2019 International Conference on Innovative Computing (ICIC)

# Analysis of Classification Techniques for Intrusion Detection

Umair Ahmad
*Department of Computer Science*
*University of Management and Technology*
Lahore, Pakistan
15026050017@umt.edu.pk

Hira Asim
*Department of Software Engineering*
*University of Management and Technology*
Lahore, Pakistan
hira.asim@umt.edu.pk

Malik Tahir Hassan
*Department of Software Engineering*
*University of Management and Technology*
Lahore, Pakistan
tahir.hassan@umt.edu.pk

Sheraz Naseer
*Department of Computer Science*
*University of Management and Technology*
Lahore, Pakistan
sheraz.naseer@umt.edu.pk

***Abstract*— In the field of machine learning, many supervised and unsupervised methods have been developed to develop an effective and efficient Intrusion Detection System (IDS). However, a comprehensive comparative analysis of different intrusion/anomaly detection methods on standard datasets is required. In this work, a standard dataset NSL-KDD is used for comparative analysis of six popular classifiers. The normal and anomalous records have been identified using following classification techniques: Decision Tree, Naïve Bayes, Ada Boost, MLP, Random Forest and Liner SVM. Implementation is done using Python based scikit-learn library. Results yielded are evaluated and compared on the basis of standard performance evaluation measures like accuracy, precision, recall, F1 score, confusion matrix and AUROC curve. The highest percentage accuracies have been achieved by MLP model, i.e., 100% and Decision Tree Model i.e., 98%. Moreover, the comparative analysis of the data mining techniques applied in last five years on two standard and publicly available datasets ISCX-IDS2012 and UNSW-NB15 has also presented in this paper.**

***Keywords*— *Intrusion Detection System, Network Security, NSL-KDD, ISCX-IDS2012, UNSW-NB15, MLP, Decision Tree, Classification, Machine Learning, Data Mining***

## I. Introduction

In recent decades, online activity has augmented because of several kinds of development that are associated with internet. The online networks and cloud systems have been developed, which not only facilitate the flow of information from one end to another end, but also store individual data over the internet. Individuals, groups and companies use various kinds of online systems to receive, send and store data. For this purpose, multiple kinds of networks are developed that ensure that information flows and stores without any undesired disturbance. Intrusion detection is a process followed to find out whether the system is intrusive or not as per user activities and system logs [1].

Unauthorized penetrations into a network have become more potent and diverse. To counter the intrusions, several systems have been developed that detect forceful intrusion(s). Such systems are known as Intrusion Detection Systems (IDS). Two major classifications of IDS are: Host Based IDS (HBIDS ) and Network Intrusion Detection System (NIDS). Some of the IDS are able to detect multiple kinds of unauthorized access whereas a very few are able to detect and counter contemporary cyber or network attacks.

A conventional Intrusion Detection System is installed on a network to protect network packets from the sniffers. A typical Intrusion Detection System analyzes packets of gathered data and alarms the system, if they detect any anomaly. Some of the Intrusion Detection Systems are able to counter attacks by following a particular protocol that intends to reduce damage in case of cyber-attack.

To address the challenges of intrusion detection, data mining techniques have been used as an instrument to develop Intrusion Detection Models, which aid in auditing large chunks of data to structurally observe and record various kinds of anomalies in the data. Some of these models are based on classification techniques, which identify and classify intrusion detection data. Some popular classification techniques are Support Vector Machines, Decision Trees, Associative Classification, k-Nearest Neighbor Classifiers, and Rule Induction Model etc. [2].

The main aim and motivation of this research initiative is to perform a comparative analysis on different classification techniques for the identification of the normal and anomalous records present in network related data. So, in this paper a publically available standard dataset NSL-KDD [3] is used for comparative analysis of six popular classifiers. For the identification of the normal and anomalous records, following six classification algorithms are applied on the processed NSL-KDD dataset named Decision Tree (DT), Naïve Bayes, Ada Boost, Multi-Layer Perceptron (MLP), Random Forest and Liner Support Vector Machine (SVM). All above mentioned classifiers are applied on the dataset using Python based scikit-learn machine learning library [4]. Results yielded are furthermore analyzed and compared using standard performance evaluation measures like accuracy, precision, recall, F1 score, confusion matrix and AUROC curve.

