JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

STUDY OF ANOMALY BASED NIDS: MACHINE LEARNING TECHNIQUES

[1]Vaishali Desai, [2]Nafisa Ansari

[1]IT and CS Co-Ordinator, [2]Assistant Professor

[1]Information Technology and Computer Science,

[1]Lords Universal College, Mumbai, India

Abstract :

Purpose: In network communications, networks are one of the most vulnerable systems where security is a major issue. The aim is to study ANIDS under which genetic- neural network architecture is proposed.

Methodology: This is a Descriptive type of research that took 8 months, and 12 research papers are studied. The observational tool is used.

Findings: The variables included in the study were selected based on the literature. As per the literature review, machine learning techniques are used to identify attacks by combining different algorithms as there is no more research done on combining Genetic and Artificial neural networks.

Contribution: There are many available misuse-based detection systems. However, most IDS lack the capability to detect previously unknown attacks. The anomaly intrusion detection system is the subset of intrusion detection systems which effectively finds both known as well as zero-day attacks. Anomaly IDS face problems such as high rate of false alarm. To overcome the problem of high false alarm rate we have proposed anomaly NIDS which uses genetic algorithm and artificial neural networks to detect intrusion and also classify the detected attacks into proper types. When there is an increase in false prediction rate the genetic-neural network algorithm is run automatically and the newer attacks are categorized into respective types such as probing, DOS, U2R, R2L. The results of this study offer guidance to the network administrator to act upon and on how to best secure their assets against attacks.

IndexTerms **- IDS, Anomaly Detection system (AIDS), Artificial Neural Network (ANN), Genetic Algorithm, ML**

I. INTRODUCTION

An intrusion detection system (IDS) is collection of tools, methods and resources to help identify intrusions. Intrusion detection system is one part of overall protection system that is installed around device. IDS can be categorized into three groups, according to the source of the audit data: Network based Intrusion Detection System (NIDS), Host based Intrusion Detection System (HIDS), Hybrid Intrusion Detection System. IDS is classified into three detection methodology as shown in figure 1.

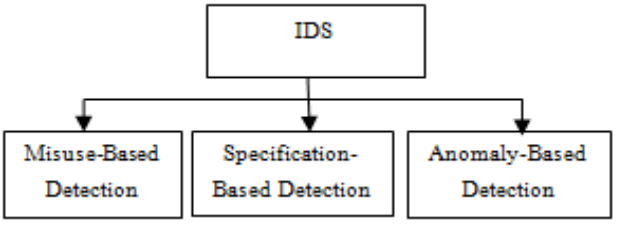

Fig 1. IDS detection methodologies

Intrusion detection system may use one of the methodologies i.e Misuse-Based detection is also called as signature-based detection method. The signatures of previously known attacks are first generated and are used as reference to detect future attacks. The disadvantage is that this method cannot detect new kind of attack, so misuse detection would not be able to detect it. The specification-based detection is the set of specification and constraints. Based on specification and constraint the program is monitored to detect anomaly. The disadvantage of specification method is time consuming as the developer has to write specification and constraint manually. So based on analysis of previous methods anomaly intrusion detection method is best for determining both known as well as unknown attacks.

The aim of this paper is to present the survey on anomaly intrusion detection system and also the survey done on proposed technique that comes under anomaly IDS are discussed. The remaining of this paper is organized as follows: section 2 represents the theoretical background of A-IDS. Section 3 represent the survey on anomaly IDS techniques. The proposed

technique is mentioned in section 4. Finally paper is concluded in section 5.

2. Anomaly network intrusion detection system

Anomaly intrusion detection system first creates the base line profile of the normal system and if any deviation from normal profile will be termed as intrusion. The anomaly detection have the capability to detect insider attack for example: if the unauthorized user is using an account and starts performing actions that are outside the normal profile then intrusion detection system will fire alarm. The anomaly intrusion detection system can also identify unknown attacks. Anomaly detection method uses different techniques to analyze anomaly. AIDS is classified in three categories as shown figure 3: Statistical Anomaly Based A-NIDS, Knowledge Base, and Machine Learning Base.

