error reduced by 24% using KDD'99 dataset.

Previously, Zargari and Voorhis [10] worked on the Corrected KDD-dataset where a combination of voting system technique and Weka feature selection technique were used to obtain the best subset of features. The results showed that the proposed subset of features was the best subset compared to the other two subsets suggested by data mining techniques. This subset of features was tested on the NSL-KDD dataset and the results produced higher detection rates. [InfoGainAttributeEval + Ranker] method showed better performance in detecting intrusions. The subset features include feature numbers [***src_bytes*** (#5), ***service*** (#3), ***count*** (#23), ***srv_count*** (#24)] + [***dst_host_srv_count*** (#33), ***dst_host_diff_srv_rate*** (#35), ***protocol_type*** (#2), ***dst_host_same_src_port_rate*** (#36), ***dst_host_same_srv_rat*e (#34) and ***dst_bytes*** (#6)].

In 2015, the new large UNSW-NB15 dataset was published by Mostafa and Slay [7] which includes different features to the ones in the KDD'99 dataset. The UNSW-NB15 and KDD'99 datasets are sharing only a few common features which makes it difficult to compare these two datasets. This study is analysing the features included in UNSW-NB15 dataset in order to reduce the number of features (curse of dimensionality) and propose a subset of features being more significant in detecting intrusions in network traffics. Also, the

results will be further analysed in relation to the KDD'99 dataset in order to determine the similarities and differences.

III. UNSW-NB15 DATASET

The UNSW-NB15 dataset [7] was published in 2015 which includes nine different moderns attack types (compared to 14 attack types in KDD'99 dataset) and wide varieties of real normal activities as well as 49 features inclusive of the class label consisting total of 2, 540, 044 records. These features are categorised into six groups called the *Flow Features*, *Basic Features*, *Content Features*, *Time Features*, *Additional Generated Features* and *Labelled Features*. The *Additional Generated Features* are further categorised into two sub-groups called *General Purpose Features* and *Connection Features*. Features numbering from 36-40 are known as General Purpose Features. Features numbering from 41-47 are known as connection features. Further, the attacks of the UNSW-NB15 dataset are categorised into 9 types known as the *Reconnaissance, Shellcode, Exploit, Fuzzers, Worm, DoS, Backdoor, Analysis and Generic*.

The UNSW-NB15 dataset has been divided into two Training datasets (#82, 332 records) and a Testing dataset (#175, 341 records) including all attack types and normal traffic records. Both the Training and Testing datasets have 45 features, (see Table I). It is important to note that the first feature (i.e. *id*) was not mentioned in the full UNSW-NB15 dataset features and also the features ***scrip***, ***sport***, ***dstip***, ***stime*** and ***ltime*** are missing in the Training and Testing dataset.

TABLE I. FEATURES LISTED IN UNSW-NB15 DATASET [7]

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

A. Common Features in UNSW-NB15 and KDD'99 Datasets

In 1999, KDD'99 dataset was created by using the recorded network traffic from DARPA 1998 dataset being summarized and pre-processed into network connections with 41-features per connection. The features in KDD'99 dataset is grouped into four categories including *Basic Features*, *Content Features*, *Time-based Traffic Features*, and *Host-based Traffic Features*. The KDD'99 consists of 4,898,430 instances which is quite larger than the UNSW-NB15 dataset. The common features in KDD'99 and UNSW-NB15 are shown in Fig. 1.

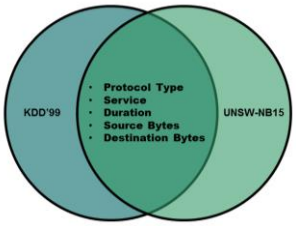

Fig. 1. *Common Features in KDD'99 and UNSW-NB15 dataset.*

As it can be seen in Fig.1, these features mainly belong to the Basic Features category apart from ***protocol type*** feature which belongs to *Flow Features* Category. Also, the ***service*** feature in KDD'99 includes different types of services (e.g. ftp, telnet, smtp, whois, klogin) [6] comparing to the ones in the UNSW-NB15 dataset (e.g. ssl & pop3) [7]. Therefore, it is difficult to apply the findings in [10] or the existing literature on the UNSW-NB15 dataset.

