WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

COMPARISON OF MACHINE LEARNING AND DEEP LEARNING ALGORITHMS FOR INTRUSION DETECTION SYSTEM

Nikita Ramnath Pandure, Smriti Bijendra Singh

MCA (IMCOST), Mumbai UniversityC-4, Wagle Industrial Estate, Near Mulund(W), Check Naka, Thane(W) – 400604

ABSTRACT

Networks are very important now a days in the world and data security has become an essential area of study. An Intrusion detection system Detect the status of the software and hardware of the network. Curing problems for current IDSs remain they improve detection precision, track unknown attacks and decrease false alarm rates after decades of development. Many Papers have focused on the development of IDSs using machine learning methods to solve the above-described problems. With the high Accuracy of computer teachings, the basic distinctions between usual and irregular data can be recognized automatically. Unknown threats may also be detected because of their generalizability via machine learning system. This paper suggests a taxonomy of IDS, which uses the primary dimension of data objects to classify and sum up IDS literatures based on and dependent on deep learning. We assume this kind of taxonomy is sufficient for researchers in Network and infrastructure security or cyber security. We selected two algorithms of deep learning (RNN, LSTM) and three algorithms from machine learning (Bayes Net, Random Forest, Neural Network) and we tested them on KDD cup 99 dataset and evaluated accuracy algorithms, and we used a WEKA Tool To calculate the accuracy for intrusion detection

Keywords*: IDS, Machine Learning, Deep Learning, Accuracy, Weka*

1. INTRODUCTION

The Internet has become a part of daily life and an essential tool today. It aids people in many areas, such as business, entertainment and education, etc. In particular, Internet has been used as an important component of business models. For the business operation, both business and customers apply the Internet application such as website and e-mail on business activities. Therefore, information security of using Internet as the media needs to be carefully concerned. Intrusion detection is one major research problem for business and personal networks. As there are many risks of network attacks under the Internet, there are various systems designed to block the Internet-based attacks. Particularly, intrusion detection systems (IDSs) aid the network to resist external attacks. That is, the goal of IDSs is to provide a wall of defence to confront the attacks of computer systems on Internet. IDSs can be used on detect difference types of malicious network communications and computer systems usage, whereas the conventional firewall cannot perform this task. Intrusion detection is based on the assumption that the behaviour of intruders different from a legal user In general, IDSs can be divided into two categories: anomaly and misuse (signature) detection based on their detection approaches (Anderson, 1995; Rhodes, Mahaffey, & Cannady, 2000). Anomaly detection tries to determine whether deviation from the established normal usage patterns can be flagged as intrusions. On the other hand, misuse detection uses patterns of well-known attacks or weak spots of the system to identify intrusions. In literature, numbers of anomaly detection systems are developed based on many different machine learning techniques. For example, some studies apply single learning techniques, such as neural networks, genetic algorithms, support vector machines, etc. On the other hand, some systems are based on combining different learning techniques, such as hybrid or ensemble techniques. In particular, these techniques are developed as classifiers, which are used to classify or recognize whether incoming Internet access is the normal access or an attack. However, there is no a review of these different machine learning techniques over the intrusion detection domain.

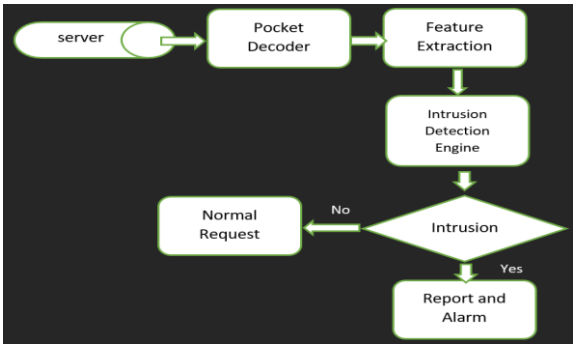

Fig.1 INTRUSION DETECTION SYSTEM PROCESS

IDS CONCEPT:

