IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 10, No. 4, August, 2023, pp. 1125-1134
Journal homepage: http://iieta.org/journals/mmep

Evaluation DDoS Attack Detection Through the Application of Machine Learning Techniques on the CICIDS2017 Dataset in the Field of Information Security

Ali Saadoon Ahmed[1*], Sefer Kurnaz[2], Arshad M. Khaleel[3]

[1] Faculty of Computer Science, Department of Computer Science, Al-Maarif University College, Al-Anbar, Ramadi 31001, Iraq
[2] Faculty of Engineering and Architecture, Department of Electrical and Computer Engineering, Altınbaş Üniversitesi, Bağcılar 34217, Turkey
[3] International Smart Card, Iraq Ministry of Education, Al-Anbar, Ramadi 31001, Iraq

Corresponding Author Email: ali.sadoon@uoa.edu.iq

https://doi.org/10.18280/mmep.100404

Received: 18 December 2022
Revised: 19 January 2023
Accepted: 25 January 2023
Available online: 30 August 2023

Keywords:
IDS threats, Python environment, DoS attacks, algorithms

ABSTRACT

Amongst network and Intrusion Detection System (IDS) threats, Distributed Denial of Service (DDoS) attacks often take precedence due to their significant potential to disrupt services, leading to financial and reputational damages for organizations. This study employs eight advanced machine learning techniques to distinguish between two types of DDoS attacks: DoS Hulk and DoS Slow HTTP Test. The applied algorithms include Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), AdaBoost, Naive Bayes (NB), Extreme Gradient Boosting (XGB), Ridge regression, and Multilayer Perceptron (MLP). Utilizing a Python environment, these methods were applied to the DDoS attacks in the CICIDS2017 dataset for classification into benign or DoS categories across two distinct experiments. The results were highly encouraging: The first experiment achieved an accuracy rate exceeding 99%, while the second experiment achieved a perfect success rate of 100%. These findings outperform those of previous studies in terms of their efficiency, demonstrating the potential of these machine learning techniques in enhancing DDoS attack detection.

1. INTRODUCTION

Over recent decades, the global internet database and international web networks have experienced substantial expansion, leading to an exponential increase in data and user engagement in cyberspace. However, this dramatic growth has concurrently amplified the prevalence of cyber-attacks, threats, and severe damage to databases and critical information systems associated with businesses and organizations connected to the internet. This widespread harm to global cyber communication protocols often results from unauthorized access to crucial data, leading to damage, theft, loss, or deletion of critical information, thereby disabling organizational networks. Unauthorized access to such valuable information often has severe, far-reaching consequences. Figure 1 illustrates an example of risk management in information security, which involves preventing or reducing the probability of inappropriate data access, misuse, disclosure, deletion, or devaluation [1].

The motivation of this article lies in addressing the research gap, reflected in the paucity of peer-reviewed papers and academic publications in the existing literature that explore the significant contributions of novel Artificial Neural Network (ANN), Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and Internet of Things (IoT) principles in detecting DDoS attacks and identifying severe cyber threats. These tools offer a higher degree of accuracy, effectiveness, and time-efficiency compared to conventional detection methods. While the global literature acknowledges the crucial benefits of these contemporary concepts, their relevance has been primarily discussed in other disciplines, such as robotic engineering, data mining, and pattern classification.

Furthermore, there is a dearth in the international literature of pivotal papers and research publications investigating the application of progressive ANN, AI, ML, DL, and IoT strategies and intelligent approaches to identify severe cyber threats and DDoS attacks in internet networks. This study aims to bridge these research gaps through numerical analysis and mathematical simulation, utilizing code development in the Python software package. In the results section, it will be observed that the application of novel ANN, AI, ML, DL, and IoT approaches can provide effective strategies and functional methods to accurately identify complex cyber threats and elusive DDoS internet attacks with superior performance and speed compared to traditional cyber threat detection methods.

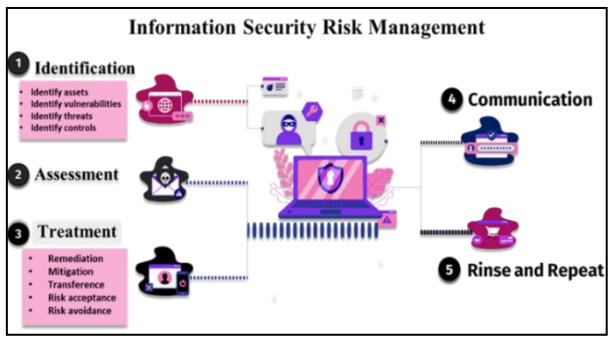

Figure 1. Information security process

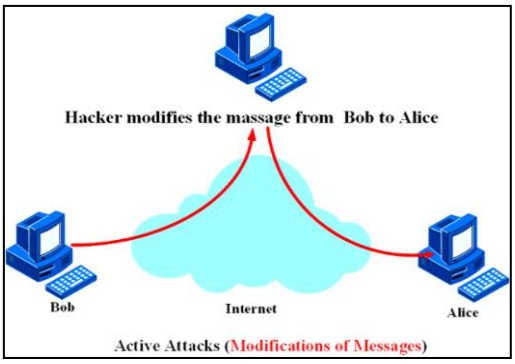

Figure 2. Active attacks in information security

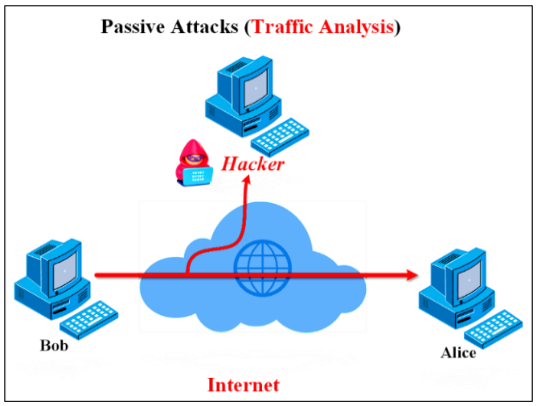

Figure 3. Passive attacks in information security

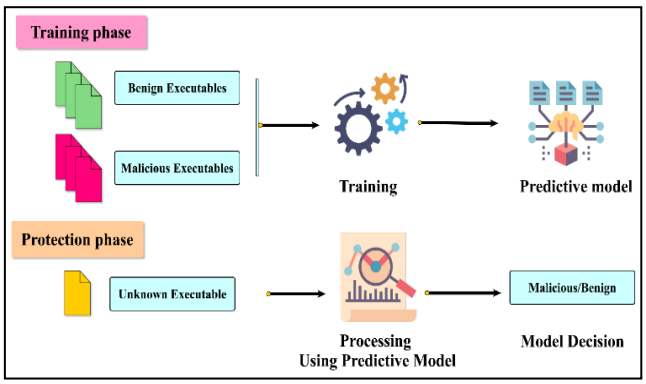

Figure 4. Machine learning in information security

Protected information, whether tangible (e.g., paper documents) or digital (e.g., data stored on computers), underscores the importance of effective policy execution without compromising organizational efficiency [2]. Any unauthorized or unwarranted attempt to access, modify, destroy, delete, implant, or disclose such information constitutes an attack, posing a significant threat to information security. Both individuals and organizations are susceptible to such threats.

