Part 2: Machine Learning

Network Capture Analysis

Marc Geggan

CMP417: Engineering Resilient Systems

2022/23

Word count: 2032

**Note that Information contained in this document is for educational purposes.*

+Contents

1 INTRODUCTION

To protect against cyber-attacks from threat actors, the security of a company's network and systems is critical. With the sophistication of cyber attacks increasing, it is crucial that a company's technical department is equipped to evaluate and assess the resilience of its systems to various types of attack. In this report, we focus on one aspect of evaluating a company's security: **Monitoring the network for unusual activity and categorize packet data according to their attack category**. The author aims to design a classifier to categorize the network packet data based on attack types and will explore two machine learning algorithms that are suitable for this task. The goal is to provide an overview of the stages of developing a suitable machine learning model and describe the stages of the data pipeline, such as data ingestion/pre-processing, modelling, analysis, and communication of results. In addition, the author will discuss at least two distinct types of evaluation metrics that can be used to test the performance of the proposed classifier. This paper will provide a thorough examination of the tools and techniques used to analyse a company's resilience to cyber-attacks, with a particular emphasis on network traffic analysis.

2 MACHINE LEARNING ALGORITHMS

2.1 RANDOM FOREST – DECISION TREE

Random Forest is a very popular and simple machine learning algorithm belonging to the supervised algorithm category, used for both regression and classification problems within machine learning. It was first proposed by Leo Breiman in 2001 *(Breiman, 2001)* and works by utilizing both 'bagging' and feature randomness to create an uncorrelated forest of decision trees. Bagging, or bootstrapping, creates subsets of original dataset with replacement to provide different data points to the model and eliminate the chance of receiving the same results.

Instead of depending on one tree, the random forest algorithm takes a prediction from each decision tree that it has created and predicts the result based on the majority outputs of each decision tree. The algorithm uses a flowchart like structure to depict prediction results and feature-based splits. Starting from the root node, the tree splits into decision nodes, and finally into leaf nodes containing the outputs. The use of multiple decision trees allows results to be repeated, combined, and gathered to improve accuracy. The diagram in figure 1 shows a random forest and its decision trees.

Although having many positives, it is also important to discuss the negatives of the algorithm so an informed decision can be made on the most appropriate to use in the model. Random forest algorithm has disadvantages such as requiring significant computational power and time for training due to the ensemble of decision trees, as well as suffering from interpretability issues and difficulty in determining the significance of each variable *(Team, 2020).*

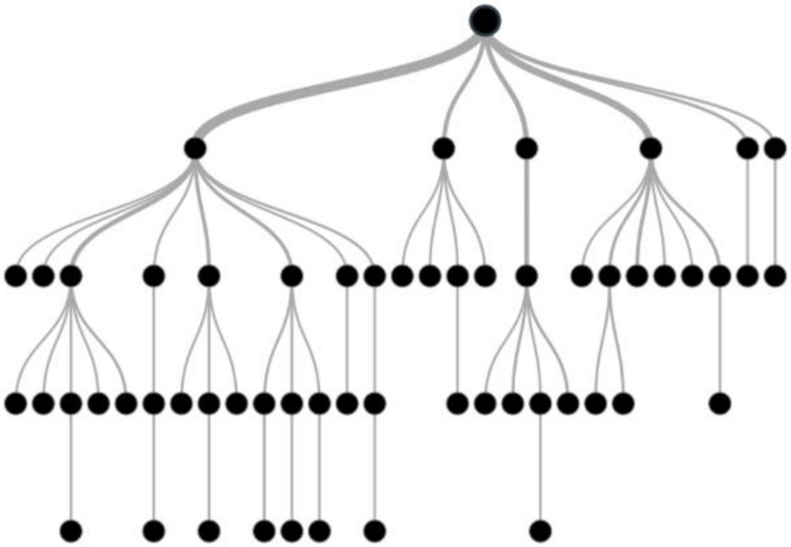


Figure 1: Random Forest graph *(Nicholson, n.d.)*

2.2 MULTI-LAYER PERCEPTRON – NEURAL NETWORK

Multi-layer perception is an artificial neural network algorithm designed around the human brain and its biological neurons, which can be used for both classification and pattern identification. The algorithm generally consists of a network of 'neurons', and a row of neurons is called a layer. The neural network can contain multiple layers, mainly named the Input layer, Hidden layer, and Output layer, which can be seen in figure 2. According to the IBM paper, neural networks operate by giving the input layer weights to judge the importance of each variable. The output is then calculated by multiplying these inputs by their corresponding weights, adding them together, and running the result through an activation function. The node fires and data is sent to the network's next layer if the output rises above a predetermined threshold. The neural network is described as a feedforward network using this approach *(IBM, n.d.).*

