JCST Journal of Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.98549733

NETWORK AND HOST BASED INTRUSION DETECTION MODEL USING VARIOUS MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

Maithili S. Deshmukh, Dr. A. S. Alvi,
Department of Information Technology, Prof. Ram Meghe Institute of
Technology & Research, Badnera
maithili11687@gmail.com, abrar_alvi@rediffmail.com

Abstract : A significant field of investigation in network security is intrusion detection. The detection of abnormalities in network data is a typical method for intrusion detection, but network threats are developing at an extraordinary speed. The system is exposed to attacks due to the discrepancy between both the development of threats and the network's existing detection reaction time. A variety of machine learning methods have been created over time to identify network breaches using packet forecasting. These methods rely on techniques that can learn automatically from data without explicitly programmed. This is especially practical given the diverse nature of the traffic. Nevertheless, despite these benefits, abuse detection still outperforms anomaly detection systems in the actual world. The principal cause of the low adoption of anomaly-based intrusion detection system is the issue of the significant false positive rate. On a network with considerable traffic, even a 1% false positive rate can result in so many false alerts that an administrator is unable to handle them. We offer suggestions for applying deep learning machine learning method to increase the accuracy of anomaly-based IDS detection for implementation on actual networks. In our research methodology we proposed intrusion detection system in two phases using various datasets namely KDDcup99, botnet, ISCX, WSNtrace, NSLKDD, NUSW-NB15 and real time twitter data. In phase 1, we proposed Intrusion detection system using Genetic Algorithm and various Machine learning techniques such as J48, Artificial Neural Network and Random Forest. In this phase, we have performed various experiments for evaluating performance of our proposed system by considering various parameters like using different population size, threshold, datasets etc. In the phase 2, we have evaluated proposed RNN-LSTM for various functions like sigmoid, tanh and ReLu using different cross validation. Our proposed system can generate its own rules. It can detect DOS, Root to login, probe, User to Root, network attack, passive attack, active attacks as well as unknown attacks. It is observed from the experimental findings that detection rate of denial of service attack is 96.9 % which is high as compared to other attacks and the accuracy rate of RNN-LSTM (ReLU) using 20-fold cross validation is 97.95 % which is high as compared to RNN-LSTM (Sigmoid) and RNN-LSTM (Tanh) when different cross validations like 10-fold, 15-fold and 20-fold are used.

Keywords : Intrusion detection system, machine learning, network attacks, DDoS, Unknown attacks, feature extraction, classification, deep learning, network security

Introduction

Intrusion detection system technologies come in a variety of forms due to the diversity of network setups. Identification, configuration, and cost-wise, every kind has benefits and drawbacks. One popular kind of intrusion detection system is a Network Intrusion Detection

System (NIDS), which examines network traffic at each of the OSI model's seven layers and looks for unusual activity. Network intrusion detection system may see traffic from multiple devices at once and are simple to implement on any network. Wireless Intrusion Prevention System (WIPS) is a recent study topic that analyzes and monitors the wireless radio spectrum in a system for invasions and takes corrective action.

The organization's overall performance is significantly influenced by the information that is kept there. The network wherein the system is located as well as the computer itself has become targets of the assault when data saved on the machine is the intended goal. Financial planning and design requirements, details about the firm's competitive position in the market, firm top confidential, such as shares and anticipated stock values, company yearly estimated advantages, and financial planning, intellectual assets, such as baseline procedures, methodologies, proprietary information, and other intangible resources, as well as individual recognition and approval data, are some of the main resources that are targeted for attack.

Using the audit data produced by the operating system, ID is the most efficient technique to find intrusions. Since practically all system operations are recorded in logs, it's possible to spot intrusions by manually reviewing these records. The audit data utilized for analysis in ID is a crucial component. Even after an assault has taken place, it is crucial to analyze the audit data to ascertain the extent of the problem. Additionally, this analysis aids in the identification of the assaults and the recording of their trends for later identification. Information systems can benefit greatly from a strong Intrusion Detection System that can analyze audit data.

Observing and analyzing things that happen in a device or networked computer system is the technique of ID. By examining user behavior that runs counter to the system's authorized use, identification was carried out. Even if a computer is not linked to the Network, each user who uses one will still run some danger of intrusion. Any attacker can try to access the system and try to take advantage of the system if it is left unattended. If the device is connected to internet, the issue is even worse. The computer can be done remotely by any user from anywhere in the world. An intrusive party can try to get access to sensitive or personal data or launch an attack to stop the system from working properly. The act of hacking into a computer system does not require human intervention. With specially designed software, it can be carried out automatically and remotely. Intrusion detection system can be installed in the network either before or after the firewall as well as router. IDS can also be used well before firewall to totally block outside incursions. In order to provide more adequate protection, the Host Intrusion detection system can be installed alongside the Network intrusion detection system on every device of the network.

There are still many detection rate, false alarm rate, feature selection, categorization, time-consuming, and skill issues even when various ID techniques are developed utilizing statistical, DM, and ANN techniques. The traditional signature-based technique is unable to identify fresh or undiscovered invasions. Updates and patches must be applied often. Because it is not built on an adaptive intelligent framework, the statistical anomaly-based technique is unable to learn from both legitimate and harmful traffic patterns. Any IDS's ultimate objective is to achieve the best detection rate while minimizing the false alarm rate. As a result, our study focuses on these issues while simultaneously attempting to use DL methods to enhance the effectiveness of IDSs. This system also proposed a hybrid feature selection approach for extract various unique features and achieves effective classification. The organization of paper in section II describes a literature review of proposed model where numerous existing authors different researches and its outcomes. In section III we describe a proposed model of system including system architecture and it its detail explanation. The section IV discusses about the results including the experimental analysis and finally section V define the conclusion of proposed model.

Literature review

A deep learning strategy for ID utilizing Recurrent Neural Networks (RNNs) is proposed by Yin Chuan-long et al. in 2017 [1]. It investigates how to develop IDS based on deep learning. Additionally, the model's efficiency is examined in binary classifier and multiclass classifier and the efficiency of the proposed approach is affected by the quantity of neurons and various learning rates. On the standard dataset, comparisons are made using the J48, Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF) and other ML techniques. The empirical findings demonstrate that RNN-IDS is very well suited for modelling high-accuracy ML classifiers, and that it outperforms them in binary classification and multiclass classification. The RNN-IDS method enhances ID accuracy and offers a new approach to ID studies.

