Network Intrusion Detection System

A Project Report

in partial fulfillment of the degree

Bachelor of Technology

in

Electronics and Communication Engineering

By

Roll.No: 17K41A0484 Name: Pothukanuri Saivenkat

Roll.No: 17K41A05E1 Name: Satram Pavan Kumar

Under the Guidance of

S Lalit Mohan

Submitted to

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

S.R.ENGINEERING COLLEGE (A), ANANTHASAGAR, WARANGAL

(Affiliated to JNTUH, Accredited by NBA) April, 2020.

SR Engineering College
Innovation . Creativity . Entrepreneurship

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Course project report entitled "Network Intrusion Detection System (NIDS)" is a record of bonafide work carried out by the student(s) Pothukanuri Sai Venkat and Satram Pavan Kumar bearing Roll No(s) 17K41A0484 and 17K41A05E1, respectively. During the academic year 2020-21 in partial fulfillment of the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering** by the Jawaharlal Nehru Technological University, Hyderabad.

Supervisor **Head of the Department**

External Examiner

ACKNOWLEDGEMENT

I wish to take this opportunity to express my sincere and deep sense of respect to our beloved principal. **Dr. V Mahesh, S R Engineering College** for providing an excellent academic atmosphere in the institution.

I express heartfelt thanks to the Head of Department, Computer Science & Engineering **Mrs. Shasikala** for providing me with necessary infrastructure and thereby giving me freedom to carry out the work.

I express heartfelt thanks to **S Lalit Mohan** for providing me guidance and support in completing project. I also express my sincere thanks to **Dr V Venkataramana** for the support throughout the course of the work.

I also thank the other staff members and friends who assisted me.

Finally, I thank my parents who inspired me always to do the best

Student Name

Pothukanuri Saivenkat

Satram Pavan Kumar

TABLE OF CONTENTS

S.NO	CONTENTS	PAGE.NO
1.	INTRODUCTION	5
2.	RELATED WORK	6
3.	BACKGROUND a. Deep Neural Network (DNN) b. Application of rectified linear units (ReLU)	8
4.	IMPLEMENTATION a. Datasets Description b. Shortcomings of KDDCup-'99' dataset c. DARPA / KDDCup-'99' dataset	9
5.	CONCLUSION	11
6.	RESULT	12
7.	REFERENCES	13

1. INTRODUCTION

In the modern world, the fast-paced technological advancements have encouraged every organization to adopt the integration of information and communication technology (ICT). Hence creating an environment where every action is routed through that system making the organization vulnerable if the security of the ICT system is compromised. Therefore, this calls for a multilayered detection and protection scheme that can handle truly novel attacks on the system as well as able autonomously adapt to the new data. There are multiple systems that can be used for shielding such ICT systems from vulnerabilities, namely anomaly detection and IDSs.

A demerit of anomaly-detection systems is the complexity involved in the process of defining rules. Each protocol being analyzed must be defined, implemented, and tested for accuracy. Another pitfall relating to anomaly detection is that harmful activity that falls within usual usage pattern is not recognized. Therefore, the need for an IDS that can adapt itself to the recent novel attacks and can be trained as well as deployed by using datasets of irregular distribution becomes indispensable. Intrusion Detect Systems (IDSs) are a range of cybersecurity-based technology initially developed to detect vulnerabilities and exploits against a target host. The sole use of the IDS is to detect threats. Therefore, it is located out-of-band on the infrastructure of the network and is not in the actual real-time communication passage between the sender and receiver of data. Instead, they solutions will often make use of a TAP or SPAN ports to analyze the inline traffic stream's copy and will try to predict the attack based on a previously trained algorithm, hence making the need of a human intervention trivial. In the field of cybersecurity, algorithms of machine learning have played an essential part. Especially, due to the incredible performance and potential of deep learning networks in recent days in various problems from a wide variety of fields which were considered unsolvable in past, the reliability of applying it for Artificial Intelligence (AI) and unsupervised challenges have increased [39]. Deep-learning is nothing but a partition of machine-learning that mimics the functions of the human brain and hence the name artificial neural network. The concept of deep learning consists of creating hierarchical representations that are complex that involve the creation of simple building blocks to solving of high-level problems. In recent days the application of deep learning methods is leveraged towards various use cases of cyber security.