B. Attack Types in KDD'99 and UNSW-NB15 Datasets

There are four main categories of attacks (total number of 24 attack types) in the KDD'99 dataset consisting of DOS, R2L (unauthorized access from a remote machine), U2R (unauthorized access to local super-user (root) privileges and Probing [6] whereas the attack categories (total number of 9 attack types) in the UNSW-NB15 dataset are *Fuzzers*, *Analysis* (e.g. port scans, email spams, HTML files), *Backdoor*, *DoS*, *Exploit*, *Generic*, *Reconnaissance*, *Shellcode* and *Worm* [12]. It is important to know that the KDD'99 Testing dataset contains more attack types than the KDD'99 Training dataset which means that the KDD'99 Training dataset includes a total of 24 attack types whereas the KDD'99 Testing dataset has an additional 14 attack types [21]. This is not true in the UNSW-NB15 Training and Testing datasets as they both include only 9 types of attacks in total.

IV. METHODOLOGY

The UNSW-NB15 dataset includes 45 features from which selecting the important features from the input data can lead to a simplification of the modelling process as well as achieving faster and more accurate detection rates. Often, the data sets contain numerous features which can be not only unimportant and redundant, but also detrimental for the results accuracy [22]. Thus, selecting proper features can significantly affect any detection method's performance (e.g. reducing overfitting). Not much research has been carried out on features selection in the UNSW-NB15 since it has been released recently [7,12]. A typical approach for performing intrusion detection using the UNSW-NB15 dataset is to employ a customized machine

learning algorithm to learn the general behaviour of this dataset in order to be able to differentiate between normal and malicious activities. This study is using Weka (version 3.8), an open source machine learning tool in order to determine the significance of features in the UNSW-NB15 dataset and propose a subset of features which can be used in anomaly detection. Theoretically, since the KDD'99 and UNSW-NB15 datasets are reporting typical flow in network traffics therefore, features with the same characteristics in both datasets should behave similarly and be of the same significance.

In recent years, many data mining algorithms have been used against KDD'99 dataset in order to detect intrusion in network traffics however, many used small samples of KDD'99 dataset in their research [11,7,18,16,15,14,13]. In 2012, Zargari and Voorhis [10] found Random Forest algorithm to be producing the best detection rates against the Corrected KDD'99 dataset (includes 311027 instances) and proposed a subset of significant features using Weka. The Random Forest algorithm is an ensemble of unpruned decision trees which tends to be more robust to noise in the training dataset being a very stable model builder. This was reported in other works such as [11]. The target is only to compare the results of applying a data mining algorithm to the datasets in order to discover the differences among the subsets of features. Therefore, this algorithm can be a good candidate to be used on UNSW-NB15 dataset even though; this was confirmed by examining many other data mining algorithms in this work in order to find the best algorithm producing the highest detection rates.

Moustafa and Slay [12] in their recent study used ARM algorithm and proposed significant features for each attack type where some of these features were repeated often in different attack types. This can be a starting point where a subset of features (*Subset 1*, see Fig.2) can be proposed based on their higher frequencies in order to find the significant features in detecting intrusions.

On the other hand, machine learning techniques (by Weka) offer different methods for attributes selection for a dataset where subsets of significant features can be proposed and evaluated. Once the best subset of features was determined (*Subset 2*, see Fig.3) then, that subset will be examined and compared with the Subset 1 and the results can be further discussed and analysed. Also, it is useful to bear in mind that the proposed subset of features in the UNSW-NB15 dataset can be analysed with the findings in the KDD'99 dataset [10].