In 2017, George Loukas et al. [2] presented the feasibility and advantages of offloading the continual deep learning-based ID process. This strategy is not constrained to a specific form of threat or the in-vehicle CAN system as previous research was, and it regularly obtains high accuracy compared to typical ML approaches. It sends time series information to Neural Network (NN) design as input, using actual information that pertains towards both virtual and real activities. The Deep Multilayer Perceptron (MLP) and a RNN framework are used in the investigation, with the latter benefiting from a hidden neurons with Long-Short Term Memory (LSTM) that is highly effective in learning the informational of various types of attacks. It uses virus, command injection, and DOS as aspects of intrusions that are relevant for autonomous vehicles. The resources that are available both onboard and remote, as well as the dependability of the channels of communication connecting them, determine how useful computation offloading is. It has created a mathematical formula to ascertain when computing offloading is advantageous considering variables based on network functioning and the processing requirements of the DL algorithm utilizing identification latency as the measure. Offloading results in a higher reduction in identification latency the more efficient the network is and the more analysis is required.

A unique 5G-oriented cyberdefense framework is proposed by Lorenzo Fernandez Maimo et al. in 2018 [3] to swiftly and effectively detect cyberattacks in 5G wireless services. By removing characteristics from network flows, the framework leverages deep learning methods to analyse network traffic. Additionally, the proposal allows for the arrangement of the cyberdefense design to be instantly adjusted in order to handle traffic fluctuations, intending to both maximize the computing resources required at any given time and fine-tune the operation and effectiveness of the evaluation and identification methods. A NN method in the anomaly detection system is shown to achieve a satisfactory level of classification accuracy rate in tests utilizing a very well-known botnet dataset. The applicability and performance of different DL techniques are analysed and determined through extensive tests using varying network traffic loads. The outcomes of the experiments demonstrate how the suggested framework may self-adapt the IDS depending on the amount of network flows collected from 5G subscribers' User Equipment (UE) in actual and optimize resource usage.

In 2018, I. Ahmad et al. [4] A crucial component of security solutions like firewalls, IDS, IPS and adaptive security appliances is ID. Different ID methods are employed; however their effectiveness is a problem. Performance of ID depends on reliability, which must advance to reduce false alarm rate and raise detection rates. Current findings have used MLP, SVM, and other methods to address performance problems. Such methods point to flaws and are ineffective when used to huge datasets, like system as well as network data. Huge amounts of traffic information are analyzed by the IDS; hence an effective classification method is able to solve the problem. This article takes into account this issue. SVM, RF, and extreme ML are the three well-known machine learning algorithms used in the proposed work. These methods are well-known for their categorization abilities. The dataset employed is the NSL-knowledge discovery and DM dataset, which serves as a baseline for assessing ID systems. The outcomes show that ELM operates better than other methods.

Sheraz Naseer et al. [5] The necessity for data network security has grown owing to the Internet's phenomenal expansion over the past ten years. An IDS is supposed to respond to the highly dynamic threat environment as the main line of defence for network framework. Experts in the fields of ML and DM have developed a variety of supervised and unsupervised algorithms to reliably identify anomalies. The application of neuron-like architecture to learning tasks is known as deep learning. It has radically changed how learning tasks are approached by making enormous strides in a variety of fields, including voice recognition, machine vision, and Natural Language Processing (NLP). It only matters that this advanced innovation be looked into for data security uses. The purpose of this paper is to look at whether DL algorithms are appropriate for anomaly-based IDS. CNN, Autoencoders, and RNNs are just a few of the DNN structures used in this investigation to construct anomaly detection methods. These deep models were tested on NSLKDDTest+ as well as NSLKDDTest21, which were given by NSLKDD. They were developed using the NSLKDD training set. On a test platform operated by a GPU, all tests in this paper are carried out. Utilizing well-known classifiers like Extreme Learning Machine, KNN, DT, RF, SVM, NB, and Quadratic Discriminant Analysis, traditional ML-based ID models were put into practice. Utilizing well-known classification criteria like Receiver Operating Principles, Area under the Curve, Extremely precise Curve, average precision, and correctness of categorization, both deep and standard ML techniques were assessed. Deep IDS algorithms' test data offered encouraging signs for their practical use in anomaly detection systems.

In 2018, Congyuan Xu et al. [6] utilized DL theory to ID and created a deep network approach with autonomous extraction of features in order to improve the effectiveness of NIDS. This study examines the behavior of time-related attack and suggests an unique IDS made up of a MLP, a softmax module, and a RNN with Gated Recurrent Units (GRU). On the well-known KDDcup99 and NSL-KDD datasets, tests show the system's greater efficiency. With false alarm rates as low as 0.05 percent and 0.84 percent, correspondingly, the total detection rate was 99.42 percent using KDD 99 and 99.31 percent using NSL-KDD. The system specifically obtained detection rates of 99.98 percent and 99.55 percent, respectively, for DOS attacks. Comparative tests demonstrated that GRU is a more efficient reduction and enhancement of LSTM and that it is better suitable as a memory unit for IDS. Additionally, when contrasted to recently released algorithms, Bidirectional GRU can achieve the best efficiency.

In 2018, Kehe Wu et al. [7] Conventional IDS have significant challenges as network traffic data continues to grow. Selected features and classification methods have a strong correlation with detection accuracy; however conventional feature selection and classifiers struggle in environments with large amounts of data. Additionally, the uneven nature of the original traffic data seriously affects the classification outcomes. In this study, CNNs are used to offer a novel NID framework. In order to address the issue of an unbalanced dataset, the convolutional neural network approach is used to autonomously extract traffic features from the original dataset. It then adjusts the cost function weight coefficient of every category dependent on its values. The method improves the precision of the class with low quantities while simultaneously lowering the False Alarm Rate (FAR). The actual traffic vector format is converted into digital form in terms of minimizing calculation costs. The effectiveness of the suggested convolutional neural network method was evaluated using the common NSL-KDD dataset. The test findings demonstrate that the suggested model outperforms conventional standard techniques in terms of accuracy, false alarm rates and computation cost. It is a dependable and efficient solution for a vast NID.