Therefore, it becomes obvious that Deep neural networks and IDSs, when combined, can work at a superhuman level. Also, since the IDSs are out-of-band on the infrastructure, common attacks like DoS which primarily aims at choking the network band to gain access of the host, cannot bottleneck the performance of it, hence this security layer cannot tamper with ease. Towards the end, the sections are organized as follows: Section II reviews the work related to IDS, different deep neural networks, and some discussions about KDDCup-'99' dataset that was published. Section III takes an in-depth look at Deep Neural Networks (DNN) and the applications of ReLU activation function. Section IV analyses the dataset used in this paper, explains the shortcoming of it and evaluates the results. Section V concludes and states a plausible workflow into the future of this research work.

2. RELATED WORK

The research on ID in network security has existed since the birth of the computer architectures. The use of ML techniques and solutions to holistic IDS has become common in recent days, but training data at hand is limited and are mostly used only for bench-marking purposes. DARPA datasets [1], are one of the most comprehensive datasets available publicly. The data of tepdump offered by the 1998 DARPA ID Evaluation network of 1998 was cleaned and utilized for the KDDCupcontest of 1999 at the 5th International Conference on Knowledge Discovery and Data Mining. The job was to organize the records of the connections that are already preprocessed into either traffic which is normal, or one of the following categories of attack: 'DoS', 'Probing', 'R2L' and 'U2R'.

The preprocessing of the KDDCup-'99' competition's data was done using the MADAMID framework [2]. The entries that used variants of decision trees showed only marginal differences in performance occupied the first three places [3, 4, 5]. The first 17 submissions of the competition were all benchmarked to perform well and are summarized [6]. The majority of published results were tested and trained with only 10% training set observing the feature reduction on the KDDCup- '99' datasets [7, 8, 9]. Few researchers used custom built datasets, with extracted from the 10% KDDCup-'99' training set [10, 11, 12].

There are a number of interesting publications where the results are indirectly compared due to the use of different training and test datasets. In a paper [13], genetic algorithm and decision trees were used for automatic rule generation for an intelligent system for enhancing the capability of an existing IDS. The integrated utilization of neural networks in IDS was suggested by [14] and [15]. [16] proposed an application of recurrent neural networks (RNNs) and [17] compared the neural network architectures' performance for statistical anomaly detection to datasets from four different scenarios.

Although the datasets of KDDCup-'99' has various issues [18], [19] argues that they are still an effective bench-marking dataset which is publicly available to compare different intrusion detection methods.

The fundamental reason for the popularity of ML-based approaches is because of its capability to attack the constantly evolving complex and diverse threats to achieve an acceptable false positive rate of ID with the reasonable computational cost. In early stages, [36] used PNrule method which is derived from P-rules and N-rules to figure out the existence and nonexistence of the class respectively. This has an advantage due to the enhancement of the detection rate in the other types of attacks except for the U2R category.

An extrapolation to traditional Feed Forward Networks (FFN) in the plane of taking inspiration from biological elements, is a network named Convolutional Neural Network (CNN). In early stages, CNN was used for processing of images by making use of normal 2D layers, pooling 2D layers and completely connected layers. [37] studied the applications of CNN for IDS with the KDDCup of '99' dataset and compared the results with several other bleeding-edge algorithms. After a broad analysis, they have concluded the superiority of CNN over the other

algorithms. The study of the utilization of the Long Short-Term Memory (LSTM) classifier was conducted by [38] with the same dataset. It has been stated that because of the capability of LSTM to see into the past and relate the successive records of connections demonstrates usefulness towards intrusion detection systems.