2.1 Statistical-based A-NIDS techniques

In statistical based anomaly IDSs, the network traffic is captured and then a profile representing its behavior is generated. As the network operates in normal conditions, a reference profile is created. After that, the network is monitored and profiles are generated periodically and an anomaly score is generated by comparing it to the reference profile. If the score passes a certain threshold, the IDS will flag an occurrence of the anomaly.

- **Univariate:** This technique model the parameters as independent Gaussian random variable, thus it defines an acceptable range of values for every variable [2].
- **Multivariate**: This technique consider the correlation between two or more metrics because by using combination of related measures better level of discrimination can be obtained instead using individual method [2].

Time series model: This technique uses interval timer, together with event counter, based on timer it can be taken into account the order and inter-arrival times of the observations and their values. Thus, an observed traffic instance will be labeled as abnormal if its probability of occurrence is too low at a given time. The A-NIDS does not require prior knowledge about normal activity of target system; instead they have ability to learn the expected behavior of system observations [2].

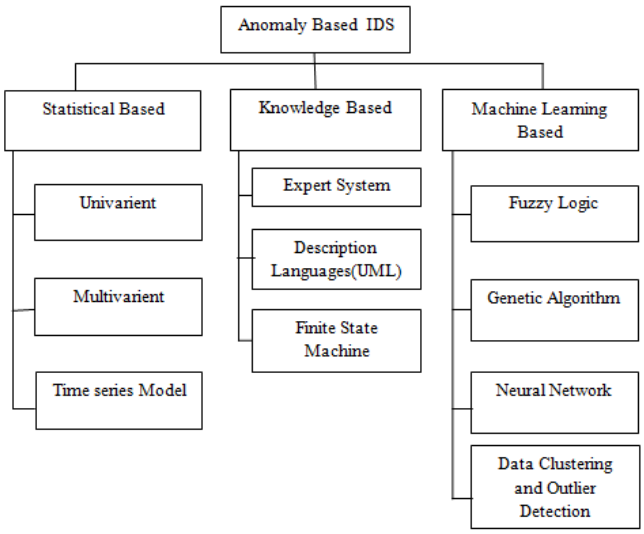

Fig 2. Classification of anomaly IDS system

2.2 Knowledge-based A-NIDS techniques

Knowledge based anomaly IDSs rely on the availability of the prior knowledge (data) of the network parameters in normal operating condition as well as the one under certain attacks.

- **The expert system:** It classify the audit data according to a set of rules involving three steps. First, from the training set different attributes and classes are identified. Second, set of classification rules, parameters are reduced. Third, classification of audit data are done accordingly.
- **UML:** Diagram are generated based on data specification in description language (UML).
- **Finite state machine**: states and transitions are defined according to available data set.

2.3 Machine learning based techniques

In machine learning based anomaly IDSs, an explicit or implicit model of the analyzed patterns is generated. These models are updated periodically, in order to improve the intrusion detection performance on the basis of the previous results.

- **Fuzzy logic**: This technique is based on approximation and uncertainty. Fuzzy techniques are used in the field of anomaly detection because the features are consider as fuzzy variable which observes as normal if it lies within given interval [2].
- **Genetic algorithm**: This algorithm finds the optimal solution for a given problem. Genetic algorithm is capable of deriving classification rules and selecting appropriate feature [2].
- **Neural network**: This technique is adopted in the field of anomaly intrusion detection system because of their flexibility and adaptability to environmental changes. Neural network technique is used to identify intrusive behavior of traffic patterns [2].
- **Clustering and outlier detection**: Observed data is grouped into clusters according to a specified similarity or distance measure. Points that do not belong to any cluster are named as the outliers [2].

3. RELATED WORK

The number of Anomaly Intrusion Detection System techniques has been proposed under MANET. Therefore, a thorough analysis of each technique is studied which will help us to better understand and drawback are analyzed.

A. Patcha *et al.*, [1] have done survey on anomaly detection system and hybrid IDS of recent and past papers. The author has discussed recent technological trends for anomaly detection and identified open problems challenges in this area. This challenges are useful for developing the intrusion detection system which helps to better understand the network behavior.

P. Garcia-Teodoro *et al.*, [2] have modified the anomaly IDS based on previous paper i.e the survey done by author A. Patcha on Anomaly based IDS. The author has mentioned issues and challenges of anomaly based intrusion detection. As per the analysis the main challenges that researches must face when trying to implement and validate a new intrusion detection method is to assess it and compare its performance with other available approaches.
After this above analysis the techniques under Anomaly IDS are studied.

