International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Hybrid-Ids: An Approach for Intrusion Detection System with Hybrid Feature Extraction Technique Using Supervised Machine Learning

Mr. Kishor P. Jadhav[1] Dr. Tripti Arjariya[2] Prof. (Dr.) Mohit Gangwar[3]

Submitted: 23/01/2023 **Revised:** 19/03/2023 **Accepted:** 14/04/2023

Abstract: At a breakneck pace, the IoT (i.e. Internet of Things) and networking technology, security has become a significant issue, such as data security, virtual machine hacking and various internal and external attacks. Conventional Intrusion Detection Systems (IDS) have a lot of limitations due to resource dependency and their complexity. Multiple researchers have implemented IDS systems with network logs or real-time network audit datasets**.** The KDDCUP99 and NSLKDD are the most popular datasets that existing authors use, but challenges persist in detecting unknown, active, passive, and others. In this paper, we proposed a heterogeneous extraction of attribute or feature and selection method for IDS, by using machine learning methodologies for the recognition of network intrusion as well as host intrusion. The numerous network log dataset has been used to detect the intruder in a vulnerable environment. The various heterogeneous feature extraction methods have been carried out for building a robust module. In the testing module, entire training rules validates the input packet data with training rules and executes the weight using a majority voting algorithm. Finally, it detects whether the current packet is normal or intruder based on majority voting values and eliminates that connection. In an extensive experimental analysis, three classifiers have been used for validation, such as ANN, SVM and RNN of different network log datasets. In observation, RNN produces the highest detection and classification accuracy over the SVM and ANN. It also reduces the time complexity and error rate with al datasets

Keywords: *NIDS, HIDS, machine learning, supervise classification, network log dataset, KDDCUP99, feature extraction, feature selection*

1. Introduction

Machine embedding and interconnection have increased network data traffic as the Internet of Things (IoT) have grown [1]. Data volume has also resulted in more significant network security concerns. With the advancement of networking technologies, malicious assaults and threat malware are becoming more prevalent and replicating at a quicker rate [2]. Network intrusion detection confronts significant problems as the primary defence against sophisticated attacks. Feature-based and anomaly-based identification are the two most prevalent detection approaches [3]. Signature-based identification is very beneficial when the attack characteristic is known. On the other hand, Exceptional situation detection may be used to identify both known and unknown assaults. Because of its simplicity and ease, the monitoring system known as feature detection is commonly used as a typical network assault detection approach. Its flaws are also readily apparent. The classification performance of the feature-based malware detection is constrained by the smaller dimensions and signatures repository update pace, and it cannot identify unknown attack types. In recent years, academics have attempted to overwhelmed this issue by introducing new methods in vulnerability scanning, particularly the recent rise of technology based on machine learning (ML). Several investigators have used ML (machine learning) methods for intrusion prevention, like SVM (support vector machine), decision trees (DT), k-nearest neighbours (KNN) and convolutional neural networks (NN) and have seen some primary results. An IDS with machine learning network model is proposed in this research. This method is an ensemble learning method that combines the benefits of numerous machine learning (ML) algorithms have been shown to improve the detection rates of possible attacks in network ID (intrusion detection). The following are the paper's key contributions:

- The numerous machine learning (ML) classifiers and deep neural network (DNN) are combined to form a hybrid learning system RNN which enhances the accuracy of ID (Intrusion Detection) technologies and provides a new route for ID (intrusion detection) investigation.

[1]*Ph.D. Research Scholar, Department of Computer Science and Engineering, Bhabha University, Bhopal, Madhya Pradesh, India*
[2]*Head, Department of Computer Science and Engineering, Bhabha University, Bhopal, Madhya Pradesh, India*
[3]*Professor, Department of CSE, B. N. College of Engineering and Technology, Lucknow, India*
[3]*Corresponding Author: Prof. (Dr.) Mohit Gangwar (mohitgangwar@gmail.com)*

- The suggested RNN combines the outputs of several basic classifier models, as well as decision data and increasing the detection model's generalisation and resilience.
- To assess our proposed system, we utilise a numerous network intrusion dataset. The experimental findings reveal that RNN outperforms conventional approaches and the majority of existing algorithms. We believe that the suggested system has promising detection accuracy on well-known dataset.