In addition, the effects of scaling on the performance of significant features can be investigated by selecting the data samples of different sizes extracted from the full Training UNSW-NB15 dataset and examine the impact of them on detection rates.

V. THE FINDINGS AND DISCUSSION

In 2016, Moustafa and Slay [12] proposed subsets of features for each attack type by using ARM algorithm. The most frequently used features in these attack types are listed in Fig. 2.

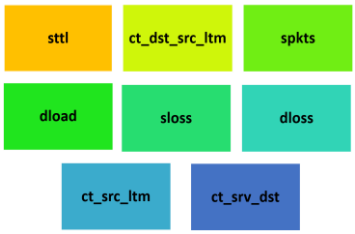

Fig. 2. *Subset 1; The most frequently used features.*

These features are defined in [12] as *Source to destination time to live* [**sttl**], *No. of rows of the same srcip and the dstip in 100 records* [**ct_dst_src_ltm**], *source to destination packet count* [**spkts**], *destination bits per second* [**dload**], *source packets retransmitted or dropped* [**sloss**], *destination packets retransmitted or dropped* [**dloss**], *No.of rows of the srcip in 100 rows* [**ct_src_ltm**] and *No. of rows of the same service and dstip in 100 rows* [**ct_srv_dst**].

Features ***sttl***, ***spkts***, ***dload***, ***sloss*** and ***dloss*** are from the category *Basic Features*. In contrast, features ***ct_dst_src_ltm***, ***ct_src_ltm*** and ***ct_srv_dst*** are from the *Additional Generated Features* category.

These proposed features form a subset of most frequently appeared features in each attack types. The result of running Random Forest algorithm in machine learning using Weka on this subset against the UNSW-NB15 Training and Testing datasets are shown in Table II and Table III, respectively.

TABLE II. UNSW-NB15 TRAINING DATASET (*SUBSET 1*)

Kappa Statistic	Correctly Classified Instance	ROC Values									
		Normal	Reconnaissance	Backdoor	DoS	Exploits	Analysis	Fuzzers	Worms	Shellcode	Generic
0.7332	80.9029%	0.985	0.929	0.909	0.909	0.943	0.937	0.941	0.650	0.777	0.990

TABLE III. UNSW-NB15 TESTING DATASET (*SUBSET 1*)

Kappa Statistic	Correctly Classified Instance	ROC Values									
		Normal	Reconnaissance	Backdoor	DoS	Exploits	Analysis	Fuzzers	Worms	Shellcode	Generic
0.6891	75.6617%	0.936	0.814	0.809	0.854	0.873	0.830	0.820	0.634	0.731	0.990

It is important to mention that many different machine learning algorithms were examined and tested on Weka employing *Subset 1* against UNSW-NB15 dataset before selecting Random Forest algorithm which performed better than the other algorithms.

As it can be seen the Kappa value for the UNSW-NB15 Training and Testing datasets for this subset are 0.7332 and 0.6891 respectively. Kappa Statistic is a measure that takes the expected figure into account by deducting it from the predictor's success and expressing the result as a proportion of the total for a perfect predictor, to yield extra success out of a possible total of predictions. The maximum value of Kappa is 1 and the expected value for a random predictor with the same column total is zero. Therefore, Kappa statistic is used to measure the agreement between predicted and observed categorisations of a dataset, while correcting for agreement that occurs by chance. The other useful statistic is ROC curve

which depicts relative trade-offs between true positives and false positives.

In order to find the best subset of features in Weka, a few methods of Attribute Selection were employed against UNSW-NB15 dataset such as CfsSubsetEval (attribute evaluator) + GreedyStepwise method and InfoGainAttibuteEval (attribute evaluator) + Ranker method. The suggested features by these methods were examined and a few machine learning algorithms (in Weka) including Random Forest algorithm were run against the UNSW-NB15 dataset where the following subset of features (Fig. 3) were performed better among the other proposed subset of features. It is important to mention that the first feature "***id***" in UNSW-NB15 dataset was not included in the calculations because the "***id***" column is actually the row numbers.

