International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 10, Issue 9, September 2021 **ISSN 2319 - 4847**

Performance Comparison of Various Supervised Machine Learning Algorithms Used in Detecting Network-Based Intrusions

Sunaina S P[1], Dr. Nagarathna N[2]

[1, 2] Department of Computer Science & Engineering, B. M. S. College of Engineering, Bengaluru, India

ABSTRACT

Machine Learning methods are most generally and effectively used to build up an intrusion detection system (IDS) to detect and distinguish at both network- and host-level in a mechanized and convenient way. Nonetheless, numerous difficulties emerge since security or cyber-attacks are consistently changing and are taking place in extremely enormous volumes that needs a versatile arrangement. Many datasets, for malicious attacks, are freely accessible for additional analysis by cyber safety organization. Attack datasets accessible freely needs to be upgraded on a methodical and standard basis because of the vital role of security attacks and that the attacking techniques are constantly changing. A proper study is necessary to learn the nature and performance of the various techniques used to develop a smart detection system for detecting and distinguishing numerous malicious attacks. Therefore, an effective comparison is made, providing a brief overview on various machine learning and deep learning algorithms used in developing a malleable and functional IDS that identifies and classifies unanticipated and unsought malicious attacks. This comparative analysis helps in finding the most suitable and finest algorithm that performs well in discovering upcoming security attacks.

Keywords: Cyberattacks, Malware, Machine Learning, Deep Neural Network, Intruders, Dataset, Network Security

1. INTRODUCTION

Information and Communication Technology (ICT) Organizations and Networks hold several delicate client information which are inclined by different attacks from inside as well as outside interlopers. Such malware might occur through automated mechanism or created manually. They are distinct in nature and are progressively confusing which lead to unrevealed information probe. For example, Bitcoin probe led to a loss of $0.7B, whereas Yahoo information probe brought about an approximation of $0.35B loss. Malicious attacks, of this kind, continuously developing with exceptionally modern techniques with growing tools and technologies, and network topologies comprising the latest advancements in the field of IoT (Internet of Things).

Security attacks constitute genuine safety issues that request the need for a novel, adaptable and much solid Intrusion Detection System (IDS). An Intrusion Detection System is a framework or mechanism employed to identify and distinguish intrusions, attacks, assaults, or violations of the safety schemes automatically at both network-level as well as host-level systems on a well-timed basis. On the basis of attack nature, detection of attacks is categorized as either network-based intrusion detection system (NIDS) or host-based intrusion detection system (HIDS).

The framework is called Network-based Intrusion Detection System (NIDS) if it is based on network behaviors. These network behaviors are gathered utilizing network tool through mirroring by networking devices, like, switches, routers, and network taps which are examined to detect attacks and potential dangers hidden inside the network traffic. A framework that utilizes system exercises in terms of several log data processing on the local host device or machine to identify malware is called as Host-based Intrusion Detection System (HIDS). Local sensors are being used to fetch such log data. The NIDS assesses every packet data relied within network traffic streams whereas, HIDS depends on the data of log documents that may comprise of sensors logs, system logs, software logs, file systems, disk assets, client account details and few other things of each system. Several institutions employ the combination of both HIDS and NIDS.

2. RELATED WORKS

Every information technology organization hold delicate client data that are vulnerable to several attacks by attackers from both inside and outside environment [4]. There exist innumerable analyses regarding the security problems associated with IDS since the introduction of computer systems. This survey provides an overview of biggest analysis to date which discovers the area of machine learning and deep learning techniques used to improve NIDS and HIDS.

2.1 NETWORK-BASED INTRUSION DETECTION SYSTEMS (NIDS)

The NIDS, for business or economical needs, utilizes either factual quantities or calculative outset values on characteristic sets like length of packet data, time of inter-arrival, size of data flow and various network data parameters to productively framework them all within a particular time interval [5].