Multi-layer perceptron neural networks have several advantages, In addition to being able to simulate non-linear relationships between inputs and outputs, MLP neural networks also offer high predicted accuracy, are resilient to noisy and incomplete data, and automatically extract features. Many different applications, including speech and image identification, natural language processing, and financial forecasting, use MLP neural networks *(Akkaya and Çolakoğlu, 2019).*

However, MLP neural networks also have some disadvantages; they are at risk of overfitting, which happens when the network becomes too complicated and starts to pick up on the noise in the training data rather than the underlying patterns. When the network is used to process fresh data, this may lead to poor performance. Additionally, MLP neural networks require extensive tuning of several kinds of hyperparameters, including the quantity of hidden layers and neurons. Another drawback of MLP neural networks is their sensitivity to feature scaling, which means the performance of the network can be impacted by the size of the input features.

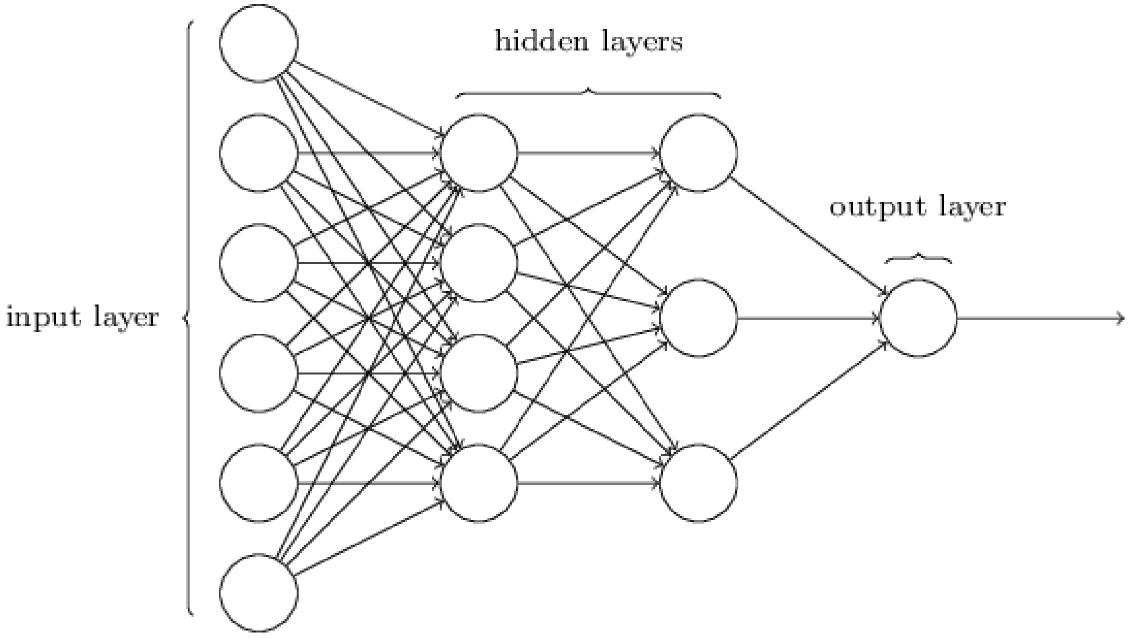


Figure 2: Multi-Layer Perceptron Neural Network *(Rcassani, 2020)*

3 BUILDING APPROPRIATE MACHINE LEARNING MODEL

There are numerous processes that must be followed while developing a suitable model in machine learning. The first phase is data ingestion/preprocessing, which involves ensuring that no data is missing and that it is shaped appropriately. This may involve cleansing, aggregation, labeling, and normalization of data, as well as removing irrelevant and duplicated data. The data used in machine learning models typically comes in the form of a csv file, so it is important that the data is properly separated, and any unrecognized characters and blank columns are removed. This is done by importing specific libraries such as scikit-learn's **sklearn, numpy, matplotlib.pyplot**, and **pandas.** The dataset is then imported and then split into training and test sets.