The ultimate motive of this paper is to exploit the possibility of randomness of the inbound cyber-attack which is unsuspicious to human sight but can be filtered by adding an artificial intelligence layer to the network. Hence, by training the neural network with the existing cyber-attacks data, it can learn to predict an inbound attack easily and can either alert the system or initiate a pre-programmed response which may abstain the attack from proceeding further. As a result, millions worth, aftershock collateral damage and expensive data leaks can be prevented just by simply adding an extra layer to the security system. The benchmarking dataset used for training the networks are bygone and for a better real-time robustness of the algorithm, more recent data must be used for retraining before deploying in the field. The obligatory of this paper is to introduce the essence of artificial neural networks into the much rapidly evolving field of cybersecurity.

3. BACKGROUND

Deep neural networks (DNNs) are Artificial Neural Network (ANN) with a multi-layered structure comprised within the input-output layers. They can model complex non-linear relationships and can generate computational models where the object is expressed in terms of the layered composition of primitives.

Below we roughly cover simple DNNs and applications of ReLU and why it is preferred over other activation functions.

A. Deep Neural Network (DNN)

While traditional machine learning algorithms are linear, deep neural networks are stacked in increasing hierarchy of complexity as well as abstraction. Each layer applies a nonlinear transformation onto its input and creates a statistical model as output from what it learns. In simple terms, the input layer is received by the input layer and passed onto the first hidden layer. These hidden layers perform mathematical computations on our inputs. One of the challenges in creating neural networks is deciding the hidden layers' count and the count of the neurons for each layer. Each neuron has an activation function which is used to standardize the output from the neuron. The" Deep" in Deep learning refers to having more than one layer which is hidden. The output layer returns the output data. Until the output has reached an acceptable level of accuracy, epochs are continued.

B. Application of rectified linear units (ReLU)

ReLu has turned out to be more efficient and have the capacity to accelerate the entire training process altogether [20]. Usually, Neural networks use a sigmoidal activation function or tanh (hyperbolic tangent) activation functions. But these functions are prone to vanishing gradient problem [21]. Vanishing gradient occurs when lower layers of a DNN have gradients of nearly null because units of higher layers are nearly saturated at the asymptotes of the tanh function. ReLU offers an alternative to sigmoidal non-linearity which addresses the issues mentioned so far [22].

4. IMPLEMENTATION

We consider Keras [23] as a wrapper on top of TensorFlow [24] as software framework. For exponentially increasing the agility of processing of data in deep-learning architectures, a GPU enabled TensorFlow in a single Nvidia-GK110BGLTesla-k40 has been used.

A. Datasets description

The DARPA's program for ID evaluation of 1998 was managed and prepared by Lincoln Labs of MIT. The main objective of this is to analyze and conduct research in ID. A standardized dataset was prepared, which included various types of intrusions which imitated a military environment and was made publicly available. The KDD intrusion detection contest's dataset of 1999 was a well-refined version of this [25].

B. Shortcomings of KDDCup-'99' dataset

ReLu has turned out to be more efficient and have the A detailed report and major shortcomings of the provided synthetic data set such as KDDCup-'98' and KDDCup-'99" were discussed by [26]. The main condemnation was that they failed to validate their data set a simulation of real-world network traffic profile. Irrespective of all these criticisms, the dataset of KDDCup-'99' has been used as an effective dataset by many researchers for bench-marking the IDS algorithms over the years. In contrast to the critiques about the creation of the dataset, [27] has revealed a detailed analysis of the contents, identified the non-uniformity and simulated the artifacts in the simulated network traffic data.

The reasons behind why the machine learning classifiers have limited capability in identifying the attacks that belong to the content categories R2L, U2R in KDDCup-'99' datasets have been discussed by [28]. They have concluded that it is not possible to get acceptable detection rate using classical ML algorithms. It is also stated the possibility of getting high detection rate in most of the cases by producing a refined and augmented data set by combining the train and test sets. However, a significant approach has not been discussed.

The DARPA / KDDCup-'88 failed to evaluate the traditional IDS resulting in many major criticisms. To eradicate this [29] used Snort ID system on DARPA / KDDCup-'98' tcpdump traces. The system performed poorly resulting in low accuracy and the impermissible false positive rates. It failed in detecting dos and probing category but contrasting performing better than the detection of R2L and U2R.

