Part 2. Machine Learning

Martin Pavlinov Zhelev

CMP417: Engineering Resilient Systems

BSc (Hons) Ethical Hacking Year 4

2023/2024

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

Contents

1 INTRODUCTION

This investigation aims to identify the most optimal Machine Learning algorithm that can be implemented to an Intrusion Detection System (IDS) to improve the cyber resilience of an organisation. To achieve its aim this paper will discuss two potential Machine Learning algorithms and their ability to categorize the network packet data provided by the IDS. During the discussion the strengths, weaknesses and limitations of each algorithm will be examined.

For the purposes of this investigation the researcher has been provided with a dataset that contains information, such as duration, protocol, state, and type of attack about packets that have been captured over a network. The dataset is from the following sources: (Moustafa and Slay, 2015) (Moustafa and Slay, 2016) (Moustafa et al., 2017) (Moustafa et al., 2017) (Sarhan et al., 2020).

Inside the dataset the traffic is categorised based on what caused it. There are ten categories, seven of which are for malicious traffic and three for not malicious. These categories can be seen in Table 1. The machine learning algorithm that gets chosen will need to be able to correctly categorise traffic based on its behaviour into one of the ten categories.

Not malicious	Malicious
Analytics	Backdoors
Generic	Denial of Service
Normal	Exploits
	Fuzzers
	Reconnaissance
	Shellcode
	Worms

Table 1: Attack categories in dataset

After the machine learning algorithm discussion, the researcher will explain how the most appropriate algorithm can be implemented to work with the provided dataset. This will include all stages of the data pipeline, including data ingestion/preprocessing, modelling, analysis, and communication of results.

In the end three different types of evaluation metrics will be analysed to determine their ability to assess the accuracy of the suggested algorithm.

2 BACKGROUND

In this section of the report the researcher will explain what Intrusion detection systems and are and their types. Alongside this they will go over what machine learning is and the several types of machine learning.

2.1 INTRUSION DETECTION SYSTEMS

“An intrusion detection system (IDS) is a network security tool that monitors network traffic and devices for known malicious activity, suspicious activity or security policy violations” (IBM, n.d.). When they detect such issues, they alert the security team, with the goal of preventing malicious actors from causing damage to the system. The most common IDS types are Network intrusion detection systems (NIDS) and Host-Based intrusion detection systems (HIDS). (Barracuda, n.d.)

2.1.1 Network intrusion detection systems (NIDS)

Network intrusion detection systems monitor the network traffic for any signs of malicious activity. Usually, they are placed behind the firewall to check any traffic that goes through it. If they detect anything malicious, they alert the security team. (IBM, n.d.)

2.1.2 Host-Based intrusion detection systems (HIDS)

Unlike NIDS Host-Based intrusion detection systems are installed on endpoints like laptops, routers, or servers. They monitor the device’s activity by taking periodic snapshots of system files and alerts the security team if any changes are detected. (IBM, n.d.)

2.2 MACHINE LEARNING

Machine learning is a type of artificial intelligence (AI) that focuses on levering data and algorithms to allow AI to mimic human learning through iterations. (IBM, n.d.)

There are four types of machine learning categories based on the type of learning method used for the training of the algorithm, which are called semi-supervised, unsupervised and reinforcement machine learning. (SAS, n.d.)

2.2.1 Supervised

This type of machine learning uses labelled datasets to train the algorithm’s ability to predict outcomes. During supervised learning there is a supervisor who trains the machine learning algorithm with well-labelled data. After the algorithm is trained on the labelled data, it is provided with new set of data that is unlabelled to see the prediction it makes as output (GeeksforGeeks, 2024). This process can be seen in Figure 1 below.