Intrusion is an illegal or undesirable attempt to obtain access to computer networks' details or damage the device. An IDS is a cybersecurity software that detects a wide range of security violations, from external attempts to interference with the insider system's intrusion. The primary functions of IDSs are hosts and networks monitored, the computer system's behaviour, alerts generated, and unusual activities responded to. IDSs are generally installed near the stable network nodes because the hosts and networks are tracked. Methods of IDS classification split into two categories: Methods focused on identification and techniques based on data source. Detection of misuse and detection of abnormality are two examples of IDS-based detection approaches. In host and network methods, IDSs can be broken into data source-based methods as shown in following figure(fig.2)

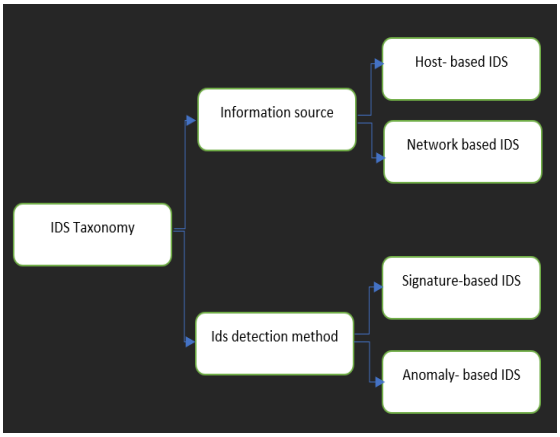

Fig.2 IDS TAXONOMY

2. MATERIALS AND METHODOLOGY

The dataset KDD CUP 99 is chosen as the basis for the planned discovery scheme. Around 4.900,000 single connection vectors are used in the KDD training data collection, each of these 41 features being labelled an attack or normal, showing the approximation to identified attacks. It is necessary to note that the experimental results are not in the same probability division as training data. In datasets there are 24 types of exercise attacks, with 14 more types of test attacks. The simulated attacks can classify into one of four groups:

- **DOS**: Various forms of attacks, such as SYN flood, are involved.
- **U2R**: unauthorized access to superuser rights on a local system.
- **R2L**: It does not have allowed remote access.
- Probing: Monitoring and examination.

In this Study we convert the KDD CUP 99 dataset from the CSV group to the ARFF position in the pre-processing stage as shown in fig.3. the KDD CUP 99 dataset pre-processed into 22 assault subcategories Instead of inflated 5 attack classes.

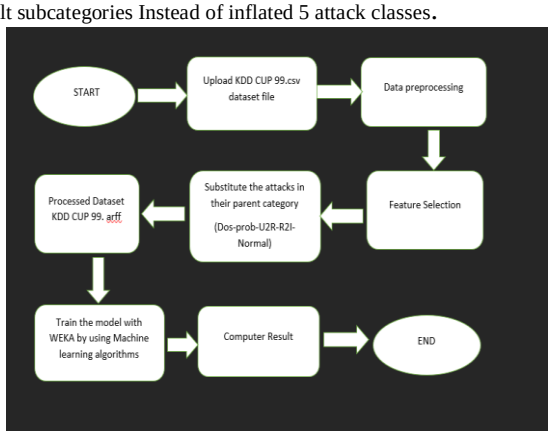

Fig.3 Research Methodology

Three machine learning algorithms were selected and implemented for IDS.

A. Bayesian Network

Bayesian networks are a type of probabilistic graphical model that uses Bayesian inference for probability computations. Bayesian networks aim to model conditional dependence, and therefore causation, by representing conditional dependence by edges in a directed graph. Through these relationships, one can efficiently conduct inference on the random variables in the graph through the use of factors

B. Random forest

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. It generates predictions without requiring many configurations in packages

C. Neural Network

A computer learning system uses a function network to recognize and translate data entries into a desired result in one output the neural network can be used as a part to transform complicated data into a format that computers can understand in different machine learning algorithms. neural network is made of artificial neurons that receive and process input data. Data is passed through the input layer, the hidden layer, and the output layer. A neural network process starts when input data is fed to it. Data is then processed via its layers to provide the desired output. A neural network learns from structured data and exhibits the output. Learning taking place within neural networks can be in three different categories this are Supervised Learning Unsupervised Learning and Reinforcement Learning.