In order to create a flexible and efficient intrusion detection to identify and categorize unplanned and unanticipated cyber-attacks, Vinaykumar R et al. [8] examined a DNN, a sort of DL method. The rapid growth of assaults and the ongoing change in network behavior necessitate the evaluation of multiple datasets that have been produced over time using both static and dynamic methods. This kind of research makes it easier to choose the optimization technique for reliably identifying upcoming cyber-attacks. A thorough analysis of deep neural network and other traditional ML classifier studies is presented on numerous standard malware datasets that are freely accessible. Using the KDDCup 99 dataset and indeed the hyper parameter selection techniques, the best network variables & layouts for deep neural networks are selected. Every deep neural network testing is performed for 1,000 epochs with a learning rate that varies between [0.01-0.5]. To perform the comparison, the deep neural network algorithm that did well enough on KDDCup 99 was applied to various datasets including Kyoto, WS-DS, NSL-KDD, UNSW-NB15 as well as CICIDS 2017. By transferring the IDS data through numerous hidden layers, the deep neural network method learns the high-dimensional, abstract characteristic representation of the information. It has been demonstrated through extensive experimentations that deep neural networks outperform traditional ML classification models. Scale-Hybrid-IDS-AlertNet (SHIA), a massively scalable and

composite deep neural network architecture, is finally suggested. It may be utilized in real time to successfully analyze network traffic & host-level activities to prevent potential cyberattacks.
For effective NID, Farrukh Aslam Khan et al. [9] offer a unique two-stage deep learning (TSDL) approach dependent on a stacked auto-encoder with such a soft-max classification model in 2018. The method has two decision phases: the first one uses a probability score value to determine if traffic flow is regular or aberrant. This is then utilized as an extra feature for identifying the normal state as well as other kinds of threats during the ultimate decision step. The proposed approach is able to classify data automatically and effectively by learning relevant feature presentations from massive quantities of unlabeled data. Numerous tests are carried out on two open datasets—the UNSW-NB15, a more recent standard dataset, as well as the KDDcup99, an older standard dataset—to assess and test the performance of the proposed scheme. Analytical and test findings indicate that the developed method greatly surpasses other approaches and methodologies and obtained good recognition accuracy, up to 99.996percent and 89.134percent for the KDDcup99 and UNSW-NB15 datasets, correspondingly. The suggested framework may eventually act as a standard for the DL and network security areas of research, it is determined.

In terms of learning streamlined and effective ID concepts in 2020, Giuseppina Andresini et al. [10] describe a novel DNN architecture by fusing an unsupervised phase for multiple-channel feature learning with such a supervised one that takes advantage of feature interconnections on cross channels. The goal is to determine whether learning and adding class-specific network flow properties to the early models could improve the accurateness. Two autoencoders are specifically learned on the normal and malicious streams, correspondingly, in the unsupervised stage. Such autoencoders could be used to produce two additional feature vectors that enable the presentation of every network flow like a multi-channel sample because the upper surface in the decoder reassembles data in the same space as the input one. To understand how one channel affects the others, a multi-channel parameterized convolution is used in the supervised step. In specific, because the samples come from 2 distinct regions (attack & normal channels), the samples designated as normal should resemble the representation created by the normal autoencoder more closely than the invasion autoencoder, and vice versa. It will be used to better separate the distinctions among normal and malicious flows by taking advantage of this anticipated reliance. On three standard datasets, the proposed NN design outperforms competing ID designs in terms of predicted accuracy.

A CPS-based IEEE 1815.1-based power system IDS is proposed by Sungmoon Kwon et al. [11] for 2020. It analyzes a power system network dependent on IEEE 1815.1 and suggests a workable application strategy for IDS. It proposes a bidirectional RNN-based IDS for an IEEE 1815.1-based network and shows how the proposed methodology can be verified using a variety of intrusion data specific to power systems, such as attacks that CPS-malware behavior, disable reassembly, false data injection and network traffic from actual power systems. The proposed method effectively identified three different false data injection and disabling attacks, as well as five different CPS malware behavior attack kinds.

The introduction of two models for an ID and classification technique is introduced by Zina Chkirbene et al. [12]. For a private network, there are two systems: the Trust-based ID and Classification System (TIDCS) and the Trust-based ID and Classification System-Accelerated

(TIDCS-A). TIDCS uses a new feature selection technique to decrease the amount of features in the data input. The attributes are primarily ordered according to their reliability ratings after being arbitrarily divided to maximize the probability of them taking part in the formation of various groupings. The categorization for any data packets received from a network node is then determined using only the highest ranking features and is stored as component of the node's past data. The TIDCS proposal calls for a periodic system cleaning in which the trust connections among participant nodes are assessed and refreshed. The TIDCS-A technique limits the nodes' exposed window and presents a dynamic technique to determine the precise timing for node cleansing stages. The machine machine learning technique and the node's prior behavior are combined to calculate the ultimate classification choice for both methods. Any discovered assault lowers the dependability of the involved nodes, causing a dynamic system cleaning. TIDCS and TIDCS-A have both been evaluated using the NSLKDD and UNSW datasets, and the results demonstrate that both systems are capable of detecting malicious activities with increased precision, detection rates, and less false alarms than cutting-edge methods. For the UNSW dataset, the precision detection rates for TICDS are 91 percent, online AODE is 83.47 percent, CADF is 88 percent, EDM is 90 percent, TANN is 90 percent, and NB is 69.6 percent. Since TICDS offers good identification and FAR along with superior accuracy identification than state-of-the-art technologies, it performs better than those latter.

In order to identify the attack aspects and carry out wireless network ID in real time, Liqun Yang et al. [13] exploit a Conditional Deep Belief Network (CDBN)-based ID technique in 2017. The majority class data sets are undersampled using the window dependent instance selection method SamSelect, as well as a Stacked Contractive Auto-Encoder (SCAE) method is proposed to minimize the aspect of the sample data in order to mitigate the effects of the extremely unbalanced dataset and data duplication on the detection performance. The suggested strategy can significantly and accurately identify the possible assault by doing this. rThe findings of the test demonstrate that CDBN can be successfully integrated with "SamSelect" and SCAE, and the suggested method has a good detection and precision rate, with an overall detection duration of 1.14 ms.

ML and DL for ID in unbalanced network traffic are being studied by Lan Liu et al. [14] in 2021. To address the issue of class imbalance, a unique Difficult Set Sampling Technique (DSSTE) method is presented. First, separate the uneven training dataset into the challenging set and the easy set using the Edited Nearest Neighbor (ENN) method. The overall samples in the challenging set will then be reduced using the KMeans technique. Zoom in and out on the constant characteristics of the minority dataset in the challenging set and create new samples to boost the minority population. Ultimately, a novel training dataset is created by combining the basic dataset, the condensed dataset of the majority in the tough, and the minority in the challenging dataset. The method evens out the initial training set's imbalance and provides targeted data augmentation for the underrepresented class that needs to learn. It allows the classification model to work better during categorization and better learn the distinctions during the training phase. Investigations were performed out using the more recent and comprehensive intrusion dataset CSE-CIC-IDS2018 as well as the venerable intrusion dataset NSL-KDD to validate the suggested technique. The following categorization methods are being used: RF, AlexNet, Mini-VGGNet, SVM, XGBoost, and LSTM. The test findings show that the suggested DSSTE algorithm is efficient than the other approaches when 24 techniques are compared.

In 2021, Qi Liu and colleagues [15] There are three main types of IDS: misuse, specification and anomaly-based. Both misuse and anomaly-based intrusion detection system are the subject of in-depth academic and commercial research. Therefore, misuse-based attack identification method will typically miss their targets since critical structures, such as smart grids (SG), may frequently be subject to advanced malicious activities in the coming years. Anomaly-based IDS can identify malicious activity, but they are not frequently used in business due to a higher false positives rates and poor accuracy of trained classifier.