One of the powerful techniques to manage the current type of attacks is Self-learning method. This comprises of basic machine learning classifiers such as supervised, semi-supervised and unsupervised learning for studying the usual and suspicious activity patterns with a wide-varied collection of usual (normal) and attack (violence) network and host-based events [6]. Deep Learning has accomplished remarkable outcomes in well-established Artificial Intelligence (AI) assessments in the area of Image processing, speech recognition, natural language processing (NLP) and much more [7]. Furthermore, these accomplishments have been changed to different network safety activities like cyber-attack recognition, categorization of android intrusions, traffic evaluation, prediction of network traffic, ransomware detection, classification of ciphered text, identifying URL attacks, anomaly detection [1].

In the paper [8], the categorization algorithm incorporates two stages. One is to anticipate the existence of the class called P-rules stage and other to anticipate the class that is not present, called as N-rules stage. This functioned well compared to other related frameworks for KDDCup 99 dataset. In the paper [9], the importance of characteristic related research was examined fir IDS for the commonly utilized dataset i.e., KDDCup 99 dataset. For every element, they had the ability to indicate that character relevance with regards to procurement of information. The paper [10] employs random forest algorithm for misuse detection that studies the patterns of malicious activities, anomaly detection using deviation identification method, and hybrid detection comprising the combination of both the misuse and anomaly detection. Paper [11] discusses about very popular Bayesian networks for recognizing cyber-attacks using Naïve Bayesian networks having a root element to depict class label and leaf elements to depict characteristics of connected nodes. According to paper [3], performance of the detection system based on recurrent neural network (RNN) surpassed other traditional machine learning models in detecting attacks and attack type for the NSL-KDD dataset.

2.2 HOST-BASED INTRUSION DETECTION SYSTEMS (HIDS)

In general, this part is restricted to identifying security attacks and intrusions at host-level. Several programming tools, for example, Metasploit, Sqlmap, Nmap, Browser exploitation give the essential architecture to analyze and assemble data from target framework weaknesses. A Few approaches have certain restrictions and intruders could attain illegal access to those systems. This is addressed by considering a classic Host-based intrusion detection system that examines and keeps track of all the activities of network traffic on the local system files, system calls and operating system [2].

As a system application interacts with the operating system through system calls, their characteristic, sequence, length and type produces a distinct trail. This helps in classifying the familiar and unfamiliar applications. It is obvious that the behavior of normal system calls is entirely different from that of the attack system calls. Therefore, evaluation of such system calls gives consequential details about the system processes. Many characteristic extraction methods are utilized for distinguishing the processes of a system on the basis of system calls. Some of those methods are N-gram which is described in the paper [12].

In the paper [13], a characteristic mechanism based model, known as N-gram is employed to analyze the information of system calls of ADFA-LD dataset and then, its characters are fed to various traditional ML algorithms like SVM and HMM for detecting and classifying assaults. In the paper [14], the reduction of the size of the system calls were done by analyzing k-means as well as k-nearest neighbors algorithm on the basis of frequency model. Paper [15] studies about the maintaining the zero-day malware and suspicious assaults in windows operating system using machine learning techniques based on characteristic mechanism models.

3. PROPOSED ARCHITECTURE

The main motive of the proposed work is to depict our system architecture and employment of state-of-the-art tools and technologies in the evaluation.

Fig. 3 illustrates the proposed system design. Here, two different kinds of supervised learning methods are used to evaluate the performance of both kinds. They are various traditional machine learning classifiers such as Support Vector Machine, Random Forest, AdaBoost, Decision Tree, K-Nearest Neighbor, Linear Regression and Naïve Bayes Algorithm, as well as artificial neural network with dense hidden layers.

In this work, one of the publically available datasets, called as KDDCup99 database is used, which contains numerous network traffic data and information about several attacks occurred over these data. This dataset contains 41 features and 5 classes. After collecting data, the next step is to preprocess the dataset to improve the performance of the model.

Preprocessing of data may include data augmentation, data normalization, and resizing. The preprocessed data will be split into train and test sets with some aspect ratios like 7:3, 8:2, 6:4, etc. However, it should be noted that the train set should be in larger number than the test dataset. This is because the systems need sufficient data during training for it to learn the features and behaviors of the seen or labelled dataset. While training, the model learns the features of the input. Thus, training model is obtained on which the predictions are made. Once trained model is generated, the new set of data can be passed to test model. This is also called as inferencing. The output of the system will be one of the two classes: Normal or attack.