We have selected two algorithms for deep learning, and they are RNN and LSTM.

A. Recurrent Neural Network (RNN)

Recurrent neural networks (RNN) are a class of neural networks that are helpful in modelling sequence data. Derived from feedforward networks, RNNs exhibit similar behaviour to how human brains function. Simply put: recurrent neural networks produce predictive results in sequential data that other algorithm can't. this algorithm are widely used in common or temporary issues like language translation and natural language processing It's in popular apps like Siri, voice, and translates with Google.

B. Long Short-Term Memory Networks (LSTMs)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. This is a behaviour required in complex problem domains like machine translation, speech recognition, and more. LSTMs are a complex area of deep learning

3. EVALUATION MATRIX

Assessment of performance metrics for IDS based on uncertainty matrix values for Machine Learning and Deep Learning approaches.

False alarm rate: It is also called false-positive and also known as the probability of wrong Prediction or false Detection.

Precision: It is Ratio of true Expected attacks to all attacks samples.

Recall: it is the ratio of number of correct positive Prediction to all attacks sample.

Accuracy Of True Rate: It is the right number of normal samples divided by the overall number of normal samples.

F- measure: it is a measure of models accuracy on a dataset by Precision and Recall.

4. EXPERIMENTAL RESULT AND DISCUSSION

We have done experimental operation on machine learning algorithms (Bayes Net, Random Forest, Neural Network) and Deep Learning Algorithms (RNN, LSTM) on the KDD CUP 99 data set using WEKA we achieved different result as follows.

Table 1 -Comparison of accuracy between ML algorithms

SR.N	**ML algorithms**	**Accuracy**	**Time in Seconds**
1	**Bayes Net**	98.7768%	17.55
2	**Random forest**	99.9723%	342.36
3	**Neural Network**	99.3482%	1505.59

The accuracy Difference between these algorithms is not substantial according to Above table 1.The Random forest algorithm shows the highest accuracy as 99.97% as compared to remaining two algorithms. The Neural Network also shows the good result but they take a long time.

Table.2 Result of Evaluation Metrics Using Random Forest

class	TP Rate	FP Rate	Precision	Recall	F-Measure
Normal	1.000	0.000	0.999	1.000	1.000
U2R	0.615	0.000	0.889	0.615	0.727
Dos	1.000	0.000	1.000	1.000	1.000
R2L	0.981	0.000	0.992	0.981	0.987
Probe	0.993	0.000	0.999	0.993	0.996

Table.3 Comparison of accuracy between DL Algorithms

SR.N.	**DL algorithms**	**Accuracy**	**Time in Seconds**
1	**RNN**	53.1757%	47.93
2	**LSTM**	64.2527%	24.98

The accuracy Difference between these two algorithms was enormous. According to the above table 2 accuracy of LSTM IS 64.25% which is more than the accuracy of RNN which is 53.17%. LSTM can detect attacks more accurately than RNN Algorithm.

Table 4 Result of all evaluation metrics using LSTM

class	TP Rate	FP Rate	Precision	Recall	F-Measure
Normal	0.113	0.210	0.113	0.113	0.113
U2R	0.000	0.000	0.000	0.000	0.000
Dos	0.783	0.880	0.773	0.783	0.778
R2L	0.000	0.000	0.000	0.000	0.000
Probe	0.000	0.000	0.000	0.000	0.000

