National College of Ireland

Evaluation and detection of cybercriminal attack type using machine learning

MSc Research Project
MSc Cyber Security

Safvaan Shadab Nakid
Student ID: 20180527

School of Computing
National College of Ireland

Supervisor: Dr Vanessa Ayala-Rivera

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	SAFVAAN SHADAB NAKID
Student ID:	X201805827
Programme:	MSC CYBER SECURITY **Year:** 2021
Module:	MSc Research Project/Internship
Supervisor:	Dr Vanessa Ayala-Rivera
Submission Due Date:	16/12/2021
Project Title:	EVALUATION AND DETECTION OF CYBERCRIMINAL ATTACK TYPE USING MACHINE LEARNING
Word Count:	6595 **Page Count** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: SAFVAAN SHADAB NAKID

Date: 16/12/2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

EVALUATION AND DETECTION OF CYBERCRIMINAL ATTACK TYPE USING MACHINE LEARNING.

Safvaan Shadab Nakid
20180527

Abstract

There has been a significant rise in data breaches and various types of computer-based attacks that result in monetary as well as infrastructural losses to both individuals and organizations alike. Law enforcement agencies and organizations often find themselves at crossroads at such times whilst dealing with unprecedented attacks or breaches. As cyber offenders or perpetrators evolve in their attack patterns rather rapidly than the organizations equipped with defensive mechanisms, it becomes virtually difficult to trace back to the attack patterns of the criminal. With its negative effects such as breach of confidentiality and integrity, data sustained by the organizations such as event logs, reports can also be used to gain further insights as to how a criminal can potentially harm a system & understand what vulnerable areas of the system infrastructure can be further identified to strengthen them.

This research paper addresses the issue of identifying such attack patterns and presents a model based on feature selection to understand the type of attack pattern employed by the offender which would then help narrow down the approach of creating the profile of the offender. A dataset of 1145 recorded data breaches and ransomware attacks from the University of Queensland was used in the research. As the dataset consisted of imbalanced columns, ROS and RUS sampling techniques were employed along with data tuning procedures and label encoding. Classification models such as Random Forest, KNN, Logistic regression were implemented on the data to identify the accurate attack-type of a given attack. Upon comparison of results, it was noted that Random Forest was able to outperform other models by achieving 95% accuracy with an average F1 score of 0.94.

1 Introduction

Computer-based crimes or Cyber-attacks as the term goes have become more recent than they had been in the last decade. It's not a farfetched theory that this has been only made possible with the digitization of data. Tech conglomerates, defence agencies, healthcare industries, financial institutions have opted for if not completely switched to this new methodology of data retention and management. This goes without saying that with such heavy reliance on computational intellect, the chances of having repercussions are not close to the bare minimum, which means that in its lifetime, an entity is more likely to be breached and compromised than ensure the security of its assets.
With estimated damages of 20B$ by ransomware alone in the year 2020, and with 70% of data breaches involving financial motivations, it is safe to say that Cybercrimes have just picked up momentum and will continue to be on the rise with new, effective, and improvised attack vectors[1]. With a projected estimation of nearly a 15.4Million in 2023 as compared to 7.9 Million in 2018[2], the rise in the denial of service attacks on critical organizations proves the argument further that the organizations and individuals alike now need to better prepare and enhance their system infrastructure to safeguard themselves from future events and breaches. Since cybercrimes can inflict damages on to the affected system or individual in any manner be it by denial of services, a malware Trojan, or a physical act of data security threat, cyberbullying, harassment, and fraud among other forms, it becomes a tedious task for law enforcement agencies to be able to pinpoint the exact occurrences of the attack.

The identification, classification, and categorization of suspected cybercriminals is a time-consuming task and often misleading due to the facts surrounding the incident (Garcia, 2018). For example, a cybercriminal who is classified as a botnet attacker can go undetected in a breach where they employed other means of intrusion such as a Ransomware Trojan, Thus, making it evident that there is a need for proper organization and management of data that can be used to critically identify the perpetrators behind the attacks.

Cybercriminal profiling is one such field of computer security that involves the collection, identification, and management of event-based data and is relied upon the evidence and artefacts discovered and documented by the digital forensic evidence team (Rogers, 2003). An accurate description of a cybercriminal consisting of their attack patterns, tools used, and attack vector exploited can prove to be a useful guide as to what amount of data assets should be stored on a network and help in defining proper planning and guidelines to secure it. This also helps in creating a strong argument in the face of law enforcement and ensures a fair and systematic prosecution and conviction of the offenders.