978-1-7281-4682-9/19/$31.00 ©2019 IEEE

The rest of the paper is organized as follows: section II is discussing the related and previous work done for intrusion detection as well as comparative analysis of data mining techniques on ISCX-IDS2012 [5] and UNSW-NB15 [6] datasets, section III is about the methodology adopted to get the desired results, section IV is analyzing and discussion the results deduced after applying the proposed methodology and finally section V is concluding the paper and discussing future directions.

## II. Literature Review

The evolution of Intrusion Detection System (IDS) is directly effected by the means data are collected and furthermore analyzed. Therefore, there is strong emphasis on data classifying techniques. However, to understand why datasets and their classification is imperative, it is important to understand intrusion detection (in-depth) and retrieving of data pertaining to data intrusion [7].

Contemporary online networks and systems are used to store and retrieve various kinds of data for several purposes. As the trust of storing and retrieving data is augmenting on these online networks and systems, so the size and complexity of these systems are growing as well. In addition, these networks and systems can become target of different kinds of cyber attacks. For instance, an intrusion attempt could be highly complex, potent and sophisticated or it could be weak and less effective [8].

In a typical Intrusion Detection System based on Neural Network, there are three layers that are not exposed and are usually installed between the evidence or information pertaining to input and output. The objective of such system is to produce information at the layer relating to output, in such a manner that size of error is extremely small [9].

In some subsequent studies, for instance in [10], ideas of such neural network based intrusion detection system are purposed which are hierarchical in functioning and they are combined with multiple statistical models in order to detect anomaly or intrusion. In these types of systems, the output produced by the selected neural network is a continuous variable. These systems consider negative value as an intrusion or anomaly (for certain) and positive value as no intrusion.

There are various intrusion detection systems that are based on statistical models and theories. One such system is proposed in [11] in which chi-square statistical theory is used to detect anomalous data. Generally, Chi-square based methods require devising of information profile that is incorporated in information system. Such systems detect abnormality when they identify fluctuations in data. These fluctuations are classified as anomalies and intrusions as these are deviations from identified normal behavior.

Kruegel, C. and Toth, T. in [12] purpose unit which uses statistical processing to detect anomalous behavior or traffic within a network. Generally, such statistical-based detection models are designed to detect rare instances of intrusion, such as Root to User attacks and User to Root attacks. In such systems, a metric is used to facilitate the detection system to identify identical requests for services. This system is also capable to score the number anomalous requests.

From the systematic study of different classification techniques like rule based detection, neural networks, fuzzy logic, hidden Marko model and random forest model etc., it has been observed that most of the classification techniques use similar instruments and processes to execute the intrusion detection process. However, some of these techniques produce more efficient results [13].

### A. *Comparative Analysis of Data Mining Techniques on ISCX-IDS2012 and UNSW-NB15 Datasets*

In this research work, the comparitive analysis of two publically available datasets i.e., ISCX-IDS2012 [5] and UNSW-NB15 [3] is also presented. The NSL-KDD dataset is based upon anomaly traces only whereas the research community have recorded the datasets like ISCX-IDS2012 and UNSW-NB15 which contain the present day normal and anomalous records. We have collected the information of techniques by which the datasets have been experimented as well as the results of accuracy, precision and false alarm rate. We have complied the summary of data mining techniques' results for both datasets in tabular form.The summary is presented in the TABLE I and TABLE II respectively.