Attacks can manifest in a myriad of forms, including but not limited to passive, active, targeted, Trojan horse, brand impersonation, botnet utilization, phishing, spamming, internal, or external [3]. An active attack seeks to alter system resources or disrupt their functions, as depicted in Figure 2. Such attacks may involve alteration of data streams or creation of false statements and can assume various forms, including masquerade, message modification, repudiation, replay, and Denial of Service (DoS).

Contrarily, passive attacks aim to extract information from the system or exploit it without impacting its resources or causing changes, modifications, or disruptions in the associated databases of businesses and organizations, as illustrated in Figure 3. When these severe cyber threats induce changes or defects in the database and web information of a particular website or business, they are classified as 'active attacks' rather than 'passive.' Such active attacks often result in vandalism and damage, causing significant harm to organization or business databases, which are typically characterized by high levels of confidentiality and privacy [4].

As shown in Figure 4, machine learning can be implemented in information security systems to analyze trends and learn from them, thereby aiding in the prevention of recurring attacks and response to evolving behaviors [5]. This can enable information security teams to be more proactive in threat prevention and real-time attack response, helping firms strategically allocate resources and reduce time spent on routine tasks. In essence, machine learning holds the potential to make information security more efficient, proactive, successful, and cost-effective [5, 6]. However, its effectiveness is contingent upon the quality of underlying data - as the adage goes, "garbage in, garbage out" [5].

The subsequent sections delve deeper into the importance of modern concepts and intelligent approaches for detecting cyber threats associated with DoS and DDoS attacks:

Section 2 reviews related works, examining recent articles discussing the critical contributions of modern concepts in DoS and DDoS threat detection.

Section 3 outlines the proposed methodology employed by the author to conduct the Python numerical analysis.

Section 4 presents the significant experimental results and discussions emanating from this research.

Finally, Section 5 provides a summary of the primary conclusions drawn from the results of this numerical research.

2. RELATED WORK

Malliga et al. [7] explored the nature and characteristics of DoS/DDoS attacks, which aim to overwhelm a target with redundant traffic. This understanding is vital as the tactics and tools utilized in such attacks are constantly evolving. The study proposed a taxonomy for identifying DoS/DDoS attacks and provided deep learning techniques for their detection. It also discussed contemporary defenses against DoS/DDoS attacks that leverage deep learning algorithms, and the factors contributing to their effectiveness. Given the importance of datasets for deep learning methods, the paper also examined both historic and current datasets documenting DDoS and DDoS-like attacks. The study concluded by emphasizing the importance of improving existing state-of-the-art strategies to counteract the unpredictable behaviors of attackers.

Le et al. [8] noted the increasing number of individuals accessing the internet due to the widespread availability of smartphones and tablets. They pointed out that most online attacks on servers, websites, and services take the form of distributed denial of service attacks. These attacks quickly deplete resources due to the massive demand from numerous attackers. Thus, it's critical to identify these attacks and prevent network security breaches before they occur. The study carried out a comparison of supervised learning-based DDoS detection methods using the CIC-IDS 2017 dataset, employing various performance metrics to compare the ability

of different methods to detect DDoS attacks.

Ahsan et al. [9] highlighted the growing significance of machine learning in cybersecurity. Machine learning aspires to enhance malware detection over existing methods by making it more actionable, scalable, and effective. However, the application of machine learning in the cybersecurity sector requires rigorous theoretical and methodological management. The study focused on the potential of applying machine learning techniques to cybersecurity data to enhance the security of existing infrastructure. It recognized the capacity of machine learning techniques to mitigate cyber threats and examined the evolution of attacks over the past decade and the limitations of these advanced models. Given the increasing concern over keeping the internet safe from malware, this research aimed to evaluate the effectiveness of current machine learning technologies in this area.

Qu et al. [10] discussed the increasing importance of implementing preventative measures to protect data, IT infrastructure, and online resources as the frequency of cyberattacks and network intrusions continues to rise. Organizations are responsible for preventing security incidents and data breaches, as well as monitoring and responding to threats or any actions that could compromise a system or network. This study delves into the intricacies of the intrusion detection system (IDS), a common perimeter security tool. Traditional security practices such as user authentication, access control, virus prevention, firewalls, cryptographic systems, and data encryption may not be sufficient in modern cyber organizations. The scalability of IDSs depends on the network size and file system. However, the most commonly used theoretical intrusion detection methods are misuse detection, signature-based IDS, and anomaly-based IDS. Given the efficacy of misuse detection against prevalent attack vectors, knowledge of intrusive activities is essential. SNORT is one of the most recognized methods for identifying abusive behavior.

Ahsan et al. [11] noted that cybersecurity dataset standards can be inconsistent, noisy, incomplete, irrelevant, or contain inconsistent instances of a specific security violation. These issues can negatively impact both learning and machine learning-based models. It is vital to address these data problems before utilizing machine learning techniques to develop a data-driven cybersecurity solution. Understanding the challenges associated with cybersecurity data and effectively addressing these issues using existing or newly developed algorithms is crucial for operations such as malware and intrusion detection. These data challenges must be rectified before using machine learning techniques to build a data-driven cybersecurity solution. Recognizing the difficulties posed by cybersecurity data and addressing these effectively using either existing or newly developed algorithms is crucial before undertaking tasks like virus and intrusion detection.

Sarker et al. [12] demonstrated that hybrid learning could enhance the performance of signature-based intrusion detection systems, a standard in the cybersecurity industry. However, these algorithms might fail to detect some attacks or events due to missing features, excessive feature reduction, or insufficient profiling. They suggest that these issues can be addressed using anomaly-based or hybrid methods, such as a combination of signature-based and anomaly-based detection strategies.