[1] https://www.comparitech.com/vpn/cybersecurity-cyber-crime-statistics-facts-trends/#:~:text=In%202020%2C%20there%20were%20a,by%2066%25%20to%20300%20million.
[2] https://www.varonis.com/blog/cybersecurity-statistics/

Machine learning, a not-so-new concept in today's field of computing has emerged in recent years that employs artificial intelligence to analyze data based on past experiences or cases to provide a relevant solution. With data produced through digital pieces of evidence and a systematic prediction with the help of machine learning, the process of cybercriminal profiling can be further simplified by predicting the type of attack that has been executed on the affected entity.
The attack type can be in any form such as denial of service attack, where the attacker disrupts critical services of the organization and affects the availability of data and resources, phishing attack where the confidentiality and integrity of assets are compromised and financial losses are incurred, Web Compromise attack which affects the integrity of the online Web site or a malware attack which can range from a Trojan to spyware.

The research in this work was based on implementing a supervised learning model employed by machine learning methodology equipped with a case retention approach that can help identify the types of attack a cybercriminal carries out and thus help furnish the process of profiling the Cyber Criminal in a more systematic and detailed manner.
The following are the research questions that this paper shall elaborate on.

- How can digital evidence collected from digital forensic investigations help in the profiling of cybercriminals?
- How does the prediction of attack patterns in cybercrimes aid in creating an offender profile?

The main objective of this research is to critically analyze and provide a better understanding of how machine learning methodologies can be used to predict the type of attacks cybercriminals carry out on assets owned by individuals and organizations and how these predictions can then help in creating an offender profile.

The report is structured in the following sections as follows Wherein, Section 2 details on the literature review conducted for the research, and Section 3 focuses on the research methodology applied along with design specifications. Moving on to Section 4 which discusses the design Specification of the model followed by the implementation process that is laid out in section 5. Finally, The evaluation shall be covered in section 6 with a conclusion and future work detailed in section 7 of the report.

2 Related Work

Criminal profiling or creating an offender sketch is a relatively new concept in the field of computer-based crimes. A literature review was conducted for this research to understand how machine learning can be incorporated into this domain.

2.1 The importance of Criminal profiling.

Offender profiling or criminal profiling can be defined as creating a behavioural sketch of a suspected individual based on their activity patterns and other physical and psychological traits present at the crime scene. Such created profiles are not only used in the conviction and apprehension of unknown suspects that may be involved in a certain crime but also help in predicting the attack patterns to identify the next target of the suspect. Behavioural analysis such as physical characteristics like gender and age, social status, and psychological traits the suspect possesses are some of the entities evaluated by an expert in this field.[3] With the recent advancements in computing and technology and the modernization of traditional processes, how an offender commits a crime has also changed with the majority of crimes nowadays happening over the internet. Some of these ways include spear-phishing which involves the impersonation of an entity intending to commit fraud using mediums such as Emails and adware, website phishing which intends to install malicious scripts on the victim's machine and steal critical data, denial of service attacks which causes disruptions in the functioning of critical services and ransomware attacks that causes the attacker to gain complete control of the system to be only "released-back" after a ransom is paid.[4] The domain of cybercrimes is not limited to such kinds of attacks and can come in many other forms which can chalk up to 105$ Trillion in annual damages by 2025[5].

The aftermath of such incidents can be referred to as virtual crime scene where the investigator examines artefacts which usually come in the form of signatures left around the crime scene or memory dumps made by the compromised system also referred to as event logs which may contain useful information such as browser cache, metadata and timestamp signatures that are collected by the digital forensic investigation team and are called as digital evidence which can then answer some questions like tools used, vulnerabilities targeted which could then help in estimating the motives behind the incident and understand the criminal. The approaches discussed by (Garcia, 2018) namely, deductive profiling which involves evidence-based analysis to understand and create a hypothesis that can then be used in the apprehension of the criminal and inductive profiling involving steps such as statistical comparison also concluded the discussion that there is a room for incorporating the techniques used in computer forensics in creating a profile of an offender.