TABLE I. Summary of Experimental Results of ISCX-IDS 2012 Dataset **[5]**

| Techniques applied | Authors | Year | Contribution | Accuracy | Precision | False Alarm Rate |
|---|---|---|---|---|---|---|
| Semi-supervised ML approach DDoS detection | M. Idhammad et al.[14] | 2018 | An online sequential semi-supervised ML approach has been presented for DDoS detection based on Co-clustering, Network Entropy Estimation, Extra-Tree Algorithm and Information Gain Ratio. | 99.88 | N/A | N/A |
| FAR + DR + feature selection | T. Hamed et al.[15] | 2018 | Based on bigram technique and feature selection method called Recursive Feature Addition (RFA), a Network Intrusion Detection Systems (NIDS) has been proposed. | 50.1 % | N/A | 0.5 |
|  | Yu et al.[16] | 2017 | Session-based detection model for network intrusion using deep learning architecture. | 99.47 % | 99.39% | N/A |
| KMC(K-Means Classifier) + NBC (Naïve Bayes Classifier) | Yassin et al.[17] | 2013 | An integrated machine learning algorithm named KMC + NBC has been proposed for significant improvement in detection rate and accuracy of the intrusion detection systems. | 99.0% | N/A | 2.2 |

TABLE II. Summary of Experimental Results of UNSW-NB15 Dataset [3]

| Techniques applied | Authors | Year | Contribution | Accuracy | Precision | False Alarm Rate |
|---|---|---|---|---|---|---|
| REPTree algorithm | M. Idhammad et al. [14] | 2018 | Revealed the online concurrent semi-supervised ML approach for network-based detection of DDoS Entropy, Co-clustering, Information Gain Ratio and Extra-Tree Algorithm. | 93.71 % | N/A | N/A |
| Decision Tree | Mustafa et al.[18] | 2016 | Examined the complexity of dataset using statistical analysis. | 85.56% | N/A | N/A |
| Naïve Bayes | Mustafa et al.[18] | 2016 | Examination of feature correlation. | 82.07 % | N/A | N/A |
| Linear Regression | Mustafa et al.[18] | 2016 | Assess complexity with regard to accuracy. | 83.15 % | N/A | N/A |
| ANN | Mustafa et al.[18] | 2016 | Assess complexity with regard to false alarm rates. | 81.34 % | N/A | N/A |
| EM clustering | Mustafa et al. [18] | 2016 | Examination of feature correlation. | 78.47 % | N/A | N/A |
| Naïve Bayes | Mustafa et al. [19] | 2015 | Comparison of the efficiency and reliability | N/A | N/A | N/A |

## III. Methodology

NSL-KDD dataset is basically an upgrade of KDD-cupp99 dataset [20] in which the redundant entries have been removed, that was causing biased classification results. As the experimented dataset is available publically, category encoding is done on the dataset in the preprocessing phase. Most of the dataset's features are nominal and some models do not work on nominal values, so category encoding method converts these features to numerical values [21]. The objective of selecting and testing this dataset is to compare existing methods and come up with an effective and pro-active intrusion detection model.

NLS-KDD dataset has 41 features which identify various kinds of anomalies, intrusions and normal data. Three features named *'protocol-type'*, *'service'* and *'flag'* are symbolic features which require quantitative conversion before testing on intrusion detection model. We have a two-class problem where network traffic is to be classified as either normal or anomalous. A number of encoding schemes can be used for feature encoding to numerical values, which can improve model's results. We have used 'LeaveOneOutEncoder' [21] scheme on three selected features, i.e., *'protocol-type'*, *'service'* and *'flag'*. The intrusion or attacks are grouped together in the NSL-KDD dataset. The forged attempts are grouped as Probing, Users to Root, Denial of Service (DoS) and Root to Local Attack, which are briefly discussed as follows.

### A. Probing

Probing is considered as a serious threat so, NLS-KDD dataset classifies it as an attack. Such attacks can be identified using their features. For instance, source-bytes and duration of connection are used to identify such attacks.

