EVALUATION OF CYBERSECURITY DATASET CHARACTERISTICS FOR MACHINE LEARNING-BASED DETECTION OF CYBERSECURITY ANOMALIES

Mr.Srinu Jatothu[1]., V.Divyanjali[2]

1 Assistant Professor, Department of ECE, Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India

2, B.Tech ECE (19RG1A0456),

Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India

ABSTRACT

Artificial Intelligence algorithms play an important role in network security and attack detection, and in some cases can provide better results than classical detection tools such as Snort or Suricata. In this sen se, the aim of this study is to evaluate the features of various mature machine learning algorithms freque ntly used in IDS scenarios. To do this, first consider splitting the data security network configuration an d separating its data into different groups. Using this classification, this task is to determine which neura l network model (multilayer or recurrent), activation function, and learning algorithm produces more res ults based on the dataset output. Finally, the results are used to determine which data in the network sec urity dataset are more relevant and representative for access detection, as well as the optimal machine le arning algorithm configuration to reduce physical computation. Keywords network security, data analys is, data, machine learning, neural networks, intrusionSearch.

Introduction

The sophistication of new computers and the development of new technologies are leading to advances i n the use of artificial intelligence (AI) and technologies in computer security. In particular, AI is more li kely to discover software issues or conflicts and interventions, creating new models that will support bet ter and more powerful decisions [1]. Among other things, this service allows human interaction to focus on more abstract tasks, such as general maintenance of the system or error detection, for example very bad. Additionally, for IT security, technology can help employees manage and analyze the big data that new data will create. . These systems process large amounts of data that need to be analyzed quickly wh ile generating various types of alerts. AlsoThe Editor responsible for reviewing and approving this articl e for publication is Luis Javier Garcia Villalba. > Committed to developing new, more effective IDS, art ificial intelligence is used as the basis for the implementation of IDS, using machine learning technolog y to classify patterns from clear and ambiguous patterns [2]. These techniques are welluited to incorpora ting and processing new information and require intelligent algorithms, such as machine learning algorit hms, to extract the content. Its implementation, especially in IDS, requires great responsibility in terms of choosing the best method and identifying potential attacks. This issue is important because there are many features in the data that can lead to overfitting of the model and lead to negative consequences of valid data [3]. Focus on research of computational models based on neural networks.

TABLE 1. Activation functions.

Function	Equation
Linear rectifier	$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$
Sigmoid	$f(x) = \frac{1}{1 + e^{-x}}$
Softsign	$f(x) = \frac{x}{1 + \lvert x \rvert}$
Hyperbolic tangent	$f(x) = \frac{2}{1 + e^{-2x}} - 1$

Private information in the context of information security. So, considering that we will have specific dat a depending on the study scenario, this study aims to show that neural network methods are useful in fin ding the research methods that give the best answer. This work focuses on the study and comparison of multilayered and recurrent neural networks, specifically focusing on information related to temporal eve nts. Finally, the novelty of this work is that it examines the behavior of different configurations of neura l networks (multilayered and recurrent) based on the proposed classification of cybersecurity datasets. I n this way, it is possible to determine which neural network configuration gives the best results in terms of accuracy for each data set. First, Chapter 2 introduces the theoretical concepts of neural networks that are of interest to this work, such as multilayer networks and recurrent networks. Chapter 3 describes the most important projects selected to use specific types of neural networks to identify cyber attacks. Chap ters 4 and 5 present the research and rationale for the selection of materials used in this study. Finally, S ections VI and VII present the results and conclusions of this study. Neural Network

Artificial neural network is a system created from simple computers called neurons, which resembles th e behavior of neurons in the brain. These neurons are interconnected by connections that control the fun ctional state of neighboring neurons. Table 1 describes the most common activities. br>
A. Neural Network Research

A neural network is a network, often called a feedforward network, in which the signal propagates in on e direction (from input to output) and there is no loop; This means that the output layer here is.

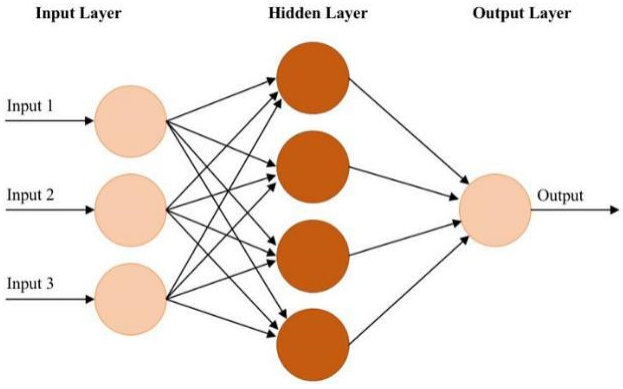

FIGURE 1. Basic structure of a multilayer neural network.