Once the data has been preprocessed, the next step is to create a model that can learn from the data it is given to make predictions. The model is essentially a file that has been trained to recognize certain types of patterns, and in this case, it will be trained to recognize network packet data. This is done by selecting the right algorithm based on the data and objectives. Random Tree is a popular algorithm that is used for modelling tasks, especially in classification and regression. The data is then provided to the algorithm, of which it will reason over and learn from *(Microsoft, 2021).* To fit the RF algorithm to the model, we must import the **RandomForestClassifier** class from the **sklearn.ensemble library**. The required number of trees must then be set, and the criterion set to entropy. This will analyze the accuracy of the split.

The third phase is analysis, which is arguably one of the most important steps, as it involves evaluating the results of the model. This can be completed using evaluation metrics to test the outcome of performance. This stage may have to be completed multiple times to achieve stable and accurate results. More information on the evaluation metrics for this stage can be found in **Section 4 – Evaluation Metrics.**

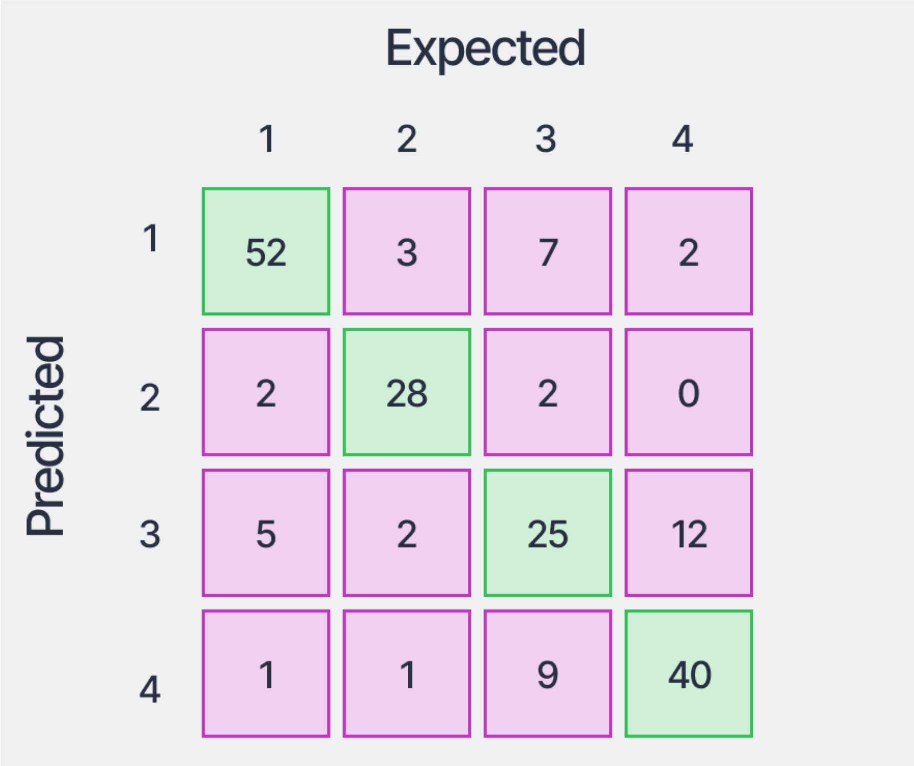
The final phase of building a machine learning mode is the communication of results. This is where the results are modified and represented in a way that is appropriate for the intended audience. One way this can be done is by turning the data into a graph or chart that can be easily interpreted. An example for this specific report would be to turn the different attack categories into a bar chart that displays the number of attacks in each category detected in the network capture. This would be achieved through python code and scikit-learn's library **sklearn.ensemble RandomForestClassifier**, and **matplotlib.pyplot**. Once the model has been trained, the feature importance must be taken and sorted into descending order. This sorted data can then be plotted into a bar chart using **plt** from **matplotlib.pyplot.**

The random forest algorithm is deemed the best fitting for analysing the network capture data as it fits into each stage of the model process. The RF method can handle huge amounts of data in forms such as CSV files. It can also handle complex data with multiple variables, making it useful for intrusion detection. During the analysis phase the RF algorithm's performance can be measured using techniques discussed in section 4. Finally, the algorithm can generate visual representations and reports for the data sets that can be used to display information to target audiences.