Despite the harsh criticisms [30], still KDDCup-'99' set is one of the most widely used publicly available bench-marking datasets reliable for studies related to IDS evaluation and other security-related tasks [31]. In the effort of mitigating the underlying problems existing with KDDCup-'99' set, a refined version of dataset named NSL-KDD was proposed by [31]. It removed the connection redundancy records in the entire train and test data. In addition to that, the invalid records were also removed from the test data. These measures prevent the classifier from being biased in the direction of the more frequent records. Even after the refinement, this failed to solve the issues reported by [32, 33], and a new dataset named UNSW-NB15 was proposed.

C. DARPA / KDDCup-’99’ dataset

The DARPA’s ID evaluation group, accumulated network-based data of IDS by simulation of an air force base LAN by over 1000s of UNIX nodes and for continuously 9 weeks, 100s of users at a given time in Lincoln Labs which was then divided into 7 and 2 weeks of training and testing respectively to extract the raw dump data TCP. MIT’s lab with extensive financial support from DARPA and AFRL, used Windows and UNIX nodes for almost all of the inbound intrusions from an alienated LAN unlike other OS nodes. For dataset, 7 distinct scenarios and 32 distinct attacks which totals up to 300 attacks were simulated.

Since the year of release of KDD-’99’ dataset [34], it is the most vastly utilized data for evaluating several IDSs. This dataset is grouped together by almost 4,900,000 individual connections which includes a feature count of 41. The simulated attacks were categorized broadly as given below:

• Denial-of-Service-Attack (DoS): Intrusion where a person aims to make a host inaccessible to its actual purpose by briefly or sometimes permanently disrupting services by flooding the target machine with enormous amounts of requests and hence overloading the host [35].

• User-to-Root-Attack (U2R): A category of commonly used maneuver by the perpetrator start by trying to gain access to a user’s pre-existing access and exploiting the holes to obtain root control.

• Remote-to-Local-Attack (R2L): The intrusion in which the attacker can send data packets to the target but has no user account on that machine itself, tries to exploit one vulnerability to obtain local access cloaking themselves as the existing user of the target machine.

• Probing-Attack: The type in which the perpetrator tries to gather information about the computers of the network and the aim for doing so is to get past the firewall and gaining root access.

5. CONCLUSION

The publicly available KDDCup-'99' dataset has been primarily used as the benchmarking tool for the study, through which the superiority of the DNN over the other compared algorithms have been documented clearly. For further refinement of the algorithm, this paper considers of DNNs with different counts of hidden layers and it was concluded that a DNN with 3 layers has been proven to be effective and accurate of all.

Since the neurons are trained with a bygone benchmarking dataset, as discussed, several times in this paper, this comes as a pitfall for this methodology. Fortunately, it can be vanquished by using a fresh dataset with the essences of the latest attack strategies before the actual deployment of this artificial intelligence layer to the existing network systems to ensure the agility of the algorithms real-world capabilities.

From the empirical results of this paper, we may claim that deep learning methods are a promising direction towards cyber security tasks, but even though the performance on artificial dataset is exceptional, application of the same on network traffic in the real-time which contains more complex and recent attack types is necessary. Additionally, studies regarding the flexibility of these DNNs in adversarial environments are required. The increase in vast variants of deep learning algorithms calls for an overall evaluation of these algorithms regarding its effectiveness towards IDSs. This will be one of the directions towards IDS research can travel and hence will remain as a work of future.

6. RESULT

The loop which generates a random index between 0 and 311029.

The predicted array consists of 0's and 1's which means:

0 = Safe (no attack detected).