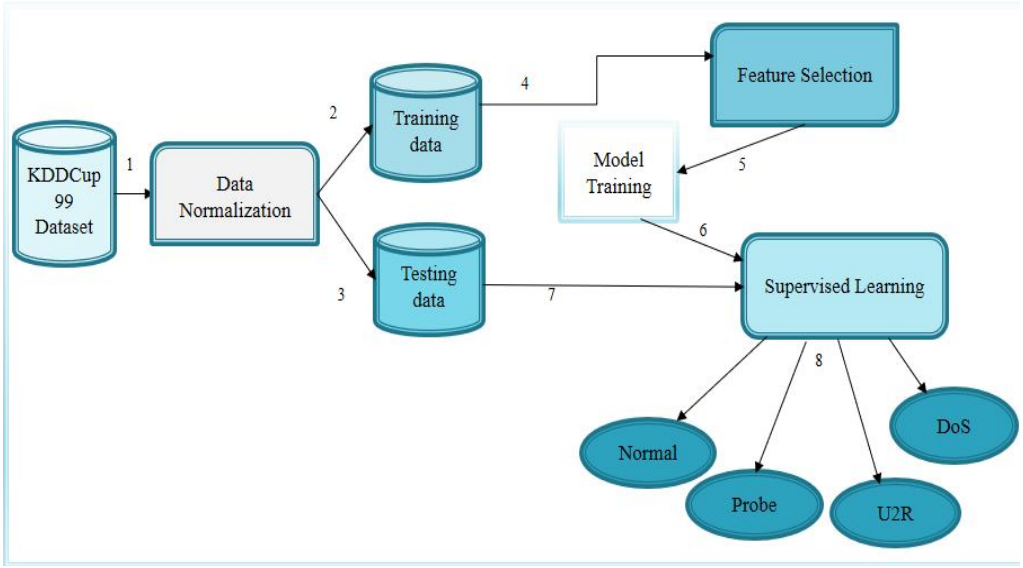

Fig 3. System Design

Source: Adapted from Vinayakumar R, M Alazab, Soman K P, Prabaharan P, Ameer Al-Nemrat & Sitalakshmi V. (2019). Deep Learning Approach for Intelligent Intrusion Detection System. IEEE Access. [1]

4. IMPLEMENTATION

4.1 DATASET

For any classification and regression models, collection of data becomes the very first and important stage. In supervised learning technique, the collected data has to be annotated properly before feeding it to the designed system. The machine will be trained over the data and it extracts significant features from it and makes accurate predictions for test set based on the learning.

In our experiment, freely-accessible KDDCup 99 is considered to compare the performance of several machine learning models in the binary classification problem. KDDCup99 was created in the year 1998, by Lincoln Lab under DARPA and was first used for the Third International Knowledge Discovery and Data Mining Tools Competition, co-occurred with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The dataset possesses

network traffic details that stores an expansive range of attacks simulated inside a military network environment. The traffic features were extracted by processing tcpdump source data which comprises, in total, 41 features and, 5 Classes- Normal, Probe, U2R, DoS, R2L.

4.2 FEATURE SELECTION

In machine learning and deep learning concepts, the feature selection or extraction technique has a prominent part in minimizing the size of the input data and a wide-varied analysis has been put through for a valid feature selection technique. In feature selection method, there are two types of methods adopted in selecting features from the dataset. They are filter method and wrapper method. In filter method, the features are chosen based on their results obtained in several experimental analyses which evaluates the significance of each feature by its dependency with the correlated target value. Wrapper method is used to get a variant of the features by computing the efficacy of a feature variant with its correlated target attribute.

The variants of the features are chosen by considering the classification model that performs the best with the selected variant. The feature variant is found by performing various search methods such as random search (RS), depth first search (DFS), hybrid search (HS), or breadth first search (BFS). On the other hand, the filter method makes use of an attribute evaluator along with a ranker to list all the features of the input data based on their rankings.

4.3 SUPERVISED MACHINE LEARNING ALGORITHMS

Supervised Machine Learning, also called as, Supervised Learning, includes different techniques and algorithms used as predictive models that can determine the output classes based on the labelled data. Some of state-of-the-art machine learning algorithms employed in building our IDS for detecting network-based attacks are described as follows.