Guezzaz et al. [13] discussed how services like telephony, Industry 4.0, and industrial IoT have been dramatically transformed with the rise of cloud computing and IoT settings. Ensuring the security of these advanced technologies is crucial. IoT security presents a significant challenge for both commercial entities and academic institutions. The role of an IDS is to monitor behavior, recognize an intrusion in real-time, and respond accordingly. Many modern IDSs leverage machine learning (ML) techniques to improve their detection rate, accuracy (ACC), and precision (DR). In their research, they introduced a hybrid IDS that operates at the network's edge and uses machine learning techniques to secure IoT devices. They combined Principal Component Analysis (PCA) and the K-Nearest Neighbor method (K-NN) to identify anomalies and instances of misuse. While PCA was used to enhance feature engineering and training, the K-NN classifier was developed to improve detection accuracy and expedite decision-making. The results demonstrated the superiority of their proposed framework over alternative approaches, achieving excellent performance on the NSL-KDD dataset with a 99.10% ACC, 98.4% DR, 2.7% False Alarm Rate (FAR), and 99.2%, and on the Bot-IoT dataset with a 98.2%.

To summarize the studies mentioned in this section, the research conducted by various scholars suggests that the practical application of ML and IoT can significantly enhance the accuracy of cyber threat detection and the identification of diverse DDoS attacks, which can be extremely challenging for traditional methods. It can also be inferred that over the past decades, the complexity of cyber threats and their detection has risen remarkably due to the rapid development of technologies and the massive growth of databases. Therefore, innovative methods and intelligent approaches for cyber threat detection should be employed for a more accurate, flexible, and highly effective identification process.

3. PROPOSED METHODOLOGY

Figure 5 depicts the proposed methodology's flowchart, which is used to describe the dataset, feature extraction methods, pre-processing approaches, and the development of eight machine learning algorithms that can detect a wide variety of network attacks that can affect any network.

3.1 Dataset description

This section presents the used dataset in this work to detect the certain Network Attacks in Information Security from a dataset, which is CICIDS2017 dataset [14]. Four days a week were used to acquire this dataset. carries out a distinctive assault and is connected to Brute force, Web-based, Infiltration, Heart-bleed, DDoS, DoS, Bot, and Scan are all forms of cyberattacks that interfere with a system's ability to function normally. According to Table 1, this dataset has 80 attributes and more than 19 million cases labeled with a classification that includes a wide variety of network attacks. That was collected from command prompt shells as shown in Figure 6.

