Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Comparative Analysis of Selected Filtered Feature Rankers Evaluators for Cyber Attacks Detection

Olasehinde Olayemi

Received: 24 June 2021 Accepted: 14 July 2021 Published: 26 July 2021

Abstract

An increase in global connectivity and rapid expansion of computer usage and computer networks has made the security of the computer system an important issue; with the industries and cyber communities being faced with new kinds of attacks daily. The high complexity of cyberattacks poses a great challenge to the protection of cyberinfrastructures, Confidentiality, Integrity, and availability of sensitive information stored on it. Intrusion detection systems monitors? network traffic for suspicious (Intrusive) activity and issues alert when such activity is detected. Building Intrusion detection system that is computationally efficient and effective requires the use of relevant features of the network traffics (packets) identified by feature selection algorithms. This paper implemented K-Nearest Neighbor and Naïve Bayes Intrusion detection models using relevant features of the UNSW-NB15 Intrusion detection dataset selected by Gain Ratio, Information Gain, Relief F and Correlation rankers feature selection techniques.

Index terms— features rankers, cyber-attacks, intrusion, classification, computer security, network packets.

1 I. Introduction

he increase in global connectivity and rapid expansion of computer usage and computer networks has made the security of the computer system an important issue; with the industries and cyber communities being faced with new kinds of attacks daily. The high complexity of intrusion poses a great challenge to the protection of cyberinfrastructure and the Confidentiality, integrity, and availability of sensitive information stored on them. The state of computer security is complicated, it is difficult to have a system that is completely free from attacks. The nature and the means of executing cyberattacks make it prevalent. Cyber-attacks are easy and cheap to execute, all that is require to stage a cyber-attacks are computer system and internet access, the nature of internet makes

Author: e-mail: olaolasehinde@fedpolel.edu.ng

The objectives of feature-ranking are threefolds: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the dataset [3]. The FS also reduces the computational time to implement an online Network Intrusion Detection System (NIDS) [4]. The efficiency of the FS methods is measured by its accuracy at removing noisy and redundant features [5]. The quality of the network traffics /dataset does not only help to build effective NIDS but also shows its potential Year 2022 launching a cyber-attack is less risky than physical attacks, and not constrained by geographical distance. [1]. Network traffics contain different types of protocols and services which accounted for the multiple features in the network packet. Some of these features are redundant or irrelevant and does not contribute the classification of the network packets as either attack or normal network packets. The redundant features are the primary causes of increasing the false alarm rate (FAR) and decrease in detection accuracy. Feature Selection (FS) Techniques are the methods used to determine the relevant features of a dataset. It is an efficient way to reduce the dimensionality of a problem [2]. Different FS techniques existed in classification and clustering problems. They are i) Filter method ii) Wrapper Method and iii) Embedded method. The filter methods are used to select the features based on the scores in various statistical correlations. Wrapper

method uses a greedy approach in feature selection. It evaluates all possible combination and produces the result for Machine learning. The embedded method combines the advantage of two models. Filtered Feature selection algorithms can be grouped into two categories from the point of view of a method's output: feature-ranking and feature-subset selection. Feature-subset selection focuses on selecting best subset of features that satisfies an evaluation criterion, feature-ranking on the other hand ranks features according to certain evaluation criterial, which measures the relevance of individual feature to the target class, and select the set of ranked features that gives the best evaluation performance, the drawback of this methods is that, a features that is not relevant to the target class on its own, can be very relevant when combined with others features.

efficiency during deployment in a real-life operating environment. NIDS analyze and monitor network traffic to detect suspicious activities and vulnerability in the system [6]. The effectiveness of NIDS is evaluated based on its ability to correctly identify network traffics as attacks traffic or benign traffics (normal) in a comprehensive dataset that contains normal and abnormal behaviors [7].