Gerhard munz *et al.*,[3] proposed novel flow based anomaly detection method based on K-means clustering algorithm to detect anomaly in traffic. This algorithm is best suited for analyzing anomaly in real traffic and generated
traffic. They have used weighted Euclidean distance for calculating distance of object from centriod. The author has combined classification and outlier detection method to overcome limitation of each individual method for analyzing anomalies in system. The method has faced limitation in feature selection phase and reliability of the system is depended on selection of feature.

Y.Dhanalakshmi *et al.*,[4] has proposed fuzzy logic and genetic algorithm to detect anomaly IDS in network. This algorithm finds the approximate value by fuzzy algorithm to detect anomaly. The author has proposed a way to include quantitative feature by using fuzzy numerical functions. By combining fuzzy and genetic algorithm the algorithm provides accuracy of generated rules.GA was used to find optimal parameters of fuzzy function as well as to select the most relevant features.

Sufyan T. Faraj Al-Janabi *et al.*, [5] have propose Back propagation Artificial neural network and designed detection module for detecting anomaly using machine learning technique i.e neural network. The module measures accurate results and the results are classified by ANN. The author has used ranking method to select appropriate features. Large training Dataset is required for best results by which its time consuming.

Kamal Kishore Prasad *et al.*,[6] They have proposed a brief overview of Intrusion Detection System, genetic algorithm (GA) and related detection techniques is presented and compared. GA is used to increase the efficiency of machine learning System and finds the optimal solution. The author has used genetic algorithm for detecting anomaly and also shown the comparison with different algorithms.

S. Selvakani Kandeeban *et al.*,[7] Author have implemented Intrusion Detection System based on Genetic algorithm using MATLAB to detect attacks using KDD data set. Based on results, author have selected nine features out of forty-one used to describe each connection of KDD99Cup dataset.

E.G Dada *et al.*,[8] This paper investigates the performances of six (6) machine learning techniques in relation to how they can effectively handle different types of attack on networks using NSL-KDD dataset. All experiments were conducted on Weka 3.8. The strengths and shortcomings of the approaches to intrusion detection called Intrusion Detection System (IDS) and different classification of IDS were explained.

Emrah Tufan *et al.*,[9] study investigated the potential of an anomaly-based ML model for IDS compared to the misuse-based models. Probing attack was studied based on machine learning techniques and used institutional data set.

M. Ring *et al.*,[10] survey of data sets for network based intrusion detection was done and describes the underlying packet and flow-based network data in detail. The paper identifies 15 different properties to assess the suitability of individual data sets for specific evaluation scenarios. The author has given recommendations for the use and the creation of network-based data sets.

A. L. Buczak *et al.*,[11] survey of machine learning (ML) and data mining (DM) methods for cyber analytics in support of intrusion detection. The complexity of ML/DM algorithms was addressed.

ANN can generalize from previous behavior to recognize future unseen behavior and the classification ability of ANN can be used to detect even slightly different intrusions. So as there is huge network, genetic algorithm is used to find the optimal solution and select proper features.

4. PROPOSED SOLUTION

Anomaly Detection techniques identify an intrusion when the observed activities in computer system demonstrate a large deviation from the normal profile built on long term normal activities. To handle dynamic profiles, learning algorithm are required to track network behavior and adapt dynamically changing concept. Neural network methods scale up much better than linear statistical models as size and complexity of learning task grows. The combination of Genetic algorithm and artificial neural network is not only able to select optimal feature set but also figures out “optimal weight values” for artificial neural network. The genetic-neural IDS system has two stages: Learning stage and detection stage.

Learning stage: The learning stage makes up the knowledge base for genetic-neural IDS system. The training is accomplished using genetic algorithm and it uses data set containing normal and abnormal sample. The learning stage consist of preprocessor, feature selection algorithm i.e. genetic algorithm, and Normalization.

4.1 PU-IDS Dataset: PU-IDS data set is the input to the system. This data set contains several weeks of attack data. KDD dataset is considered as a standard benchmark for intrusion detection evaluations. PU-IDS attack include (DOS, PROB,U2R, R2L).

