Network Anomaly Detection Using Machine Learning | A Review Paper

Syed Atir Raza F2019108005@umt.edu.pk SST department University of management and technology, Lahore Pakitan	Sania Shamim F2019108018@umt.edu.pk SST department University of management and technology, Lahore Pakitan

Abstract:

Every day billions of people and million of institutions communicate with each other over the Internet. In the past two years, the number of people using the Internet has increased very fast, today this number has exceeded 4 billion and this increase is continuing rapidly. Opposite to this development, the number of attacks made on the Internet are increasing day by day. Against these attacks, there are two basic methods used to detect the attacks in order to ensure the CIA(confidentiality, integrity, availability) of information security; identification based on signature, and detection based on anomaly.

In this study, it is aim to review the literature and to conclude that which IDS is quick and effectively by means of machine learning methods, reviewing machine learning algorithms that can be used to detect network anomalies, to check which dataset(Darpa98, KDD99, CAIDA,NSL-KDD, ISCX 2012, CICISD2017) will be best enough by comparing other datasets, to determine the success level of the study by comparing the results obtained in it with the studies conducted in this area, reviewing and suggested suitable algorithms by conducting extensive research on machine learning algorithms, To combine two or more machine learning algorithms to make a new one which should be more powerful or efficient and have low error rate, suggesting the appropriate dataset by performing comprehensive research on the alternatives to the dataset.

Introduction:

Every day billions of people and million of institutions communicate with each other over the Internet. In the past two years, the number of people using the Internet has increased very fast, today this number has exceeded 4 billion and this increase is continuing rapidly[1] .

Opposite to this development, the number of attacks made on the Internet are increasing day by day. Against these attacks, there are two basic methods used to detect the attacks in order to ensure the CIA(confidentiality, integrity, availability) of information security; identification based on signature, and detection based on anomaly.

Signature-based methods use the database which are specially created to detect attacks. Research showed that this method is much successful, but the databases need to be constantly updated and new attack information processed. Moreover, even if the databases are up-to-date, they are vulnerable to the zero-day (previously unseen) attacks. If these attacks are not in the database, they cannot prevent these attacks. The anomaly-based approach focuses on detecting unusual network behaviours by examining network flow. This method, which has been successful in detecting attacks that it has not encountered before, so is effective against zero-day attacks. Upon looking into further details it is found, more than half of today's internet usage is encrypted using SSL / TLS (Secure Sockets Layer / Transport Layer Security) protocols, and this rate is increasing day by day [2]. Because of the inability to observe the contents of the encrypted internet stream, signature based methods do not work effectively on this type of data. However, the anomaly-based approach analyses data by its general properties such as size, connection time, and number of packets. So, it does not need to see the message content and it can also do the analysis of encrypted protocols. Due to all these advantages, the anomaly-base detection method is being used intensively to detect and prevent network attacks [2].

In this study, it is aim to contribute to the literature by developing a system that detects network anomaly quickly and effectively by means of machine learning methods, examination of machine learning algorithms that can be used to detect network anomalies, to determine the success level of the study by comparing the results obtained in it with the studies previously conducted in this area, choosing suitable algorithms by conducting extensive research on machine learning algorithms, selecting the appropriate dataset by performing comprehensive research on the alternatives to the dataset, Choosing the suitable hardware/equipment platform, deciding on the right evaluation criteria, selecting the appropriate software platform.

Literature Review Table:

Researcher Name	Year	Technique	Remarks
Catania, C	2013	Author trying to convey that how signature based intrusion detection system can be help in future and also how automatic intrusion detection system can be helpful in future to improve security in better way.	It's a good contribution in intrusion detection systems but some points need to be more researched specially on Automatic NIDS. However, it's a good contribution.
O. Y. Al-Jarrah	2014	Author trying to convey that how enterprise networks can be protected from intrusions, how to improve defense mechanism against cyber attacks, and discussed (RandomForest-Forward Selection Ranking), KDD99 NSL-KDD99 compared and efficiency calculated	It's a good contribution as author successfully evaluated and compared performance of each feature set fairly, voting algorithm with forward selection and backward selection helped a lot in understanding the anomaly more easily.
Johann Stanek	2015	Author Tried to convey that how to protect valuable information of end users over the internet and detect the anomaly using DARPA98, NSL-KDD.	A good contribution as end users are also important part of any network and their data security is also important.
Nour Moustafa	2015	Author described the way to apply DARPA 99 data set for network anomaly detection using machine learning, use of decision trees and Naïve base algorithms of machine learning, artificial neural network to detect the attacks signature based.	Author successfully made his point clear that these approaches are enough capable in NIDS. However as per my view, these techniques are not enough good for real world traffic
Ravi Kiran Varma	2016	Author proposed a set of network traffic features that can be extracted for real time intrusion detection. also proposed fuzzy entropy based heuristic for ant colony optimization, proposed featured IDS algorithm for	Author successfully defined and proposed a system that is much enough able to detect anomaly and proposed a system that is real world intrusion detection system. As other datasets are not enough able for real time