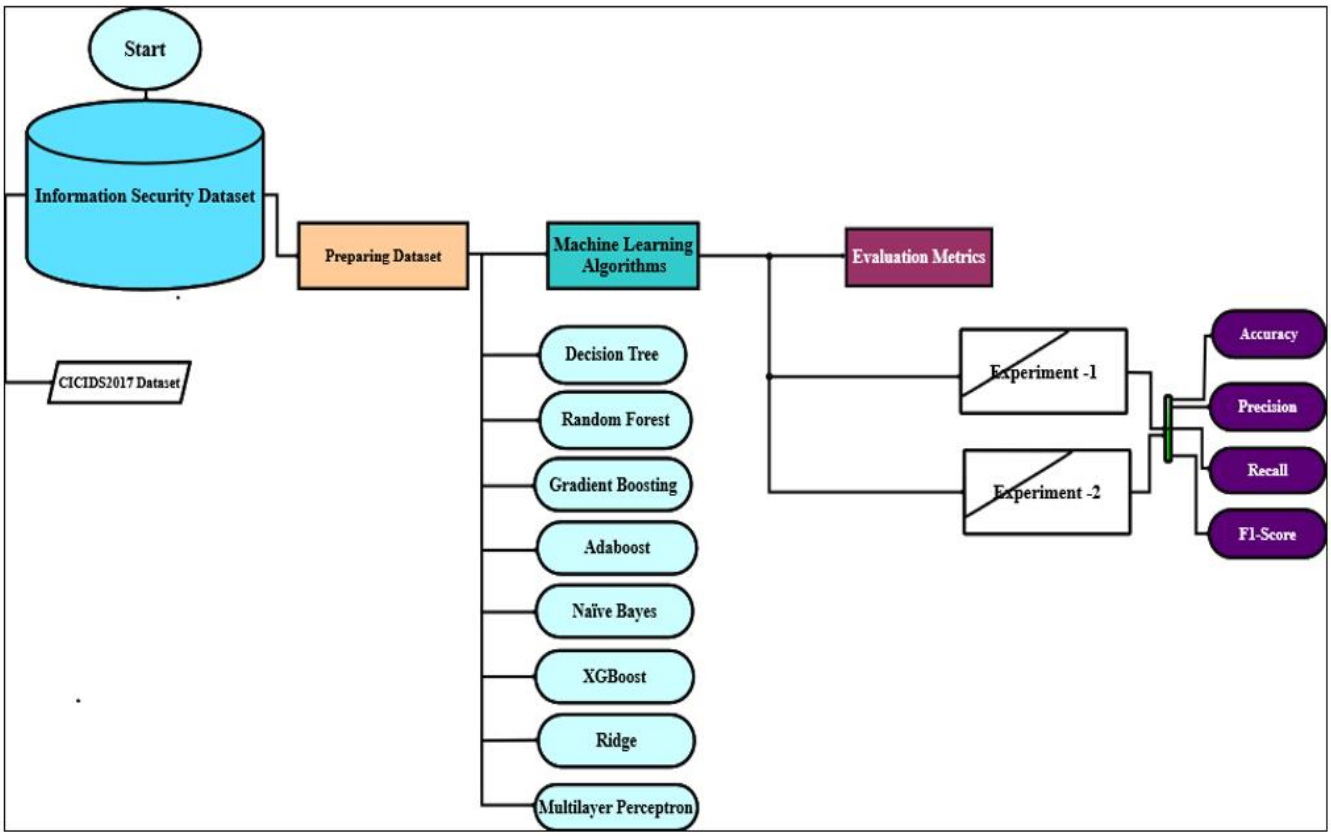

Figure 5. Proposed methodology flow-chart

```
scaler = MinMaxScaler()
cols_to_norm = ['Dst Port', 'Protocol', 'Flow Duration', 'Tot Fwd Pkts', 'Tot Bwd Pkts',
       'TotLen Fwd Pkts', 'TotLen Bwd Pkts', 'Fwd Pkt Len Max',
       'Fwd Pkt Len Min', 'Fwd Pkt Len Mean', 'Fwd Pkt Len Std',
       'Bwd Pkt Len Max', 'Bwd Pkt Len Min', 'Bwd Pkt Len Mean',
       'Bwd Pkt Len Std', 'Flow Byts/s', 'Flow Pkts/s', 'Flow IAT Mean',
       'Flow IAT Std', 'Flow IAT Max', 'Flow IAT Min', 'Fwd IAT Tot',
       'Fwd IAT Mean', 'Fwd IAT Std', 'Fwd IAT Max', 'Fwd IAT Min',
       'Bwd IAT Tot', 'Bwd IAT Mean', 'Bwd IAT Std', 'Bwd IAT Max',
       'Bwd IAT Min', 'Fwd PSH Flags', 'Bwd PSH Flags', 'Fwd URG Flags',
       'Bwd URG Flags', 'Fwd Header Len', 'Bwd Header Len', 'Fwd Pkts/s',
       'Bwd Pkts/s', 'Pkt Len Min', 'Pkt Len Max', 'Pkt Len Mean',
       'Pkt Len Std', 'Pkt Len Var', 'FIN Flag Cnt', 'SYN Flag Cnt',
       'RST Flag Cnt', 'PSH Flag Cnt', 'ACK Flag Cnt', 'URG Flag Cnt',
       'CWE Flag Count', 'ECE Flag Cnt', 'Down/Up Ratio', 'Pkt Size Avg',
       'Fwd Seg Size Avg', 'Bwd Seg Size Avg', 'Fwd Byts/b Avg',
       'Fwd Pkts/b Avg', 'Fwd Blk Rate Avg', 'Bwd Byts/b Avg',
       'Bwd Pkts/b Avg', 'Bwd Blk Rate Avg', 'Subflow Fwd Pkts',
       'Subflow Fwd Byts', 'Subflow Bwd Pkts', 'Subflow Bwd Byts',
       'Init Fwd Win Byts', 'Init Bwd Win Byts', 'Fwd Act Data Pkts',
       'Fwd Seg Size Min', 'Active Mean', 'Active Std', 'Active Max',
       'Active Min', 'Idle Mean', 'Idle Std', 'Idle Max', 'Idle Min',
       'Timestamp_year', 'Timestamp_month', 'Timestamp_week', 'Timestamp_day',
       'Timestamp_hour', 'Timestamp_minute', 'Timestamp_second']
data[cols_to_norm] = scaler.fit_transform(data[cols_to_norm])
data

[12]  ✓ 9.8s
```

Figure 6. Dataset features

Table 1. CICIDS2017 features

No.	Feature	No.	Feature
1	Dst Port	21	Flow IAT Max
2	Protocol	22	Flow IAT Min
3	Timestamp	23	Fwd IAT Tot
4	Flow Duration	24	Fwd IAT Mean
5	Tot Fwd Pkts	25	Fwd IAT Std
6	Tot Bwd Pkts	26	Fwd IAT Max
7	TotLen Fwd Pkts	27	Fwd IAT Min
8	TotLen Bwd Pkts	28	Bwd IAT Tot
9	Fwd Pkt Len Max	29	Bwd IAT Mean
10	Fwd Pkt Len Min	30	Bwd IAT Std
11	Fwd Pkt Len Mean	31	Bwd IAT Max
12	Fwd Pkt Len Std	32	Bwd IAT Min
13	Bwd Pkt Len Max	33	Fwd PSH Flags
14	Bwd Pkt Len Min	34	Bwd PSH Flags
15	Bwd Pkt Len Mean	35	Fwd URG Flags
16	Bwd Pkt Len Std	36	Bwd URG Flags
17	Flow Byts/s	37	Fwd Header Len
18	Flow Pkts/s	38	Bwd Header Len
19	Flow IAT Mean	39	Fwd Pkts/s
20	Flow IAT Std	40	Bwd Pkts/s
41	Pkt Len Min	61	Bwd Byts/b Avg
42	Pkt Len Max	62	Bwd Pkts/b Avg
43	Pkt Len Mean	63	Bwd Blk Rate Avg
44	Pkt Len Std	64	Subflow Fwd Pkts
45	Pkt Len Var	65	Subflow Fwd Byts
46	FIN Flag Cnt	66	Subflow Bwd Pkts
47	SYN Flag Cnt	67	Subflow Bwd Byts
48	RST Flag Cnt	68	Init Fwd Win Byts
49	PSH Flag Cnt	69	Init Bwd Win Byts
50	ACK Flag Cnt	70	Fwd Act Data Pkts
51	URG Flag Cnt	71	Fwd Seg Size Min
52	CWE Flag Count	72	Active Mean
53	ECE Flag Cnt	73	Active Std
54	Down/Up Ratio	74	Active Max
55	Pkt Size Avg	75	Active Min
56	Fwd Seg Size Avg	76	Idle Mean
57	Bwd Seg Size Avg	77	Idle Std
58	Fwd Byts/b Avg	78	Idle Max
59	Fwd Pkts/b Avg	79	Idle Min
60	Fwd Blk Rate Avg	80	Label

Table 2. Attack classes in CICIDS2017 dataset

Attack	Label	Frequency
Benign	Normal	446,772
DoS attacks-Hulk	Malicious	461,912
DoS attacks-SlowHTTPTest	Malicious	139,890

To provide more illustration on the datasets shown in Table 1 and Figure 6, it is vital to note that those types of datasets are designed and formulated to help employ in cyber threats detection and information security depending on well-known websites associated with various datasets in several fields. Some of those websites are called the Kaggle website, which comprises the CICIDS2017 dataset.The morning and afternoon hours of days Tue., Wed., Thu., and Fri. were used to collect the data for this study. Different attacks with related Brute force, Web-based, and Denial of Service (DoS) incidents, Heartbleed, bots, Distributed Denial of Service (DDoS), infiltration, and scanning occur every day. As a result, this dataset, which is displayed in Figure 6, has 80 features, more than 19 million events, and a category label that encompasses several types of network attacks. Because of the huge number of the instances of this dataset, a simple represented sample was selected of the dataset that contains 1,048,574 instances and 80 features with DoS attack as a main type of attack. This huge number of instances will take a lot of time in days to execute and need a Personal Computer of Laptop with high requirements in terms of high RAM, GPU, etc. As shown in Table 2, the class label has two types of DoS attack (DoS Attacks-Hulk, and DoS attacks-SlowHTTPTest) and benign (Normal).

3.2 Requirements run the proposed model

Because of the huge number of instances of this dataset, we selected a represented sample of the dataset that contains 1,048,574 instances and 80 features with a DoS attack as the main type of attack. This huge number of instances will take a lot of time in days to execute. In this proposed system, we recommend using these requirements shown in Table 3. So, you can run the model, in this study, two experiments were used to identify distinct forms of network attacks, and the results were presented in a timeframe that allowed for their consideration.