### B. User to Root

In this attack, the intention of the attacker is to access administrative control and privileges by identifying and exploiting existing vulnerabilities of a system or network. Any unauthorized access to super local privileges is a form of an attack and when such attack is detected, system must be alarmed.

### C. Denial of Service (DoS)

Denial of Service is a serious threat and a form of intrusion which aims to reduce the existing resources of a network or a system. Therefore, NLS-KDD dataset categorizes DoS as a kind of attack or anomaly.

### D. Root to Local Attack

In such attacks, a remote machine is used to penetrate into a system or network. In NLS-KDD data, such attacks can be identified from the source of origin, duration of connection and service requested.

For the experimentation, the NLS-KDD dataset is divided into training and testing portions. The training data *KDDtrainPlus* consists of 125973 records whereas the testing data *KDDtestPlus* has 22543 tuples.

## IV. Results and Discussion

As per the proposed methodology, we have applied multiple classification techniques using the scikit-learn Python's library [4]. TABLE III is showing the accuracy, precision, recall and F1- score computed against the applied classifier on the training (*KDDTrainPlus*) and testing (*KDDTestPlus*) portions of the NSL-KDD dataset.

TABLE III. Accuracy, Precision, Recall and F1-Score of the classifiers applied on the *KDDTrainPlus* and *KDDTestPlus* datasets.

| Classifier | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| **MLP** | 1.0 | 1.0 | 1.0 | 1.0 |
| **Decision Tree** | 0.986 | 0.99 | 0.99 | 0.99 |
| **Naïve Bayes** | 0.765 | 0.84 | 0.77 | 0.77 |
| **Ada Boost** | 0.743 | 0.84 | 0.74 | 0.75 |
| **Random Forest** | 0.743 | 0.81 | 0.74 | 0.75 |
| **Linear SVM** | 0.974 | 0.97 | 0.97 | 0.97 |

As per the statistics present in the TABLE III, it can be clearly observed that against each classifier, we got reasonably good results. However, MLP and Decision Tree classifiers have achieved the highest accuracies, i.e., 1.0 and 0.986.

The computational time taken by each classifier is also recorded. Fig. 1 is showing the computational time taken by each of the classifier applied on the training as well as testing datasets.

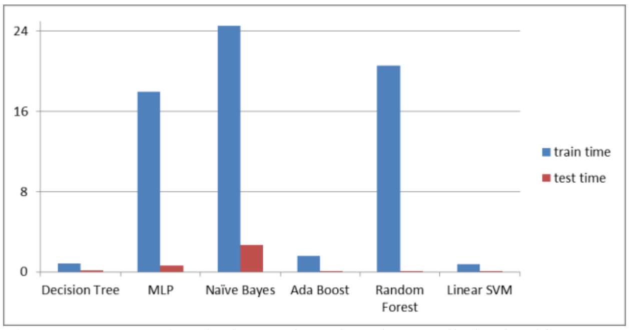

Fig. 1. Computational time taken by the applied classifiers on *KDDtrainPlus* and *KDDtrainPlus* datasets

It is clear from the Fig. 1, Naïve Bayes classifier has taken the maximum computational time i.e., 24.52 seconds training time and 2.68 seconds testing time. On the other hand, Linear SVM and Decision Tree have taken the smallest training time.

## *A. Evaluation of the Results*

For the evaluation of the correctness of the results yielded by the classifiers, following evaluation metrics are taken into consideration.

*1) Confusion Matrix:* The confusion matrix is used to show the correct and incorrect instances of a classifier model on a set of test data for which the true values are known. The matrix is N x N format, where N is number of classes (target). The performance of the classifier can be evaluated using the data in the confusion matrix *[22]*. A confusion matrix consists of four different combinations of actual and predicted values. As per our methodology, we have defined these labels as follows:

*a) True Positive (TP):* When an anomolous record is classified as anomaly, the outcome will be a TP.

*b) False Positive (FP):* When a normal instance is considered as anomaly. The result will be considered as FP.