Fig. 3. *Subset 2; Significant features suggested by machine learning techniques.*

These features are defined in [12] as *Service type (e.g. http, ftp, smtp, …etc)* [**service**], *Source to destination bytes* [**sbytes**], *Source to destination time to live* [**sttl**], *Mean of packet size transmitted by the srcip* [**smean**] and *No. of rows of the same dstip and the sport in 100 rows* [**ct_dst_sport_ltm**].

Features ***service***, ***sbytes***, and ***sttl*** are from the *Basic Feature* category. Feature ***smean*** is from the *Content Features* category and feature ***ct_dst_sport_ltm*** is from *Additional Generated Features category*.

It is observed in Figs. 2 and 3 that ***sttl*** feature is a significant feature in both *Subsets 1* and *2* selected from UNSW-NB15 dataset. On the other hand, ***service*** and ***sbytes*** features are common in KDD'99 and UNSW-NB15 datasets. The output of Weka running the proposed subset of features (*Subset 2*) against the UNSW-NB15 Training and Testing datasets are shown in Figs. 4 and 5, respectively.

The Kappa value for the UNSW-NB15 Training and Testing datasets are 0.7567 and 0.7639, respectively, Figs 4 and 5. The results demonstrate better detection rates than the corresponding results of the *Subset 1* (Fig.1), extracted from the literature [12]. Also, the number of employed features in *Subset 2* are only 5 compared to the ones in *Subset 1*, being 8 features in total. (lower dimensions), Figs. 4 and 5.

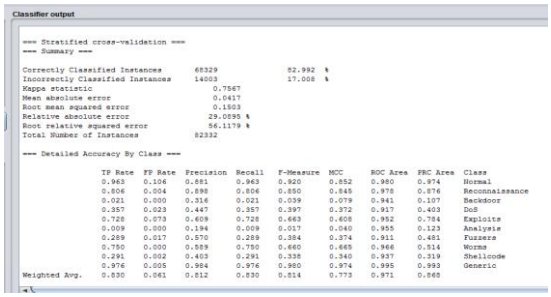

Fig. 4. *UNSW-NB15 Training Dataset (Subset 2).*

It is interesting to note that the Kappa values in the Weka outputs are less than the ones in the Training KDD'99 dataset when different algorithms are used against the Testing KDD'99 In Fig.4 the Kappa value is higher than the Kappa value in Fig.5. This might be due to the fact that the UNSW-NB15 Testing dataset includes the same amount of attack types as the UNSW-NB15 Training dataset whereas this is not true in case of KDD'99 dataset. The KDD'99 Testing dataset has 14 more attack types than the Training KDD'99 dataset [10].

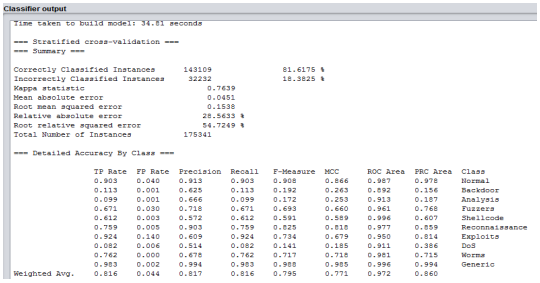

Fig. 5. *UNSW-NB15 Testing Dataset (Subset 2).*

The results of running *Subset 2* against UNSW-NB15 Training and Testing datasets are summarized below in Tables 4 and 5 where indicates that *Subset 2* produces better results than *Subset 1*. It is noticeable that *Subset 2* includes ***service*** and ***sbytes*** features which are in common with the proposed subset by [10] in KDD'99 dataset.