[3] https://www.linkedin.com/pulse/criminal-profiling-use-support-digital-forensic-jim%C3%A9nez-serrano/?articleId=6665961806869123072

[4] https://data.world/qambait/11-ways-cyber-criminals-can-attack-cyber-security-article/workspace/file?filename=11+Ways+Cyber+Criminals+Can+Attack+%7C+Cyber+Security+Article+%7C+Qamba+IT

[5] https://www.comparitech.com/vpn/cybersecurity-cyber-crime-statistics-facts-trends/

2.2 The limitations surrounding Criminal profiling.

There can be various driving factors that can affect the validity of the created criminal profile as most of the characteristics detailed in the profile are based on evidence and the methods used in classifying or evaluating these evidence may not be permissible or ethical in the eyes of a judicial proceeding. One of these reasons may itself be the motivation of the investigator to convict an individual for a certain case that may have similar patterns to one of the previous cases which can be considered as the psychological influence of the investigator (Nakid, 2021). Evidence, as discussed earlier comes in a grey area as the results obtained through deductive profiling can sometimes be wrong or misleading if the attacker has improvised his attack patterns and has learned to change or mask his steps such as using spoofed IP Address, change in tools used for exploitation, or difference in targeted attack vectors (Garcia, 2018). Another major concern that affects the authenticity of the criminal profile is privacy, wherein the investigator can be subjected to data that may not be related to the case itself and thus cause an infringement of privacy rendering the evidence itself impermissible in court which may be due to improper data handling by the user and lack of education and awareness (Aminnezhad, 2012). Phishing websites created by offenders were successful in masking their identity as GDPR enforced privacy rights to individuals and organizations after its introduction in 2018 which caused access to essential data such as WhoIS information impossible for the investigators (Ferrante, 2018).

2.3 Machine Learning and Criminal profiling

Machine learning can be defined as the process of making predictions using stored data. It involves the incorporation of Artificial Intelligence and mathematical algorithms to make viable predictions regarding a certain condition[6]. Various researches in the field of machine learning have been made which have furnished the application of the science in various fields of day-to-day processes such as in medical fields where it is used to predict death rates and risk of various diseases, in businesses to predict the best-selling product and most effective business model and so on. One such analysis was done by (Chanjin, 2015) using hierarchical clustering where GPS information was extracted from Windows 8 OS and other various instant-messaging applications to obtain user information such as school history, nearest data-transmission point, etc, which helped locate the position of the suspects using K-means algorithm and narrow down the perimeter of the investigation. Another work published by (Baumgartner, 2008) details the process of profiling cybercriminals using an algorithmic approach. The research focuses on the importance of machine learning in creating decision-aid tools for police investigations and introduces a Bayesian network approach wherein the behavioural characteristics of an offender are extracted from a dataset and then implementing the extracted features in an inference engine to predict the profile. The research was successful in predicting the characteristics of an unknown offender such as employment, gender, etc with 80% accuracy. The attack method used by a cybercriminal and how can this

[6] https://www.ibm.com/cloud/learn/machine-learning

method be associated with a suspect offender was identified by (Bilen, Abdulkadir & Özer, A. B., 2021) using datasets consisting of various characteristic information. Prediction models such as the Support vector machine was used on a dataset consisting of attacks that occurred in a certain province that was able to produce 95.02% accuracy in predicting 8 attack types, followed by logistic regression with 65.42% accuracy. The research concluded that the probability of being under attack decreased as the education levels and awareness amongst the individual increased. The similarity in the attack patterns of defacements of websites was analyzed by (Mee Lan Han, 2019) using a case-based reasoning approach which helped in finding out the type of hackers behind these attacks and create a characteristic pattern of the offender such as the encoding used in the attack, the regions the hackers mostly attack in, the background of the attacker, among others. (Rasmi, 2013) proposed a new algorithm called Similarity of attack intentions(SAI) based on attack intentions algorithm (AIA) with an accuracy score of 0.68 wherein the algorithm model was divided into three sections which identified the intentions of the attack. A similarity metric was then generated and the most feasible attack intentions from the result were selected by assigning a probability value to each previous attack. The AIA algorithm (Jantan, 2011) was based on D-S evidence theory that predicted intentions of an attack based on a set of recorded evidence. The intentions were nothing but motives such as "gaining root privilege" as described by (W. Peng, Z. Wang, and J. Chen, 2009) wherein the prediction of intention was based on an Intrusive intention recognition algorithm based on D-S evidence theory.