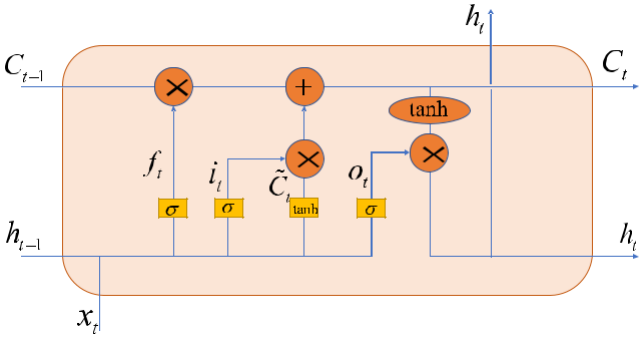

FIGURE 2. Basic scheme of an LSTM unit.

The data used for the problem discussed in this article should include information about the differences between connections in the network and the connection label indicating whether the connection is disru ptive and its type or normal. Any object needs to be recorded and classified as the algorithm used for det ection will use a learning process. stability [41] [44] and is considered evidence [45] [40]. The selection of this data is supported by several factors: the effectiveness of the attack, the recording of the attack, an d the distribution of data, similar to the instructions in the past section. 47 of these are related to the pro ducts in the profile; The last two points concern the type of attack and the behavior in the profile (norma l or strict). This information is availabl

TABLE 2. Classification of Data set UNSW-NB15 features.

Group	Feature
Basic characteristics	srcip, sport, dstip, dsport, proto, state, dur, sbytes, dbytes, sttl, dttl, sloss, dloss, service, sload, dload, spkts, dpakts
Content characteristics	swin, dwin, stcpb, dtcpb, smeansz, dmeansz, trans_depth, res_bdy_len
Traffic characteristics based on time	sjit, djit, stime, ltime, intpkt, dintpkt, tcprtt, synack, ackdat

To implement the neural network, Python was used as the programming language and the TensorFlow li brary, an open source library developed by the Google Brain team, was used. This library provides all th e tools needed to design, train and evaluate the performance of neural networks. Learn algorithms and di fferent groups. The activation and optimization of the neuron were varied in different experiments. Onc e defined, results regarding the accuracy of different tasks can be obtained. Based on this important fact, optimization was chosen and various tests were performed using the optimizer. To check the accuracy, we compare the measured data of the corresponding predictions with the actual results of these texts to obtain all predictions predicted by the algorithm and obtain the correct percentage. If necessary, the valu e focuses on the measurement error of the test of the added estimator, and the crossentropy of an expone ntial function contains the true value of the label. Once this error is achieved, it is averaged and takes a r educed value in the next training session. Finally, the weights of the neural network are initialized with values. Multi-Layer Neural Network Analysis

The multilayer neural network structure consists of three interconnected layers: input layer, hidden layer

and output layer. The distribution of neurons in each layer of this network follows the rules defined in [48] and is described in Table 3.

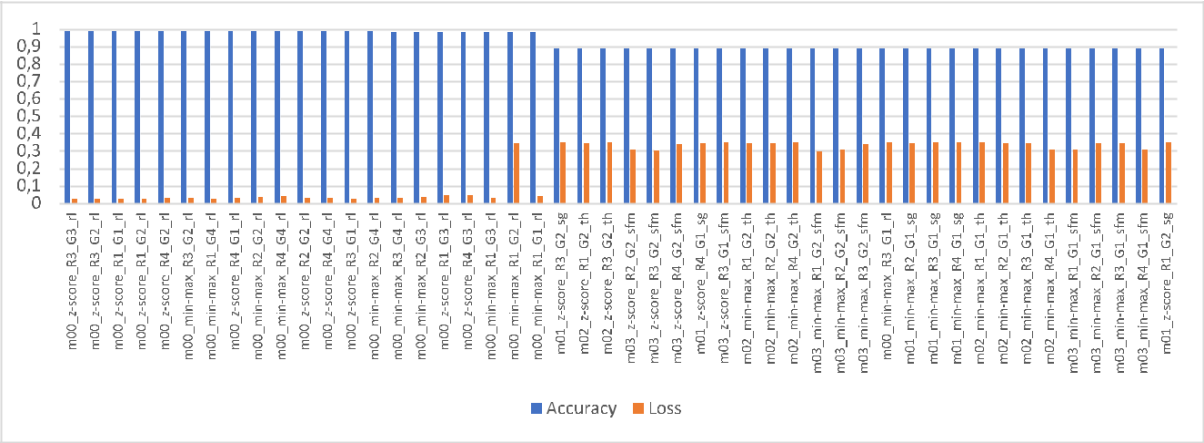

FIGURE 3. Feedforward neural network results in terms of loss and accuracy. The vertical axis shows a normalized value of accuracy and loss performance; while horizontal is to point out the most representative feed forward neural networks configurations of the proposed following the next pattern: <configuration code>_<normalization function>_<rule selected for the number of nodes in the hidden layer>_<group of characteristics proposed>_<activation function>.

TABLE 3. Rules of Calculation of Nodes in Hidden layers.

Rule code	Method
R_1	$H = 0.75\ x\ Input + Output$
R_2	$H = \frac{(Input + Outpt)}{2}$
R_3	$H = 0.70\ x\ Input$
R_4	$H = 0.90\ x\ Input$

TABLE 4. Testing performed for each group of characteristics using the multilayer neural network

TABLE 8. Comparison between multiples researches approaches.