*c) True Negative (TN): When the classifier detects a normal trafic instance as normal, it is called a TN.*

*d) False Negative (FN): When anomalous instances are considered as normal. The outcome will be known as FN.*

From TABLE IV to TABLE IX, we have presented the results of the classifier models in a confusion matrix form.

TABLE IV. Confusion Matrix of Ada Boost Classifier Model for *KDDTestPlus* dataset

| **Ada Boost** | | **Actual Labels** | | **Total** |
|---|---|---|---|---|
| | | **Anomalous** | **Normal** | |
| **Predicted Labels** | **Anomalous** | TP = 9433 | FN = 277 | **9710** |
| | **Normal** | FP = 5507 | TN = 7326 | **12833** |
| | **Total** | **14940** | **7603** | **22543** |

TABLE V. Confusion Matrix of Decision Tree Classifier Model for *KDDTestPlus* dataset

| **Decision Tree** | | **Actual Labels** | | **Total** |
|---|---|---|---|---|
| | | **Anomalous** | **Normal** | |
| **Predicted Labels** | **Anomalous** | TP = 9534 | FN = 176 | **9710** |
| | **Normal** | FP = 418 | TN = 12415 | **12833** |
| | **Total** | **9961** | **12591** | **22543** |

TABLE VI. Confusion Matrix of MLP Classifier Model for *KDDTestPlus* dataset

| **MLP** | | **Actual Labels** | | **Total** |
|---|---|---|---|---|
| | | **Anomalous** | **Normal** | |
| **Predicted Labels** | **Anomalous** | TP = 9710 | FN = 0 | **9710** |
| | **Normal** | FP = 0 | TN = 12833 | **12833** |
| | **Total** | **9710** | **12833** | **22543** |

TABLE VII. Confusion Matrix of Random Forest Classifier Model for *KDDTestPlus* dataset

| **Random Forest** | | **Actual Labels** | | **Total** |
|---|---|---|---|---|
| | | **Anomalous** | **Normal** | |
| **Predicted Labels** | **Anomalous** | TP = 9101 | FN = 609 | **9710** |
| | **Normal** | FP = 6285 | TN = 6548 | **12833** |
| | **Total** | **15386** | **7157** | **22543** |

TABLE VIII. Confusion Matrix of Support Vector Machine Classifier Model for *KDDTestPlus* dataset

| **SVM** | | **Actual Labels** | | **Total** |
|---|---|---|---|---|
| | | **Anomalous** | **Normal** | |
| **Predicted Labels** | **Anomalous** | TP = 9570 | FN = 140 | **9710** |
| | **Normal** | FP = 3 | TN = 12830 | **12833** |
| | **Total** | **9573** | **12970** | **22543** |

TABLE IX. Confusion Matrix of Naive Bayes Classifier Model for *KDDTestPlus* dataset

| **Naive Bayes** | | **Actual Labels** | | **Total** |
|---|---|---|---|---|
| | | **Anomalous** | **Normal** | |
| **Predicted Labels** | **Anomalous** | TP = 9670 | FN = 40 | **9710** |
| | **Normal** | FP = 5368 | TN = 7465 | **12833** |
| | **Total** | **15038** | **7505** | **22543** |

From the results of the classifier shown in the tables (TABLE IV to TABLE IX) above, as well as accuracy, precision, recall and F1- score presented in the TABLE III, it can be clearly observed that the classification results yielded by the MLP and Decision Tree models are the best for the *KDDTestPlus* dataset.

*2) Accuracy and F1- Score:* The computed Accuracy and F1-score percentage of each classifier applied on the *KDDTestPlus* dataset is shown in the Fig. 2. It is clear from the Fig. 2 that the Accuracy and F1 Score of MLP is 1.0, Decision Tree is 0.98 and Linear SVM is 0.97 respectively in the case of *KDDTestPlus* dataset.