The structure of the paper is broken down into the following sections: Part II provides a demonstration of the literature review of a suggested model, describing how different existing systems have been analysed, and Section III lists all of the existing systems. The research methods of the provided system as well as the implementation details are discussed in Part III, whilst the suggested algorithm is explained in depth in Section IV. In part V, the findings and discussion of the experimental set-up are reviewed, as well as the different experimental results. Finally, in section VI of the suggested model, the conclusion and suggestions are discussed.

2. Literature Survey

Machine learning (ML) algorithms based on network monitoring have been employed in numerous sectors. A monitoring method for social networks for analysing and detection of traffic collisions is provided utilising neural networks (NN) with bi-direction long-short-term memory [1]. The recommended approach employs query-based crawling to obtain facts about traffic from social media: this method collects phrases associated with traffic occurrences such as heavy traffic, street closures, and so on. To organize the obtained data, several pre-processing strategies are used, including stem, tokenize, POS tag, and categorization. By using a LDA (i.e latent Dirichlet allocation) methodology, the data is then automatically classified as 'traffic' or 'non-traffic'. Traffic-labelled information is divided into positive, negative, and neutral categories. It provides a statement tagged with if it is a traffic or non-traffic statement, and the polarities (good, bad, or neutral) of that traffic statement (neutral, negative or positive). The BoW (i.e bag-of-word) method is then used to convert each sentence into a single-hot encoding format that can be input into the (Bi-LSTM) Bi-directional LSTM NN(neural network). Following the learning procedure, the MM (neural network) employs the softmax layer to perform multi-class classification of the language where geography, traffic incidence, and polarity categories are all factors to consider. The recommended strategy associates many classic ML (machine learning) as well as advanced deep learning methods in perspective of F-score, correctness, as well as other measures.

Machine learning (ML) has been used to create and expand medical systems through a number of experiences and workshops [2]. Different recommended machine procedures were in use, including KNN (K Nearest Neighbours), LR (logistic regression), RF (Random Forest), K means clustering and several others, as well as deep learning techniques like CNN and Recurrent neural network fully - connected layers, an automatic-encoder. These various methods enable scientists to deal with a large number of different databases, such as diagnostic imaging, background, health history, video files, and so on. As more than just a consequence, a wide lot of factors and techniques are covered, including causality, Covid-19 researches, and diagnose illnesses, also including ailments and heart ailments.

Healthcare surveillance is implemented in order to predict heart attacks utilising clever ensemble deep learning algorithms [3]. Real-time health monitoring may help predict and prevent heart attacks. The recommended ensemble deep learning technique for illness prediction scored 98.5 percent accuracy. The approach employs various types of data stored in a database in the cloud. It also includes information obtained from devices placed on the body to gather over ten distinct types of medical data and also doctors' daily computerised medical records which includes data like smoking history and family diseases. The Framingham Risk Factors approach is used to apply feature fusion which then combines data and extracts a fused and useful fraction from the information. Various preprocessing procedures including normalisation, missing value filtering, and feature weighting are used to organise and prepare the data. The information is then used by an ensemble DL (deep learning) system to forecast whether or not a heart illness would arise.

A NIDS is an inline or passive intrusion detection system. It only detects at both the network as well as host levels. Only the managed network design works with NIDS. Adopting NIDS saves money and time since no sensor programming is required at the host level. NIDS can detect attacks in near-real-time traffic monitoring. However, it has the following flaws. It can't tell whether the detection of attack will work or not since it can't see the host machine. The attack cannot be detected in encrypted network traffic [4]. Also, NIDS may not detect all messages in a large or active network. As a consequence, it's possible that it'll miss an attack conducted during a busy moment.

With Software defined networking, NIDS becomes a powerful security defensive tool. At the network

entrance point, there is a system for monitoring network attacks [5]. For ages, NIDS have been researched and used. It refers to a software or device that scans network data for suspicious or malicious activity [6]. Malware assaults, fraudulent users, data leaks, and DDoS attacks are just a few possibilities. To be successful, NIDS must be properly incorporated into security system.

The SVM classifier [7] was utilized in a PCA-based ID (intrusion detection) system. The NSLKDD dataset is often used to learn and enhance the aberrant pattern detection algorithm. A Minimum-Maximum normalisation method handled the least misclassification errors [3]. The PCA method reduces the complexity of the NSLKDD dataset, lowering the set of parameters that can be trained is required. For Support vector machine, a nonlinear radial basis function kernels was being used. Utilizing detection performance, false alarms and correlation coefficient measurements, the presented method has 95 percent accuracy. An aggressive gradient-boosting classifier is used to distinguish between conventional and DoS assaults [9]. A free and open-source SDN framework for experimenting with and designing SDN-based approaches was used to analyse and test the detection strategy.