Specification-based intrusion detection system can be viewed as the most effective detection machine for cyber-physical systems (CPS), along with SG, due to misuse-based intrusion detection system' attempt to detect new types of attacks, necessity for routinely mechanically creating and modifying signatures, and anomaly-based intrusion detection system' poor reputation for high FAR. According to some, rule learning-based specification-based IDS may end up being the most effective intrusion detection system for SG. To account for changing system behaviour in SG, ID rules are routinely updated automatically using rule learning methods. Due to the symbolic depiction of learned rules, rule learning oriented intrusion detection system can ultimately not only identify recently discovered assaults but also gain higher interpretability. The current work offers a thorough and methodical analysis of rule learning methodologies and their applicability to intrusion detection system in SG. Additionally, it summarizes the most crucial standards for understanding ID rules and judging their value. This research not only provides an overview of several crucial rule learning approaches, but also the first survey of their use in IDS, indicating prospective uses in SG security.

Network attacks on the railway ECN, including IP Scan, DoS and Man in the Middle, Port Scan are anticipated in 2021, according to Chuan Yue et al. [16]. A unique dataset is created by extracting 34 features of various protocol components from the original data produced by our ECN testbed. The dataset is optimized using a temporal sequence construction techniques and a data imaging technique. Six fundamental classifiers—LeNet-5, AlexNet, long short term memory, VGGNet, SimpleRNN and GRU—are constructed using a variety of common CNNs and RNNs. The integration of all the basis classifiers is presented using a dynamic weight matrix voting mechanism. On the basis of the dataset, the proposed methodology is assessed. The experiment's findings demonstrate that the suggested approach excels at combining the benefits of all basic classifiers and delivers a higher detection accuracy of 0.975.

In 2021, Ulya Sabeel et al. [17] suggest combining a defensive artificial intelligence engine with a two-step feature selection method and hyperparameter artificial intelligence model optimization. The suggested approach is used in this study to identify binary attack flows, and the CICIDS2017 dataset is used to train and validate the AI models. The system is then tested using synthetic attack flows that are designed to resemble real-world circumstances. Numerou DL and ML models, including deep neural network, Linear-SVC, and Layered DT Classification models, are used to show the efficacy of the suggested a typical attack flow identification strategy. The suggested defensive artificial intelligence engine greatly raises the

True Positive Rate (TPR) of artificial intelligence models on a variety of unusual attacks, according to simulation data.

Jan Lansky et al. [18] present an extensive survey and categorization of deep learning-based ID techniques in 2021 with an emphasis on these methods. It begins by outlining the key underlying ideas on the intrusion detection system framework and several DL approaches. Then, it categorizes these schemes based on the specific deep learning techniques each one uses. It explains how effective ID is achieved by using DL networks in the ID process. Final thoughts and future trends are highlighted after a thorough examination of the examined intrusion detection systems is given.

Gustavo De Carvalho Bertoli et al. [19] predict that in 2021. NIDs have challenges from design to operation due to the rise in devices connected and the hackers' ongoing evolution of their tactics. As a consequence, ML methods are increasingly being used in NIDS. These research' dataset, unfortunately, is no longer relevant in terms of attack and baseline traffic. In order to allow the full deployment of the system, this article examined the AB-TRAP architecture, which permits the usage of current network traffic and takes operational considerations into account. The AB-TRAP architecture consists of five steps: (i) creating the attack dataset; (ii) creating the genuine dataset; (iii) training ML methods; (iv) realizing (implementing) the approaches; and (v) assessing the effectiveness of the realized model after implementation. To stop TCP port scanning assaults, the AB-TRAP is tested in both local (LAN) and remote (internet) contexts. A DT with minimum CPU and RAM use in kernel space produced a f1-score of 0.96 and an area under the ROC curve of 0.99 for the LAN study scenario. With a mean f1-score of 0.95, a mean area under the ROC curve of 0.98, and a mean overhead of 1.4 percent CPU and 3.6 percent RAM on user-space in a single-board computer, the online case uses 8 ML methods. The reproducibility, usage of the most recent network activity, assaults, and addressing of the problems regarding the method's reality and deployment are some of this architecture's most important attributes.

A unique IDS design that incorporates data gathering, analysis, and extraction of features by fusing data reconstruction characteristics, reconstruction errors, auto-encoder variables, and GMM is proposed by Wang Hui et al. in 2021 [20]. After training and testing on several ID data sets, the system beats previous unsupervised learning-based identification methods in regards of reliability, recall, F1-score, as well as other evaluation metrics.

Proposed system design

Our study's approach recommends a system for detecting intrusions by using a Recurrent Neural Network with Long Short-Term Memory. Using ML/DL Techniques, the harmful behaviour has previously been researched and discovered by a number of current methodologies. However, problems such as a high rate of false alarms and inadequate classification accuracy persist in such systems. Take a look at figure 1, which illustrates the proposed RNN-LSTM IDS. The technique we offer can be used with data from a wide range of SNS.At first, the system pulls information from the user's most recently seen tweet comments through the Twitter API connection. The biggest issue with social networking apps is that they can't spot fake or suspicious profiles.