4 Evaluation Metrics

To assess the performance of the machine learning model, it is important to use a form of evaluation metric. These metrics are used to compare the predicted outcomes with the actual outcomes to determine how accurate the model is. Different models can be used determined on the type of problem being solved and the goals of the model used. Since Random Tree algorithm is probability based, then confusion matrix and AUC-ROC evaluation metrics may be used.

The confusion matrix presents a table of different outcomes of the prediction and results of the classification to visualize and categorise outcomes (Figure 3). These categories are named 1) True Positive: Number of times positive results = predicted positive result. 2) False Positive = number of times negative values are predicted positive. 3) True Negative = When negative values are predicted negative accurately. 4) False negative = Number of times a positive value is predicted as negative. As the matrix scales with number of variables, the true positives will be visible along the diagonal *(Simplilearn, 2023).*



		Expected			
		1	2	3	4
Predicted	1	52	3	7	2
	2	2	28	2	0
	3	5	2	25	12
	4	1	1	9	40

Figure 3: Basic layout of a Confusion Matrix *(Kundu, 2022)*.

From the confusion matrix, multiple calculations can be performed to understand the accuracy and precision of the model. For example, the accuracy can be determined by the formula below:

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN}$$

Figure 4: Accuracy formula

AUC is an abbreviation for Area Under ROC Curve, and ROC is an abbreviation for Receiver Operating Characteristics. It is known as a scale-invariant evaluation metric since it ranks model predictions rather than providing absolute values. The ROC represents a probability curve, plotted between two parameters, false positive rates (FPR) against true positive rates (TPR). The AUC, or area under curve, is a measure of the overall performance in the model's ability to correctly identify differences between positive and negative results. It scores in a range from 0 to 1, with 0.5 representing random classifier, 1 representing perfect classifier and 0 representing worst performance classifier *(Hoo, Candlish, and Teare, 2017)*.