4.3.1 Support Vector Machine (SVM)

Support Vector Machine could be known as one of the popular machine learning algorithms implied for solving classification problem. Although, it can be used for both regression and classification challenges, SVM is best-known for binary-classification. It analyses annotated data provided after plotting them in an n-dimensional space where n is the number of classes considered. Then, a hyper-plane is drawn, for which, each data falls onto either side of the plane. Therefore, the prediction can be done easily by looking at data-point coordinates and their corresponding labels, concluding the classes to which they belong.

4.3.2 Random Forest (RF)

Random Forest Algorithm is a supervised machine learning technique where it combines multiple decision trees on several subsets of the dataset provided. This type of learning is also called as Ensemble Learning. This method will help in enhancing the performance of the model by increasing its predictive accuracy. The final output can be calculated by taking average of predicted values of each of the tree. This technique is a best example to show that, with combination of two or more classifiers, the model can attain higher accuracy.

4.3.3 Ada Boost (AB)

Adaptive Boosting, often abbreviated to AdaBoost, is a type of ensemble method that is used to build a stronger classifier by combining many weak classifiers. It is an end-to-end algorithm, where each of the weak classifiers are employed for different purposes at different stages of the classification process. For example, the model is built by training the data using one classifier, then second classifier will be used for checking and correcting the errors from training the first classifier. More and more classifiers will be combined such that the model attains the maximum accuracy. In other words, the weak classifiers will be added at each end of the process until the classifier predicts the trained data well-enough.

4.3.4 Decision Tree (DT)

Decision trees, as the name suggests, are a type of supervised learning method, graphically represented as a tree-like structure. It contains two kinds of nodes, they are Decision node and leaf node. The decision nodes are responsible for making decision or rules based on certain conditions. These rules form the branches. The leaf nodes are nothing but the output of those decisions and are said to be last-level nodes of the tree. The decisions are made based on the information of features present in the given dataset.

4.3.5 K-Nearest Neighbor (KNN)

The k-nearest neighbor is one of the supervised learning methods that is found to be easier and simpler to implement compared to all other traditional machine learning algorithms. It is used for both classification and regression problems. KNN classifies new data into one of the existing categories by comparing similarities between new data and the exiting one. This is why KNN is referred as non-parametric algorithm as it does not make any assumptions regarding classification of the provided data.

4.3.6 Linear Regression (LR)

Linear Regression is another classical machine learning algorithm based on supervised learning method that is used for regression task. In this algorithm, a target variable (y) will be set, for which multiple predictor variables (x) are directly proportional to it. It is called the linear regression as it represents a liner connection between the dependent variable (y) and one or more independent variables (x). Since the algorithm shows linear relationship between x and y, a graph can be plotted that provides a slopped straight line based on how the value of dependent variable changes with that of the independent variable.

4.3.7 Naïve Bayes (NB)

Naïve Bayes algorithm is a combination of different machine learning algorithms which is used as classification technique for both binary as well as multi-classification problems based on Bayes theorem. The theorem states that "a hypothesis (h) may be the class to assign for a new data instance (d)". This can be briefed out as an assumption made bout the independence among predictor variables.

4.3.8 Deep Neural Network (DNN)

An artificial neural network with dense hidden layers, known as Deep Neural Network, is implied as the computational model to compare its performance with classical machine learning algorithms. In which, Feed Forward Neural network (FFN) is employed to transmit data from one layer to another. The depth of DNN or number of hidden layers to be used is chosen by performing Hyper Parameter Selection technique. A Multilayer Perceptron (MLP), one of the FFNs, is used that contains multiple layers including an input layer at the initial stage to feed data to the network.

Fig 4.3.8. shows the architectural diagram of deep neural network having one input layer, one output layer, and multiple hidden layers where the information of the data is transformed from one layer to the other in a forward flow. The neurons, also called as nodes, are fully-connected to one another, at each layer.