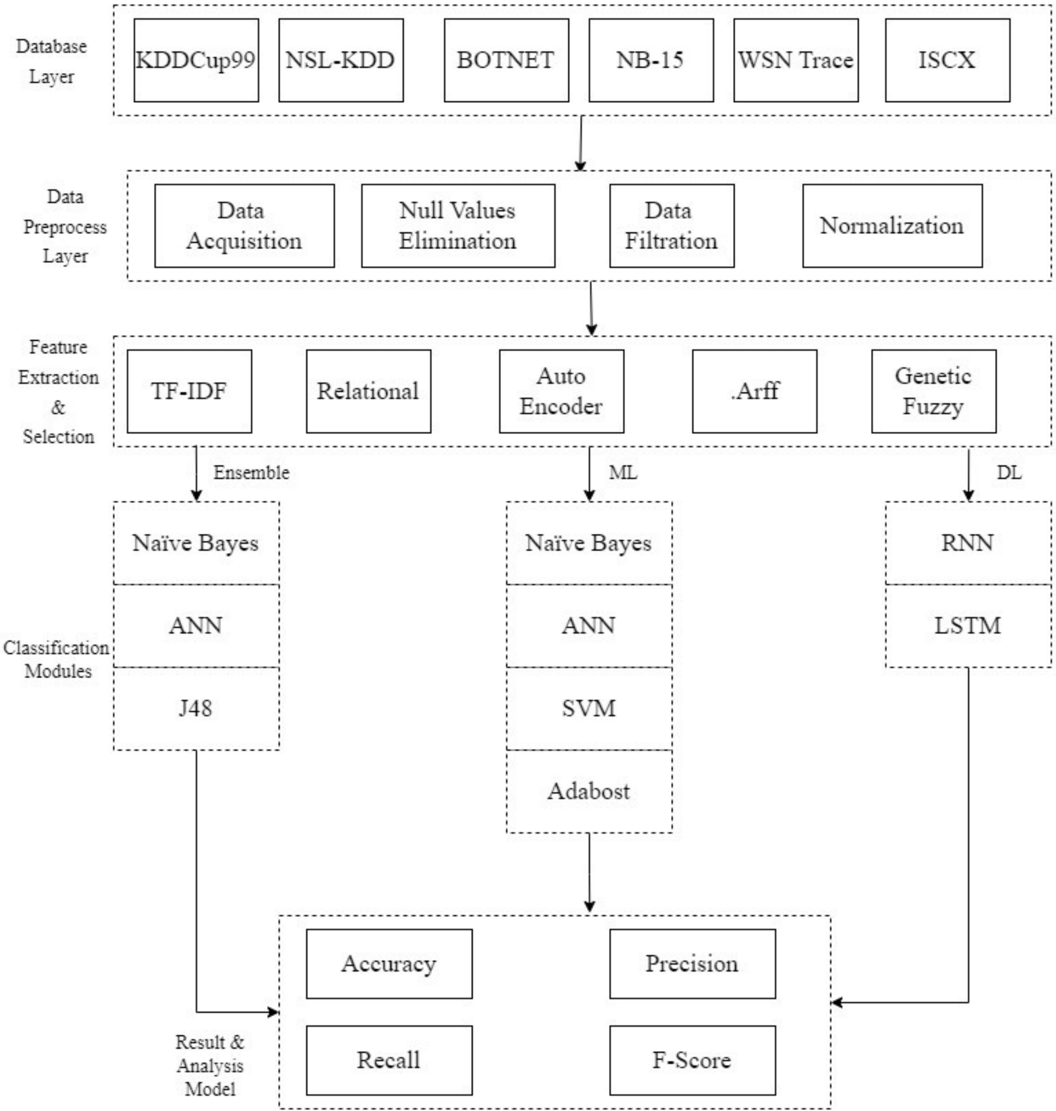

Figure 1 : Proposed system design architecture for IDS using machine learning and deep learning

We combined an NLP system with a machine learning method to fix this problem in current software. To begin, we collect information from several social media platforms. When information arrives, it is filed away in data sets and data repositories. The information gathered from social media platforms like Twitter and others may not be in a standard format. Pre-processing this sort of data using a targeted sampling strategy and data filtering methods is essential. The filter has been used to remove misclassified occurrences, and the systematic sampling method is utilized to separate the data. After a data cleanup, the next steps are feature extraction and feature selection. In order to categorize data, we used a Recurrent Neural Network with Long Term Memory. Our proposed approach uses a machine learning method to quickly and accurately identify suspicious items in real-time streams of data, whether they are synthetic or coming from the real world. It has a low error rate and may be used on both homogeneous and heterogeneous datasets in a row. In the following, we'll go through the specifics of our suggested framework.

Data Collection : System uses well known datasets namely KDDCUP 99, Botnet, WSNtrace, NSLKDD, ISCX, NUSW-NB 15 and it collects real-time data from the Twitter connection using Twitter API, which extracts the data from tweet comments recently viewed by users. This collected data is pre-processed using various pre-processing techniques.

Data pre-processing : There may be a lot of useless information and gaps in the data. Data preparation is carried out to handle this portion. Numerous data pre-processing techniques, such as data cleansing, data transformation, and data reduction, have been utilized at this step.

Data cleaning- It addresses noisy data, missing information, etc. Different strategies have been adopted when some data in the information is incomplete, such as filling in the gaps or disregarding the tuples. Data may contain null values that are incomprehensible to machines. This noisy data may result from poor data collecting, incorrect data input, etc. Regression, clustering, and the binary approach are used to address it.

Data transformation- This technique is used to change the data into the form that is suited for the mining procedure. Normalization, attribute choice, and discretization are involved in this technique. The process of normalization involves scaling the data values to fall between a range of -1 and +1. To aid in the mining procedure, additional features are built from the available set of characteristics in the feature selection procedure. The discretization technique substitutes periodic levels for the numerical attribute's raw values.

Feature extraction, generation and evaluation : From the data input, this procedure retrieves a variety of features. The extracted features are then standardized using a feature selection threshold, which eliminates redundant and unnecessary features for training. The normalized data with relational characteristics is used to extract a variety of hybrid attributes, and training is carried out by selecting an optimization strategy. The detailed explanation of the procedures of feature extraction, feature generation, and feature evaluation are explained as follows

Feature Extraction - This dimensionality-reduction technique divides the original information into identifiable categories based on their relationships. The fact that these enormous datasets include a large number of parameters and that processing those variables takes a lot of computational power is one of their distinguishing characteristics. Hence, in this situation, extraction of features can be effective for choosing specific variables and integrating some relevant variables, which in a sense would lessen the quantity of data. Precision and recall measurements would be used to assess the outcomes. Among the most popular methods for reducing the number of linear dimensions is PCA. It is a method for unsupervised learning.

Feature generation - It is the process of creating brand-new features from pre-existing ones. It is impossible to handle the larger datasets because of how widely the dataset sizes fluctuate. Thus, the feature generating procedure might be extremely useful in facilitating the task. We employ several mathematical formulas and statistical methods to improve precision and clarity while avoiding the generation of meaningless characteristics. This technique typically increases the model's knowledge to improve its accuracy. Increasing model accuracy can therefore be accomplished through this technique. By identifying relevant interactions, this approach sort of ignores the pointless interactions.

Feature evaluation - To complete the task in a well-organized way, it is crucial to prioritize the characteristics at first and thus, feature evaluation is a crucial technique for this. Here, all features are being assessed in order to grade them accurately and afterwards use them in accordance with the necessities of the moment. The irrelevant ones can be avoided. Therefore, feature evaluation is a crucial operation to carry out in order to obtain a correct final output from the model by minimizing bias and incomplete data.

Linear and Non-Linear Feature Extraction - Linear & non-linear extraction of features fall under two main groups. PCA is one illustration of linear feature extraction. A dataset's primary features are combined in a normalized linear manner to form a principal component. In essence, PCA is a technique for extracting essential variables from a wide set of variables present in a data set. Data is usually orthogonally transformed via PCA into a lower-dimensional space, maximizing the distribution of the data. As anomalies and outliers are regarded as noise or useless data over the entire dataset, it can be used for abnormality and anomaly analysis.