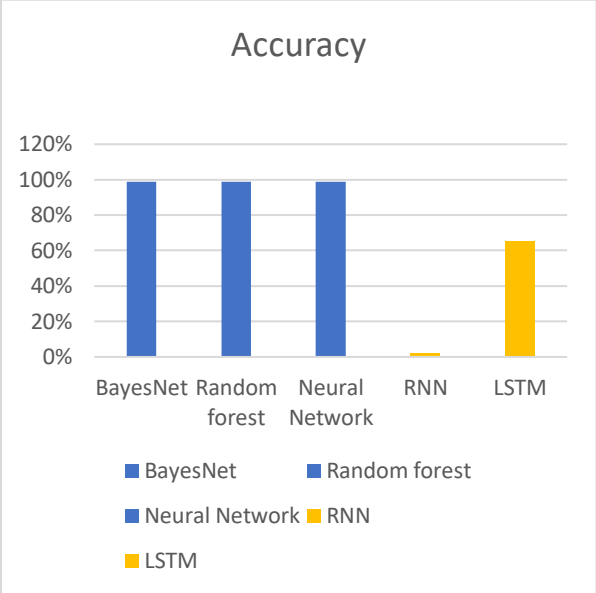

Fig.4 Accuracy analysis of ML and DL algorithms

When Using the ML and DL methods with KDD cup 99 datasets, The Result shows that machine learning performed better on this dataset comparing to deep learning.

5. CONCLUSION

Intrusion Detection System is a very important in the field of network security. The performance of the classifier is degrading by using intrusive patterns, accuracy and also, it's time-consuming. Bayes Net, Random Forest, Neural Network, RNN, and LSTM are among the ML and DL algorithms considered for IDS. With the KDD cup 99 dataset's we proposed a DL and ML approach. By looking at the results that depend on accuracy, the powerful classifier can be identified. The conclusion of this paper is that the random forest classifier achieves better result with accuracy of 99.97% based on the analysis.

6. FUTURE WORK

For future work, other classifiers can be considered. Further, all other algorithms can be used with the different dataset. WEKA supports supervised and unsupervised classifiers so it can be used to implement other algorithms and the results compared to those in this report in future.

REFERENCES

[1] Abdullah SMSA, Ameen SYA, Sadeeq MA, Zeebaree S. "Multimodal emotion recognition using deep learning," Journal of Applied Science and Technology Trends. 2021;2:52-58.

[2] Shrestha A, Mahmood A. "Review of deep learning algorithms and architectures," IEEE Access. 2019;7:53040-53065.

[3] R. Prasad and V. Rohokale, Cyber Security: The Lifeline of Information and Communication Technology. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-31703-4.

[4] A. Pal Singh and M. Deep Singh, “Analysis of Host-Based and NetworkBased Intrusion Detection System,” IJCNIS, vol. 6, no. 8, pp. 41–47, Jul. 2014, doi: 10.5815/ijcnis.2014.08.06.

[5] O. Ahmed and A. Brifcani, “Gene Expression Classification Based on Deep Learning,” in 2019 4th Scientific International Conference Najaf (SICN), Al-Najef, Iraq, Apr. 2019, pp. 145–149. doi: 10.1109/SICN47020.2019.9019357.

[6] C. Yang, G. N. Odvody, C. J. Fernandez, J. A. Landivar, R. R. Minzenmayer, and R. L. Nichols, “Evaluating unsupervised and supervised image classification methods for mapping cotton root rot,” vol. 16, no. 2, pp. 201– 215, 2015, doi: 10.1007/s11119-014- 9370-9.

[7] N. Asaad Zebari, D. Asaad Zebari, D. Qader Zeebaree, and J. Najeeb Saeed, “Significant features for steganography techniques using deoxyribonucleic acid: a review,” IJEECS, vol. 21, no. 1, p. 338, Jan. 2021, doi: 10.11591/ijeecs.v21.i1.pp338-347

[8] M. E. Aminanto and K. Kim, “Deep Learning in Intrusion Detection System: An Overview,” p. 12

[9] M. U. (2019) Masoodi, F. S., & Bokhari, “Symmetric Algorithms I,” , no. 79, 2019

[10] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, “Building an efficient intrusion detection system based on feature selection and ensemble classifier,” vol. 174, no. April, 2020, doi: 10.1016/j.comnet.2020.107247.

[11] C. Ambikavathi and S. K. Srivatsa, “Predictor Selection and Attack Classification using Random Forest for Intrusion Detection,” , vol. 79, no. 05, pp. 365–368, 2020

[12] M. A. Jabbar, R. Aluvalu, and S. S. Reddy, “RFAODE: A Novel Ensemble Intrusion Detection System,” , vol. 115, pp. 226–234, 2017, doi: 10.1016/j.procs.2017.09.129