TABLE IV. UNSW-NB15 TRAINING DATASET (SUBSET 2)

Kappa Statistic	Correctly Classified Instance	ROC Values									
		Normal	Reconnaissance	Backdoor	DoS	Exploits	Analysis	Fuzzers	Worms	Shellcode	Generic
0.7567	82.992%	0.980	0.978	0.941	0.917	0.952	0.955	0.911	0.966	0.937	0.995

TABLE V. UNSW-NB15 TESTING DATASET (SUBSET 2)

Kappa Statistic	Correctly Classified Instance	ROC Values									
		Normal	Reconnaissance	Backdoor	DoS	Exploits	Analysis	Fuzzers	Worms	Shellcode	Generic
0.7639	81.6175%	0.987	0.977	0.892	0.911	0.950	0.913	0.961	0.981	0.996	0.996

In order to investigate the effects of scaling on the features proposed in *Subsets 1* and *2*, three data samples were randomly taken from UNSW-NB15 dataset by the sizes of 6841 (Sample 1), 10291 (Sample 2) and 20582 (Sample 3). The total number of samples are determined by dividing the total number of instances in the UNSW-NB15 dataset by 12, 8, and 4, creating Samples 1, 2, and 3 respectively. Fig. 6 demonstrates the effects of scaling on the performance of the proposed subsets of features in detecting intrusions. As it can be seen the *Subset 2* performs better than the *Subset 1* indicating that the *Subset 2* includes more significant features.

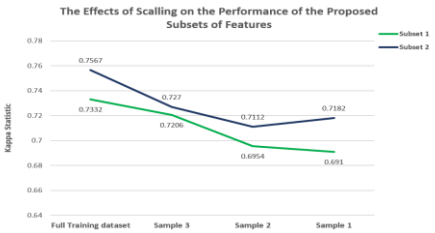

Fig. 6. *The Effects of Scaling on Features Performance.*

VI. CONCLUSIONS

Since 1999, KDD'99 dataset has been the most popular and employed dataset by researchers despite its inherent problems. Some of the work were focused on exploring significant features in this dataset and proposing subsets of significant features to be used instead [10]. In 2015, new large UNSW-NB15 dataset consisting of current network threats was made available for research purposes [7]. This study employed data mining and machine learning techniques on UNSW-NB15 dataset in order to explore significant features in detecting network intrusions. The UNSW-NB15 dataset includes even more number of features than KDD'99 dataset (curse of dimensionality) and most of the features are not the same in these datasets. Therefore, it is hard to compare these datasets. The results then were compared with the findings of previous works using KDD'99 [10] and UNSW-NB15 datasets [12].

Two subsets of best features were examined, one being extracted from work in [12] and the other one was proposed by this work using machine learning techniques. The results indicate that Subset 2 proposed by this work improves Kappa statistic which means better intrusion detection rates. This subset includes two common features with the proposed subset of significant features in [10] where KDD'99 dataset was employed. It is important to mention that the KDD'99 Testing dataset contains 14 more types of attacks than the KDD'99 dataset [10] which is not true in case of UNSW-NB15 Training and Testing datasets. Both the UNSW-NB15 Training and Testing datasets include the same amount of the attack types which can be the reason to why the Kappa values for the Training and Testing datasets in UNSW-NB15 and KDD'99 are not following the same trend.

REFERENCES

[1] McHugh John, 2000, "Testing intrusion detection systems: a critique of the 1998 & 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory," ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262–294.

[2] McHugh John., 2001, "Intrusion and intrusion detection", IJIS (2001), 1, pp. 14-35, [Online] available:
http://www.icir.org/vern/cs261n/papers/ijis-published.pdf

[3] Ghorbani A., Lu W., and Tavallaee M., 2010, "Network Intrusion Detection and Prevention: Concepts and Techniques", Springer Science, LLC.

[4] 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.

[5] Mahoney M. V. and Chan P. K., 2003, "An analysis of the 1999 darpa/lincoln laboratory evaluation data for network anomaly detection," in Proc. 6th Int. Symp. Recent Adv. Intrusion Detection. Berlin, Germany: Springer Verlag, pp. 220–237.

[6] 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.

[7] 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.