		real time IDS to produce promising result.	intrusion detection.
Soo-Yeon Ji	2016	A multilevel intrusion detection method for abnormal network traffic was introduced using NSL-KDD which gave 96% accuracy in detecting the attacks.	It is a good contribution and no doubt it is giving 96% accurate result in detecting attacks, but the world needs a system that able to protect end-users information and in real time which NSL-KDD data set can not.
Mehdi Hosseinzadeh Aghdam	2016	Author defined feature selection for intrusion detection systems using ant colony optimization and DARPA98, proposed methodology has low computational complexity. provided high accuracy and low number of features.	DARPA98 is an oldest data set used in IDS, it is simple and can not stay for a long time as world is moving towards advanced technology day by day so not agreed with this approach. However a good contribution provide researchers a point to find more best way.
M A Jabbar	2017	Author defined and introduced new approaches in detection anomalies using ML and Data mining(DM) techniques, author used CAIDA dataset, and defined ensemble classifier (RFAODE) for intrusion detection system and Average One-Dependence Estimator (AODE) resolved the attribute dependency issue in Naïve bayes classifier with the accuracy of 90%, author combines Forest(FR) and (AODE) in this method.	A good contribution in IDS, this approach reduces error rates and improved accuracy. As combining two ML and DM algorithms and techniques improved the efficiency and results as well.
Wesam Bhaya	2017	Author proposed a methodology to detect DDOS attacks using efficient cluster analysis in big data, using DARPA 2000, CAIDA2007, CAIDA2008 datasets	A good contribution in detecting DDOS attack on large scale on big data.
Suleman Khalid	2017	Author described that how to protect network from intrusion via distributed machine learning on a smart gateway network, using ISCX-2012, and Linear and Sigmoid Kernel functions.	A good contribution detecting network anomalies at small level and over cyber-physical network and detecting anomalies which may cause using WI-Fi.
Tarfa Hamed	2018	Author proposed NIDS based feature selection method called recursive feature addition and bigram techniques, model implemented , developed , and tested over ISCX-2012 dataset, tested various metrics	A good contribution in detecting a network anomaly in a new way and efficient way.

		by using bigram technology	
Christopher B. Freas	2018	Author estimated high attack performance in large scale network flows, identified major threats to a big data using QAD machine learning algorithm and CICIDS2017 dataset and KDD99 dataset.	A very helpful model and method that can be used to protect a big data and also worked on real-time network traffic. A good contribution in network anomaly detection.
Nasrin Sultana	2018	The author tried to proof that software defined networking technology is much effective in detecting and monitoring anomaly in a network	The author successfully made her point clear, and it's a good contribution in this field.
Saddam Hossen	2018	Author defined analyzing network detection system with machine learning algorithm deep reinforcement learning algorithm.	The author successfully made his point clear, and it's a good contribution in this field that how to secure a data and confidentiality of the data over network.
Sidney C Smith	2018	Author explained the use of packet header anomaly detection in Lossy Network Traffic Compression for Network Intrusion Detection Applications.	A good contribution while understanding anomaly in network applications that are being used for intrusion detection.
Leandros Maglaras	2019	Author proposed a novel intrusion detection system that combines different classifiers approaches based on decision tree and rule based models, using Jrip algorithm, REP tree, and forest PA, proposed the system over CICIDS2017 data set.	A very helpful NIDS that plays an important role in understanding Hierarchical Intrusion.
Novian Anggis	2019	Author explained the methodology in improving ada-boost based intrusion detection system performance on CICIDS2017, using Synthetic Minority Oversampling Technique Principal Component Analysis, and Ensemble Feature Select.	A good contribution in improving Adaboost-based IDS.
Kazi Abu Taher	2019	Author proposed a system to classify network traffic whether it is malicious or benign. It is found that Artificial Neural Network (ANN) based machine learning with wrapper feature selection outperform support vector machine (SVM) technique while classifying network traffic and can	A good contribution towards use of Supervised Machine Learning Technique with Feature Selection and make the data more secure.