2.4 Case-based Reasoning approach

Case-based reasoning or CBR is a machine learning approach that uses a database of problem-solutions to solve new problems, this is done by storing these problem sets as tuples which are then checked with a new case as it arrives. Upon a positive match, the solution accompanied with that case is affixed with the current case and in cases with no match with previous sets, the new case is saved and then compared with its nearest neighbours thus providing a suitable solution[7]. The work published by (Mee Lan Han, 2019) was also based on this approach and employed CBR as a methodology which was implemented using clustering and case vector approach. Meta-learning for data processing is yet another application of CBR which helps in selecting accurate features required in prediction along with novelty-prediction which also is based on feature prediction but a large scale of data[8].

2.5 Summary of Findings

With the review of literature carried out to analyze and understand current trends in the field of criminal profiling, it was seen that a majority of the research conducted was based on criminal cases that did not involve computers as an attack medium. With that being said, and keeping the aforementioned issues in mind, it is safe to say that the application of criminal

[7] https://www.geeksforgeeks.org/ml-case-based-reasoning-cbr-classifier/

[8] https://www.researchgate.net/publication/336821744_Case-Based_Reasoning_-_Methods_Techniques_and_Applications

profiling for cybercrimes is vast and is hugely dependent on data and has scope for implementation in areas such as detection of virus-writers from a group of other hackers. The motivation for this research is based on the said review and focuses on approaching the problem of identification of attack types used by the cyber attackers on various organizations.

3 Research Methodology

For this research, the methodology that is used for predicting the attack type in cybercrime is based on the case-based reasoning methodology discussed by (Rasmi, 2013) which comprises of four stages are namely Reuse, Revise, Retain and Retrieve phases which are as shown in the figure below:

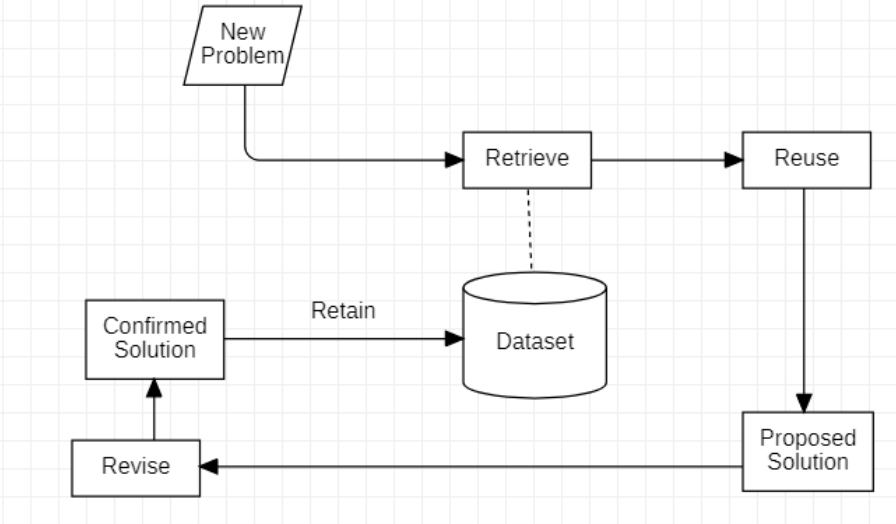


Figure 1: Case-based reasoning methodology

Following an experience-based approach, case-based reasoning aims to solve problems by providing a solution from a previous case-based that is adaptable to the current problem. One advantage that can be taken from using CBR is that the solutions produced are based on past experiences and not general estimation of the prediction model thus enhancing the accuracy of results produced (Bernstein, 2019). With this in mind, the functionality of the CBR methodology is divided into four phases as stated above.

A) Retrieve phase.

Once a target problem is passed on to the model, the retrieval phase is responsible to compare the current problem with stored cases and give out relevant cases that may be used as a probable solution for the current problem. This is done based on the similarity assessment of surface features. Now, surface features are nothing but the attribute-value pairs that describe the stored case itself. The retrieval of the most similar type of attack type was done on the test case discussed in section 6 of this report.

B) Reuse and Revise Phase.

Once similar cases have been retrieved from the database the case with the best probable solution is then selected as a new outcome for the current problem. The reuse phase then comes into action which is responsible for selecting the probable solution from an array of cases. Adaptation is a part of the Reuse phase which is carried out in the event where a new problem has significant differences in terms of characteristics. The two ways this occurs are by Substitution, which does as the name suggests, changes some part of the solution to a probable alternative, and transformation which alters the solution entirely. The result is then pushed onto the reuse phase which is then amended in the revising phase. The revise phase can be simply defined as a process in which the current solution is updated and is a collective effort of reuse and adaptation processes.