Table 3. Configuration parameters

Parameter	Value
OS	Win10/MAC/Ubuntu
Programming language	Python version 3.0
Size Dataset Used	1,048,574 rows×80 columns
CPU	Core™ i7-1165G7 Processors
RAM	8 GB
GUI	GeForce RTX 40 Series

3.3 Machine learning types

The aforementioned Network Attacks assaults are identified by first preparing the dataset and then fitting it to eight machine learning techniques. If the sample size is set to.1, We used the hold-out approach to separate the data into a training and testing set. Therefore, the training dataset constitutes 90% of the total datasets while the testing dataset constitutes 10%. It is the purpose of the testing dataset to assess the efficacy of the machine learning models developed with the aid of the training dataset.

The field of machine learning makes use of computational approaches to transform raw data into usable models [15]. As a branch of artificial intelligence, machine learning has its roots in traditional statistical analysis [16, 17]. Thanks to the work of these companies over the past decade, machine learning has risen to prominence as one of the most talked about areas of computer science. In this area, you'll find giants like Google, Microsoft, Facebook, Amazon, and many others. Their operations are already generating massive amounts of data, and this rate is only expected to increase [16]. This has opened the door for the revival of computational and statistical methods for producing robust models automatically. Many machine learning techniques can be used with either non-programmatic apps or API invocations [17], thanks to the availability of open-source implementations. Python, R, Weka, Orange, and Rapid Miner are a few examples of such implementations [18]. Using visual analytics tools like Tableau and Spot fire, the results of these algorithms may be turned into dashboards and operational pipelines [19].

To put it another way, a machine learning program or learning program is a computer program that learns from its own experiences. There are four broad classes of machine learning implementations, each corresponding to a certain type of learning "signal" or "response" that can be received by a

learning system [20] as shown in Figure 7.

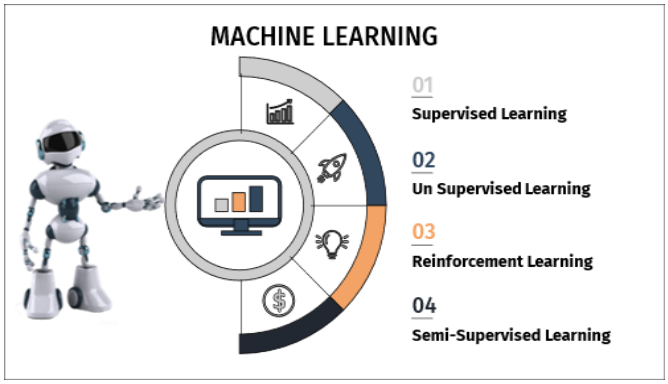

Figure 7. Machine learning types

3.3.1 Supervised learning

This type sometimes called supervised machine learning, is a subset of ML and AI [21]. It stands out from similar approaches because it instructs computers to correctly classify data or predict events using labeled datasets. The weights are changed as additional input data are added to the model until the model is correctly fitted [22], during cross-validation. Companies can scale real-world solutions such as removing spam from authentic emails using supervised learning [23]. The loss function evaluates the algorithm's precision, and iterations are carried out until the error is sufficiently reduced. Classification and regression problems can be separated into two categories when employing data mining for supervised learning [24, 25] (Figure 8).

3.3.2 Unsupervised learning

Using machine learning techniques, unsupervised machine learning classes tagged datasets. Figure 9 [26] illustrates this point. These algorithms find groups or trends in data with help from a human operator. Due to its ability to recognize patterns, it is ideal for exploratory analysis, cross-selling, consumer and segmentation [27]. Models of unsupervised learning can perform operations such as clustering, associating, and dimension reduction. Below, descriptions of each instructional approach are indicated [28, 29]. Clustering is a method used in data mining that involves classifying data sets without labels based on their similarities or differences. Clustering techniques are used to sort disparate data points into related groups so that underlying patterns and structures can be better understood. Clustering algorithms come in various Flavors, including those that are mutually exclusive, overlapping, hierarchical, and probabilistic [30].

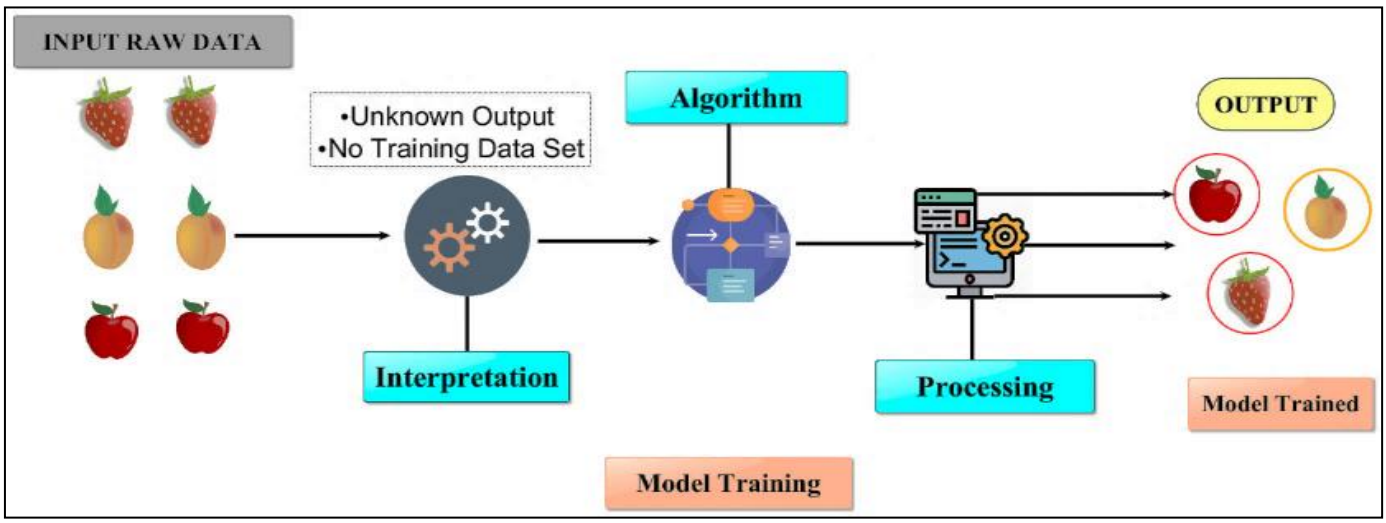

Figure 8. The labeled datasets employing ML methods

3.3.3 Reinforcement learning

Reinforcement learning is one of the three basic machine learning paradigms, along with supervised learning and unsupervised learning, and is a subfield of artificial intelligence that investigates how intelligent agents should act to earn a constant reward. Contrary to supervised learning, reinforcement learning does not call for the explicit correction of undesirable behavior or the presentation of labeled input-output pairings [31]. To the contrary, we need to strike a balance between mining and exploring undiscovered regions. Benefits of both supervised and RL algorithms are incorporated in partially supervised algorithms [32].