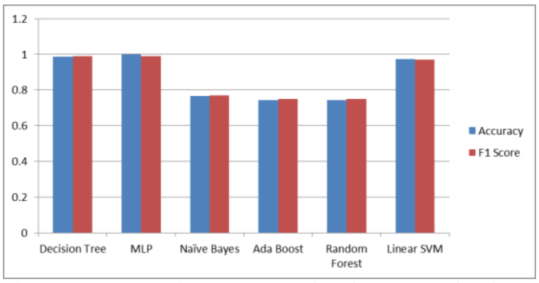

Fig. 2. Accuracy and F1 Score Results of KDDTestPlus dataset

*3) AUROC (Area Under the Receiver Operating Characteristics) Curve:* AUROC curve depicts the capability of a classifier model in distinguishing between classes. For checking any classification model's performance at various thresholds settings, AUROC is one of the most important evaluation metrics. Higher the AUC, better the model is at predicting 1s as 1s and 0s as 0s *[23]* The AUROC curve is showing the comparison of the applied classifiers on *KDDTestPlus* dataset in the Fig. 3.

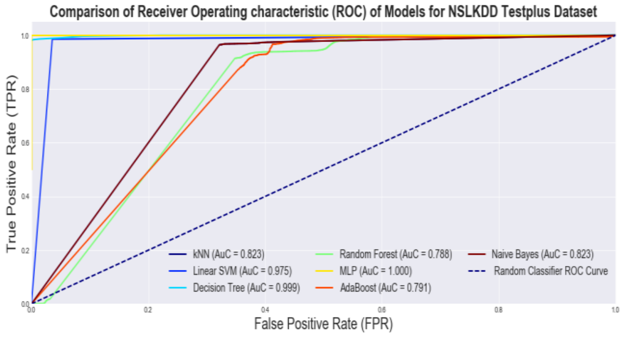

Fig. 3. AUROC Curve for the *KDDTestPlus Dataset*

As per the results of the AUROC curve shown in the Fig. 3, it can be clearly viasualized that the MLP, Decision Tree and Linear SVM have shown the best performances.

## V. Conclusion and Future Directions

In our experimental research, six classification models: Decision Tree, Naïve Bayes, MLP, Support Vector Machine, Random Forest and Ada boost are trained and evaluated on a standard dataset named NSL-KDD. The normal and anomalous records have been identified using these classification techniques. The category encoding method is done on the dataset in the pre-processing phase, which showed improved results. The percentage scores of classification evaluation metrics like accuracy, precision, recall, F1-Score and AUROC of MLP, Decision Tree and Linear SVM are comparatively higher as compared to others.

The comparative analysis of two dataset named ISCX-IDS2012 and UNSW-NB15 based on recent literature is presented as well. These methods are latest development in the anomaly detection domain. The majority of the experiments are done using almost same classifiers by quoted papers. These datasets and methods can establish the baseline for further investigation of anomaly detection in network traffic.

In future, preparation of realistic datasets based upon real world network traffic are needed which can test the effectiveness of the IDS models. To improve the information and network security infrastructure, deep learning can be explored to devise next-generation intrusion detection systems for more accurate and instant detection of network threats. The development of an IDS model for multi-class data problems using multiple networks in deep learning is possible as well. The building of a system which could manage and classify large scale network traffic in real time is still an active area of research.

## References

[1] V. Singh and S. Puthran, "Intrusion detection system using data mining a review," in *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*, 2016, pp. 587–592.

[2] S. A. Mulay, P. R. Devale, and G. V. Garje, "Decision tree based support vector machine for intrusion detection," in *2010 International Conference on Networking and Information Technology*, 2010, pp. 59–63.

[3] "NSL-KDD | Datasets | Research | Canadian Institute for Cybersecurity | UNB." [Online]. Available: https://www.unb.ca/cic/datasets/nsl.html. [Accessed: 05-Jul-2019].

[4] "scikit-learn: machine learning in Python — scikit-learn 0.21.2 documentation." [Online]. Available: https://scikit-learn.org/stable/. [Accessed: 05-Jul-2019].