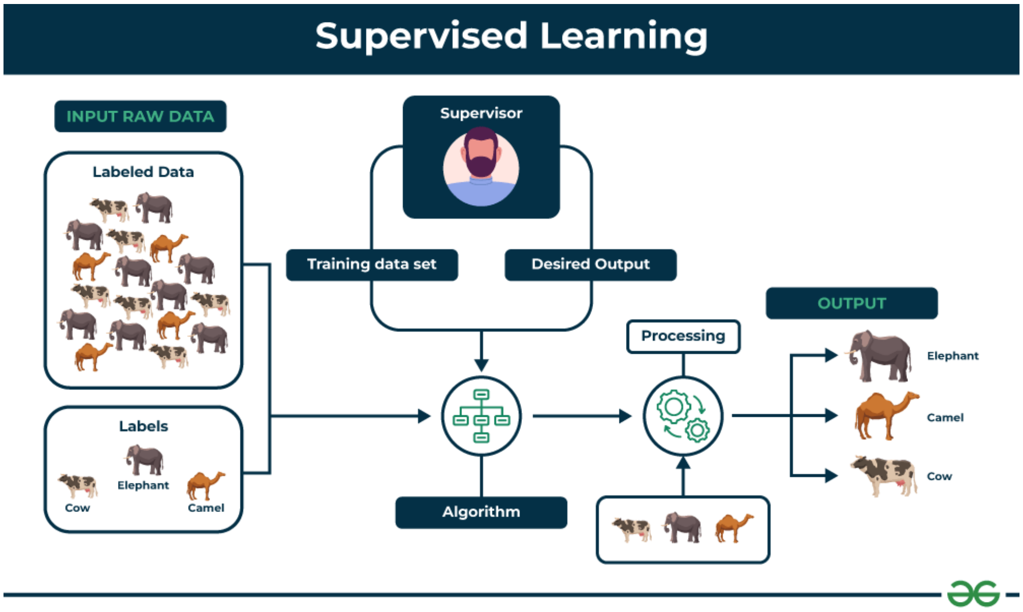

Figure 1: Supervised machine learning diagram. (GeeksforGeeks, 2024)

2.2.2 Semi-Supervised

Semi-supervised machine learning involves the use of a small size of labelled data alongside the unlabelled data for the training of the model. This allows it to make predictions about unlabelled data based on the partially labelled data. This process can be seen in Figure 2 below.

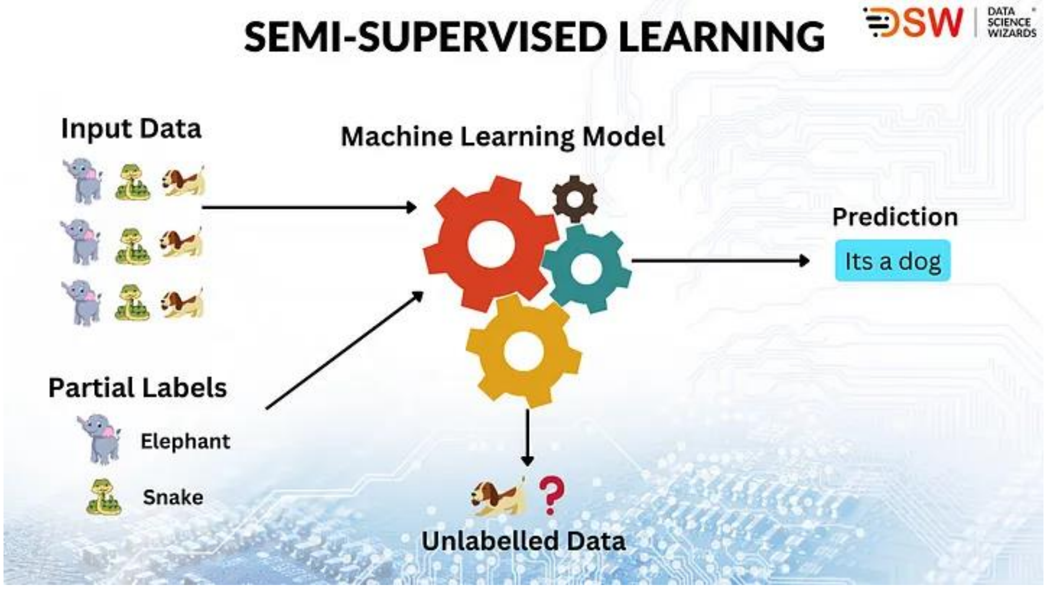

Figure 2: Semi-supervised machine learning diagram. (Data Science Wizards, 2023)

2.2.3 Unsupervised

Unlike in supervised learning during unsupervised learning the machine learning model learns from unlabelled data. Because of that the model will discover patterns and relationships in the data without any guidance (GeeksforGeeks, 2024). This process can be seen in Figure 3 below.