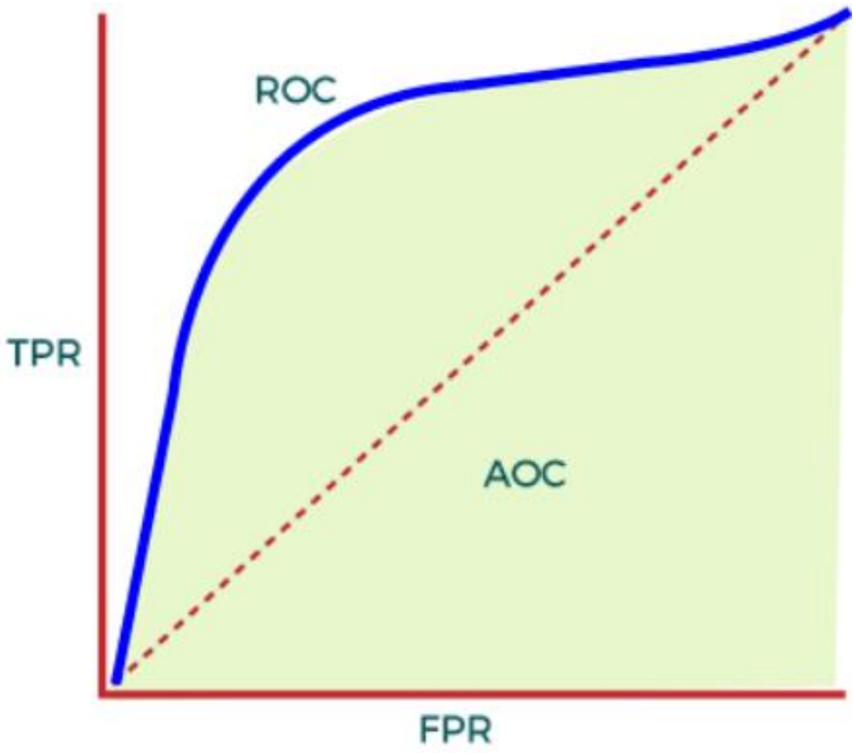


Figure 5: AOC-ROC Curve Diagram *(JavaPoint, n.d.)*

5 CONCLUSION

This research has examined the process of creating an appropriate machine learning model for monitoring network activity and classifying packet data according to the sorts of attacks. Two machine learning techniques—Random Forest and Multi-Layer Perceptron (MLP) neural networks—were highlighted. Random Forest is a popular algorithm that combines multiple decision trees to create a more accurate model. It works by building many decision trees during training and then takes the mean of predictions from the trees to decide an output. Compared to Multi-layer perceptron, random forest provides several advantages, including a lower risk of overfitting and a lower requirement for parameters, making it less complex to run.

Random Forest was the preferred algorithm to MLP in this scenario. Firstly, there were concerns with overfitting the model with a complicated neural network design due to the datasets short size. Secondly, Random Forest is simpler to understand, which has benefits when attempting to understand the predictions made by the model, Finally RF can handle missing values and categorical variables better than MLP and is faster to train.

After the preferred algorithm was decided, the report then outlined the development of a suitable machine learning model. This involved several processes, including data ingestion/pre-processing, where data is shaped appropriately, and duplicated or irrelevant data is removed. The dataset is then split into training and test sets, and once pre-processed, the data will be passed into the random forest algorithm after all correct parameters are set. The results are then evaluated with the use of evaluation metrics. Two types of metrics were discussed in this report – Confusion matrix, and AUC ROC. After the results have been evaluated, the data is displayed in a suitable fashion for the intended audience.

Overall, it was determined that the Random Forest algorithm was most suitable for ScottishGlen's technical staff and their issue of classifying network packet data to understand and categorize different types of attacks. The implementation of this machine learning algorithm within their categorization model will allow ScottishGlen's technical staff to respond more effectively to network security threats and improve overall resilience.

REFERENCES

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), pp.5–32. doi:https://doi.org/10.1023/a:1010933404324.

Hoo, Z.H., Candlish, J. and Teare, D. (2017). What is an ROC curve? Emergency Medicine Journal, 34(6), pp.357–359. doi:https://doi.org/10.1136/emermed-2017-206735.

IBM (n.d.). *What are Neural Networks? | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/neural-networks.

JavaPoint (n.d.). AUC-ROC Curve in Machine Learning - Javatpoint. [online] www.javatpoint.com. Available at: https://www.javatpoint.com/auc-roc-curve-in-machine-learning.

Kundu, R. (2022). Confusion Matrix: How To Use It & Interpret Results [Examples]. [online] www.v7labs.com. Available at: https://www.v7labs.com/blog/confusion-matrix-guide.

Nicholson, C. (n.d.). Decision Tree. [online] Pathmind. Available at: https://wiki.pathmind.com/decision-tree.

QuinnRadich (2021). What is a machine learning model? [online] learn.microsoft.com. Available at: https://learn.microsoft.com/en-us/windows/ai/windows-ml/what-is-a-machine-learning-model.

rcassani (2020). rcassani/mlp-example. [online] GitHub. Available at: https://github.com/rcassani/mlp-example.

simplilearn (2023). What is a Confusion Matrix in Machine Learning? [online] Simplilearn.com. Available at: https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning#:~:text=A%20confusion%20matrix%20presents%20a.

Team, G.L. (2020). Random forest Algorithm in Machine learning: An Overview. [online] Great Learning Blog: Free Resources what Matters to shape your Career! Available at: https://www.mygreatlearning.com/blog/random-forest-algorithm/#advantages-and-disadvantages-of-random-forest

Walch, K. (2021). How to build a machine learning model in 7 steps. [online] SearchEnterpriseAI. Available at: https://www.techtarget.com/searchenterpriseai/feature/How-to-build-a-machine-learning-model-in-7-steps.