3.3.4 Semi-supervised learning

Machine learning method that combines features of both supervised and unsupervised approaches [31]. During the training phase, it makes use of both labelled and unlabeled datasets. Since semi-supervised learning employs pseudo labelling during model training, it requires a smaller labeled training dataset than supervised learning. The procedure allows for the integration of several neural network models and training techniques [31]. To better understand how semi-supervised learning works, we will look at its overall structure and function, which will be detailed below [32].

3.4 ML techniques and algorithms used

As this dataset includes labels, we were able to employ eight supervised machine learning methods. Ridge, Multilayer Perceptron, Gradient Boosting, and Adaptive Boosting are some examples of these techniques, Machine Learning Algorithms such as Naive Bayes, eXtreme Gradient Boosting, Decision Tree, and Random Forest are depicted in Figure 9.

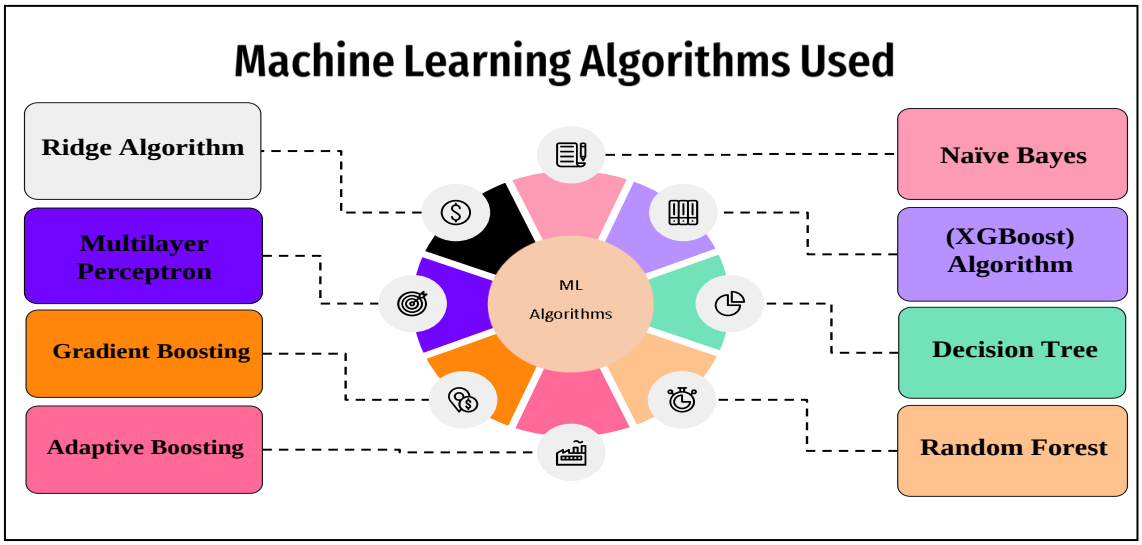

Figure 9. The used of machine learning algorithms

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

In this chapter, we covered the precision, accuracy, recall, and f1-score metrics used to assess the performance of machine learning algorithms. Further, we detailed the results of each machine learning algorithm's experiments (binary classification experiment and Multiclass classification experiment) using the aforementioned four assessment criteria. Finally, we compared our results with those of previous studies with regards to the dataset employed, the preprocessing processes taken, the treatment of categorical and numeric data, the machine learning algorithms employed, and the evaluation metrics utilized to gauge the effectiveness of the algorithms.

4.2 Evaluation metrics

Accuracy, precision, recall, and f1-score are used to evaluate machine learning algorithms.

These metrics' formulas are as follows:

TP=True Positives, TN=True Negative, FP=False Positives, and FN=False Negative.

Following are the equations for these metrics:

Accuracy: The ratio of correctly predicted samples to all samples, or simply the ratio of correctly predicted samples to all samples, is the most intuitive performance metric.

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

Precision can be viewed as the proportion of confirmed positive samples relative to the total number of confirmed positive samples projected.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall is the proportion of accurately anticipated positive samples to the total number of positive samples predicted.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-score is the weighted average of Precision and Recall.

$$F1\text{-}score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

4.3 Our results

The experimental findings from two tests on each machine learning method are presented in this section: determine between a normal attack and a DoS attack in the first experiment and between two DoS attacks in the second experiment. In this dataset, we applied two experiments to detect several types of attacks, Figure 10 shows Total Number of Samples in Each Category in the First Experiment Binary classification (normal and malicious attacks).

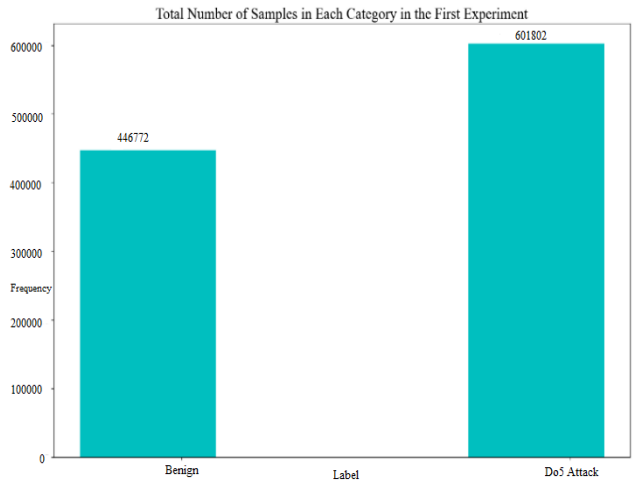

Figure 10. Total number of samples in each category in the first experiment

4.4 Binary classification experiment

Machine learning methods are used in this experiment to determine if the sample in this dataset represents a legitimate attack or a malicious one. The performance results for the machine learning methods utilized are based on four metrics: precision, accuracy, f1-score, and recall, as shown in Table 4

and Figure 11. In these metrics, all algorithms fared better during the detecting procedure. These findings suggest that machine learning algorithms are capable of effectively detecting and differentiating between standard and DoS attacks.