Feature Selection

For several purposes, feature selection is the most important phase in creating a better model. One is being limiting the amount of features that can be taken into account while developing a model through feature selection involves some level of cardinality reduction. The majority of the time, data is either inaccurate or includes more data than is necessary to solve the model. We have gathered real- time data of twitter, which has several features but it has few features which doesn't obtain any advantage after adding. There are some redundant columns, and using them might affect the model. Feature selection not only boosts the model's performance, but it also speeds up the modeling process. When creating a model, if unnecessary columns are included, more CPU and storage are needed for training, and more memory space is required for the produced model. Even if resources were not a concern, it is still crucial to do feature selection and choose the optimal features because unnecessary columns might harm the model's performance in a number of ways, including identifying useful patterns in distorted or redundant information is more challenging, and most DM methods need substantially bigger training data sets if the set of data is highly dimensional.

The technique actively chooses or excludes features during the feature selection process depending on how valuable it is for evaluation. Obtaining too much information that is of low value or not enough data that is of great value are two issues that feature selection aids in resolving. Finding the smallest number of data source attributes that are important for creating a model is our aim while choosing features

Classification: Finally, the system detects each record, either attack or normal using a supervised classification technique. We used RNN and LSTM as a supervised classification algorithm. Then supervised machine learning is applied in order to train the classifier.

A type of artificial neural network called a recurrent neural network has interconnections between its units that create a directed graph along a series. Time series data handling is its principal usage. Sequential data is to be utilized by RNN. Because they complete the same task for each element of a sequence, RNNs are known as recurrent networks. RNNs can also be explained by the fact that they contain a "memory" that stores data about previous calculations. An RNN's construction is graph

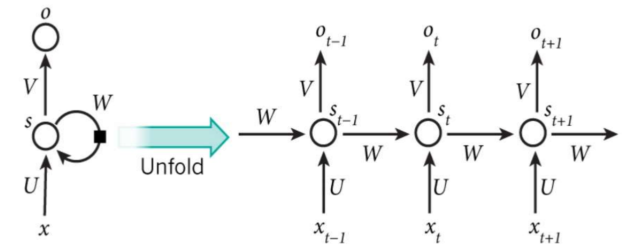

Figure 2 Framework of recurrent neural network

With the exception that activation is transferred to the hidden layer from both the existing external input & the hidden layer activations first stage backwards in time, the forward pass of the recurrent neural network is nearly identical to that of an MLP with one hidden layer. We have the following as input to hidden units. Consider following equation.

$$a_h^t = \sum_{i=1}^{I} \mathrm{w}_{ih}\mathrm{x}_i^t + \sum_{\acute{h}=1}^{H} \mathrm{w}_{\acute{h}h}\mathrm{b}_{\acute{h}}^{t-1} \tag{3.2}$$

$$b_h^t = \theta_h(a_h^t) \tag{3.3}$$

For the output unit: The recurrent neural network's back propagation is merely regular back propagation. While beginning at t = T and iteratively applying the following functions, reducing t at every step, it is possible to determine the entire series of delta values.

$$a_k^t = \sum_{h=1}^{H} w_{hk} b_h^t \qquad 3.4$$

δj T+1= 0, ∀j, because no error is obtained from over the termination of the sequence.

$$\delta_h^t = \theta^j(a_h^t)(\sum_{k=1}^{K} \delta_k^t w_{hk} + \sum_{h'=1}^{H} \delta_{h'}^{t+1} w_{hh'} \qquad 3.5$$

$$3.6 \qquad \delta_h^t = \frac{\partial O}{\partial a_j^t}$$

In this section we have discussed our proposed methodology of intrusion detection system using machine and deep learning techniques. We have proposed our research work in two phases. In phase 1, using Genetic algorithm and ML classifiers, IDS is modeled and in phase 2, IDS is designed using RNN-LSTM. Various experiments are performed to evaluate the performance of our propose system.

Results and Discussion

During the experimental study, we generate the matrices so that the system performance evaluation would be as accurate as possible. The computer has a 2.8 GHz Intel i3 processor and 4 gigabytes of random access memory (RAM). The system was developed using an open-source Python architectural framework. Following the completion of the system's installation, a number of operating systems that are now in use were contrasted with the system that was presented. The specifics of both trials are described in further depth below.

In this experiment we demonstrate performance of RNN-LSTM (ReLu) using twitter dataset with various cross validation and results are illustrated in table 5.7. According to this analysis we conclude 20 fold cross validation provides highest classification accuracy of 97.95 % using RNN with Sigmoid function.

Consider the following Figure 3 which depicts the validation of model with 20 fold cross validation using RNN-LSTM (ReLu) classifier.

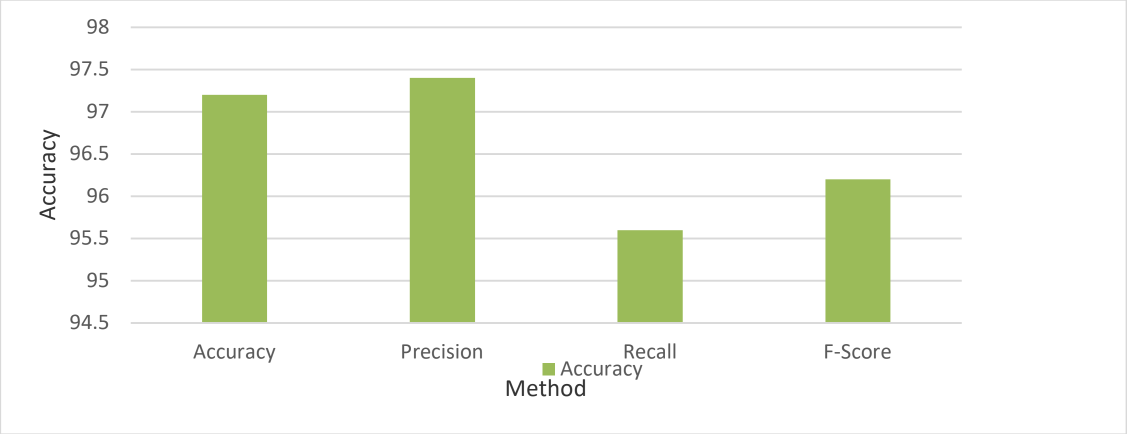

Figure 3: Performance of RNN-LSTM (ReLu) with 10-fold cross validation

Experimental findings of figure 3 shows by using 10 fold cross validation, the accuracy, precision, recall and f-score of RNN-LSTM (ReLu) model is 97.2, 97.4, 95.6 and 96.20 respectively. Consider the following figure 4 which depicts the validation of model with 15 fold cross validation using RNN-LSTM (ReLu) classifier.