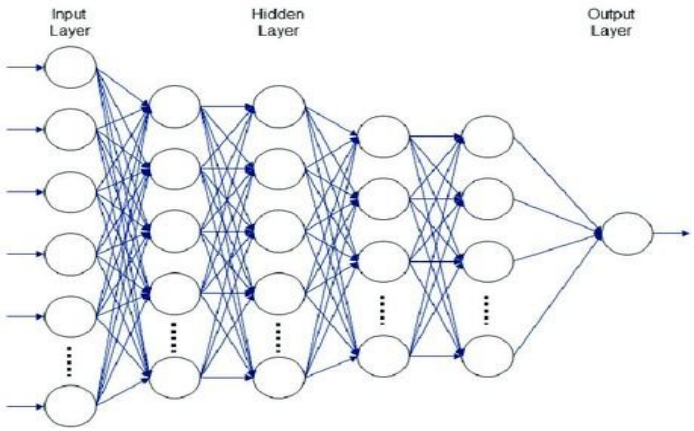

Fig 4.3.8. Architecture of Deep Neural Network

Source: Adapted from Hyun Kim, Kun Yeun Han. (2020). Urban Flood Prediction Using Deep Neural Network with Data Augmentation. Water 2020, 12, 899. [2]

For a multi-class problem, the network employs Softmax as a non-linear activation function. Whereas, for a binary classification problem, we use ReLU activation function to enhance the learning of the model in depth. Among many linear and non-linear activation functions, Rectified Linear Units (ReLU) has given great results in terms of accuracy. It is even capable of increasing the speed and performance of the model at training stage. ReLU had been a buzzword ever since it was found to have reduced the gradient vanish problem. Compared to many traditional non-linear function like sigmoid, softmax and tangent function, it seen that ReLU is more useful and best in training large amount of data. It is

most-suitable approach towards training and testing huge data as it takes less computational time and cost. Mathematically, ReLU is termed as below:

$$f(x) = \max(0, x)$$

5. EXPERIMENTAL ANALYSIS AND RESULTS

As discussed earlier, the proposed architecture is only designed for network-based intrusion detections and for which, it collects network-traffic data in real-time. The output of our proposed system is classification of requested data into either normal or attack signature, if any.

5.1 ATTRIBUTE MEASURES

The resultant four attributes values obtained for various traditional machine learning classifiers and DNN model over our KDDCup99 dataset is given in Table 1.

Table 1: Binary Classification test results of different traditional machine learning algorithms for KDDCup 99 Dataset.

Algorithm	Accuracy	Precision	Recall	F1-Score
DNN-1	0.929	0.998	0.915	0.954
DNN-2	0.929	0.998	0.914	0.954
DNN-3	0.930	0.997	0.915	0.955
DNN-4	0.929	0.999	0.913	0.954
DNN-5	0.927	0.998	0.911	0.953
SVM	0.811	0.994	0.770	0.868
Random Forest	0.927	0.999	0.910	0.953
Naïve Bayes	0.929	0.988	0.923	0.955
AdaBoost	0.925	0.995	0.911	0.951
Decision Tree	0.928	0.999	0.912	0.953
KNN	0.929	0.998	0.913	0.954
Linear Regression	0.848	0.989	0.821	0.897

With our test analysis, it is found that the DNN model as well as the different machine learning classifiers perform better when more number of minimal feature sets are used.

5.2 PERFORMANCE EVALUATION

Comprehensive representation of test results obtained for each of the algorithms, namely, AdaBoost, Decision Tree, K-Nearest Neighbour, Linear Regression, Naïve Bayes, Random Forest, Support Vector Machine as well as the Deep Neural Network model with utmost 5 layers, are plotted as bar graphs, in comparison with mAP attributes such as Accuracy, precision, recall and F1-score, in figures a, b, c and d, respectively. By looking at the Map values generated for the test dataset, most of the algorithms perform better with respect to one or more attributes, and the same have poor performance with respect to other attributes when compared with rest of the classifiers.