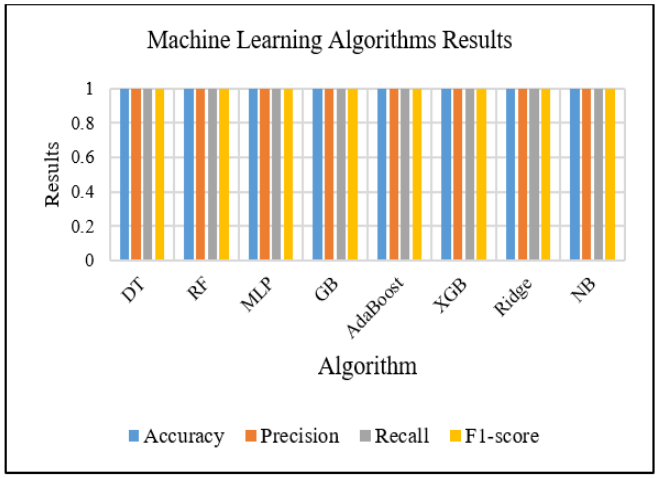

Figure 11. Performance results of binary classification

The experimental results for each machine learning technique utilizing a cybersecurity dataset are summarized in this section based on four evaluation metrics.

Table 4. Performance results of binary classification

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DT	1.0	1.0	1.0	1.0
RF	1.0	1.0	1.0	1.0
MLP	1.0	1.0	1.0	1.0
GB	1.0	1.0	1.0	1.0
AdaBoost	1.0	1.0	1.0	1.0
XGB	1.0	1.0	1.0	1.0
Ridge	1.0	1.0	1.0	1.0
NB	1.0	1.0	1.0	1.0

4.5 Multiclass classification experiment

In this study, machine learning methods were applied to the DoS Attacks-Hulk and DoS Attacks-SlowHTTPTest datasets to identify harmful activities. Precision, accuracy, f1-score, and recall are the four measures that were used to evaluate the performance of the machine learning algorithms, as indicated in Table 5. In terms of these detection metrics, all algorithms performed at or near the top. According to these findings, machine learning algorithms can efficiently identify and differentiate between two types of DoS attacks.

Table 5. Performance results of multiclass classification

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DT	0.99	0.99	0.99	0.99
RF	0.99	0.99	0.99	0.99
MLP	0.99	0.99	0.99	0.99
GB	0.99	0.99	0.99	0.99
AdaBoost	0.99	0.99	0.99	0.99
XGB	0.99	0.99	0.99	0.99
Ridge	0.99	0.99	0.99	0.99
NB	0.99	0.99	0.99	0.99

4.6 Discussions

In this paper, we applied eight ML algorithms to detect the DoS attacks with two types: DT, RF, MLP, GB, AdaBoost, XGB, Ridge, and NB. Then, the findings of two trials conducted for each machine learning algorithm are done to determine between a normal attack and a DoS attack in the first experiment and between two DoS attacks in the second experiment. In this dataset, we applied two experiments to detect several types of attacks: 1) Binary classification (normal and malicious attacks); 2) Multiclass classification (malicious attacks types). These experiments are done to know if each algorithm is able to distinguish between the DoS attack types in the dataset. The results in both experiments show as following: in the first one, all algorithms achieved better performance results in these metrics in the detection process. These results indicate that the machine learning algorithms can detect and distinguish between normal and DoS attacks effectively. While in second experiment, all algorithms achieved the higher performance results in these metrics in the detection process. These results indicate that the machine learning algorithms can detect and distinguish between two DoS attacks effectively.

It is worth mentioning that the major the managerial implications and practical contributions of this research are reflected in helping computer decision-makers, cyber-attacks policy actors, communication experts, cyber security professionals, and network specialists choose the optimum method and the most suitable mechanisms that can be implemented to detect highly complex and challenging DDoS detection threats that cannot be identified easily using conventional approaches. Thus, quick responses and active decisions could be selected with the help of ML and IoT concepts to facilitate the detection of severe DDoS attacks and harmful cyber threats to accomplish higher extents of privacy and security for organizations and individuals' databases.

5. CONCLUSIONS

In this study, a comprehensive analysis was conducted with a focus on two distinct forms of distributed denial of service (DDoS) attacks: DoS Attacks-Hulk and DoS Attacks-Slow HTTP Test. Eight differing machine learning methodologies were utilized for the identification process, including Decision Trees (DT), Random Forest (RF), Gradient Boosting (GB), AdaBoost, Naive Bayes (NB), eXtreme Gradient Boosting (XGB), Ridge, and Multilayer Perceptron (MLP).

The dataset employed for the study was CICIDS2017, accessible via the Information Security category on the Kaggle website. This dataset, comprising millions of instances, represents 80 attributes across 15 different attacks.

These algorithms were subsequently utilized to categorize the instances into one of two categories: DoS attacks or Benign. This classification was performed in two separate experiments. The first experiment centered on binary classification, distinguishing between normal and malicious attacks. The second experiment delved into multiclass classification, discerning various types of malicious attacks. The primary objective of these experiments was to ascertain the capacity of each algorithm to distinguish between different types of DoS attacks in the dataset.

To facilitate the application of diverse machine learning algorithms to this dataset, non-numerical features within the

dataset were converted into numerical features using a popular encoding. The Timestamp feature, containing detailed temporal information for each instance, was processed to extract numerical features due to the inherent incompatibility of machine learning algorithms with time formats. Punctuation marks such as dots (.), dashes (-), and colons (:) were used as separators for each temporal unit, including year, month, week, hour, minute, and second. Following this extraction, the Timestamp feature was removed, resulting in an increase in the total number of features from 80 to 85.

The outcomes of this study were encouraging, with high accuracy rates achieved in both experiments: a perfect 100% in the first, and over 99% in the second. These results demonstrate a significant improvement over previous related work.

In conclusion, this research utilized a range of machine learning techniques to identify two types of DDoS attacks, namely DoS Attacks-Hulk and DoS Attacks-SlowHTTPTest, with an impressive degree of accuracy. The analysis of the CICIDS2017 Dataset, along with the application of various machine learning models, has contributed to the broader field of information security by enhancing our understanding of network attacks and their detection. This study highlights the potential of advanced machine learning approaches in the ongoing fight against increasingly sophisticated cyber threats.