Mininet was being used to perform Software defined network based cloud identification in real - time basis. To prevent overfitting, we employed logistic regression with a regularisation term penalty. The XGBoost term was created to enhance computations by constructing structure trees. The training set consisted of 400 K samples from the KDD Cup 1999 dataset. Two normalisation methods were used: logarithmic and Min-Max. Total accuracy of XGBoost was 98 percent compared to RF and SVM's which was 96 and 97 percent. The Mininet and floodlight platforms were used to simulate an SVM detection system based on DDoS attack characteristics [10]. Based on the packet network, the items are divided into six tuples named, FES (flow entry speed), SSIP (source IP) speed, SDFP (flow packet standard deviation) pair-flow ratio and source port (SP) speed. On the basis of support vector machine classifier's 6 features, the present network status is normal or attacked. AF (Attack flow), DR (Detection rate) and FAR (False alarm rate) were chosen for 95 percent accuracy.

According to TSDL [11], In Intrusion detection system, a classifier could be a layered auto-encoder using softmax in the output nodes. In order to detect assaults, TSDL was built. Various datasets were pre-processed to boost detection and monitoring efficiency. UNSW-NB15 detection accuracy was 89%. Many NIDS neural network models were published [34], including variation auto-encoders, seq2seq structures using LSTM, and fully connected networks.

Several datasets were utilised to construct and test the recommended approach for discriminating between attack and regular packets in the network. In order to prepare data for neural network training, preprocessing techniques including one-hot encoding and normalisation have been used. Because of these characteristics, neural networks (NN) may learn complicated features from the many scopes of a small packet. A deep NN (neural network) design [12] for evaluating intrusion attempts was developed using 4 hidden neurons. To analyse and decrease data, feature scale and encode were utilized. This project required over 50 features from many datasets. To handle more features and reduce training time, powerful GPUs were used.

Network intrusion detection system (NIDS) employs an unsupervised stochastic auto-encoder neural net. An auto-encoder was used with GANS to generate it. In GANS, the generator and discriminator are two competing NN (neural networks). The goal of the competition is to use the Jensen-Shannon approach to solve the optimization problem. The generator attempts to construct phony data packets, whereas the discriminator determines whether or not the information is genuine. The suggested approach also includes a regularisation penalty for overfitting control behaviour. Others had lower detection rates than U2RL and R2L. [14] AE, CNN, two convolutional layers, and the soft - max classification output were used to illustrate multiple channel deep learning of Network Intrusion detection system features. The three datasets were examined: CICIDS, KDDcup99, UNSWNB15. The recommended model works well, but the assault's structure and attributes are not well defined.

To ensure that only qualified individuals are appointed to public positions, Saikat Bose et al. [15] propose using a cutting-edge data security approach. The first step of the approach included hiding sensitive data in the e-initial mail part of the commission's server. Private share pieces and the intervals in their hosting matrices are hash-operated into a circular orientation. Any digitally signed messages downloaded from the specified site may be verified using the same hash operations and public sharing. For each part of the electronic letter, on-the-spot fingerprints are concealed in two separate places using the same methods. To safeguard the copyright signature of the posting place, the fourth segment of each area is encrypted using a hash function. To guarantee that the whole of the certified electronic letter reaches its destination, the commission's server checks the legality of the appointment and the candidate's signatures. The better test results from more perspectives demonstrate

the superiority of the proposed process over the prior methods.

Incorporating machine learning over SDN enhances NIDS implementation. It uses machine learning for network monitoring on the SDN's central controller. Anomaly detection has been enhanced using tree-based ML (machine learning). A multiple class categorization job is done with only five characteristics by detecting an attack and classifying it.