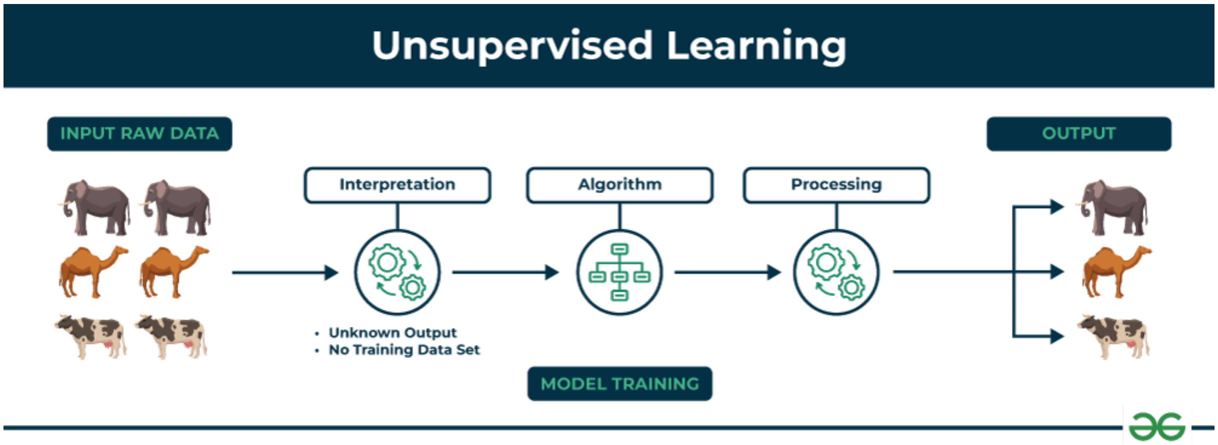

Figure 3: Unsupervised machine learning diagram. (GeeksforGeeks, 2024)

2.2.4 Reinforcement

Reinforcement machine learning is a type of machine learning where the model gets trained through trial and error, without having any instructions provided. This is done by training an agent that learns what are the best possible interactions with an environment by receiving feedback in the form of positive and negative feedback. (PythonGeeks, n.d.)

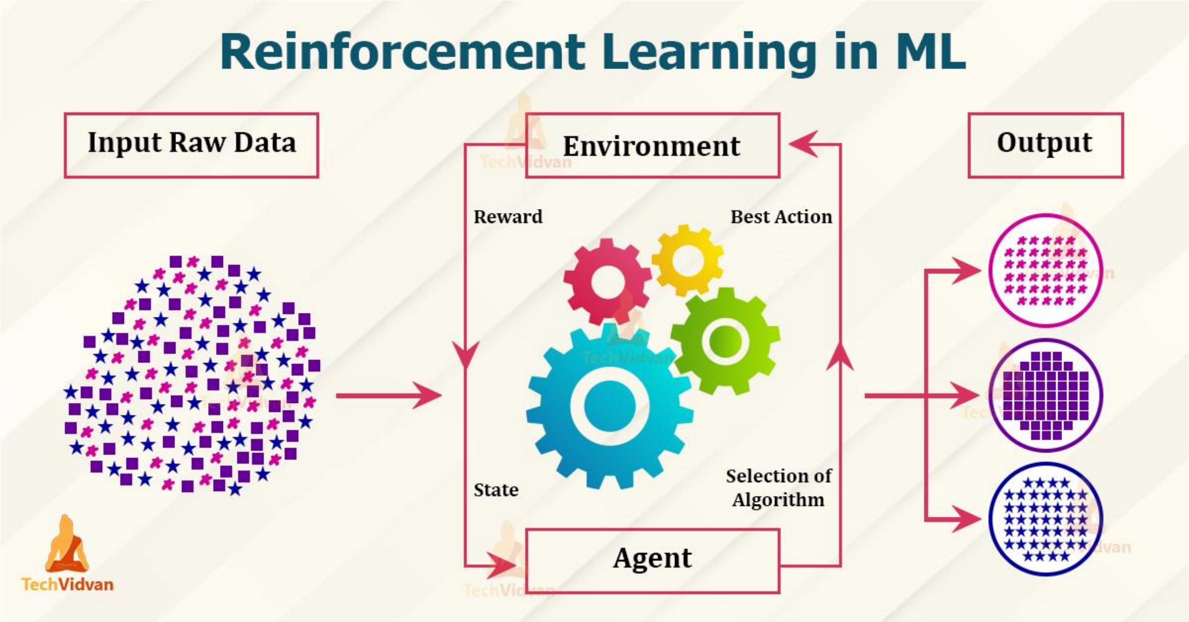

Figure 4: Reinforcement machine learning diagram (Kadari, 2024)