1= Not Safe (Attack Detected).

```
for i in range(1000):
    predict_rand = rd.randint(0, 311029)
    print("Checking for %d index in predicted array"% predict_rand)
    if predicted[predict_rand] == 0:
        print("The network is SAFE to use.")
        break
    else:
        print("The Network is not at all SAFE.")
        break
```

```
...
Checking for 189610 index in predicted array
The network is SAFE to use.

>>>
```

```
...
Checking for 269228 index in predicted array
The Network is not at all SAFE.

>>>
```

7. REFERENCES

[1] R. Lippmann, J. Haines, D. Fried, J. Korba and K. Das. ”The 1999 DARPA off-line intrusion detection evaluation”. Computer networks, vol. 34, no. 4, pp. 579 595, 2000. DOI http://dx.doi.org/10.1016/S1389- 1286(00)00139-0.

[2] W. Lee and S. Stolfo. ”A framework for constructing features and models for intrusion detectionsystems”. ACM transactions on information and system security, vol. 3, no. 4, pp. 227261, 2000. DOI http://dx.doi. Org/10.1145/382912.382914.

[3] B. Pfahringer. ”Winning the KDD99 classification cup: Bagged boosting”. SIGKDD explorations newsletter, vol. 1, pp. 6566, 2000. DOI http://dx.doi.org/10. 1145/846183.846200.

[4] M. Vladimir, V. Alexei and S. Ivan. ”The MP13 approach to the KDD’99 classifier learning contest”. SIGKDD explorations newsletter, vol. 1, pp. 76 77, 2000. DOI http://dx.doi.org/10.1145/846183. 846202.

[5] R. Agarwal and M. Joshi. ”PNrule: A new framework for learning classier models in data mining”. Tech. Rep. 00-015, Department of Computer Science, University of Minnesota, 2000.

[6] C. Elkan. ”Results of the KDD’99 classifier learning”. SIGKDD explorations newsletter, vol. 1, pp. 63 64, 2000. DOI http://dx.doi.org/10.1145/846183. 846199.

[7] S. Sung, A.H. Mukkamala. ”Identifying important features for intrusion detection using support vector machines and neural networks”. In Proceedings of the symposium on applications and the Internet (SAINT), pp. 209216. IEEE Computer Society, 2003. DOI http: //dx.doi.org/10.1109/saint.2003.1183050.

[8] H. Kayacik, A. Zincir-Heywood and M. Heywood. ”Selecting features for intrusion detection: A feature relevance analysison KDD 99 intrusion detection datasets”. In Proceedings of the third annual conference on privacy, security and trust (PST). 2005.

[9] C. Lee, S. Shin and J. Chung. ”Network intrusion detection through genetic feature selection”. In Seventh ACIS international conference on software engineering, artificial intelligence, networking, and parallel/distributed computing (SNPD), pp. 109114. IEEE Computer Society, 2006

[10] S. Chavan, K. Shah, N. Dave, S. Mukherjee, A. Abraham and S. Sanyal. ”Adaptive neuro-fuzzy intrusion detection systems”. In Proceedings of the international conference on information technology: Coding and computing (ITCC), vol. 1, pp. 7074. IEEE Computer Society, 2004. DOI http://dx.doi.org/ 10.1109/itcc.2004.1286428.

[11] S. Chebrolu, A. Abraham and J. Thomas. ”Feature deduction and ensemble design of intrusion detection systems”. Computers & security, vol. 24, no. 4, pp. 295307, 2005. DOI http://dx.doi.org/10.1016/j. Cose.2004.09.008.

[12] Y. Chen, A. Abraham and J. Yang. ”Feature selection and intrusion detection using hybrid flexible neural tree”. In Advances in neural networks (ISNN), vol. 3498 of Lecture notes in computer science, pp. 439 444. Springer Berlin / Heidelberg, 2005. DOI http: //dx.doi.org/10.1007/11427469 71.

[13] C. Sinclair, L. Pierce and S. Matzner. ”An application of machine learning to network intrusion detection”. In Proceedings of the 15th annual computer security applications conference (ACSAC), pp. 371377. IEEE Computer Society, 1999. DOI http://dx.doi.org/ 10.1109/csac.1999.816048.

[14] H. Debar, M. Becker and D. Siboni. ”A neural network component for an intrusion detection system”. In Proceedings of the IEEE computer society symposium on research in security and privacy, pp. 240250. IEEE Computer Society, 1992. DOI http://dx.doi.org/ 10.1109/risp.1992.213257.

