Received Date: 3rd November, 2022
Accepted Date: 23rd April, 2023

Evaluation of Network Intrusion Detection with Feature Selection using Random Forest and Deep Neural Network

Nabin Pakka[1*], Krishna Rauniyar[2], Sagar Dangal[3] & Rupan Chaulagain[4]

[1]*Department of Electronics Engineering, Thapathali Campus, E-mail: nabin.732419@thc.tu.edu.np*
[2] *Department of Electronics Engineering, Thapathali Campus, E-mail: krishna.732419@thc.tu.edu.np*
[3]*Department of Electronics Engineering, Thapathali Campus, E-mail: sagar1.732419@thc.tu.edu.np*
[4]*Department of Electronics Engineering, Thapathali Campus, E-mail: rupan.732419@thc.tu.edu.np*

***Abstract*— Modern era relies heavily on the internet for communication and confidential data exchange. Integrity, validity, and security of data transmitted should not be compromised. Intrusion detection plays a role of paramount importance in secure transmission. However, Network intrusions are evolving. Itraises the necessity for a more robust and evolving detection system. The main objective of this paper is to compare and contrast the performance of Random Forest and Deep Neural Networks on the CICIDS-2017 dataset to build a robust Network Intrusion Detection System. Data of DDoS, DoS, and PortScan attacks from CICIDS-2017 are considered for analysis. The data is preprocessed then feature selection algorithms are applied and the best split is selected for classification. The paper compares theperformance of Random Forest and Deep Neural Networks. It was observed that DNN performs best on CICIDS-2017 data to classifybetween different attacks on the network.**

***Keywords* — Canadian Institute for Cyber Security Intrusion Detection System (CICIDS), Denial of Service (DoS), Distributed Denial of Service (DDoS)**

I. INTRODUCTION

Much information is shared daily via the internet. Information shared ranges from normal chat to confidential reports of government and big companies. Privacy, integrity, and validity of such information are of paramount importance. Network Intrusion Detection System (IDS) comes into play to mitigate situations such as theft of data, integrity violations, etc. Network IDS monitors analyze the internet traffic and look out for any anomalies. Most traditional IDS are signature-based. Hence, it cannot cope with the ever-evolving attack patterns and threats. To counter such problems, there arises the necessity ofrobust, modifiable, and extensible IDS. A network IDS utilizing the prediction capability of machine learning algorithms and neural networks is a viable solution. The paper utilizes the Canadian Institute of Cyber-security Intrusion DetectionSystem (CICIDS)-2017 dataset [1] to preprocess network data and build models. The data is first preprocessed using a null check, redundancy check, normalization, etc. The preprocessed data is subjected to an ANOVA test, correlation-based measures, and an Extra tree classifier to select the most relevant features. Random Forest [2] and Deep Neural Networks [3] is used on the pruned dataset to build models and their performances are compared to unveil the best algorithm for thedata. The rest of the paper is organized as follows: Section II discusses the architecture of the project. Section II.A gives insights into the dataset. Section II.B is about data preprocessing and II.C discusses feature selection techniques. Section II.D discusses the algorithms to be compared. Section

II.E introduces evaluation techniques used to compare the models. Section III discusses the performance of the algorithms. Section IV compares the models. Section V contains the conclusion of the paper.

II. ARCHITECTURE

The CICIDS-2017 dataset is preprocessed then important features are selected. The selected feature is utilized to generate a model using Random Forest and Deep Neural Network after which the performance is compared and the best one is chosen.