4.2 Data Preprocessing: The PU-IDS dataset is converted into machine readable form.

4.3 Data set normalization: Normalization is the process to achieve maximum accuracy, and each numerical value in the data set is normalized by deciding particular range for example: range between [0.05,0.95] according to the following method i.e mean and standard deviation:

$$x = (x - Min)/ (Max - Min)$$

therefore x is the numerical value, *Min* is the minimum value and *Max* is the maximum values for the attribute that x belongs to.

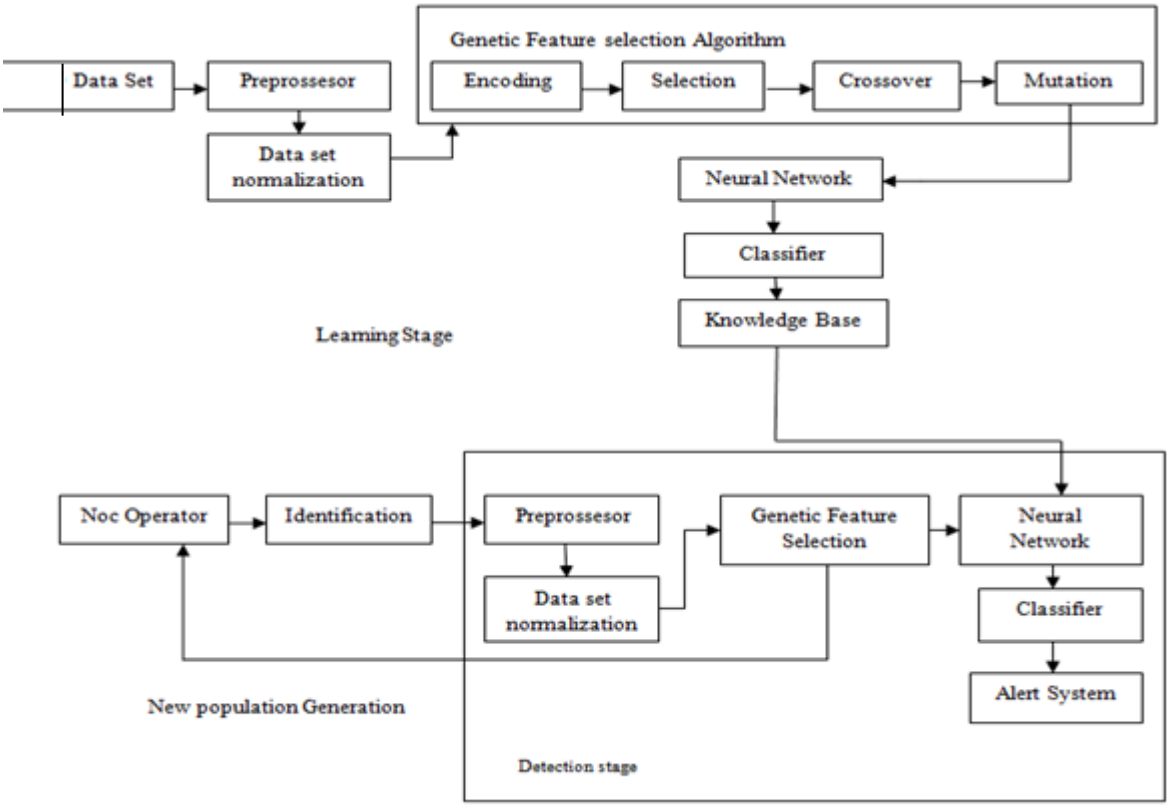

Fig 3: Genetic-neural network algorithm

4.4 Genetic feature selection: Genetic algorithm is based on the principles of natural selection and genetics The genetic feature selection algorithm will select appropriate number of features from PU-IDS data set. This selection method will use genetic algorithm i.e selection, crossover and mutation and also calculate the fitness value to identify best chromosomes.