Proposed System Design

A comprehensive explanation of the proposed system is shown in Figure 1, which may be seen below. In the past, several different methods that are still in use today to detect the potentially dangerous behaviour or incursion described in [15] and [6] were created. Yet, these kinds of systems continue to have problems, such as a high rate of false alarms and a low degree of accuracy in categorization. The primary function of the system is to collect information from the connected Twitter account by making use of the Twitter API and incorporating recently seen comments on Twitter. The failure of social networking programmes to recognise accounts that aren't regular or invaders is by far the most significant problem with these programmes. In our research, we combined natural language processing (NLP) and machine learning (ML) techniques to address the challenges posed by existing computer systems. Initially, we visit a number of different social networking websites in order to gather information. After being received, it is stored in the data sources where it originated as well as in the dataset files that were provided. It is possible that the data is unstructured at times due to the fact that it was collected from a variety of internet sources, such as Twitter. It is necessary to do this kind of pre-processing on the data using a certain sampling strategy and data filtering techniques. Both the bloom filter and the systematic sampling strategy were used to separate the data. The bloom filter was used to exclude instances that were incorrectly categorised. Electrical analysis need to be used for the purpose of providing sentence recognition and tokenization. Words that have been tokenized should be stored in a string array, since this makes string verification much simpler.. For the purpose of classifying the whole network, we made use of a variety of machine learning (ML) and deep learning strategies. These strategies included various machine learning techniques and Recurrent neural network. With the help of the proposed system, it is possible to identify potentially malicious entries during the system's run time.

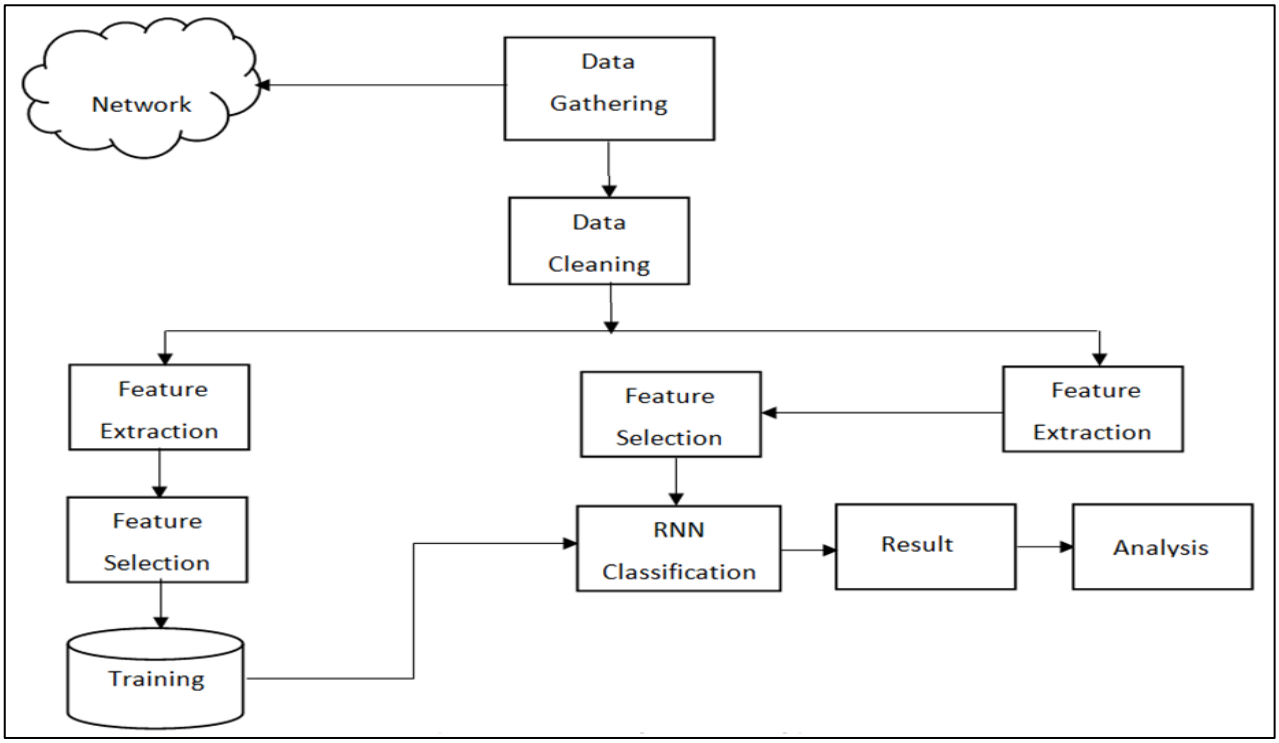

Fig. 1: System architecture for IDS using machine learning technique

Pre-processing: The data has been checked in accordance with the standards and guidelines that were established during the preprocessing stage. When one of these values exceeds or breaches the boundaries, the system immediately terminates the instance since it has violated the limits that have been set for it by the property's lower and upper limits for specified values. The phases of collecting data, aggregating them, cleaning them, filtering them, and normalising them are all included in the pre-processing phase.