Feature-ranking techniques ranked features independently without involving any learning algorithm based on statistics, information theory, or some functions of classifier's outputs [8]. It consists of scoring each feature according to a particular evaluation criterion [9]. Several authors have proposed various features selection methods. In the work of Wang and Gombault [9], IG and Chi-squared were applied to extract nine most important features from the forty one features to build Bayesian Network and C 4.5 decision tree classifiers to detect DDoS attack in the network. Results obtained shows that the detection accuracy remains the same while the overall efficiency improved. Authors in [10] proposed a multi-filter feature selection techniques that combines the results four filter selections methods on NSL-KDD intrusion network dataset to achieve an optimum selection. C4.5 decision tree evaluation of the thirteen optimal selected features out of forty one features shows a high detection rate and classification accuracy when compared to the forty-one features and other classification techniques. [11] Proposed a feature selection method based on Decision Dependent Correlation (DDC). Mutual information of each feature and decision is calculated and top 20 important features {feature no.: 3, 5, 40, 24, 2, 10, 41, 36, 8, 13, 27, 28, 22, 11, 14, 17, 18, 7, 9 and 15} are selected and evaluated by SVM classifier. The classified result is 93.46% detection accuracy. [12] Applied Information Gain (IG), Correlation-based (CFS), Gain Ratio (GR) feature selection to reduce the dimensionality of NSL-KDD dataset, and built a decision tree classifiers of the three feature selection methods. The three classifier recorded an improved performance than the classifier built with the whole NSL-KDD dataset.

[13] Proposed a feature selection method that combined three filter methods; Gain ratio, Chi-squared and Relief F (triple-filter) in a cluster-based heterogeneous Wireless sensor network (WSN) for attacks classification. 14 important features of the NSL-KDD intrusion detection benchmark dataset out of the 41 original features were extracted for intrusion detection classifier. Results obtained show that the proposed method can effectively reduce the number of features with a high classification accuracy and detection rate in comparison with other filter methods.

2 II. Methodology

The proposed architecture of the Comparative Analysis of Selected Filtered Feature Rankers Evaluators for Cyber Attacks Detection is depicted in Figure **??**. The discretization of the UNSW-NB15 dataset was first carried out to make it suitable for machine learning. Four Filtered Feature Rankers Evaluators algorithms; (Information Gain, Relief F, Gain Ration, and Correlation) rankers were used to rank and select the optimal relevant features of training and testing UNSW-NB15 intrusion datasets. The training dataset with the all it feature and the reduced features of the training datasets were used to train the K Nearest Neighbors (KNN, and Naive Bayes' algorithms. The testing dataset with the all it features and the reduced features of the testing dataset were used to evaluate the two classifiers. The model's training is depicted in black arrow lines while the model's evaluation is depicted in red arrow lines in the figure. The results of the evaluation for each reduced dataset were analyzed.

3 a) Description of UNSW-NB15 Dataset

The UNSW NB-15 dataset was developed using the IXIA Perfect Storm tool by the cybersecurity research group at the Australian Center for Cyber Security [14]. It is a fusion of normal network traffic packets, and synthetic modern-day network traffics attacks. The training and testing contain 82,332 and 174,341 records with 49 features each, respectively [14]. The dataset comprises nine attack categories and normal traffic, and it is suitable for the effective detection of existing and new attacks [14]. The details of both attack and normal traffic, coupled with the records in the training and testing categories, are presented in Table 1.

4 i. Description of Attributes Selection Evaluators

Attributes Selection Evaluator ranks features based on their relevant to the target class, ranking is a way of evaluating relevant features and selecting a minimal set of features based on given criteria in order to build simple models, that take less time to compute and become more understandable Feature ranking evaluation criterion compute the score S(fi) of feature (fi) of the training dataset. By convention a high score implies important (relevant) of the feature to the target class and select the k highest ranked features according to S. This is usually not optimal, but computationally efficient and often preferable to other, more complicated feature

selection methods that involve searching through the entire search space. In this study, we use four feature-ranking techniques; Correlation Attribute Evaluator (CAE), Gain Ratio Attribute Evaluator (GAE), Information Gain attribute Evaluator (IGAE) and Relief F Attribute Evaluator (RFAE).

5 a. Correlation Attribute Evaluator (CAE)

Correlation Attribute Evaluator (CAE), evaluate Attribute using correlation analysis. The correlation between each attributes x and the target class Y, can be measured by finding correlation coefficient. A good feature is expected to have a higher correlation coefficient between it and target class. In correlation attribute evaluator method the attributes are considered based on their values where each value is treated as an indicator. CAE handles only nominal attributes input for evaluation and it uses Pearson's formula for computing correlation coefficient. for a candidate feature xi ? X and regression target Y the Pearson correlation coefficient is given by (1) where cov designates the covariance and var the variance.