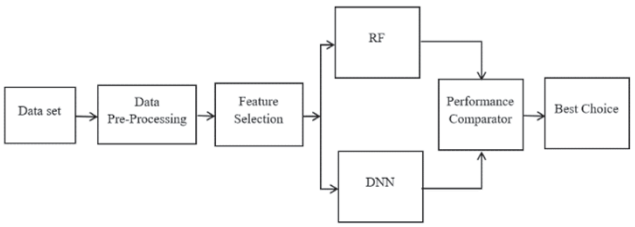


Fig. 1. Architecture of Project

A. Dataset

The Canadian Institute of Cybersecurity Intrusion Detection System CICIDS-2017 dataset was used [1]. The dataset spanned over eight different files containing five days of normal and attacks traffic data from the Canadian Institute of Cybersecurity. The files were merged and a dataset containing 741595 instances and 79 features with 4 class labels was formed. The 4 class labels that were going to be predicted by the generated models were DDoS, DoS, Port Scan, and Benign.

* *Corresponding Author*

B. Data Preprocessing

It is a fact that the best detection model should be able to detect attacks of any type. So, to design such a typical IDS the data should have relevant attributes and features. Theoretically, it is shown that having a large number of attributes and features results in higher accuracy and performance but practically, many machine learning algorithms have shown that it is not always the case. Redundant data directly creates many problems during the training phase and leads to problems like overfitting and degradation of the overall performance of the system. Random forest algorithms can be affected due to the presence of redundant features during the training of the model. The random forest can overfit by having large amounts of redundant features while performing the ensemble technique. Even DNN algorithms overfit to redundancy. So, to solve all the issues discussed above data preprocessing and feature selection algorithm is used. Data Preprocessing is a data miningtechnique used for data cleaning, handling missing values, and reduction and transformation of data. After applying data preprocessing techniques, feature selection algorithms are used for the removal of redundant and irrelevant features.

C. Feature Selection

It refers to the process of selecting a subset of relevant features that fully describes the given problem with minimum degradation of performance [4]. The following feature selectionmethods were used.

1) Univariate ANOVA Test

The simplest type of data analysis is called univariate analysis. Uni means "one" thus your data has only one variable. Data is taken, summarized, and patterns are discovered within the data.In a univariate analysis, a variable is just a category or subset into which your data falls.

Analysis of variance (ANOVA) [5] uses F-tests to statistically assess the equality of means when you have three or more groups. The term F-test is since these tests use the F-statistic totest the hypotheses. An F-statistic is the ratio of two variances.Variances quantify how evenly distributed the data points are around the mean. When the individual data points tend to deviate further from the mean, there are more variances. The total of the squared departures from the mean represents a variation. The formula for the variance is:

$$S^2 = \frac{\sum(x_i - x)^2}{n-1} \qquad (1)$$

2) Correlation-based Feature Selection

Correlation is a well-known similarity measure between two features. If two features are linearly dependent, then their correlation coefficient is ±1. The correlation coefficient is 0 if there is no correlation between the features. The correlation approach is used to determine the relationship between the features. Two broad categories can be used to measure the correlation between two random variables. The other is based on information theory, while the first is based on classical linear

correlation. The most well-known of these two metrics is the linear correlation coefficient. CFS attribute evaluator evaluatesthe features which are highly correlated with class, yet uncorrelated with each other [6]. As per the standard literature,for a pair of variables (X, Y), the linear correlation coefficient 'r' is given by:

$$r = \frac{\sum(x_i - x)(y_i - y)}{\sqrt{\sum(x_i - x)^2 \sum(y_i - y)^2}} \qquad (2)$$

3) Feature Importance using Extra Tree Classifier

Extremely Randomized Trees Classifiers (also known as Extra Trees Classifiers) are a form of ensemble learning technique that combine the findings of various de-correlated decision trees gathered in a "forest" to get their classification results [7].Conceptually, it is extremely similar to a Random Forest Classifier and only differs from it in how the decision trees in the forest are built. The initial training sample is used to build each decision tree in the Extra Trees Forest. Then, each tree is given a random sample of k features from the feature set at eachtest node, from which it must choose the best feature to divide the data according to certain mathematical criteria (typically the Gini Index). There are numerous de-correlated decision trees produced as a result of this random sampling of features. For each feature, the normalized total reduction in the mathematical criteria used in the decision of feature split (Gini Index if the Gini Index is employed in the building of the forest) is computed to perform feature selection during the construction of the forest. The Gini Importance of the feature is the name given to this value. The process of feature selection involves ranking each feature according to its Gini Importance, with theuser selecting the top k features that appeal to them.