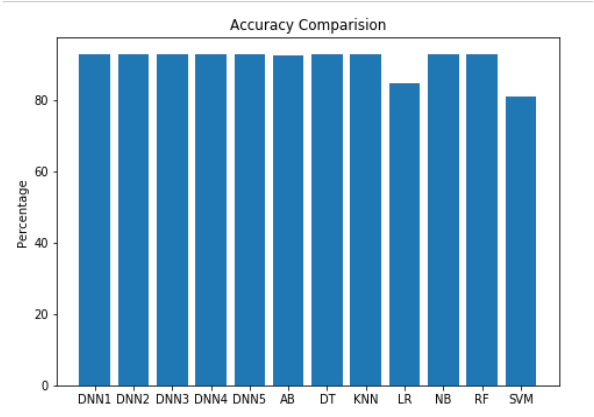

Fig 5.2.1. Accuracy Evaluation of Algorithms Employed.

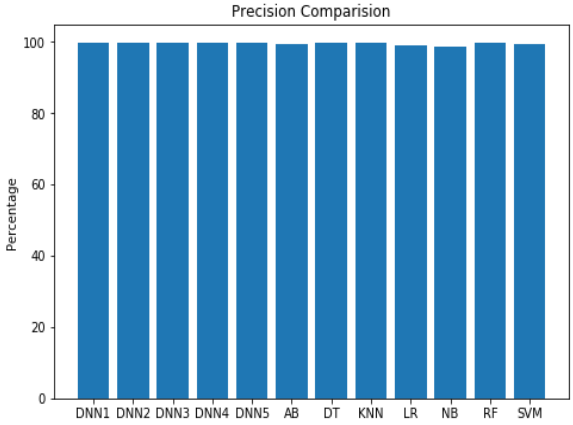

Fig 5.2.2. Precision Evaluation of Algorithms Employed.

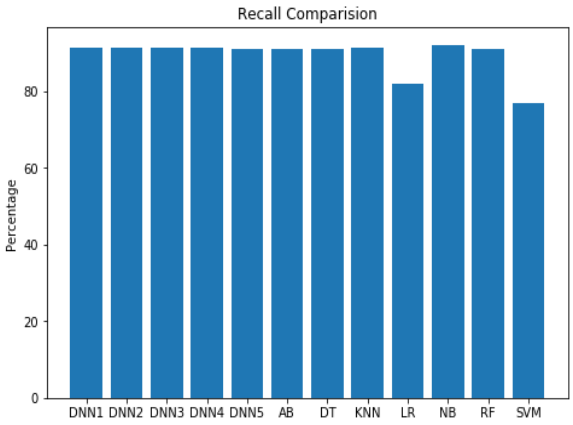

Fig 5.2.3. Recall Evaluation of Algorithms Employed.

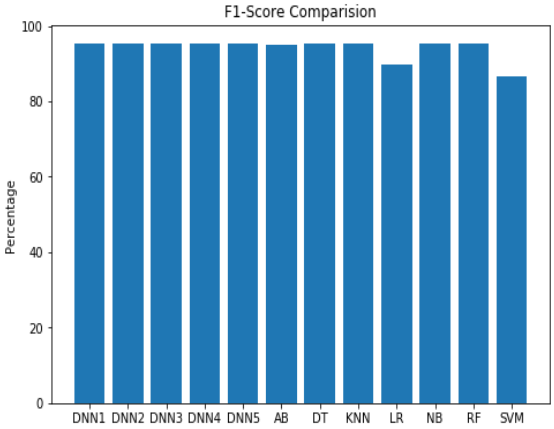

Fig 5.2.4. F1-Score Evaluation of Algorithms Employed.

With reference to Table 1, algorithms such as DT, NB, KNN and DNN have better accuracy in detecting the attacks when compared with other models. Likewise, SVM and RF models' performance values are precise enough than the rest of the algorithms. However, DNN has better performance in terms of accuracy, precision, recall as well as F1-score. Thus making it the mostly preferable method for binary as well as the multi-classification applications.

Moreover, another reason why DNN is best-suited for classification and regression problem is, it transmits the information present in the dataset from one hidden layer to the next, so that more and more important features are extracted from each layer. This will enable the system to learn all the features of the network data and help the same in detecting and classifying the requested data into either normal or attack signature. If the data contains any attack signature, then that attack will be categorized into its predicted label.