6 b. Information Gain attribute Evaluator (IGAE)

Information gain (IG) measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. Concretely, it measures the expected reduction in entropy (uncertainty associated with a random feature) [15], it is given by equation 2.

7 Info Gain (Class, feature) = H (target class (Y)) -H (target class(Y) | feature (X))

8 ???? = ??(??) ? ??(??|??) ? ??(??) ? ??(??|??)

(2)

9 Where H (Y); t he e ntropy of t he t arget c lass H (Y) and ??(?? | ??) is t he e ntropy of t arget class gi ven a c ertain attribute ??.

The entropy of the target class Y is given by equation (3).

10 ??(??)

= ? ? ??(??) log 2 ???(??)? ?????**(3)**

Equation (**??**) gives the entropy of target class Y after observing feature X.

11 ??(??|??)

= ? ? ??(??) ? ??(??|??) log 2 ???(??|??)? ????? ????? (4)

12 c. Gain Ratio Attribute Evaluator (GRAE)

Gain ratio (GR) is a modification of the information gain that reduces its bias. It considered the number and size of branches in choosing an attribute. It assess the value of an attribute by measuring its gain ratio with respect to the target class [16]. the root attribute is the attribute of the UNSW-NB15 with the highest gain ratio, the gain ratio is the ratio of the information gain and the split information for the attribute as presented in equation 5.

13 Gain Ratio = ????????????????????? ????????(??) ?????????? ?????????????????????(??)

The information gain of attribute X is given by equation 2.

The Split information value of an attribute is chosen by taking the average of all the values in the domain of current attribute. It is given by equation 6.

14 ??????????(??)

= ? ? |??| |??| ? log 2 |??| |??| ?????? d. Relief Attribute Evaluator (RFAE)

Relief Attribute Evaluator (RFAE) sample an instance recurrently using distance function taking into consideration the value of the given attribute for the nearest instance of the same and different class [13]. The original Relief algorithm, proposed by Kira and Rendell [8], is a two-class filtering algorithm for features normalized to [0, 1]. Each feature is initially assigned a zero weight. An A-dimensional training example R is chosen randomly and the Euclidean distance to all other instances calculated. Denote the nearest hit in the same class H, and the nearest miss in a different-class M. Since a good feature R[A] should be able to separate class

values, it should have a small distance to H and a large distance to M. Hence W[A] is adjusted to reward good features and penalize poor ones. The final selection of features is made by selecting those large W[A], (that is . those that exceed a given threshold.)

15 ii. Description of Machine learning techniques

Two machine learning algorithms, namely; KNN and Naïve Bayes were used in this study to build the intrusion detection system.

16 a. K-Nearest Neighbor

Let p i and qt represent the instance to be classified and the other instances in the dataset having the same number of features as P respectively, Knearest neighbor Euclidean distance between p i and q t is defined in equation 7.2 1 (,q) ( ) n i t i i i d p p q = = ? ?**(7)**

From equation (3), a given instance will be classified as the attack categories having majority attacks among top k closest instance to the given instance.

17 b. Naïve Bayes

Given the UNSW-NB15 intrusion detection dataset that have X number of attributes called the predictors (X = x 1 , x 2 ,...,x n) and another attribute y called the class label, with ten members y 1 ,....y 10 , the Naive Bayes probability that a class y j will be assigned to a given unlabelled instance X is given in equation 8.

18 ??(??

?? | ?? 1 , ? . , ?? 43) = ??(?? ??)??(?? ?? |?? ??) ??(?? ??) (? ?? = 0,1, ? ? ,9)**(8)**

Maximum posterior probability for classifying a new instance attack categories is given in Equation 9.?? = arg ?????? ?? ???? ?? ? ?? 9 ?? =0 ??? ?? ??? (?? 1 , ?? 2 , ? ?? 43 | ?? ??)**(9)**

19 c) Performance Evaluation Metrics

Performance evaluation metrics play significant roles in assessing the predictive performance of the model and determining the model's fitness for the classification purpose. The confusion matrix, also known as the error matrix, is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model.