3 ALGORITHMS

In this section three prominent machine learning algorithms called “Decision Trees”, “Random Forest” and “K-Means Clustering” will be discussed. An analysis of their respective strengths and weakness will be performed and the most suitable machine learning algorithm for the current investigation will be chosen.

3.1 DECISION TREES AND RANDOM FOREST

Decision trees are a supervised learning algorithm that can predict variables for regression problems or split the provided data into categories for classification problems. This is done using root nodes, branches, internal nodes, and leaf nodes. The root node is the dataset. It is split into internal nodes that make a split depending on different conditions into further internal nodes or a final decision called a leaf node (IBM, n.d.). The structure of a decision tree is similar to a flow chart and can be seen in Figure 5 below.

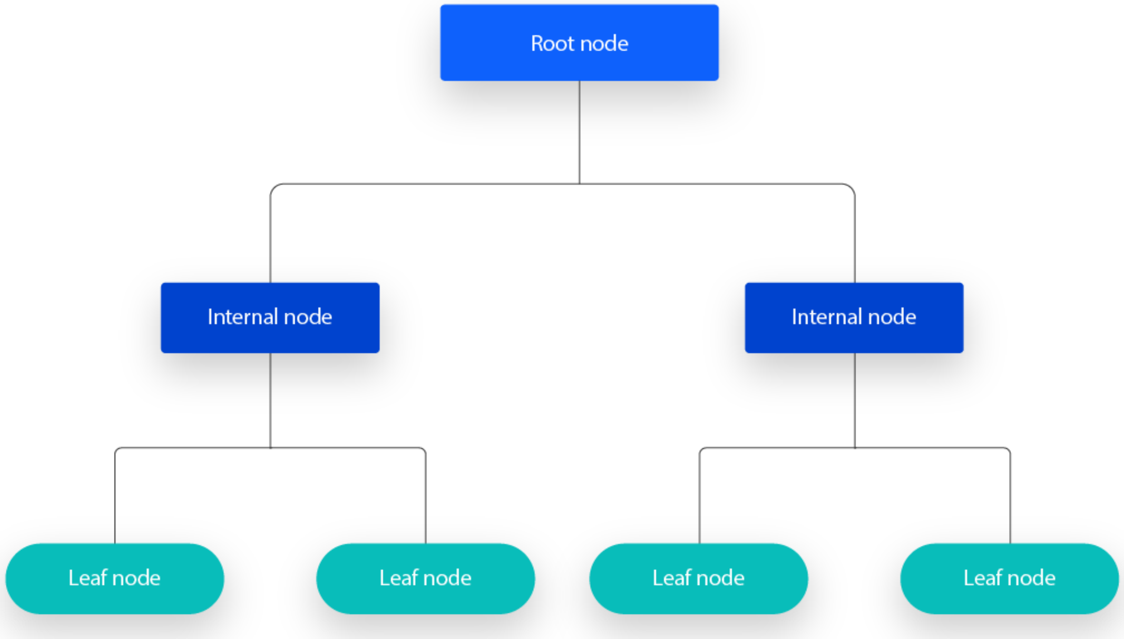

Figure 5: Decision tree diagram. (IBM, n.d.)

Random forest is a type of supervised machine learning algorithm that uses the mean output of multiple “decision trees” to reach the desired result. When using multiple decision trees that are uncorrelated to each other in the algorithm the machine can achieve more accurate results (IBM, n.d.). The structure of a random forest can be seen in Figure 6 below.