The formula for the entropy is:

$$E = -\sum_i^c P_i \log_2 P_i \qquad (3)$$

Here C is the number of classes p_i is the probability of randomlypicking an element of class i.

The formula for information gain (IG) is:

$$IG(T, X) = E(T) - \sum_T^{S_i} E(S_i) \qquad (4)$$

Here T is the target population before the split T=∑ {All Splits}, the total number of observations before splitting. Entropy (T) is the measure of the disorder before the split or level of uncertainty. S_i is the number of observations on thei^{th} split. Entropy (S_i) is a measure of the disorder for the target variable on a split.

D. Algorithms

1) Random Forest

A supervised learning approach called random forest is appliedfor both classification and regression. But it is mainly used for classification problems. Its fundamental concept is “the wisdom of crowds”. Uncorrelated decision trees are built using a random forest technique on data samples. Unlike decision trees, random forest selects from a random subset of features rather than selecting features with the most accurate split to lower correlation across trees. Each created decision tree from the forest predicts its class. Then the class with the maximum vote is deemed as valid output. It is an ensemble method, which reduces over-fitting by averaging the outcome, making it superior to a single decision tree. The working of Random Forest is explained below:

- Begin by choosing random samples from a predetermineddataset.
- After that, this algorithm builds a decision tree for each sample. The predicted result for each decision tree will then be discovered.
- Each anticipated outcome will be voted on in this stage.
- Finally, choose the prediction result that received the mostvotes as the final prediction result.

2) Deep Neural Network

A supervised learning algorithm for binary classifiers is called a perceptron. This enables neurons to learn and processes elements in the training set one at a time. It was introduced by Frank Rosenblatt in 1957 [8]. It is a basic unit of an artificial neural network. An artificial neural network is a computer system designed to simulate the way the human brain analyzes and processes information. ANN learns using backpropagation[9]. Backpropagation is the set of learning rules used to guide artificial neural networks. Backpropagation propagates the difference between the actual class and the predicted class using a cost function. The propagated error is differentiated and used to update the weights of nodes of the neural network. A deep Neural Network is a multi-layered ANN. It consists of multiple hidden layers. The paper uses DNN with three hidden layers. The used model is shown in the figure below.

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	2048
batch_normalization (BatchNo	(None, 64)	256
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 32)	2080
batch_normalization_1 (Batch	(None, 32)	128
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 16)	528
batch_normalization_2 (Batch	(None, 16)	64
dropout_2 (Dropout)	(None, 16)	0
dense_3 (Dense)	(None, 4)	68

Total params: 5,172
Trainable params: 4,948
Non-trainable params: 224

Fig. 2. DNN model

E. Evaluation Techniques

For evaluation of the estimation of different algorithms following terms are used:

- True Positive (TP)- The numbers of estimation estimated as a class and correctly classified.
- True Negative (TN) - The numbers of estimation estimatedas not a class but found to be of that class.
- False Positive (FP) - The numbers of estimation estimated as a class but classified incorrectly.
- False Negative (FN) - The numbers of estimation estimatedas not a class and found do not belong to that class.

Based on the above terms, the following are the most usedmetrics:

- Accuracy: It is the ratio of the number of correctly classified data of the test set to the total amount of data. Itranges from 0 to 1.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

- Precision: It is the ratio of the TP of a class to the sum ofTP and FP of the same class. It ranges from 0 to 1.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

- Recall: It is the ratio of the TP of a class to the sum of TP and FN of the same class. It ranges from 0 to 1. It is also called True Positive Rate (TPR).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

- F1-Score: It is the harmonic mean of Precision and Recall.A model is better if it has a higher F1-score.

$$\text{F1- Score} = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right) \quad (8)$$

- False Positive Rate: It is the ratio of FP to the sum of FP and TN. It ranges from 0 to 1.