[5] "IDS 2012 | Datasets | Research | Canadian Institute for Cybersecurity | UNB." [Online]. Available: https://www.unb.ca/cic/datasets/ids.html. [Accessed: 15-Jul-2019].

[6] "The UNSW-NB15 data set description." [Online]. Available: https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/. [Accessed: 05-Jul-2019].

[7] A. Adebowale, S. A. Idowu, and A. Amarachi, "Comparative study of selected data mining algorithms used for intrusion detection," *Int. J. Soft Comput. Eng. IJSCE*, vol. 3, no. 3, pp. 237–241, 2013.

[8] A. K. Sood, P. Greko, and R. J. Enbody, "Abusing Glype proxies: attacks, exploits and defences," *Netw. Secur.*, vol. 2012, no. 12, pp. 8–15, 2012.

[9] Y. Abuadlla, G. Kvascev, S. Gajin, and Z. Jovanovic, "Flow-based anomaly intrusion detection system using two neural network stages.," *Comput Sci Inf Syst*, vol. 11, no. 2, pp. 601–622, 2014.

[10] D. M. Farid *et al.*, "An adaptive ensemble classifier for mining concept drifting data streams," *Expert Syst. Appl.*, vol. 40, no. 15, pp. 5895–5906, 2013.

[11] N. Ye and Q. Chen, "An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems," *Qual. Reliab. Eng. Int.*, vol. 17, no. 2, pp. 105–112, 2001.

[12] C. Kruegel and T. Toth, "A survey on intrusion detection systems," in *TU Vienna, Austria*, 2000.

[13] S. Ganapathy, K. Kulothungan, S. Muthurajkumar, M. Vijayalakshmi, P. Yogesh, and A. Kannan, "Intelligent feature selection and classification techniques for intrusion detection in networks: a survey," *EURASIP J. Wirel. Commun. Netw.*, vol. 2013, no. 1, p. 271, 2013.

[14] M. Idhammad, K. Afdel, and M. Belouch, "Semi-supervised machine learning approach for DDoS detection," *Appl. Intell.*, vol. 48, no. 10, pp. 3193–3208, 2018.

[15] T. Hamed, R. Dara, and S. C. Kremer, "Network intrusion detection system based on recursive feature addition and bigram technique," *Comput. Secur.*, vol. 73, pp. 137–155, 2018.

[16] Y. Yu, J. Long, and Z. Cai, "Session-based network intrusion detection using a deep learning architecture," in *International Conference on Modeling Decisions for Artificial Intelligence*, 2017, pp. 144–155.

[17] W. Yassin, N. I. Udzir, Z. Muda, and M. N. Sulaiman, "Anomaly-based intrusion detection through k-means clustering and naives bayes classification," in *Proc. 4th Int. Conf. Comput. Informatics, ICOCI*, 2013, pp. 298–303.

[18] N. Moustafa and J. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the

comparison with the KDD99 data set," *Inf. Secur. J. Glob. Perspect.*, vol. 25, no. 1–3, pp. 18–31, 2016.

[19] N. Moustafa and J. Slay, "The significant features of the UNSW-NB15 and the KDD99 data sets for network intrusion detection systems," in *2015 4th international workshop on building analysis datasets and gathering experience returns for security (BADGERS)*, 2015, pp. 25–31.

[20] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 1–6.

[21] "Category Encoders — Category Encoders latest documentation." [Online]. Available: http://contrib.scikit-learn.org/categorical-encoding/. [Accessed: 15-Jul-2019].

[22] S. Narkhede, "Understanding Confusion Matrix - Towards Data Science," *Medium*, 26-May-2019. [Online]. Available: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62. [Accessed: 11-Jul-2019].

[23] S. Narkhede, "Understanding AUC - ROC Curve - Towards Data Science," *Medium*, 26-May-2019. [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. [Accessed: 15-Jul-2019].