REFERENCES

[1] Breda, G., Kiss, M. (2020). Overview of information security standards in the field of special protected industry 4.0 areas & industrial security. Procedia Manufacturing, 46: 580-590. https://doi.org/10.1016/j.promfg.2020.03.084

[2] Singh, M. (2021). An overview of automotive vehicles and information security. Information Security of Intelligent Vehicles Communication: Overview, Perspectives, Challenges, and Possible Solutions, 1-13. https://doi.org/10.1007/978-981-16-2217-5_1

[3] Alkhudhayr, F., Alfarraj, S., Aljameeli, B., Elkhdiri, S. (2019). Information security: a review of information security issues and techniques. In 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), IEEE, 1-6. https://doi.org/10.1109/CAIS.2019.8769504

[4] Ning, J.T., Xu, J., Liang, K.T., Zhang, F., Chang, E.C. (2018). Passive attacks against searchable encryption. IEEE Transactions on Information Forensics and Security, 14(3): 789-802. https://doi.org/10.1109/TIFS.2018.2866321

[5] Lee, J.H., Shin, J., Realff, M.J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. Computers & Chemical Engineering, 114: 111-121. https://doi.org/10.1016/j.compchemeng.2017.10.008

[6] Fedin, F.O., Trubienko, O.V., Chiskidov, S.V. (2020). Machine learning model of an intelligent decision support system in the information security sphere. In 2020 International Russian Automation Conference (RusAutoCon), IEEE, 215-219. https://doi.org/10.1109/RusAutoCon49822.2020.9208122

[7] Malliga, S., Nandhini, P.S., Kogilavani, S.V. (2022). A comprehensive review of deep learning techniques for the detection of (distributed) denial of service attacks. Information Technology and Control, 51(1): 180-215. https://doi.org/10.5755/j01.itc.51.1.29595

[8] Le, D.T., Dao, M.H., Nguyen, Q.L.T. (2020). Comparison of machine learning algorithms for DDoS attack detection in SDN. Информационно-управляющие системы, 3 (106): 59-70. https://doi.org/10.31799/1684-8853-2020-3-59-70

[9] Ahsan, M., Nygard, K.E., Gomes, R., Chowdhury, M.M., Rifat, N., Connolly, J.F. (2022). Cybersecurity threats and their mitigation approaches using machine learning-a review. Journal of Cybersecurity and Privacy, 2(3): 527-555. https://doi.org/10.3390/jcp2030027

[10] Qu, X.F., Yang, L., Guo, K., Ma, L.R., Sun, M., Ke, M.X., Li, M. (2021). A survey on the development of self-organizing maps for unsupervised intrusion detection. Mobile Networks and Applications, 26: 808-829. https://doi.org/10.1007/s11036-019-01353-0

[11] Ahsan, M., Gomes, R., Denton, A. (2018). Smote implementation on phishing data to enhance cybersecurity. In 2018 IEEE International Conference on Electro/Information Technology (EIT), IEEE, 0531-0536. https://doi.org/10.1109/EIT.2018.8500086

[12] Sarker, I.H., Abushark, Y.B., Khan, A.I. (2020). Contextpca: predicting context-aware smartphone apps usage based on machine learning techniques. Symmetry, 12(4): 499. https://doi.org/10.3390/sym12040499

[13] Guezzaz, A., Azrour, M., Benkirane, S., Mohy-Eddine, M., Attou, H., Douiba, M. (2022). A lightweight hybrid intrusion detection framework using machine learning for edge-based IIoT security. The International Arab Journal of Information Technology, 19(5): 822-830. https://doi.org/10.34028/iajit/19/5/14

[14] https://www.kaggle.com/datasets/cicdataset/cicids2017.

[15] Ayodele, T.O. (2010). Machine learning overview. New Advances in Machine Learning, 2: 9-18. https://doi.org/10.5772/9374

[16] Alzubi, J., Nayyar, A., Kumar, A. (2018). Machine learning from theory to algorithms: an overview. In Journal of Physics: Conference Series, IOP Publishing, 1142: 012012. https://doi.org/10.1088/1742-6596/1142/1/012012

[17] Muhamedyev, R.I. (2015). Machine learning methods: an overview. Computer Modelling & New Technologies, 19(6): 14-29.

[18] Padarian, J., Minasny, B., McBratney, A.B. (2019). Machine learning and soil sciences: a review aided by machine learning tools. Soil, 6(1): 35-52. https://doi.org/10.5194/soil-6-35-2020

[19] Belousov, K., Baranov, D., Galinskaia, T., Ponomarev, N., Zelyanskaya, N. (2019). Using machine learning and visualization tools to monitor national media. In Innovation in Medicine and Healthcare Systems, and Multimedia: Proceedings of KES-InMed-19 and KES-IIMSS-19 Conferences, Springer Singapore, 481-490. https://doi.org/10.1007/978-981-13-8566-7_44

[20] Simons, J., Adams Bhatti, S., Weller, A. (2021). Machine learning and the meaning of equal treatment. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 956-966. https://doi.org/10.1145/3461702.3462556.

[21] Cunningham, P., Cord, M., Delany, S.J. (2008). Supervised learning. In Machine Learning Techniques for Multimedia: Case Studies on Organization and

Retrieval, Berlin, Heidelberg: Springer Berlin Heidelberg, 21-49. https://doi.org/10.1007/978-3-540-75171-7_2

[22] Van Engelen, J.E., Hoos, H.H. (2020). A survey on semi-supervised learning. Machine Learning, 109(2): 373-440. https://doi.org/10.1007/s10994-019-05855-6

[23] Zhu, X.J., Goldberg, A.B. (2022). Introduction to semi-supervised learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 3(1). https://doi.org/10.1007/978-3-031-01548-9

[24] Castelli, M., Vanneschi, L., Largo, Á.R. (2018). Supervised learning: classification. Reference Module in Life Sciences, Encyclopedia of Bioinformatics and Computational Biology, 1: 342-349. https://doi.org/10.1016/B978-0-12-809633-8.20332-4

[25] Stöter, F.R., Chakrabarty, S., Edler, B., Habets, E.A. (2018). Classification vs. regression in supervised learning for single channel speaker count estimation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 436-440. https://doi.org/10.1109/ICASSP.2018.8462159.

[26] Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. New York: Springer, 2: 1-758. https://doi.org/10.1007/978-0-387-21606-5

[27] Mishra, N.K., Celebi, M.E. (2016). An overview of melanoma detection in dermoscopy images using image processing and machine learning. arxiv Preprint arXiv: 1601.07843. https://doi.org/10.48550/arXiv.1601.07843

[28] Greene, D., Cunningham, P., Mayer, R. (2008). Unsupervised learning and clustering. Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval, 51-90. https://doi.org/10.1007/978-3-540-75171-7_3

[29] Shi, T., Horvath, S. (2006). Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics, 15(1): 118-138. https://doi.org/10.1198/106186006X94072.

[30] Dayan, P., Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. Current Opinion in Neurobiology, 18(2): 185-196. https://doi.org/10.1016/j.conb.2008.08.003.

[31] François-Lavet, V., Henderson, P., Islam, R., Bellemare, M.G., Pineau, J. (2018). An introduction to deep reinforcement learning. Foundations and Trends® in Machine Learning, 11(3-4): 219-354. http://doi.org/10.1561/2200000071

[32] Wiering, M.A., Van Otterlo, M. (2012). Reinforcement learning. Adaptation, Learning, and Optimization, 12(3): 729. https://doi.org/10.1007/978-3-642-27645-3