20 i. Accuracy

Accuracy (ACC) is the ratio of all correctly classified network packets to the total number of instances in the intrusion test dataset, it is given by equation.1. An accuracy of 1 implies error rate of 0 and an accuracy of 0 indicate error rate of 10.

21 TP TN ACC FN FP TN TP

+ = + + +**(10)**

ii. False Positive Rate (FPR) or False Alarm Rate (FAR) False Positive Rate (FPR) or False Alarm Rate (FAR) is the proportion of actual network attacks cases that were predicted as Normal packets by the model. FPR should be as low as possible to avoid unwanted false alarms. it is given by equation 11.

22 FP FPR FAR

TN FP = = + (11) iii. Precision Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. it is given by equation 12.

23 Pr

TP ecision TP FP = +

24 III. Experimental Setup and Results Discussion

Four feature selection rankers were used to select the relevant features of the UNSW-NB15 intrusion dataset to build Intrusion Detection System. Two classification models (Naïve Bayes and KNN) were used to build the Intrusion detection system for the cyberattacks detection and classification of Network traffic a computer network. The relevant features of the UNSW-NB15 intrusion detection dataset selected by the four (4) filter features rankers are presented in Table 2. Relief F features ranker selected thirteen (13) features, Information Gain features ranker selected fourteen (15) features, Gain ration selected fifteen (**??**4) features, while correlation ranker selected eleven (11) features. It was observed that Proto, Service and Ct_dst_sport_ltm were the only features that were commonly selected by the feature selection algorithms. Thus, they were the features observed

to be the most relevant based on the four methods of evaluating the relevance and having the greatest importance in the detection and classification of attack packets in the network traffics.

These reduced selected features with the complete features were used to build Intrusion detection systems of Naïve Bayes and KNN. The UNSW-NB15 testing dataset was used to evaluate all the classifiers. The confusion matrix and the performance of the KNN and Naïve Bayes classifiers with each of the selected features of the ranking feature technique is presented in Table **??** and 4 respectively. From tables 3 and 4, it shows that KNN and Naive Bayes intrusion detection models of Relief F selected features that identified thirteen (13) features recorded the best performance in terms of detection accuracy, classification precision and false alarm rate. The Intrusion detection models of KNN and Naïve Bayes of the fourteen (14) features identified by the Gain Ratio recorded the second best performance in terms of the selected performance metrics. Correlation and information Gain recorded the third and the fourth performances among the Rankers Features Selection Techniques respectively. Intrusion detection models of the two classifier with all of features of the UNSW-NB15 intrusion detection network dataset recorded the least and poorest performance, this result shows the importance and ability of the Rankers Features Selection Techniques to improve the performance of intrusion detection models.

The comparison analysis of the two classifiers shows that, KNN intrusion detection models recorded better detection accuracy, precision and false alarm rate than the Naïve Bayes model in the classification of UNSW-NB15 intrusion detection network dataset, it can be further deduced that the Relief F features selection method with KNN is the best-performing algorithm for the detection of network packets of UNSW-NB15 intrusion detection dataset. The comparison analysis of the selected Rankers Features Selection Techniques with each machine learning algorithms, based on the selected performance metrics is illustrated in Figure **??**.