Algorithm	Number of features	Accuracy
Decision Trees [45]	22	89.86%
Decision Trees [45]	13	89.76
Decision Trees[50]	47	85.41
J48 Classifier [42] (Worms attacks)	25	99.94
Our proposed FFNN	19	98.8%
Our proposed RNN	19	98%

Configuration Code	Rule	Activation Function	Optimizer
m00	R_1,2,3,4	Linear rectifier	Adam
m01	R_1,2,3,4	Sigmoid	Adam
m02	R_1,2,3,4	Hyperbolic tangent	Adam
m03	R_1,2,3,4	Softsign	Adam
m04	Best rule	Best activation function	Gradient descent
		Best activation	

D. Correlation Between Data Sets

After descriptive experiments using multi
layer neural networks on different data sets, it was found that gambling results can be obtained using o nly one layer. However, in an experiment using all the data for each feature group, it can be seen that methods 1 and 2 have better performance and accuracy with a value of 99%, while groups 3 and 4 hav a rating of 98%. For recurrent networks, the highest accuracy achieved in the experiment was close to 98%. Architecture of Neural Networks and Normalization of Input Values
Each neural network can provide different results depending on its configuration; In this example, the rule that provides the most accuracy in each population profile group is R_1. Comparison of Various Research Methods

Research Methods

Different research related to our study is presented. In [44], the authors analyzed UNSW-
NB15 data from various machine learning methods and proposed a data prioritization method to reduc e data ingestion. This study investigates an analysis that reduces features beyond the 47 features curre ntly presented in the literature. The best performance was achieved using 22 features giving 89.86% a ccuracy, followed by 13 features using decision trees giving 89.76% accuracy. Additionally, some rec ent studies have explored the existing literature, mainly evaluating machine learning methods, such as [49] where UNSWNB15 Si and support removal vector machine achieved the best accuracy of 85.41 % based on decision tree (C5.0) [49] . Also in [41], the authors proposed specific options for network attack and proposed various algorithms in the classification range and achieved an accuracy of 99.94 % in the process. The best data (Table 8) shows a representative of different studies and their results w ith those presented in the comparison report.

CONCLUSION

This study examines various machine learning approaches employed in the field of cybersecurity for the purpose of anomaly detection. This study investigates the utilization of diverse machine learning techniques, encompassing supervised learning, unsupervised learning, deep learning, and rule-based methods. This study examines the practical applications of cybersecurity in several domains. Furthermore, this underscores the significance of feature selection, engineering, and assessment metrics in the development of resilient anomaly detection models. In conclusion, a comprehensive evaluation of the strengths and limits of different methodologies is essential in order to facilitate the selection of relevant approaches by both present and future researchers.

REFERENCES

[1]. "Machine Learning Techniques in Cybersecurity." Encyclopedia. Retrieved from https://encyclopedia.pub/entry/25675.

[2]. Abdullah, A. H., Ahmed, M. H., & Wahab, M. H. A. (2021). A Comparative Study of Network Intrusion Detection Techniques Using NSL-KDD Dataset. IEEE Access, 9, 91924-91942.

[3]. Akhtar, S., Faisal, M., Ahmad, S., & Rho, S. (2020). Machine learning-based ransomware detection: State-of-the-art and future research directions. Journal of Network and Computer Applications, 153, 102539.

[4]. Akinyele, J. R., Gao, K., & Zhu, S. (2015). Insider threat detection using log analysis and machine learning. International Journal of Information Security, 14(5), 403-415.

[5]. Alawami, A. K., Khan, M. K., & Kiong, T. E. (2020). Insider threat detection: A review and research directions. Journal of Network and Computer Applications, 153, 102531.

[6]. Alazab, M., Hobbs, M., & Abawajy, J. (2018). A survey of botnet detection techniques. Journal of Network and Computer Applications, 110, 60-71.

[7]. Alzahrani, B., Zulkernine, M., & Alazab, M. (2020). Machine learning-based intrusion detection techniques for securing industrial control systems: A review. Computers & Security, 88, 101628.

[8]. Bhattacharya, S., Gupta, P., & Chatterjee, J. (2021). A comparative study of machine learning algorithms for malware detection. Multimedia Tools and Applications, 80(10), 14935-14957.

[9]. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.

[10]. Chiong, R., Lee, V. C., & Zhou, L. (2017). Anomaly detection in cyber security: A machine learning approach. In Machine learning paradigms: Advances in data analytics (pp. 81-112). Springer, Cham.

[11]. Demertzis, K., & Karampelas, P. (2020). A review of anomaly detection techniques in financial markets: An application to emerging markets. Expert Systems with Applications, 146, 113172.

[12]. Dhamecha, T. I., & Thakkar, P. (2020). A Comprehensive Review on Anomaly Detection Techniques using Machine Learning. International Journal of Advanced Research in Computer Science, 11(4), 44-51.

[13]. Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2018). Data augmentation using synthetic data for time series classification with deep residual networks. arXiv preprint arXiv:1808.08467.