		improve the efficiency as much as possible and reduce the ratio of cyber attacks	
Sara Mohammadi	2019	Author explained and proposed that combining datasets can improve the efficiency of IDS, author combined KDD-CUP99 and CICIDS2017 and using decision tree and cuttlefish algorithm which helped in improving the efficiency and obtained result was 95% with a low false rate 1.65%.	A good idea and system to combine the most reliable datasets and machine learning algorithms to proposed a new system in detecting network anomaly.
Selvakumar B	2019	Author explained Firefly algorithm based Feature Selection for Network Intrusion Detection.	A good contribution that improve the methods of detecting an anomaly in a network with a maximum efficiency and low error rate.

Discussion:

Data Sets:

To detect network anomaly by using machine learning methods, there should be a need of large amount of harmful and harmless network traffic for training and testing purposes. Which is not possible for a real network traffic to be used publicly because of some privacy issues. To fulfill this need, many datasets have been produced and continue to be produced. So, here is a discussion of some popular datasets and then they will be compare and evaluate and then decide that which one will be perfect to use either single dataset will enough or should there is a need to compare two or more datasets to generate a new one to detect network anomaly.

Darpa98:

A dataset created by MIT Lincoln laboratory with DARPA funding, it is aimed to create a training and testing environment for Intrusion Detection Systems. In this dataset, the United States Air Force's local computer network was simulated. The data stream consists of processes such as file transfer via FTP, internet browsing, sending and receiving e-mail and IRC messages. In addition to Benign/Normal network traffic, it includes 38 attacks that can be grouped under attack types such as Denial of Service (DoS), User to Remote (U2R), Probe, and Remote to Local(R2L)[3].

But DARPA98 has received a lot of criticism, especially not including the fact that it does not reflect real world network traffic, it is no longer up-to-date and not include flows that can be classified as false positives (the benign data classified as attack, false alarm). However, the DARPA98 dataset is still important because it was used as a source for the creation of commonly used datasets such as KDD Cup 99 and NSL-KDD[4][5].

KDD 99:

This dataset was created by the University of California, Irvine for use by intrusion detection systems in The Third International Knowledge Discovery and Data Mining Tools Competition (The KDD Cup '99). The data packets that make up the DARPA98 dataset are used. 21 properties have been created by applying feature extraction process to be used by machine learning methods[6].

It was divided into two parts as training part and test part to detect cyber attacks and then to control. The training section consists of 4898431 and the test section consists of 311029 data streams. KDD99 contains 38 attack types. Of these attacks, 14 are only specific to the test section and represent unknown attacks. Thus, the detection of unknown attacks on the test section can also be controlled[7][6].

But It is divided into two parts as training part and test part. The training section consists of 4898431 and the test section consists of 311029 data streams. KDD99 contains 38 attack types. Of these attacks, 14 are only specific to the test section and represent unknown attacks. Thus, the detection of unknown attacks on the test section can also be controlled[5], [8].

Compared to the DARPA99 dataset KDD99 which is more suitable for machine learning methods with both the new feature system and training and data parts has been preferred in many studies[5].

CAIDA:

CAIDA (Centre of Applied Internet Data Analysis) is an organization engaged in internet data analysis. The dataset provided by the facilities of this organization is referred to by the same name. The dataset that makes up this dataset comes from a few hours of data flow recording of the OC48 backbone connection over San Jose city. This dataset also contains a section that simulates an hourly DDoS attack of ICMPV6 based and detect and defend against malicious traffics[9].