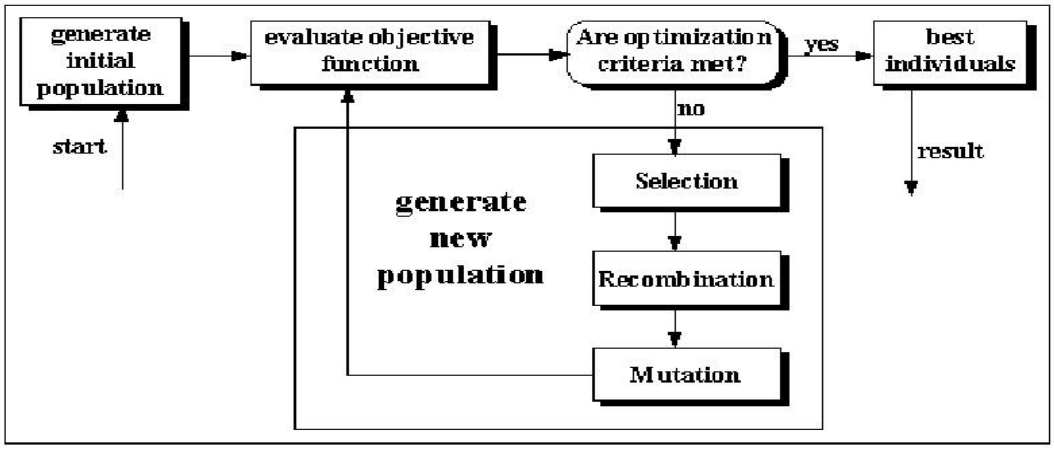

Fig 4: Genetic Algorithm Flow

4.4.1 Feature selection encoding: In encoding scheme the chromosome is a bit string whose length is determined by the number of features. Each feature is associated with one bit in the string. If the i-th bit is 1, then the i-th feature is selected, otherwise, that component is ignored. Each chromosome thus represents a different subset of features.

4.4.2 Initial population: In genetic algorithm the initial population is generated randomly. The number of 1's are generated randomly and scattered into chromosomes.

4.4.3 Fitness Evaluation: The feature subset selection method is used to reduce the number of features for better performance. Each feature subset contains number of features. The features are selected based on the accuracy achieved.

4.4.4 Crossover: It is a process where each pair of individuals selects randomly participates in exchanging their parents with each other, until a total new population has been generated [6].

4.4.5 Mutation: It flips some bits in an individual, and since all bits could be filled, there is low probability of predicting the change.

4.5 Artificial neural network: After the features are selected for the system the input is given to artificial neural network. Artificial neural network is employed to learn the input and output relationship of an intrusion prediction model using genetic algorithm. Neural network has one input, one hidden and one output layer. Artificial neural network has specific number of inputs based on selected number of features. The number of hidden layer is chosen by back propagation computation process. The advantage of using neural network would be the ability to generalize from past behavior to detect novel attacks. The accuracy of classification by ANN benefits from its classifier algorithm. The classifier algorithm determines the best solution by trying to minimize the number of incorrectly classified cases during the training process. A neural network might be trained to recognize known suspicious behaviors with a high degree of accuracy. ANN learning can solve problems with the noisy and complicated training data. ANN learning is robust to errors in the training dataset.

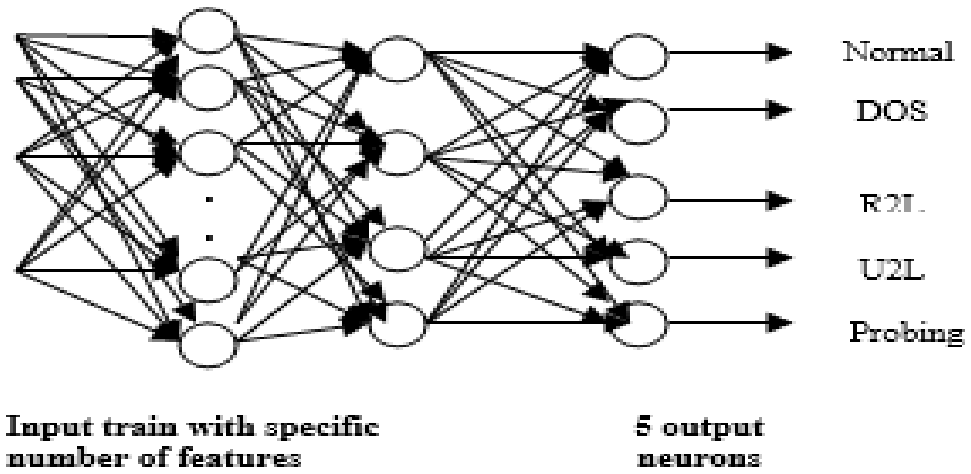

Fig 5: Artificial neural network

4.6 Classification module: For classifying the output ANN is used to classify into 5 categories i.e normal, DOS, U2R, R2L, probing. After classification the data is stored in knowledge base for referring the future attacks.