( ) E

$$\mathcal{R}(i) = \frac{cov(X_i, Y)}{\sqrt{var(X_i)var(Y)}}$$

Figure 1: T

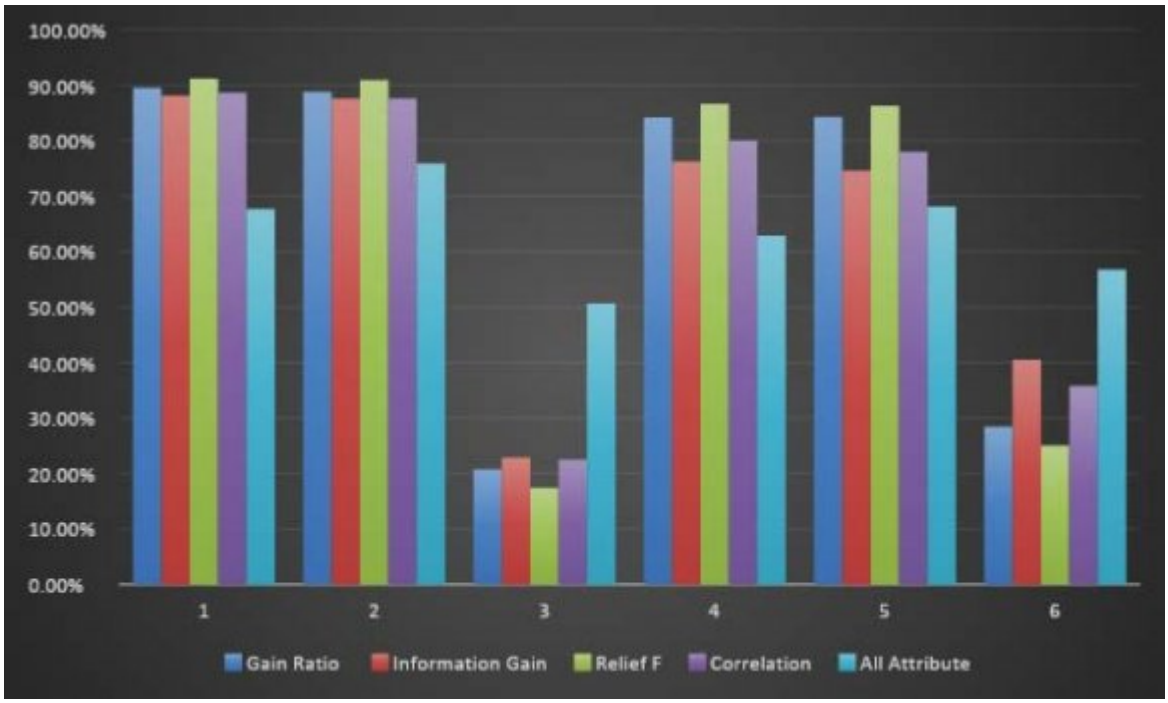

Figure 2:

1

Models
Evaluation

[Note: b) Data Munging and AnalyticThis section outlines the Feature Rankers Evaluators and the machine learning techniques used for this study.]

Figure 3: Table 1 :

2

Year 2022				
Relief F (13)		Gain Ratio (14)	Information Gain (15)	Correlation Ranker (11
proto, state, ct_dst_src_ltm	service, smean,	Proto, service, smean, ct_state_ttl, ct_dst_sport_ltm,	proto, service, state, smean, ct_state_ttl,	swin, stt, dwin,
Sttl, ct_state_ttl,		ct_dst_dport_ltm,	ct_dst_sport_ltm,	Dwin, ct_dst_sport_ltm,
ct_srv_src, ct_dst_sport_ltm, ct_srv_dst, ct_dst_ltm, ct_src_ltm	dttl,	ct_srv_dst, dbytes, rate, dmean ,dpkts , dur, sload	Sbytes, ct_src_dport_ltm, Sbytes, stcpb, dtcpb	ct_src_dport_ltm, dttl, cprtt,

Figure 4: Table 2 :

relevant features to model intrusion detection systems of KNN and Naïve Bayes machine learning algorithms. The Results of the features ranking shows that Relief F features ranker selected thirteen (13) features, Information Gain features ranker selected fourteen (15) features, Gain ration selected fifteen (**??**4) features, while correlation ranker selected eleven (11) features. Features selected by Relief F recorded the best performance, Gain Ratio recorded the second best performance. Correlation and information Gain recorded the third and the fourth performances respectively, while the use of all the features recorded the least and poorest performance, this result shows the importance and ability of the Rankers Features Selection Techniques to improve the performance of intrusion detection models. All the KNN models recorded better performance than all Naïve Bayes models. The models' performance were measured in terms of Classification /detection accuracy, precision and false alarm rate. The results further shows models of KNN with the reduced features of Relief F features selection method recorded the best overall performance for the detection of network packets of UNSW-NB15 intrusion detection dataset.

.1 a) Ethical Standard Funding

This research work is self-funded research undertaken by the authors at the Department of Computer Science, School of Applied sciences, Federal Polytechnics, Ile-Oluji, Nigeria.

.2 Conflict of Interest

The corresponding author states that there is no conflict of interest.

[Gogoi et al.] , P Gogoi , M H Bhuyan , D K Bhattacharyya .