C) Retain Phase.

Once a feasible solution that fits the current problem is selected, the new case is then stored in the case-base along with its solution either manually by authorized personnel or automatically thus retaining the new case and completing the CBR cycle (Mantaras, 2005). The way this information stored is stored shall differ based on the system developed and the classification models used.

3.1 Dataset Information

For this research, a dataset of data breaches and ransomware attacks on organizations was taken from the University of Queensland repository (Ko, R., Tsen, E. and Slapnicar, S., 2020). The dataset consisted of 1146 records with 26+ features which were released as a part of a study focused on exploring cyber resilience in organizations. describes the attack-type feature which is used as a target variable in the prediction. As seen in Table 1, the data across the dataset is distributed within these four attack types with the highest among them being malware injection attacks or incidents where there have been confirmed cases of malware being installed.

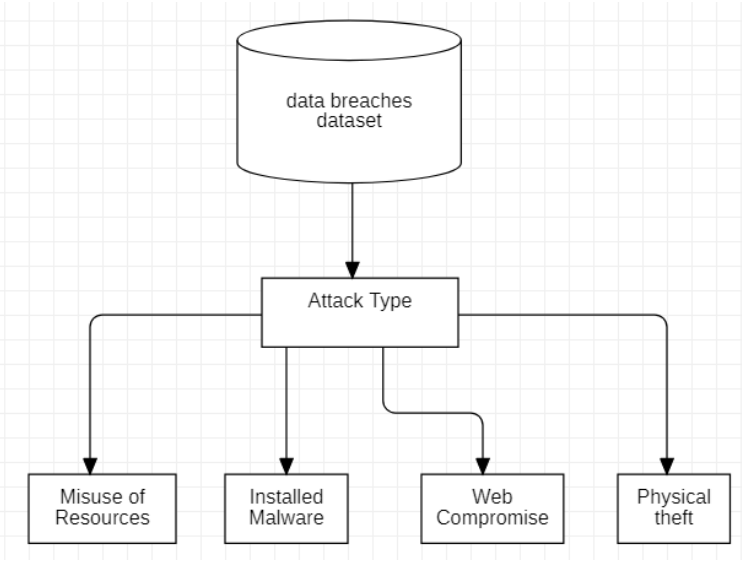


Figure 2: Feature information

a. Misuse of resources

Misuse of resources is nothing but using peripherals provided by the company in an unacceptable or unethical manner such as downloading suspiciously large files or sharing malicious links unintentionally which can lead to a breach of policies and cause theft of critical stored data, or other monetary damages.

b. Installed malware

Malware such as ransomware, encryption-based locker trojans, or adware that are used for attacks can be directly installed onto the system through various means thus affecting the confidentiality of a system.

c. Physical theft

Physical threat refers to the event wherein an external entity infiltrated an organisation or gained access to a restricted zone within the organisation to steal critical information, asset or data such as confidential plans portfolios or disk drives.

d. Web Compromise

Web compromise refers to the incidents that occurred over websites managed and owned by the organisations wherein the affected users were misled to a suspicious URL to gain access to the companies' resources and mainframe server using backdoors.

Table 1: Attack types

ATTACK TYPE	NUMBER OF ATTACKS
Installed Malware	259
Physical Theft	214
Web Compromise	74
Misuse of Resources	200

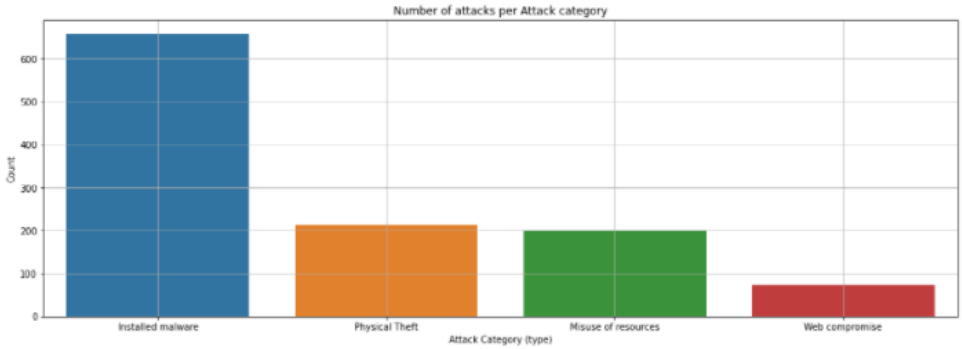


Figure 3: Attacks per Category

3.2 Data Processing

The data cleansing procedure was carried out on the data to get rid of all Null values and to create an equalised sample space to perform prediction. The features that weren't necessary for prediction were dropped from the prediction data frame and the categorical attributes within the set were changed to numerical ones to perform mathematical computations. Finally, classification algorithms or models were applied to the transformed dataset which was able to produce evaluation metrics and accuracy scores.