6. CONCLUSION AND FUTURE SCOPE

In this work, we have presented different ways to identify cyber-attacks using traditional machine learning classifiers as well as feature selection techniques to build an ideal intrusion detection system. Our proposed framework is able to examine the activities at network-level and classifying attacks/intrusions, if any, by employing the suitable model. The experimental analysis and results show that the IDS developed using DNN and wrapper feature selection technique performs better compared to all other methods in terms of detecting the traffic data with the highest accuracy rate of 93.02%.

We suppose that the generated results shall be useful in future study for developing a detection model which may identify unknown intrusions along with the detection of known intrusions. This is because, none of the existing approaches is capable of detecting novel attacks. Detection and classification of new intrusions is still a topic to be

researched on, since every IDS exists today gives high false positive rate. Additionally, the proposed model can be improved more by building a system the surveils the events occurring at network-level such as DNS and BGP activities.

References

[1] Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., ... & Wang, C., "Machine Learning and Deep Learning Methods for Cybersecurity." IEEE Access, 2018.

[2] Saracino, A., Sgandurra, D., Dini, G., & Martinelli, F., "Madam: Effective and efficient behavior-based android malware detection and prevention." IEEE Transactions on Dependable and Secure Computing, 15(1), 83-97, 2018.

[3] Yin, C., Zhu, Y., Fei, J., & He, X., "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks." IEEE Access, 5, 21954-21961, 2017.

[4] Mukherjee, B., Heberlein, L. T., & Levitt, K. N., "Network intrusion detection." IEEE network, 8(3), 26-41, 1994.

[5] Azab, A., Alazab, M. & Aiash, M., "Machine Learning Based Botnet Identification Traffic." The 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (Trustcom 2016), Tianjin, China, 23-26 August, pp. 1788-1794, 2016.

[6] Staudemeyer, R. C., "Applying long short-term memory recurrent neural networks to intrusion detection. South African Computer Journal," 56(1), 136-154, 2015.

[7] Lecun, Y., Bengio, Y., & Hinton, G., "Deep Learning." Nature, 521(7553), 436, 2015.

[8] R. Agarwal, and M. V. Joshi, PNrule: A New Framework for Learning Classifier Models in Data Mining, Technical Report TR 00-015, Department of Computer Science, University of Minnesota, 2000.

[9] H. Kayacik, A.N. Zincir-Heywood, and M.I. Heywood, "Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets," Proceedings of the third annual conference on privacy, security and trust 2005, PST 2005, DBLP.

[10] Zhang, Jiong, Mohammad Zulkernine, and Anwar Haque. "Randomforests-based network intrusion detection systems." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38, no. 5: 649-659, 2015.

[11] Amor, N. Ben, Salem Benferhat, and Zied Elouedi. "Naive bayesian networks in intrusion detection systems." In Proc. Workshop on Probabilistic Graphical Models for Classification, 14th European Conference on Machine Learning (ECML) and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), 23rd September, in CavtatDubrovnik, Croatia, p. 11. 2003.

[12] Hofmeyr, S. A., Forrest, S., & Somayaji, A., "Intrusion detection using sequences of system calls." Journal of computer security, 6(3), 151180, 1998.

[13] Aghaei, E., & Serpen, G., "Ensemble classifier for misuse detection using N-gram feature vectors through operating system call traces." International Journal of Hybrid Intelligent Systems, (Preprint), 1-14, 2017.

[14] Xie, M., Hu, J., Yu, X., & Chang, E., "Evaluating hostbased anomaly detection systems: Application of the frequency-based algorithms to adfa-ld." In International Conference on Network and System Security (pp. 542-549). (2014, October). Springer, Cham.

[15] Haider, W., Creech, G., Xie, Y., & Hu, J., "Windows based data sets for evaluation of robustness of host based intrusion detection systems (ids) to zero-day and stealth attacks." Future Internet, 8(3), 29, 2016.

AUTHORS

Sunaina S P, Master of Technology, Department of Computer Science and Engineering, B. M. S. College of Engineering, Bengaluru, Karnataka – 560 019.

Dr. Nagarathna N, Professor, Department of Computer Science and Engineering, B. M. S. College of Engineering, Bengaluru, Karnataka – 560 019.