Data Cleaning : While cleaning and mending inaccurate or false information from papers, records, or datasets, it is necessary to locate and change (or remove) any lost mistaken, incorrect, or nonsensical information. Moreover, it is necessary to replace, update, or remove any filthy or sensitive information. In order to cleanse the data in an interactive manner, they could employ scripting software or transaction processing. We employed consistent sampling strategies to achieve data balance, and we filtered the standardised dataset in order to exclude the occurrences of things that were wrongly categorised.

Extraction of feature: In the feature extraction phase, we pull out the typical and numerical values from the text data. Many techniques exist for extracting features like TF-IDF and Co-relation co-occurrence from larger datasets, as well as relational features and dependency features.

Feature Selection: After feature extraction is finished, feature selection is performed to optimize the feature set based on a handful of quality criteria. In order to fine-tune the features, the weighted term frequency approach was used, and the resulting data was sent on to the training component.

Classification: Finally, the system employs a supervised classification technique to ascertain whether or not a record represents an assault. In addition, the system displays a normal-time assault classification that has not been seen before.

We used a supervised classification technique (RNN and LSTM) in this study. The classifier is then trained using supervised machine learning. In this case, there is an initial presence of class-labeled data. Multiple decision trees are constructed using randomly picked features from the feature set, and the RNN's output is the majority output class of all the decision trees. This method is used to detect identity fraud on social media.

Algorithm

System testing algorithm

Input: Train dataset TrainingDB-Lits [], Test dataset TestingDB-Lits [] and Threshold Th.

Output: Whose weight is heavier than Th is determined by Reslt-set <cls name, SimWt>

Step 1: As shown in the following equation, it works in a convolutional layer with invader h training and test data for each testing data.

$$\text{testingFeature(k)} = \sum_{m=1}^{n} (.\ \text{featSet[A[i] … … . . A[n]} \leftarrow \text{TestingDBLits})$$

Step 2: By using below code, generate a feature vector from testing Feature (m).

$$\text{Extract\_FeatSet\_x [t…..…n]} = \sum_{x=1}^{n}(t) \leftarrow \text{testingFeature (k)}$$

The net convolutional layer receives Extract FeatSet x[t] from each pooling layer. Each layer stores every instance's characteristics from the testing dataset.

Step 3: For each train instance, use the function below,

$$\text{trainingFeature(l)} = \sum_{m=1}^{n} (.\ \text{featSet[A[i] … … . . A[n]} \leftarrow \text{TrainingDBList})$$

Step 4: Generate new feature vector from trainFeature(m) using below function

$$\text{Extract\_FeatSet\_Y[t……n]} = \sum_{x=1}^{n}(t) \leftarrow \text{TrainingFeat (l)}$$

The net convolutional layer receives Extract FeatSet Y[t] from each pooling layer. Each layer contains derived features from each training dataset occurrence.

Step 5: In dense layer, calculate all testing records using train dataset.

$$\text{weight wt} = \text{calculateSim (FeatSetx} \,||\, \sum_{i=1}^{n} \text{FeatSety[y])}$$

Step 6: Return Weight wt

3. Results and discussion

The prediction performance of our suggested RNN algorithm is shown to be superior to that of all other algorithms evaluated in a separate study using the same evaluation measures. On top of that, due to limitations in data completeness and quantity, not all available datasets accurately represent the true extent of software problems and new defects. The proposed fix may be put through further testing when implemented in working software systems. Ten-, fifteen-, and twenty-fold cross-validation are all possible using the three-split data technique.

Table 1: Extracted source code descriptions from all dataset files

Total Size	100000
Training Samples	70000
Testing Samples	30000

The system provides four comparisons between the findings of this study and those of other systems that have been computed using the same or comparable datasets. In the following Figure 2 we see a contrast between the state-of-the-art and more traditional machine learning (ML) techniques.