[Mitchell ()] , T Mitchell . *Machine Learning. McGraw-Hill* 1997.

[Int. J. Mach. Learn. Cyber ()] , 10.1007/s13042-015-0469-8. https://doi.org/10.1007/s13042-015-0469-8 *Int. J. Mach. Learn. & Cyber* 2017. 8 p. .

[Qu et al. (2005)] 'A new dependency and correlation analysis for features'. G Qu , S Hariri , M Yousif . 10.1109/TKDE.2005.136. *IEEE Transactions on Knowledge and Data Engineering* Sept. 2005. 17 (9) p. .

[Chandrashekar and Sahin ()] 'A survey on feature selection methods'. Girish Chandrashekar , Ferat Sahin . *Computers and Electrical Engineering* 0045-7906. 2014. 40 p. .

[Isabelle ()] 'An Introduction to Variable and Feature Selection'. G Isabelle , AndreE . *Journal of Machine Learning Research* 2003. 3 p. .

[Tariq et al. ()] *Analysis of Security Techniques for Detecting Suspicious Activities and Intrusion Detection in Network Traffic. Semantic Scholar. 2012https:// www.semanticscholar.org/paper/Analysis-of-Securit y-Techniques-for-Detecting-and-Tariq-Arshad*, W Tariq , M Arshad , M Saqib , N Gul . 2012. (a993 dab8bcd79ec8468c36489e2acabf957b71d0#citing -papers)

[Aldehim ()] *Determining appropriate approaches for using data in feature*, G Aldehim , WangW . 2017.

[Wang ()] 'Efficient detection of DDoS attacks with important attributes'. W Wang , GombaultS . *the 3rd IEEE International conference on Risks and Security of Internet and Systems (CRiSIS'08)*, (Tozeur, Tunisia) 2008. p. .

[Opeyemi et al. ()] *Ensemble-based Multi-Filter Feature Selection Method for DDoS Detection in Cloud Computing, journal of wireless communication and networking*, O Opeyemi , R C Kim-Kwang , D Ali , X Zheng , Mqhele . 10.1186/s13638-016-0623-3. 2016. 2016.

[Duch et al. ()] 'Feature Ranking, Selection and Discretization'. W Duch , T Winiarski , J Biesiada . *Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP)*, 2003. 2003. p. .

[Sang-Hyun and Hee-Su (2014)] 'Feature Selection using Attribute Ratio in NSL-KDD data'. C Sang-Hyun , C Hee-Su . *International Conference Data Mining, Civil and Mechanical Engineering (ICDMCME'2014)*, (Bali (Indonesia) 2014. Feb 4-5, 2014.

[Schreier (2015)] *On Cyberwarfare, DCAF Horizon*, F Schreier . https://www.dcaf.ch/sites/default/files/publications/documents/OnCyberwarfare-Schreier.pdf 2015. 2015. Accessed 2 nd April, 2021.

[Kalita ()] 'Packet and Flow Based Network Intrusion Dataset'. J K Kalita . 10.1007/978-3-642-32129-0_34. https://doi.org/10.1007/978-3-642-32129-0_34 *Communications in Computer and Information Science* 2012. 2012. p. .

[Olasehinde et al. ()] 'Reduced Features Intrusion Detection Systems Classification Accuracy Improvement'. O O Olasehinde , K Williams , B O Adegoke . 10.1023/A:1006624031083. http://dx.doi.org/10.1023/A:1006624031083 *International Journal of Scientific & Engineering Research* 2019. 2019. 10 (12) p. .

[Kavita and Pranjali (2014)] 'Survey on Data Mining Techniques for Intrusion Detection System'. P Kavita , D Pranjali . *International Journal of Research Studies in Science, Engineering and Technology* April 2014. 1 (1) p. .

[Munson and Khoshgoftaar ()] 'The dimensionality of program complexity'. J C Munson , T M Khoshgoftaar . *Proceedings of the 11th international conference on Software engineering, ICSE '89*, (the 11th international conference on Software engineering, ICSE '89) 1989. ACM. p. .

[Moustafa and Slay J ()] *UNSW-NB15: A Comprehensive DataSet for Network Intrusion Detection Systems Military Communications and Information Systems Conference*, N Moustafa , Slay J . 2015. p. .