[15] J. Cannady. ”Artificial neural networks for misuse detection”. In Proceedings of the 1998 national information systems security conference (NISSC), pp. 443456. Citeseer, 1998.

[16] H. Debar and B. Dorizzi. ”An application of a recurrent network to an intrusion detection system”. In International joint conference on neural networks, 1992. IJCNN., vol. 2, pp. 478 483 vol.2. jun 1992. DOI http://dx.doi.org/10.1109/ijcnn. 1992.226942. [17] Z. Zhang, J. Lee, C. Manikopoulos, J. Jorgenson and J. Ucles. ”Neural networks in statistical anomaly intrusion detection”. Neural network world, vol. 11, no. 3, pp. 305316, 2001.

[18] J. McHugh. ”Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory”. ACM transactions on information and system security, vol. 3, no. 4, pp. 262294, 2000. DOI http://dx.doi.org/10.1145/382912.382923.

[19] M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani. ”A detailed analysis of the KDD CUP 99 data set”. In IEEE symposium on computational intelligence for security and defense applications, Cisda, pp. 16. IEEE, Jul. 2009. DOI http://dx.doi.org/10.1109/cisda. 2009.5356528.

[20] X. Glorot, A. Bordes, and Y. Bengio, ”Deep sparse rectifier neural networks,” in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 315323.

[21] Bengio, Y., Simard, P. and Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), pp.157-166.

[22] Maas, A.L., Hannun, A.Y. and Ng, A.Y., 2013, June. Rectifier nonlinearities improve neural network acoustic models.

In Proc. icml (Vol. 30, No. 1, p. 3).
[23] F. Chollet, ”Keras (2015),” URL http://keras. Io, 2017.
[24] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., ”Tensorflow: A system for large-scale machine learning.” in OSDI, vol. 16, 2016, pp. 265283.
[25] Stolfo, S., Fan, W. and Lee, W., KDD-CUP-99 Task Description. 1999- 10-28)[2009-05-08]. http://KDD. ics. uci. edu/databases/kddcup99/task, html.
[26] J. McHugh. ”Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory”. ACM transactions on information and system security, vol. 3, no. 4, pp. 262294, 2000. DOI http://dx.doi.org/10.1145/382912.382923.
[27] M. Mahoney and P. Chan. ”An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection”. In Recent advances in intrusion detection, vol. 2820 of Lecture notes in computer science, pp. 220237. Springer Berlin / Heidelberg, 2003.
[28] Sabhnani, Maheshkumar, and Gursel Serpen.” Why machine learning algorithms fail in misuse detection on KDD intrusion detection data set.” Intelligent Data Analysis 8, no. 4 (2004): 403-415.
[29] S. Brugger and J. Chow. ” An assessment of the DARPA IDS evaluation dataset using snort”. Tech. Rep. CSE-2007-1, Department of Computer Science, University of California, Davis (UCDAVIS), 2005.
[30] J. McHugh. ” Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory”. ACM transactions on information and system security, vol. 3, no. 4, pp. 262294, 2000. DOI http://dx.doi.org/10.1145/382912.382923.
[31] Tavallaee, Mahbod, Ebrahim Bagheri, Wei Lu, and Ali-A. Ghorbani.” A detailed analysis of the KDD CUP 99 data set.” In Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defense Applications 2009. 2009.
[32] Moustafa, Nour, and Jill Slay.” The evaluation of Network Anomaly Detection Systems: Statistical analysis of the U [8] H. Kayacik, A. ZincirHeywood and M. Heywood.” Selecting features for intrusion detection: A feature relevance analysison KDD 99 intrusion detection datasets”. In Proceedings of the third annual conference on privacy, security, and trust (PST). 2005.
[33] Moustafa, Nour, and Jill Slay.” UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data).” Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.
[34] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/database
[35] McDowell, M. (2013). Understanding Denial-of-Service Attacks USCERT. United States Computer Emergency Readiness Team.