3.3 Classification Models

For the basis of this research, CBR methodology was applied using the following classifiers as these supported the nearest neighbouring classification.

a. Random Forest Classifier (RFC)

Using decision tree classification methodology, RFC generates a sample space of multiple decision trees consisting of multiple parent nodes and ***n*** child nodes and uses an averaging function to predict the solution.

b. Logistic Regression

Logistic regression is a prediction mechanism used to predict categorical outcomes of a given problem set analysing the relationship between the independent and dependent variables and giving out the probabilistically accurate outcome for a given problem.

c. KNN algorithm

The K-nearest neighbour algorithm works on the principle of closeness and works by computing the distance between the current problem towards its nearest problem that has the same characteristic features.

d. Support Vector Classification (SVC)

Using the support vector machine module SVC, classification amongst the data is computed by the transformation of data and then segregating it based on the parameters specified in the problem. One advantage of using SVC is that the problem of a non-linear relationship between two nodes is ignored thus achieving higher accuracy rates.

4 Design Specification

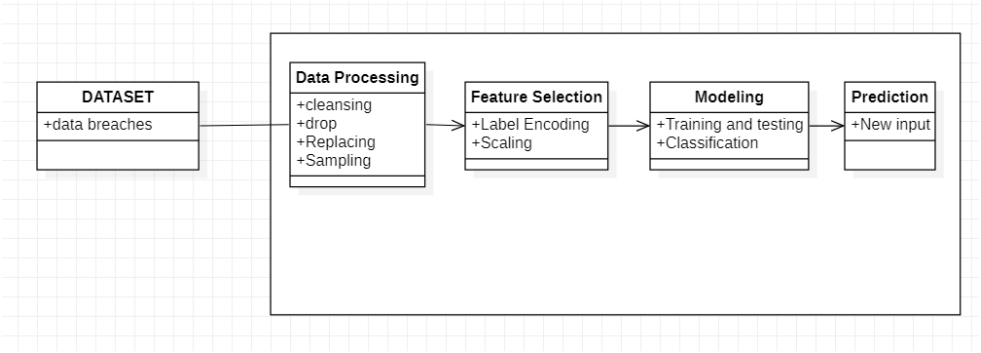


Figure 4: Design Specification

The design principle followed for developing the prediction model is shown as per figure [4]. Each of the steps taken therein is detailed as follows:

a. Data Processing

For the prediction algorithm to work, the dataset to be used must be ready to be able to produce mathematically accurate results. Since the dataset used in the research consisted of various columns that were not useful and contained incomplete or unequal data, data cleansing procedures were carried out. The procedures included removal of null values from the data, changing the data type of various features from ***object to float64*** and dropping of certain columns which would have produced inefficient results if used in the prediction and classification. Once this was done, the *Nan* values were then replaced with '0' to make the datasets numerically identifiable. The sampling process i.e., oversampling was carried out on the cleansed data to make the values in the features equal to each other and balanced.

b. Feature selection

The feature selection phase was used to assess and ascertain which of the columns or features from the dataset can be used for prediction considering that these features produce accurate results. Using ***ANOVA*** or analysis of variance on the target ***x_test*** axis, the following output was produced.

```
array([False, False, False, False, False, False, False, False, False,
       False, False,  True,  True, False, False, False, False, False,
       False,  True, False, False,  True,  True, False, False, False,
       False,  True, False, False, False, False, False, False, False,
       False, False])
```

```
x.columns
```

```
Index(['Organisation', 'Critical Industry', 'Organisation size',
       'Level of digital intensity', 'Sector', 'Country',
       'Cyber security role', 'Cyber security frameworks',
       'Education and awareness policy', 'Policy',
       'Prevention, Detection and Recovery', 'Improper network segmentat
ion',
```