[8] Harriet Taylor, 2015, "Biggest Cybersecurity Threats in 2016", CNBC, [Online] available:
http://www.cnbc.com/2015/12/28/biggest-cybersecurity-threats-in-2016.html

[9] Mnar Saeed Alnaghes, Fayez Gebali, 2015, "A Survey on Some Currently Existing Intrusion Detection Systems for Mobile Ad Hoc Networks", 2nd International Conference on Electrical and Electronics Engineering, Clean Energy and Green Computing (EEECEGC2015), pp. 12-18.

[10] Zargari S. and Voorhis D., "Feature Selection in the Corrected KDD-dataset," 2012 Third International Conference on Emerging Intelligent Data and Web Technologies, Bucharest, 2012, pp. 174-180.

[11] Ghorbani A., Gharibian F., 2007, "Comparative Study of Supervised Machine Learning Techniques for Intrusion Detection", 5th Conference on Communication Networks and Services Research (CNSR'07), pp. 350-358.

[12] Moustafa N. and Slay J., "The significant features of the UNSW-NB15 and the KDD99 sets for Network Intrusion Detection Systems", the 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS 2015), collocated with RAID 2015, 2016, [Online] available:
http://handle.unsw.edu.au/1959.4/unsworks_41254

[13] Engen, Vegard. Machine learning for network based intrusion detection: an investigation into discrepancies in findings with the KDDCUP'99 data set and multi-objective evolution of neural network classifier ensembles from imbalanced data. Diss. Bournemouth University, 2010.

[14] Pfahringer Bernhard, "Winning the KDD99 classification cup: bagged boosting", ACM SIGKDD Explorations Newsletter,V.1, Issue 2, 2000, [Online] available: http://dl.acm.org/citation.cfm?id=846200

[15] 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, [Online] available: http://neuro.bstu.by/ai/Todom/My_research/Papers-0/For-research/D-mining/Anomaly-D/KDD-cup-99/mlmta03.pdf

[16] Kumar, G. Sunil, and C. V. K. Sirisha, "Robust Preprocessing and Random Forests Technique for Network Probe Anomaly Detection.," International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-6, January 2012. [Online] available:
http://www.academia.edu/9521473/Robust_Preprocessing_and_Random_Forests_Technique_for_Network_Probe_Anomaly_Detection

[17] Bajaj and Arora, "Improving the Intrusion Detection using Discriminative Machine Learning Approach and Improve the Time Complexity by Data Mining Feature Selection Methods", International Journal of Computer Applications (0975-8887), Volume 76-No.1, August 2013. [Onine] available:
http://research.ijcaonline.org/volume76/number1/pxc3890587.pdf

[18] Pervez M. S. and Farid D. M., "Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs," The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014), Dhaka, 2014, pp. 1-6.

[19] Ingre B. and Yadav A., "Performance analysis of NSL-KDD dataset using ANN," 2015 International Conference on Signal Processing and Communication Engineering Systems, Guntur, 2015, pp. 92-96.

[20] Aghdam Hosseinzadeh M. and Kabiri, "Feature Selection for Intrusion Detection System Using Ant Colony Optimization," International Journal of Network Security, Vol 18, No.3, May 2016, pp.420-432. [Online] Available : http://ijns.jalaxy.com.tw/contents/ijns-v18-n3/ijns-2016-v18-n3-p420-432.pdf

[21] Stolfo S., Fan W., Lee W., and Prodromidis A., "Cost-based Modelling and Evaluation for Data Mining with Application to Fraud and Intrusion Detection: Results from the JAM Project", Computer Science Department, Columbia University, 1999, [Online] available: https://pdfs.semanticscholar.org/cd61/d9fda32950fb7c1510c1cba5a45ac69497f4.pdf

[22] Sung A. and Mukkamala S., "Identifying important features for intrusion detection using support vector machines and neural networks", Symposium on Applications and the Internet, IEEE, pp. 209-216, 2003.