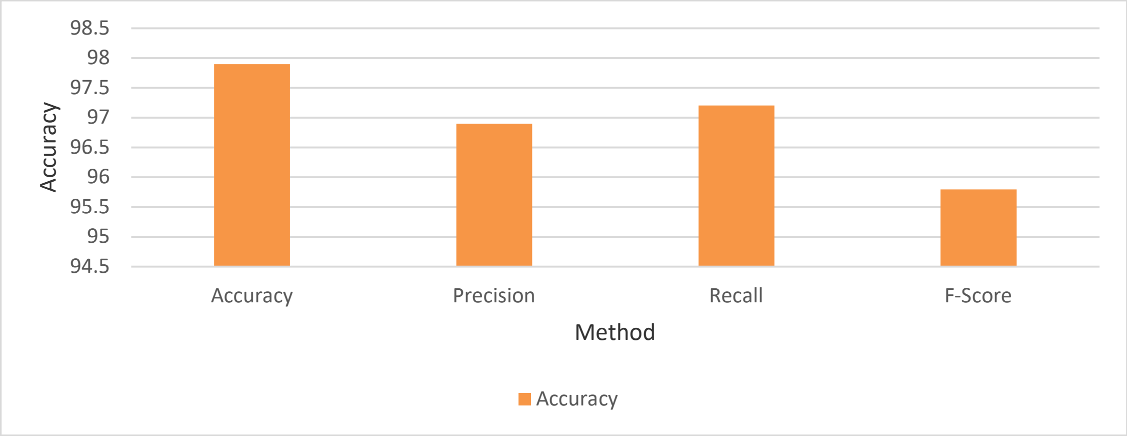

Figure 4: Performance of RNN-LSTM (ReLu) with 15-fold cross validation

Experimental findings of figure 5.19 shows by using 15 fold cross validation, the accuracy, precision, recall and f-score of RNN-LSTM (ReLu) model is 97.9, 96.9, 97.2 and 95.80 respectively. Consider the following figure 5 which depicts the validation of model with 20 fold cross validation using RNN (ReLu) classifier.

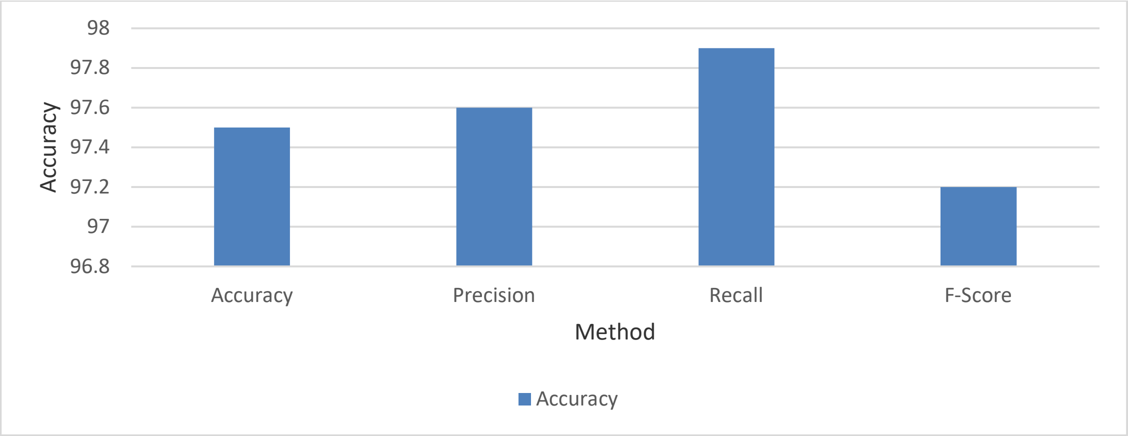

Figure 5: Performance of RNN-LSTM (ReLu) with 20-fold cross validation

Experimental findings of figure 5 shows by using 20 fold cross validation, the accuracy, precision, recall and f-score of RNN-LSTM (ReLu) model is 97.95, 97.6, 97.90 and 97.20 respectively. The system can detect the all kind of attacks like DOS, Probe, U2R and R2L respectively. It can be applicable for online as well as offline environment like NIDS and HIDS. System can work on different network dataset e.g. KDD Cup 99, NSL KDD, ISCX and normal time network sniffer dataset. The experiment analysis works with different analysis tools. This research has an immense potential to open doors for improving the intrusion detection applications domain as it offers a better detection rate in catching any attack before network security is compromised.

We have performed various experiments for the evaluation of performance of our research methodology. We have proposed our research methodology in two phases. In phase 1, we proposed Intrusion detection system using Genetic Algorithm and various Machine learning techniques such as J48, Artificial Neural Network and Random Forest. In this phase, we have performed total 6 experiments for evaluating performance of our proposed system by considering various parameters like using different population size, threshold, datasets etc. We have concluded that,

- Detection rate of Denial of Service attack is 96.9 % which is high as compared to other attacks.

- As the size of population increases, the total number of GA rules also increases for various datasets. That is number of GA rules are directly proportional to population size.

- Accuracy rate of Denial of Service attack and Error rate of Root to Login attack is high as compared to other attacks using various datasets.

- For detecting different known as well as unknown attack, 0.50 threshold value produces optimum result.

- When the performance of J48, ANN and RF algorithm is evaluated using various Datasets like KDDcup99, botnet, NSLKDD, ISCX, NUSW NB-15, it is observed that accuracy rate of Artificial Neural Network classification is 79.94 % which is high as compared to other algorithms.

In the phase 2, we have evaluated proposed RNN-LSTM for various functions like sigmoid, tanh and ReLu using different 10-fold, 15-fold and 20-fold cross validation. It is observed that accuracy rate of RNN-LSTM (ReLU) using 20-fold cross validation is 97.95 % which is high as compared to RNN-LSTM (Sigmoid) and RNN-LSTM (Tanh) when different cross validations like 10-fold, 15-fold and 20-fold are used.

Conclusion

Information security is greatly aided by intrusion detection, and the fundamental technology is the ability to precisely identify different network threats. In our research, we investigate the design of an intrusion detection system using ML and DL methods. We have proposed our research methodology in two phases using various datasets like KDDcup99, botnet, NSLKDD, ISCX, NUSW NB-15, WSNtrace and real-time Twitter data. In phase 1, we proposed Intrusion detection system using Genetic Algorithm and various Machine learning techniques such as J48, Artificial Neural Network and Random Forest. In this phase, we have performed various experiments for evaluating performance of our proposed system by considering several parameters like using different population size, threshold, datasets etc. We have concluded that, Detection rate of Denial of Service attack is 96.9 % which is high as compared to other attacks. As the size of population increases, the total number of GA rules also increases for various datasets. That is number of GA rules are directly proportional to population size. Accuracy rate of Denial of Service attack and Error rate of Root to Login attack is high as compared to other attacks using various datasets. For detecting different known as well as unknown attack, 0.50 threshold value produces optimum result. When the performance of J48, ANN and RF algorithm is evaluated using various Datasets like KDDcup99, botnet, NSLKDD, ISCX, NUSW NB-15, it is observed that accuracy rate of Artificial Neural Network classification is 79.94 % which is high as compared to other algorithms.