But In the CAIDA dataset data flows are exemplified only by specific applications and specific attacks. Therefore, the variety of sampling is quite limited. In addition, in this data set, data streams are not labelled. The fact that the data are unlabelled makes it very difficult to use this dataset in machine learning applications[10], [11]

NSL-KDD:

As KDD data set was performing well and was much more efficient than DARPA however, as the world is progressing day by day and use of internet is also increasing and as if a thing has a positive side then it has a negative side too as it is observed that now a days cyber attacks are becoming more and more frequent. So, intrusion detection system should also be enough strong to keep the network flow normal and also it was becoming difficult to maintain the manage or protect the big data. So for this purpose NSL-KDD invented which was combination of the various java adaptive techniques and using some deep learning concepts which includes auto encoder techniques. NSL-KDD is predecessor of KDD99. In other terms we can say that all the mistakes that were in KDD99 was eliminated and a new data set NSL-KDD was proposed[12],[13],[14].

As it was an improved version of KDD99 to eliminate mistakes in KDD99 but the issue with this dataset is that the NSLKDD dataset consists of 4 parts under two main headings as training and testing training data (KDD Train+), 20% of the training data (KDD Train+ 20Percent), test data (KDDTest+) and a smaller version of the test data with all difficulty levels(KDDTest-21) which make it difficult to use[15].

ISCX 2012:

As many data sets introduced but they all were creating problems and were not enough good for the real world traffic or for anomalies detections. To overcome this factor ISCX(Intrusion detection evaluation dataset) was created using the seven-day Internet stream on the test bed created by the Canadian Institute for Cybersecurity. It was developed using real devices normal and malicious streams including FTP, HTTP, IMAP, POP3, SMTP and SSH protocols were created. All data was labeled in this and attack variety was very high and includes different types of attack (Infiltrating, DOS, DDOS and Brute Force SSH)[16].

But the ISCX 2012 data set does not include SSL / TLS (Traffic Sockets Layer / Transport Layer Security) traffic, which accounts for more than half of today's Internet traffic so giving the impression that it will be inadequate to meet today's needs so this dataset will also not enough helpful[17], [18].

CICIDS 2017:

Another dataset introduced CICIDS 2017 (Intrusion Detection Evaluation Dataset) created by the Canadian Institute for Cybersecurity at the University of New Brunswick. This data set used various real time traffic and using the techniques it was enough helpful in using to detect anomalies in real world. This dataset consists of a “4-day (3rd July- 6th July 2017) data stream on a network created by computers using up-to-date operating systems such as Windows Vista / 7 / 8.1 / 10, Mac, Ubuntu 12/16 and Kali Linux”. Details and initial working of this dataset was as follow:[19],[20].

(Working Hours)	**pcap File size**	**Duration**	**CSV File Size**	**Attack Name**	**Flow Count**
Monday	10GB	All Day	257 MB	No Attack	529918
Tuesday	10GB	All Day	166 MB	FTP-Patator, SSH-Patator	445909
Wednesday	12GB	All Day	272 MB	DoSHulk,DoS GoldenEye,DoS slowloris,Heartbleed	692703
Thursday	7.7GB	Morning	87.7 MB	Web Attacks (Brute Force, XSS, Sql Injection), Infiltratio	170366
		Afternoon	103 MB		288602

However, on a critical analysis it is found that CICIDS 2017 is much more efficient then all other above mentioned dataset, and has following advantages[19][20][12], [14]:

A. To obtained data is the real-world data; was obtained from a testbed consisting of real computers.
B. Data streams are collected from computers with the up-to-date operating system. There is operating system diversity (Mac, Windows, and Linux) between both attacker and victim computers.
C. Data sets are labelled. In order to apply the machine learning methods, the feature extraction, which is a critical step, was applied and 85 features) were obtained.
D. Both raw data (pcap files, captured network packets files) and processed data (CSV files, separated data files) are available to work on.
E. In the course of deciding which attack to take place, the 2016 McAfee security report was used, so there is a wide and up-to-date assortment of attacks.
F. It is more abundant than other data sets in terms of protocols used. It also includes the HTTPS (Hypertext Transfer Protocol Secure) protocol in addition to FTP, HTTP, SSH and e-mail protocols.[2], [20]–[22].

But this dataset have also some disadvantages[20] such as :

- Raw data files and processed data files are very large (37.9 GB and 885.7 MB respectively).
- Unlike the KDD99 and NSL-KDD datasets, CICIDS2017 does not have separate files dedicated to training and testing. These sections should be created by users. How to do this is handled in the Creation of Training and Test Data section[20].