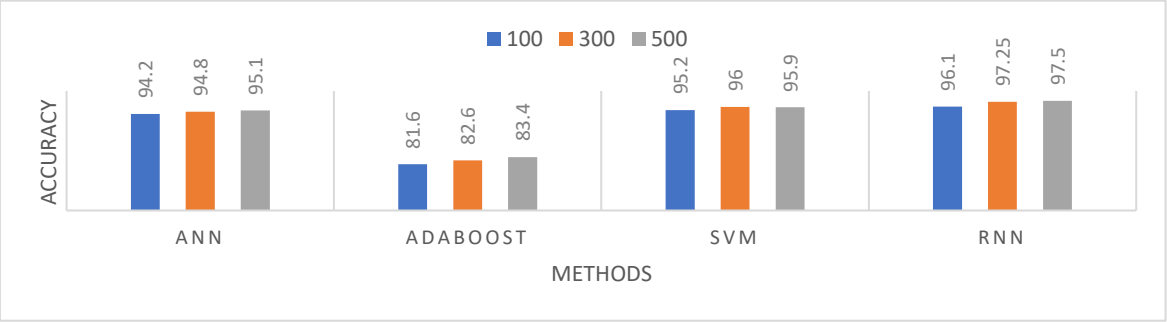

Fig. 2: Comparative analysis of proposed vs existing classification for attack detection

For attack detection, Figure 2 compares the accuracy of the proposed methods with that of two state-of-the-art machine learning techniques. This graphic demonstrates the superior accuracy achieved by the proposed RNN compared to that of machine learning techniques.

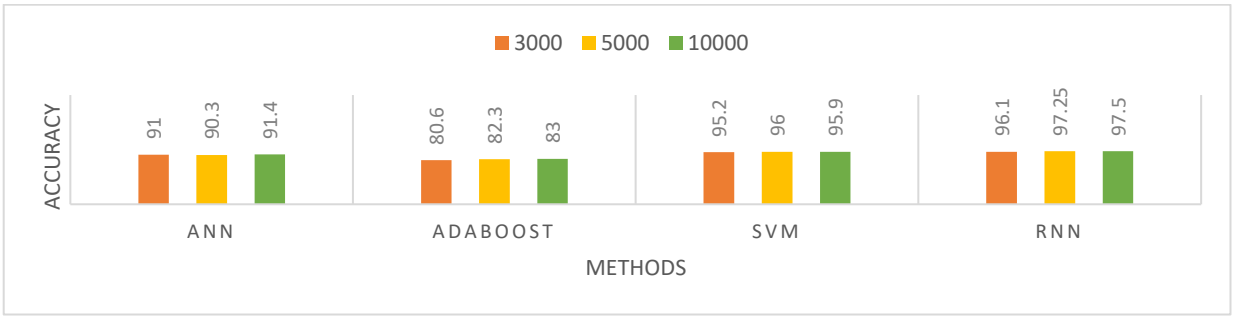

Fig. 3: Comparative analysis of proposed vs existing classification for vulnerability detection

Figure 3 compares the classification accuracy of the proposed techniques for vulnerability identification with that of two current machine learning algorithms. Evidence shows that the suggested RNN outperforms state-of-the-art machine learning methods in terms of detection accuracy.

4. Conclusion

After finishing experimental analysis, we can infer that different detection methods and soft computing and classification methodologies may identify various threats. Anomaly detection using signatures and the construction of multiple rules for both test and train were explored, using the KDD- Cup dataset used. Each stage represents the highest possible precision in attack detection, although none of them are dedicated to pinpointing the source of an unanticipated attack or misuse. The effectiveness of deep learning in assessing network security has inspired the development of a novel use case. As an added bonus, the technology has finished a detailed and exhaustive investigation of data safety. It's worth noting that traditional machine learning approaches for network security become less efficient as the volume of processed data increases. On the other side, deep learning methods have significantly changed how cyber threats are evaluated. The system uses a wide range of methods, such as vulnerability scanning and flow characterization, to identify network anomalies. Nonetheless, the system has constraints, such as the reliability of the input and output data. The need for faster and more relevant data evaluation has led to the rise in popularity of cutting-edge deep learning

algorithms. Machine learning (ML) methods and deep learning categorization using recurrent neural networks are used in the suggested implementation. As compared to more traditional machine learning (ML) techniques like support vector machines, random forests, naive bayes, and J48, the accuracy provided by deep learning algorithms is much higher. This system can detect both active and passive network threats from the outside world. Our 15-fold cross-validation on the KDDCUP99 and NSLKDD datasets shows that the proposed model achieves a maximum accuracy of 96.00% when employing RNN.