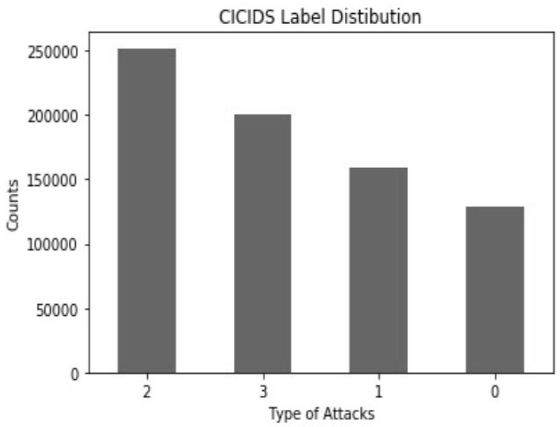
$$FPR = \frac{FP}{FP + TN} \quad (9)$$

- Receiver Operating Characteristics (ROC) curve: ROC curve is the plot based on the trade-off between the TPR onthe y-axis to FPR on the x-axis across different thresholds.The area under the ROC curve (AUC) is the area under the ROC curve. It is used as a comparison metric for machine learning models. Higher AUC corresponds to a higher quality model.

$$AUC = \int^{1} Recall\ d(FPR) \quad (10)$$

F. Performance

In the formed dataset, 970 instances were missing values hencethey were pruned. Out of 79 features, under inspection, 10 features were redundant with only 0 as a value for every instance so they were also removed. Normalization was performed on numeric features from the range 0-1. High-class imbalance [10] is a situation in a dataset where the dataset is used for training a classifier or detector, in such a case the detector is biased towards the majority class. The obtained dataset was class imbalanced as shown in the image below.




```
#### {'DDoS':0, 'PortScan':1, 'dos':2, 'BENIGN':3}
```

Fig. 3. CICIDS Label Distribution

Therefore, to balance the data we applied an under-sampling method and finally formed a dataset containing 512100 instances and 69 features. After Applying feature selection techniques, the 40 most relevant features were selected to trainthe models.

G. Random Forest

Random Forest Classifier was trained using n_estimators is the number of trees as 200 and max_depth that is the depth of the tree as 2. The confusion matrix and evaluation metrics were observed as shown in the table below.

TABLE 1

CONFUSION MATRIX OF RANDOM FOREST

Predicted Attack	Actual Attack				
	Labels	0 DDoS	1 PortScan	2 DoS	3 BENIGN
	0 DDoS	25531	0	14	29
	1 PortScan	0	25347	111	100
	2 DoS	3506	1	21757	216
	3 BENIGN	1399	1155	2860	20394

The above shows the confusion matrix of the random forest model. The model predicted well with most of the values in thediagonal. Only some values were predicted wrong. Test data appears to be well balanced with almost equal instances.

TABLE 2

EVALUATION METRICS OF RANDOM FOREST

Labels	TP	FP	FN	TN
0 DDoS	25531	43	4905	71941
1 PortScan	25347	211	1156	75706
2 DoS	21757	3723	2985	73955
3 BENIGN	20394	5414	345	81681

Here type I error for DDoS was found to be 43 out of 30436 instances and the type II error was 4905 out of 30436. For the port scan, the type I error was found to be 211 out of 26503 andthe type II error was 1156 out of 26503. For DoS, type I error was found to be 3723 out of 24742 and type II error was 2985 out of 24742. For benign, type I error was found to be 5414 outof 20739 and type II error was 345 out of 20739.