However, after a analysis and studies it can be concluded that CICDS2017 is much more efficient than others datasets as it is providing all the features that a IDS required, moreover this datasets enables to monitor a real-world traffic and detect anomaly on that traffic and any abnormal behavior over the network.

<u>Machine Learning Algorithms:</u>

Machine learning is a science and art that enables the programmed computers to learn from the data given to them. As machine learning is much more important thing in detecting anomaly and abnormal behavior of a network and analyzing the traffic, there are some machine learning algorithms that are enough able with different efficiency ratio and playing an important role in IDS. However, some important ML algorithms are as follow:

Naïve Bayes:

A machine learning algorithm that is simplified with the addition of the independence condition on Bayes' theorem. It is a network of probabilities consisting of a parent node representing the unobserved state and multiple child nodes representing the observed states[23]. Its efficiency is 86% [24].

Decision Tree:

Decision trees are one of the popular classifiers used in machine learning methods. In this approach, the rules used are fairly straightforward and understandable. Each decision tree consists of nodes (root-node and sub-nodes), branches and leaves. Within each node, there is a decision statement.

Applies the divide-and-conquer strategy. It makes very large and meaningless data smaller and group them into a meaningful one. Decision tree algorithm, ID3(Iterative Dichotomiser 3) is used. This algorithm is suitable for situations where the training set contains many features. It also stands out with its remarkable features such as giving a reasonable value without doing too much computation, and connecting more than two branches to decision nodes[24], [25] Its efficiency is 95%[26].

Random Forest:

Random forest is a machine learning approach that uses decision trees. In this method, a "forest" is created by assembling a large number of different decision tree structures which are formed in different ways. This algorithm can work well with very large and complicated datasets. The over fitting problem frequently encountered by decision trees is very rare in this algorithm, calculates and uses the importance level of the variables when making the classification and also helpful in feature selection in machine learning having an efficiency of 94%[27].

AdaBoost:

AdaBoost(adaptive boost) a machine learning algorithm developed to improve classification performance, we can say that it’s a boosting method. In this algorithm the data first divided into groups with rough draft rules. Whenever the algorithm run, new rules are added to this rough draft rules. In this way many weak and low performance rules called "basic rules" are obtained these weak rules are combined into a single rule that is much stronger and more successful which are helpful in detecting abnormal behavior of any network[28]. Its efficiency is 94%[28], [29].

MLP

MLP (Multi-Layer Perceptron) is a genre of artificial neural networks. Artificial neural networks (ANN) is a machine learning method that takes inspiration from the way the human brain works. The intention of this method is to imitate the properties of the human brain, such as learning, decision making, and deriving new information. While the human brain is made up of interconnected cells called neurons, artificial neural networks are made up of interconnected hierarchical artificial cells. It is good at coping with complicated problems, can work with missing data, But it is difficult to build the network structure, user should decide the appropriate network structure, Overfitting problems may be encountered[30]. Its efficiency is 83%[31].

QDA:

QDA (Quadratic Discriminant Analysis) is a discriminant analysis method. Discriminant Analysis is a statistical technique for assigning a measured data to one group among many groups. Observed data must be assigned to the group to which it belongs. If it is assigned a group that does not belong to it, an error occurs which is enough helpful in understanding and detecting abnormal behavior of network. To apply the Quadratic Discriminant Analysis, the number of samples observed must be greater than the number of groups[32]. Its efficiency is 86%[33], [34].

Conclusion:

After a deep study it is found that there should be need of more improvement in IDS, as the world is making progress day by day no doubt that technology is also getting advance day by day with respect to time but as technology is making progress on the other hand cyber terrorism also increasing as everything with positive side has a negative side. After a deep study and analysis of datasets using for implementing various models of IDS CICIDS2017 is much more efficient and stable with higher accuracy rate and lower failure rate among the other datasets and is applicable in monitoring and anomaly detection over the real-world traffic, but there is a need of more improvement in it as it is a latest dataset there should be a more study on this and to make its performance more efficient, However, It is also observed that if two or more datasets with good and higher efficiency can combine then a resulting new can be more efficient and reliable. However, Random forest and ID3 are much better approach in network anomaly detection using machine learning as they have good efficiency, less error rate and can overcome overfitting issues. Not only this, after reviewing and studying machine algorithms a point should came here that all the algorithms are working well in their own but there is a error rate while applying a single algorithm, but what if we combine two or more machine learning algorithms with higher efficiency ? which may lead to a result of a new AdaBoost algorithm generation and can be helpful in IDS in a more and efficient rate and can have much more efficiency , accuracy and as well as more success rate than all other ML algorithms which are in use now a days in IDS. There is a need to work more in this field to improve it more as possible as a time is coming where it is very important to protect CIA(confidentiality, integrity, availability) from each and every aspect.