4.7 Detection Stage: The detection stage is the entry point in the system. The NOC operator is the agent which analyze the traffic and traffic is passed to system, the identification unit identifies the traffic pattern and passes further to preprocessor unit.

The traffic pattern is analyzed properly based on requirement of system and this is done through the genetic feature selection algorithm. The neural network will analyze the traffic pattern and appropriate action is taken accordingly. The Artificial neural network have been used in classification of data. The results can only be obtained after completing both of training and testing phases. The intrusion data have been classified into five categories. The first category represents normal data and the other four are attack types. These attack types are probing, denial-of-service (DOS), remote-to-local (R2L), and user-to-root (U2R) attacks. Each of these categories indeed contains sub-types of attacks, as shown in table 1 below.

Table 1: Classification of attack types

PROBING	DOS	U2R	R2L
Ipsweep, mscan, nmap, portsweep, sendsaint, satan	Apache2, mailbomb, Neptune, pod,smurf, teardrop, udpstrom.	Buffer_overflow, httptunnel, rootkit, sqlattack.	ftp write, guess_passwd, imap, worm, sendmail, xlock.

4.8 Alert system: This is the final stage of the proposed system. This stage involves identifying the events that occurred whether abnormal or not, then sending the required signals to alert administrator (or user) accordingly.

5. CONCLUSION AND FUTURE WORK

Anomaly Intrusion Detection Technique is best method to detect anomaly by using machine learning techniques such as Fuzzy logic, genetic algorithm, artificial neural network and K-means clustering algorithm using Knowledge Base technique. The future work can be implementation of above approach i.e genetic-neural network and we can also combine different machine learning techniques and data mining techniques.

REFRENCES:

[1] A. Patcha and J.M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends", Elsevier J. Computer Networks, volume 51, number 12, pages 3448-3470, 2007.
[2] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez and E. Vazquez, "Anomaly-Based Network Intrusion Detection: Techniques, Systems and Challenges", Elsevier J. Computers and Security, vol. 28, num. 1-2, pp. 18-28, 2009.
[3] Gerhard munz, Sa Li, Georg Carle "Traffic Anomaly Detection Using K-Means Clustering" Computer Networks and Internet Wilhelm Schickard Institute for Computer Science University of Tuebingen, Germany,2007.
[4] Y.Dhanalakshmi and Dr.I. Ramesh Babu, "Intrusion Detection Using Data Mining Along Fuzzy Logic and Genetic Algorithms," IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2,2008.
[5] Sufyan T. Faraj Al-Janabi and Hadeel Amjed Saeed, "A Neural Network Based Anomaly Intrusion Detection System",2011.
[6] Kamal Kishore Prasad and Samarjeet Borah, "Use of Genetic Algorithms in Intrusion Detection Systems: An Analysis", International Journal of Applied Research and Studies (IJARS) ISSN: 2278-9480 Volume 2, Issue 8 (Aug - 2013).
[7] S. Selvakani Kandeeban and R. S. Rajesh, "A Mutual Construction for IDS Using GA", International Journal of Advanced Science and Technology Vol. 29, April, 2011.
[8] E.G Dada, J.S Bassi and O.O Adekunle, " An investigation into the effectiveness of machine learning techniques for intrusion detection", Arid Zone Journal of Engineering, Technology and Environment, December, 2017.
[9] Emrah Tufan et al, "Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network", IEEE, March 2021.
[10] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, ''A survey of network-based intrusion detection data sets,'' Comput. Secur., vol. 86, pp. 147–167, Sep. 2019.
[11] A. L. Buczak and E. Guven, ''A survey of data mining and machine learning methods for cyber security intrusion detection,'' IEEE Commun. Surveys Tuts., vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.
[12]] D. K. Bhattacharyya and J. K. Kalita, Network Anomaly Detection: A Machine Learning Perspective. Boca Raton, FL, USA: CRC Press, 2013