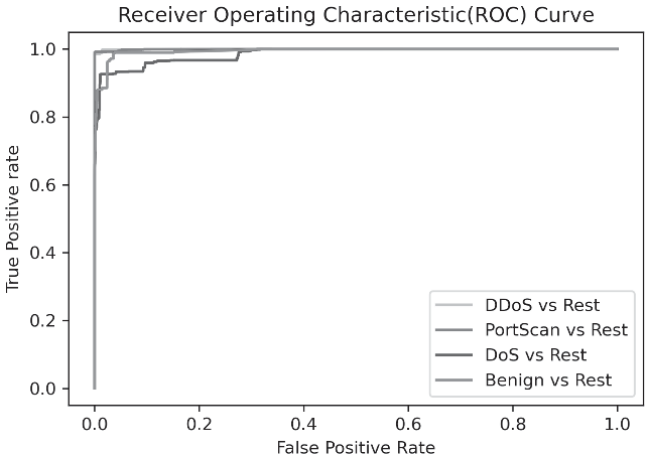


Fig. 4. ROC curve of Random Forest

The figure above shows the Receiver operating characteristic (ROC) curve of the multilabel CICIDS-2017 dataset. The OneVsRest algorithm was used to generate the ROC curve. Thearea under the curve (AUC) of DDoS was 0.99, PortScan was 1.0, DoS was 0.99 and Benign was 0.99.

H. Deep Neural Network

The DNN architecture was trained for a maximum of 50 epochs. An early stopping mechanism was used as a callback with validation error as a parameter in minimum mode. For a more generalized model, patience of 3 was added to the callback andthe model was trained to obtain the following results. The confusion matrix and evaluation metrics were observed asshown in the table below.

TABLE 3 CONFUSION MATRIX OF DNN

		Actual Attack			
	Labels	0 DDoS	1 PortScan	2 DoS	3 BENIGN
Predicted Attack	0 DDoS	9799	0	95	82
	1 PortScan	0	10021	9	2
	2 DoS	19	0	10004	67
	3 BENIGN	285	10	256	9351

The above shows the confusion matrix of the DNNs model. The model predicted well with most of the values in the diagonal. Only some values were predicted wrong. Test data appears to be well balanced with almost equal instances.

TABLE 4 EVALUATION METRICS OF DNN

Labels	TP	FP	FN	TN
0 DDoS	9799	177	304	29720
1 PortScan	10021	11	10	29958
2 DoS	10004	86	360	29550
3 BENIGN	9351	551	151	29947

Here type I error for DDoS was found to be 177 out of 10103 instances and the type II error was 304 out of 10103. For the port scan, type I error was found to be 11 out of 10031 and typeII error was 10 out of 10031. For DoS, type I error was found to be 86 out of 10364 and type II error was 360 out of 10364. For benign, type I error was found to be 551 out of 9502 and type II error was 151 out of 9502. The above results show that

port scan was predicted best while BENIGN was predicted worst.