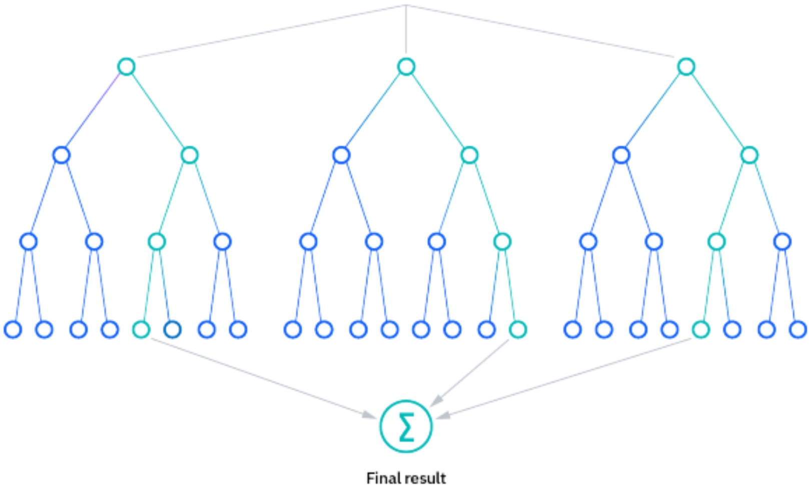

Figure 6: Random Forest diagram (IBM, n.d.)

The random forest algorithm has several benefits such as its high accuracy, versatility, and efficiency in working with large data sets, which would make it suitable for an intrusion detection machine learning model. Alongside this, it is a model that is resistant to outliers, can scale up easily, and, as such, will be able to grow alongside the organization. However, they are negatives in this model such as depending on the size of dataset used it requires a significant amount of time and computation power to compute the numerous decision trees. (Great Learning, 2023)

3.2 K-Means Clustering

K-Means clustering is an unsupervised machine learning algorithm that works by grouping similar objects into K number of clusters in such a way that ”the sum of the squared distances between the objects and their assigned cluster mean is minimized” (Sharma, 2024)

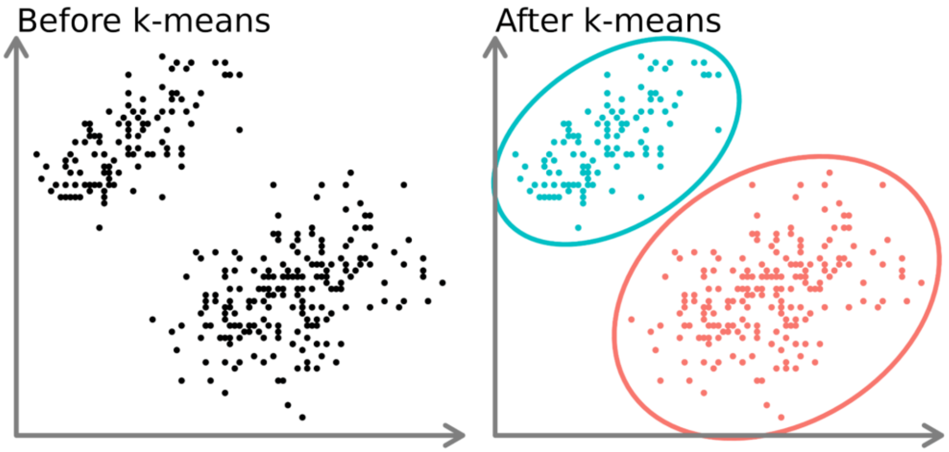

Figure 7: Before and after K-Means clustering (DataCamp, 2023)

This machine learning model also has the advantages of being able to scale up as the size of the organisation increases, but unlike random forest it is not that computationally expensive and as such could run closer to real time. However, this algorithm unlike random forest has the negative that it is sensitive to outliers which can end up creating their own clusters. This is an issue because outliers may be present in a real-world IDS alert dataset. (Banoula, 2023)

3.3 Summary

Decision trees work by using structures similar to flowcharts to get results. They are a straight-forward way of predicting values or categorising data but are prone to inaccuracy.

On the other hand, random forest is also a supervised machine learning algorithm, but they use multiple decision trees to get results, which makes the machine learning model more robust, better performing with large datasets and more resistant to outliers. However, they have the negative of requiring high amount of computational power.

Finally, K-Means clustering is an unsupervised machine learning algorithm, which works by grouping data into clusters. It is a lot of less computationally expensive, which can a lot it to run much closer to real time compared to random forest, however it is sensitive to outliers.