References:

- [1] "94a715dfc1e4b7cce6d79c8d451c2450a893dec9 @ dailywireless.org." .
- [2] R. Abdulhammed, M. Faezipour, H. Musafer, and A. Abuzneid, "Efficient network intrusion detection using pca-based dimensionality reduction of features," in *2019 International Symposium on Networks, Computers and Communications (ISNCC)*, 2019, pp. 1–6.
- [3] S. C. Smith, I. I. Hammell, and J. Robert, "The use of Snap Length in Lossy Network Traffic Compression for Network Intrusion Detection Applications," *J. Inf. Syst. Appl. Res.*, vol. 12, no. 1, p. 17, 2019.
- [4] D. A. Cieslak, N. V Chawla, and A. Striegel, "Combating imbalance in network intrusion datasets.," in *GrC*, 2006, pp. 732–737.
- [5] R. Bala and R. Nagpal, "A REVIEW ON KDD CUP99 AND NSL-KDD DATASET," *Int. J. Adv. Res. Comput. Sci.*, vol. 10, no. 2, p. 64, 2019.
- [6] V. Kumar, D. Sinha, A. K. Das, S. C. Pandey, and R. T. Goswami, "An integrated rule based intrusion detection system: Analysis on UNSW-NB15 data set and the real time online dataset," *Cluster Comput.*, pp. 1–22, 2019.
- [7] T. Merino *et al.*, "Expansion of cyber attack data from unbalanced datasets using generative adversarial networks," in *International Conference on Software Engineering Research, Management and Applications*, 2019, pp. 131–145.
- [8] H. P. Vinutha and B. Poornima, "Analysis of NSL-KDD Dataset Using K-Means and Canopy Clustering Algorithms Based on Distance Metrics," in *Integrated Intelligent Computing, Communication and Security*, Springer, 2019, pp. 193–200.
- [9] O. E. Elejla, M. Anbar, B. Belaton, and S. Hamouda, "Labeled flow-based dataset of ICMPv6-based DDoS attacks," *Neural Comput. Appl.*, vol. 31, no. 8, pp. 3629–3646, 2019.
- [10] R. Bian, S. Hao, H. Wang, A. Dhamdere, A. Dainotti, and C. Cotton, "Towards passive analysis of anycast in

global routing: unintended impact of remote peering," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 49, no. 3, pp. 18–25, 2019.