References

[1] Ali, F.; Ali, A.; Imran, M.; Naqvi, R.A.; Siddiqi, M.H.; Kwak, K.S. Traffic accident detection and condition analysis based on social networking data. Accid. Anal. Prev. 2021, 151, 105973.

[2] Sarkar, S.K.; Roy, S.; Alsentzer, E.; McDermott, M.B.A.; Falck, F.; Bica, I.; Adams, G.; Pfohl, S.; Hyland, S.L. Machine Learning for Health (ML4H) 2020: Advancing Healthcare for All. Proc. Mach. Learn. Res. 2020, 136, 1–11.

[3] Ali, F.; El-Sappagh, S.; Islam, S.R.; Kwak, D.; Ali, A.; Imran, M.; Kwak, K.S. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf. Fusion 2020, 63, 208–222.

[4] Modi, C.; Patel, D.; Borisaniya, B.; Patel, H.; Patel, A.; Rajarajan, M. A survey of intrusion detection techniques in Cloud. J. Netw. Comput. Appl. 2013, 36, 42–57.

[5] Wang, P.; Chao, K.; Lin, H.; Lin, W.; Lo, C. An Efficient Flow Control Approach for SDN-Based Network Threat Detection and Migration Using Support Vector Machine. In Proceedings of the 2016 IEEE 13th International Conference on e-Business Engineering (ICEBE), Macau, China, 4–6 November 2016; pp. 56–63.

[6] Ikram, S.T.; Cherukuri, A.K. Improving Accuracy of Intrusion Detection Model Using PCA and optimized SVM. J. Comput. Inf. Technol. 2016, 24, 133–148.

[7] Zolotukhin, M.; Hämäläinen, T.; Kokkonen, T.; Niemelä, A.; Siltanen, J. Data Mining Approach for Detection of DDoS Attacks Utilizing SSL/TLS Protocol. In Internet of Things, Smart Spaces, and Next Generation Networks and Systems; Balandin, S., Andreev, S., Koucheryavy, Y., Eds.; Springer: Cham, Switzerland, 2015; pp. 274–285.

[8] Mehr, S.Y.; Ramamurthy, B. An SVM Based DDoS Attack Detection Method for Ryu SDN Controller. In Proceedings of the 15th International Conference on Emerging Networking EXperiments and Technologies, Orlando, FL, USA, 9–12 December 2019; pp. 72–73.

[9] Dey, S.K.; Rahman, M.M. Effects of Machine Learning Approach in Flow-Based Anomaly Detection on Software-Defined Networking. Symmetry 2020, 12, 7.

[10] Khan, F.A.; Gumaei, A.; Derhab, A.; Hussain, A. A Novel Two-Stage Deep Learning Model for Efficient Network Intrusion Detection. IEEE Access 2019, 7, 30373–30385.

[11] Malaiya, R.K.; Kwon, D.; Suh, S.C.; Kim, H.; Kim, I.; Kim, J. An Empirical Evaluation of Deep Learning for Network Anomaly Detection. IEEE Access 2019, 7, 140806–140817.

[12] Yang Jia, M.W.; Wang, Y. Network intrusion detection algorithm based on deep neural network. IET Inf. Secur. 2019, 13, 48–53.

[13] Yang, Y.; Zheng, K.; Wu, B.; Yang, Y.; Wang, X. Network Intrusion Detection Based on Supervised Adversarial Variational Auto-Encoder With Regularization. IEEE Access 2020, 8, 42169–42184.

[14] Andresini, G.; Appice, A.; Mauro, N.D.; Loglisci, C.; Malerba, D. Multi-Channel Deep Feature Learning for Intrusion Detection. IEEE Access 2020, 8, 53346–53359.

[15] Saikat Bose, Tripti Arjariya, Anirban Goswami, Soumit Chowdhury Multi-Layer Digital Validation of Candidate Service Appointment with Digital Signature and Bio-Metric Authentication Approach International Journal of Computer Networks & Communications (IJCNC) Vol.14, No.5, September 2022 DOI:10.5121/ijcnc.2022.14506