For the purposes of this investigation the researcher decided to focus on the random forest machine learning model, because they were provided with prelabelled data. Alongside this while working slower, random forest can provide higher accuracy and is more resilient to outliers which are highly likely to be present in the alert output from an IDS environment.

4 IMPLEMENTATION

In this section the steps required to implement a random forest machine learning model will be described.

4.1 DATA INGESTION/PREPROCESSING

Data Ingestion/Preprocessing is the first step that needs to be taken for implementing a machine learning algorithm. To do this the data from the dataset needs to be analysed to decide which variables are important and necessary for training the algorithm. This is done so only independent variables such as protocol, state, date, service and etc. remain alongside the attack category which is the dependant variable. Independent variables are the cause, while depended on variables are the outcome.

To finish preprocessing the dataset needs to be organised into data for testing and data for training, however that was already done in the dataset provided for this research.

4.2 ALGORITHM TRAINING

The second step in this process is fitting the algorithm to the dataset to allow it to train effectively. To do this the number of trees that will be used for the random forest needs to be determined as well as their maximum depth and number of features for splitting at each node. Finally, the criterion used to determine the quality of each split in the output needs to also be added.

4.3 ALGORITHM ANALYSIS

The third step is testing the algorithm with the provided testing dataset to determine its accuracy of detecting malicious traffic and classifying its attack category. This step is especially important for the successful implementation of the algorithm because it determines its viability. The evaluation metrics used for this will be described in Section 5 - Evaluation Metrics.

4.4 COMMUNICATION OF RESULTS

The last step is the communication of the results from running the algorithm. This can involve sending the output to the IDS to allow it to perform certain actions such as changing firewall rules, based on the results. However, how such a process will be implemented is out of scope for this research.

5 Evaluation Metrics

Evaluation metrics are used for figuring out the performance of machine learning models. They different statistical data that help with the the selection of models and the tuning their parameters. (DeepAI, n.d.)

There is a lot of various kinds of evaluation metrics, however for the purposes of these research three of them called “Confusion Matrix,” “Accuracy Rate” and “F1 Score” will be discussed.

5.1 Confusion Matrix

Confusion matrix is a performance measuring method of the output of the machine learning algorithm (Kundu, 2022). An example confusion matrix can be seen in Table 2 below.

		Actual Values	
		Positive	Negative
Predicted Value	Positive	TP	FP
	Negative	FN	TN

Table 2: Confusion Matrix

The metrics present in the matrix are the following:

- TP – True positives
- TN – True negatives
- FP – False positives
- FN – False negatives

It holds the information about the amount of each one of those four metrics from the output of the model. Using this matrix, various evaluation metrics such as “Accuracy rate” and “F1 score” can be calculated".

5.2 Accuracy Rate

The accuracy rate is determined by calculating what percentages of the total number of evaluations the model got right. This provides us a percentage of the model’s successful predictions (Agrawai, 2024). The formula used for doing this calculation can be seen below.

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

5.3 F1 SCORE

The F1 score metric provides an average of the scores from the pression and recall scores of the machine learning model. Precision is the percentage of how many of the positive predictions are correct, while recall is the percentage of true positives from data, that the model was able to identify (Kundu, 2022). The formulas for both can be seen below.

$$Precision\% = \frac{TP}{TP + FP} \quad Recall\% = \frac{TP}{TP + FN}$$

Because the F1 score takes both the precision and recall into consideration, to get a high F1 score the model needs to score a high percentage in both. This makes it useful in research context, because it provides better means of evaluation a models performance compared to simple Accuracy rate (Kundu, 2022). The formula for the F1 score can be seen below.

$$F1 = 2\ x\ \frac{Precision\ x\ Recal}{Precision + Recall}$$

In the case of this machine learning model "F1 score" would be more suitable because it provides a more correct metric of the machine learning model's performance and thus would give more insight into any issues during training.

6 References

Agrawai, S.K. (2024) *Evaluation Metrics For Classification Model - Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/ (Accessed: 18 March 2024).

Banoula, M. (2023) *K-means Clustering Algorithm: Applications, Types, and Demos [Updated] | Simplilearn*. Available at: https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm (Accessed: 17 March 2024).