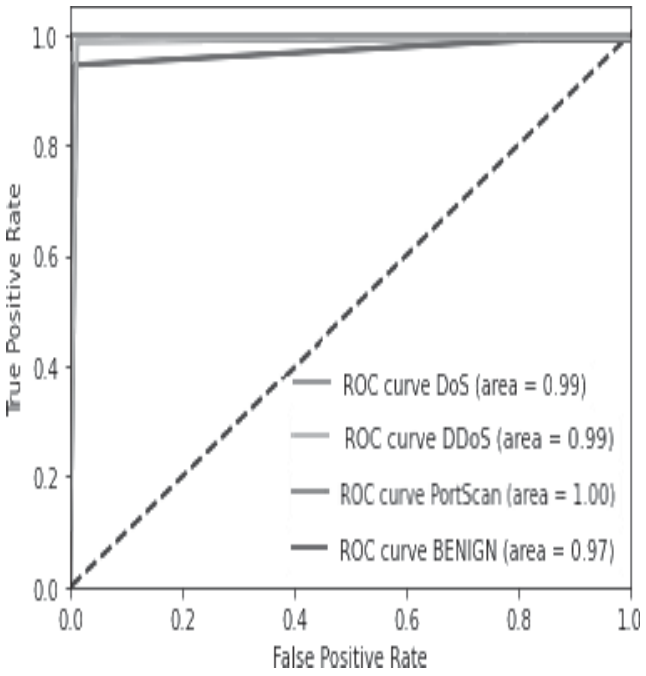


Fig. 5. ROC curve of DNN

The figure above shows the Receiver operating characteristic (ROC) curve of the multilabel CICIDS-2017 dataset. The OneVsRest algorithm was used to generate the ROC curve. Thearea under the curve (AUC) of DDoS was 0.99, PortScan was 1.0, DoS was 0.99 and Benign was 0.97.

I. COMPARISON

Both models were trained and tested using the dataset. The models were compared based on the evaluation metrics such as AUC, F1-Score, Precision, and Recall. Random Forest performed slightly better than Deep Neural Network accordingto the AUC of ROC curves. Both had AUC nearly equal to 1.0. But precision and recall of DNN out-performed Random Forest. F1-score also advocated for DNN as a better performer. Thus, Deep Neural Networks performed better than Random Forests at predicting attacks. The detailed comparison between the two models is shown in the table below with their scores.

TABLE 5

Performance Comparison Between DNN and Random Forest

Attacks / Non-Attack	Random Forest				Deep Neural Network			
	Precision	Recall	F1 Score	AUC	Precision	Recall	F1 Score	AUC
DDoS	0.84	1.00	0.91	1.00	0.97	0.98	0.98	0.99
PortScan	0.96	0.99	0.97	1.00	1.00	1.00	1.00	1.00
DoS	0.88	0.85	0.87	0.99	0.97	0.99	0.98	0.99
BENIGN	0.98	0.79	0.88	0.99	0.98	0.94	0.96	0.97

CONCLUSION

The experiment has demonstrated that there is no single machine learning algorithm that can handle all types of attacks effectively and efficiently, but the Deep Neural Network provides more accuracy for the given CICIDS-2017 dataset. The accuracy of the model can further be increased

by training on multiple data sources. There are only a few datasets for intrusion detection. Training on new datasets in the future and using more advanced algorithms will surely generate a robust, adaptable network Intrusion Detection System.

REFERENCES

[1] I. Sharafaldin, A. H. Lashkari and A. A. Ghorba-ni, “Towards Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization,” *Fourth Internation Conference on Information System Security and Privacy,* 2018.

[2] L. Breiman, “Random Forests,” *Machine Learning,* vol.45, pp. 5-32, 2001.

[3] Y. Bengio, I. J. Goodfellow and A. Courville, Deep Learning, MIT press, 2015.

[4] S. Thaseen and C. A. Kumar, “An Analysis of Supervised Tree Based Classifiers for Intrusion Detection System,” in *International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME)* , 2013.

[5] L. S and S. Wold, “Analysis of variance (ANOVA),” *Chemometrics and Intelligent Laboratory Systems,* vol. 6, no. 4, pp. 259-272, 1989.

[6] M. A. Hall, “Correlation-based Feature Selection forMachine Learning,” 1999.

[7] P. Geurts, Ernst and L. D. & Wehenkel, “Extremelyrandomized trees,” *Mach Learn,* vol. 63, pp. 3-42, 2006.

[8] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” *Psychological Review,* vol. 65, no. 6, pp. 386-408, 1958.

[9] A. Ng, K. Katanforoosh and Y. B. Mourri,“Coursera,” [Online]. Available: https://www.coursera.org/learn/neural-networks-deep- learning? specialization=deep-learning. [Accessed 12 05 2020].

[10] C. Mera and J. W. Branch, “A Survey on Class Imbalance Learning on Automatic Visual Inspection,” *IEEE Latin America Transactions,* vol. 12, no. 4, pp. 657-667, 2014.