- [11] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Futur. Gener. Comput. Syst.*, vol. 100, pp. 779–796, 2019.
- [12] S. S. Panwar and Y. P. Raiwani, "Performance Analysis of NSL-KDD Dataset Using Classification Algorithms with Different Feature Selection Algorithms and Supervised Filter Discretization," in *Intelligent Communication, Control and Devices*, Springer, 2020, pp. 497–511.
- [13] T. Bhaskar, T. Hiwarkar, and K. Ramanjaneyulu, "Adaptive Jaya Optimization Technique for Feature Selection in NSL-KDD Data Set of Intrusion Detection System," *Available SSRN 3421665*, 2019.
- [14] C. Zhang, F. Ruan, L. Yin, X. Chen, L. Zhai, and F. Liu, "A Deep Learning Approach for Network Intrusion Detection Based on NSL-KDD Dataset," in *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, 2019, pp. 41–45.
- [15] P. Negandhi, Y. Trivedi, and R. Mangrulkar, "Intrusion Detection System Using Random Forest on the NSL-KDD Dataset," in *Emerging Research in Computing, Information, Communication and Applications*, Springer, 2019, pp. 519–531.
- [16] S. Dwivedi, M. Vardhan, S. Tripathi, and A. K. Shukla, "Implementation of adaptive scheme in evolutionary technique for anomaly-based intrusion detection," *Evol. Intell.*, pp. 1–15, 2019.
- [17] N. Sultana, N. Chilamkurti, W. Peng, and R. Alhadad, "Survey on SDN based network intrusion detection system using machine learning approaches," *Peer-to-Peer Netw. Appl.*, vol. 12, no. 2, pp. 493–501, 2019.
- [18] W. A. H. M. Ghanem and A. Jantan, "Training a Neural Network for Cyberattack Classification Applications Using Hybridization of an Artificial Bee Colony and Monarch Butterfly Optimization," *Neural Process. Lett.*, pp. 1–42, 2019.
- [19] R. Panigrahi and S. Borah, "A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems," *Int. J. Eng. Technol.*, vol. 7, no. 3.24, pp. 479–482, 2018.
- [20] V. Gustavsson, "Machine Learning for a Network-based Intrusion Detection System: An application using Zeek and the CICIDS2017 dataset." 2019.
- [21] N. Bakhareva, A. Shukhman, A. Matveev, P. Polezhaev, Y. Ushakov, and L. Legashev, "Attack Detection in Enterprise Networks by Machine Learning Methods," in *2019 International Russian Automation Conference (RusAutoCon)*, 2019, pp. 1–6.
- [22] Z. Chiba, N. Abghour, K. Moussaid, A. El Omri, and M. Rida, "An Efficient Network IDS for Cloud Environments Based on a Combination of Deep Learning and an Optimized Self-adaptive Heuristic Search Algorithm," in *International Conference on Networked Systems*, 2019, pp. 235–249.
- [23] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Syst.*, p. 105361, 2019.
- [24] K. J. Mathai, "Performance Comparison of Intrusion Detection System Between Deep Belief Network (DBN) Algorithm and State Preserving Extreme Learning Machine (SPELM) Algorithm," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019, pp. 1–7.
- [25] M. Aloqaily, S. Otoum, I. Al Ridhawi, and Y. Jararweh, "An intrusion detection system for connected vehicles in smart cities," *Ad Hoc Networks*, vol. 90, p. 101842, 2019.
- [26] Z. El Mrabet, H. El Ghazi, and N. Kaabouch, "A Performance Comparison of Data Mining Algorithms Based Intrusion Detection System for Smart Grid," in *2019 IEEE International Conference on Electro Information Technology (EIT)*, 2019, pp. 298–303.
- [27] P. S. Chaithanya, M. R. G. Raman, S. Nivethitha, K. S. Seshan, and V. S. Sriram, "An Efficient Intrusion Detection Approach Using Enhanced Random Forest and Moth-Flame Optimization Technique," in *Computational Intelligence in Pattern Recognition*, Springer, 2020, pp. 877–884.
- [28] M. Mazini, B. Shirazi, and I. Mahdavi, "Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms," *J. King Saud Univ. Inf. Sci.*, vol. 31, no. 4, pp. 541–553, 2019.
- [29] Q. Liu, Y. Si, L. Li, and D. Wang, "ECG Identification Based on PCA and Adaboost Algorithm," in *International Conference on Human-Computer Interaction*, 2019, pp. 50–62.
- [30] N. E. Jackson, M. A. Webb, and J. J. de Pablo, "Recent advances in machine learning towards multiscale soft materials design," *Curr. Opin. Chem. Eng.*, vol. 23, pp. 106–114, 2019.
- [31] A. Shokoohsaljooghi and H. Mirvaziri, "Performance improvement of intrusion detection system using neural networks and particle swarm optimization algorithms," *Int. J. Inf. Technol.*, pp. 1–12, 2019.
- [32] W. Książek, M. Abdar, U. R. Acharya, and P. Pławiak, "A novel machine learning approach for early detection of hepatocellular carcinoma patients," *Cogn. Syst. Res.*, vol. 54, pp. 116–127, 2019.
- [33] M. Alrowaily, F. Alenezi, and Z. Lu, "Effectiveness of machine learning based intrusion detection systems," in *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*, 2019, pp. 277–288.
- [34] R. Abdulhammed, H. Musafer, M. Faezipour, and A. A. Abuzneid, "Towards Efficient Features Dimensionality Reduction for Network Intrusion Detection on Highly Imbalanced Traffic," 2019.