Barracuda (n.d.) *What is an Intrusion Detection System? | Barracuda Networks*. Available at: https://www.barracuda.com/support/glossary/intrusion-detection-system (Accessed: 15 March 2024).

Data Science Wizards (2023) *What is Semi-Supervised Learning? A Guide for Beginners | by Data Science Wizards | Medium*. Available at: https://medium.com/@datasciencewizards/what-is-semi-supervised-learning-a-guide-for-beginners-a7452a597b8c (Accessed: 16 March 2024).

DataCamp (2023) *Introduction to k-Means Clustering with scikit-learn in Python | DataCamp*. Available at: https://www.datacamp.com/tutorial/k-means-clustering-python (Accessed: 17 March 2024).

DeepAI (n.d.) *Evaluation Metrics Definition | DeepAI*. Available at: https://deepai.org/machine-learning-glossary-and-terms/evaluation-metrics (Accessed: 18 March 2024).

GeeksforGeeks (2024) *Supervised and Unsupervised learning - GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/supervised-unsupervised-learning/ (Accessed: 16 March 2024).

Great Learning (2023) *Random forest Algorithm in Machine learning | Great Learning*. Available at: https://www.mygreatlearning.com/blog/random-forest-algorithm/ (Accessed: 17 March 2024).

IBM (n.d.) *What is a Decision Tree? | IBM*. Available at: https://www.ibm.com/topics/decision-trees (Accessed: 16 March 2024).

IBM (n.d.) *What is an intrusion detection system (IDS)? | IBM*. Available at: https://www.ibm.com/topics/intrusion-detection-system (Accessed: 15 March 2024).

IBM (n.d.) *What is machine learning (ML)? | IBM*. Available at: https://www.ibm.com/topics/machine-learning (Accessed: 16 March 2024).

IBM (n.d.) *What is Random Forest? | IBM*. Available at: https://www.ibm.com/topics/random-forest (Accessed: 17 March 2024).

Kadari, P. (2024) *What is Reinforcement Learning and How Does It Work*. Available at: https://www.analyticsvidhya.com/blog/2021/02/introduction-to-reinforcement-learning-for-beginners/ (Accessed: 16 March 2024).

Kundu, R. (2022) *Confusion Matrix: How To Use It & Interpret Results [Examples]*. Available at: https://www.v7labs.com/blog/confusion-matrix-guide (Accessed: 19 March 2024).

Kundu, R. (2022) *F1 Score in Machine Learning: Intro & Calculation*. Available at: https://www.v7labs.com/blog/f1-score-guide (Accessed: 19 March 2024).

Moustafa, N., Creech, G. and Slay, J. (2017) 'Big Data Analytics for Intrusion Detection System: Statistical Decision-Making Using Finite Dirichlet Mixture Models', *Data Analytics and Decision Support for Cybersecurity: Trends, Methodologies and Applications* pp.127-56.

Moustafa, N. and Slay, J. (2015) 'UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set).' *2015 military communications and information systems conference (MilCIS)*. Canberra IEEE, pp.1-6.

Moustafa, N. and Slay, J. (2016) 'The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set', *Information Security Journal: A Global Perspective*, 25(1-3), pp.18-31.

Moustafa, N., Slay, J. and Creech, G. (2017) 'Novel Geometric Area Analysis Technique for Anomaly Detection Using Trapezoidal Area Estimation on Large-Scale Networks', *IEEE Transactions on Big Data*, 5(4), pp.481-94.

PythonGeeks (n.d.) *Reinforcement Learning in Machine Learning - Python Geeks*. Available at: https://pythongeeks.org/reinforcement-learning-in-machine-learning/ (Accessed: 16 March 2024).

Sarhan, M., Layeghy, S., Moustafa, N. and Portmann, M. (2020) 'NetFlow Datasets for Machine Learning-Based Network Intrusion Detection Systems.' *Big Data Technologies and Applications: 10th EAI International Conference, BDTA 2020, and 13th EAI International Conference on Wireless Internet, WiCON 2020*. Virtual Event Springer International Publishing., pp.117-35.

SAS (n.d.) *A guide to the types of machine learning algorithms | SAS UK*. Available at: https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html (Accessed: 16 March 2024).

Sharma, P. (2024) *The Ultimate Guide to K-Means Clustering - Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/ (Accessed: 18 March 2024).