In the phase 2, we have evaluated proposed RNN-LSTM for various functions like sigmoid, tanh and ReLu using different 10-fold, 15-fold and 20-fold cross validation. It is observed that accuracy rate of RNN-LSTM (ReLU) using 20-fold cross validation is 97.95 % which is high as compared to RNN-LSTM (Sigmoid) and RNN-LSTM (Tanh) when different cross validations like 10-fold, 15-fold and 20-fold are used. Our proposed system can detect denial of service attack, root to login attack, probe attack, active attack, network attack, passive attack, user to root attack. Proposed system can generate its own rules so there is no need of manual energy for monitoring packets. It can detect Network-IDS as well as Host-IDS effectively. As real-time twitter data is used by using actual tweets, our system can detect whether the twitter account is real or fake. In future we will try to reduce the training time required to train the system.

References

[1] Yin Chuan-long, Zhu Yue-fei, Fei Jin-long and He Xin-zheng. “A Deep Learning Approach for Intrusion Detection using Recurrent Neural Networks”, 2017, IEEE.

[2] George Loukas, Tuan Vuong, Ryan Heartfield, Georgia Sakellari, Yongpil Yoon and Diane Gan. “Cloud-Based Cyber-Physical Intrusion Detection for Vehicles Using Deep Learning”, 2018, IEEE.

[3] Lorenzo Fernandez Maimo, Angel Luis Perales Gomez, Felix J. Garcıa Clemente, Manuel Gil Perez and Gregorio Martınez Perez. “A Self-Adaptive Deep Learning-Based System for Anomaly Detection in 5G Networks”, 2018, IEEE.

[4] I. Ahmad, M. Basheri, M.J. Iqbal and A. Raheem. “Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection”, 2018, IEEE.

[5] Sheraz Naseer, Yasir Saleem, Shehzad Khalid, M Khawar Bashir, Jihun Han, M Munwar Iqbal and Kijun Han. “Enhanced Network Anomaly Detection Based on Deep Neural Networks”, 2015, IEEE.

[6] Congyuan Xu, Jizhong Shen, Xin Du AND Fan Zhang. “An Intrusion Detection System Using a Deep Neural Network with Gated Recurrent Units”, 2018, IEEE.

[7] Kehe Wu, Zuge Chen and Wei Li. “A Novel Intrusion Detection Model for a Massive Network Using Convolutional Neural Networks”, 2018, IEEE.

[8] Vinaykumar R, Mamoun Alazab, Soman Kp, Prabaharan Poornachandran, Ameer Al-Nemrat and Sitalakshmi Venkatraman. “Deep Learning Approach for Intelligent Intrusion Detection System”, 2018, IEEE.

[9] Farrukh Aslam Khan, Abdu Gumaei, Abdelouahid Derhab and Amir Hussain. “TSDL: A Two-Stage Deep Learning Model for Efficient Network Intrusion Detection”, 2018, IEEE.

[10] Giuseppina Andresini, Annalisa Appiice, Nicola Di Mauro, Corrado Loglisci, Donato Malerba. “Multi-Channel Deep Feature Learning for Intrusion Detection”, 2020, IEEE.

[11] Sungmoon Kwon, Huunguk Yoo and Tashik Shon. “IEEE 1815.1-Based Power System Security With Bidirectional RNN-Based Network Anomalous Attack Detection for Cyber-Physical System”, 2020, IEEE.

[12] Zina Chkirbene, Aiman Erbad, Ridha Hamila, Amr Mohamed, Mohsen Guizani and Mounir Hamdi. “TIDCS: A Dynamic Intrusion Detection and Classification System Based Feature Selection”, 2016, IEEE.

[13] Liqun Yang, Jianqiang Li, Zhonghao Sun, Yufei Zhao and Zhoujun Li. “Real-Time Intrusion Detection in Wireless Network: A Deep Learning-Based Intelligent Mechanism”, 2017, IEEE.

[14] Lan Liu, Pengheng Wang, Jun Lin and Langhou Liu. “Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning”, 2021, IEEE.

[15] Qi Liu, Veit Hagenmeyer and Hubert B. Keller. “A Review of Rule Learning-Based Intrusion Detection Systems and Their Prospects in Smart Grids”, 2021, IEEE.

[16] Chuan Yue, Lide Wang, Dengrui Wang, Ruifeng Duo and Xiaobo Nie. “An Ensemble Intrusion Detection Method for Train Ethernet Consist Network Based on CNN and RNN”, 2021, IEEE.

[17] Ulya Sabeel, Shahram Shah Heydari, Khalid Elgazzar and Khalil El-Khatib. “Building an Intrusion Detection System to Detect Atypical Cyberattack Flows”, 2021, IEEE.

[18] Jan Lansky, Saqib Ali, Mokhtar Mohammadi, Mohammed Kamal Majeed, Sarkhel H. Taher Karim, Shima Rashidi, Mehdi Hosseinzadeh and Amir Masoud Rahmani. “Deep Learning-Based Intrusion Detection Systems: A Systematic Review”, 2021, IEEE.

[19] Gustavo De Carvalho Bertoli, Lorenco Alves Pereira Junior, Osamu Saotome, Aldri L. Dos Santos, Filipe Alves Neto Verri, Cesar Augusto Cavalheiro Marcondes, Sidnei Barbieri,

Moises S. Rodrigues and Jose M. Parente De Oliveira. “An End-to-End Framework for Machine Learning-Based Network Intrusion Detection System”, 2021, IEEE.

[20] Wang Hui, Wang Dongming, Li Dejian, Zeng Lin and Wang Zhe. “A Framework For Network Intrusion Detection Based on Unsupervised Learning”, 2021, International Conference on Artificial Intelligence and Industrial Design (AIID), IEEE.

[21] Deshmukh, M.S., Alvi, A.S. (2022). Detection and Prevention of Malicious Activities in Vulnerable Network Security Using Deep Learning. In: Gunjan, V.K., Zurada, J.M. (eds) Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications. Lecture Notes in Networks and Systems, vol 237. Springer, Singapore. https://doi.org/10.1007/978-981-16-6407-6_29

[22] Deshmukh, M.S. and Alvi, A.S., 2022. Network Intrusion Attack Detection and Prevention using Various Soft Computing and Deep Learning Techniques. JOURNAL OF ALGEBRAIC STATISTICS, 13(3), pp.1505-1514.