Figure 5: Feature selection

Figure [5] shows the output of ANOVA in Boolean True and False values. Each of these values corresponds to the Index columns in the dataset and as can be seen from the output produced, ***improper network segmentation*** can be used as one of the predictor columns for the classification algorithm to work on. Label encoding was done by splitting the dataset into float and numeric values and assigning each categorical attribute a numerical position which was then followed by scaling of the entire dataset to make values positively and negatively equal to each other.

c. Modelling and Prediction

This phase comprises of applying classification models to the now cleaned dataset. But before this, the models were needed to be divided into a training set of 70% data and a testing set of 30% data to ensure that the model can produce accurate output. The models chosen for this stage were Random Forest, KNN, Logistic regression and SVC respectively. Once the classification was complete, the model with the best F1-Score and accuracy was selected for making a prediction which in this case was Random Forest Classifier.

5 Implementation

The implementation discusses the final steps taken in the research to create the prediction model and finally progresses towards the evaluation phase.

5.1 Coding the model

a. Data processing:

Since this was a robust dataset consisting of categorically different values, data cleansing was needed to be done to make the dataset fit for prediction.

```
df.isnull().sum()
```

```
Year                                                          0
Organisation                                                  0
Critical Industry                                             0
Organisation size                                           205
Level of digital intensity                                    0
Sector                                                        0
Country                                                       0
Cyber security role                                           0
Cyber security frameworks                                     0
Education and awareness policy                                0
Policy                                                       17
Prevention, Detection and Recovery                            0
Improper network segmentation                               517
Inappropriate remote access                                 443
Absence of encryption                                       371
Detector                                                     19
Restructuring after attack                                  140
Bribe/ransom paid                                             0
Free identity or credit theft monitoring                    118
Additional disclosure of information                        151
Number of users affected                                      0
Overall nature of attack                                    336
Attack type                                                 398
Attacker                                                      0
Attack vector                                               397
Impact on data                                                0
Aspect of Confidentiality-Integrity-Availability triad affected    0
Individual(s) name(s) leaked/exposed                          7
Address(es) leaked/exposed                                   11
Other personally identifiable information (PII) leaked/exposed    11
Track 1 - Credit card details leaked/exposed                 17
Track 2 - Credit card details leaked/exposed                295
Social security number/tax number leaked/exposed             11
Subsequent fraudulent use of data                             2
Investigation                                                 0
Undertook investigation                                       0
Litigation by public                                          0
Penalties/settlement paid or actions imposed                  0
Imposed penalties or actions on organisation                  0
Fines issued by government or relevant body                   0
Settlement paid                                               0
Effect on share price                                      1032
Summary                                                       0
Unnamed: 43                                                1145
dtype: int64
```

Figure 6: Null Values

Figure [6] shows all the null values present in the dataset which increased the need of eliminating such values. To do this each of the attributes with null values was called first to check for null values and then using the ***fillna()*** method these null values were replaced with an integer "0".

```
df["Policy"].value_counts()
```

```
Yes    1123
No        5
Name: Policy, dtype: int64
```

```
df["Policy"].fillna("Yes",inplace=True)
```

Figure 7: Replacing Null Values

Figure [7] shows the code snippet for one of the null categories. The same process was followed to tackle the rest of the null values. Upon completion, the Null values were eliminated as seen in figure [8].

```
profiles.isnull().sum()
```

```
Year                                                          0
Organisation                                                  0
Critical Industry                                             0
Organisation size                                             0
Level of digital intensity                                    0
Sector                                                        0
Country                                                       0
Cyber security role                                           0
Cyber security frameworks                                     0
Education and awareness policy                                0
Policy                                                        0
Prevention, Detection and Recovery                            0
Improper network segmentation                                 0
Inappropriate remote access                                   0
Absence of encryption                                         0
Detector                                                      0
Restructuring after attack                                    0
Bribe/ransom paid                                             0
Free identity or credit theft monitoring                      0
Additional disclosure of information                          0
Number of users affected                                      0
Overall nature of attack                                      0
Attack type                                                   0
Attacker                                                      0
Attack vector                                                 0
Impact on data                                                0
Aspect of Confidentiality-Integrity-Availability triad affected    0
Individual(s) name(s) leaked/exposed                          0
Address(es) leaked/exposed                                    0
Other personally identifiable information (PII) leaked/exposed    0
Track 1 - Credit card details leaked/exposed                  0
Social security number/tax number leaked/exposed              0
Subsequent fraudulent use of data                             0
Investigation                                                 0
Undertook investigation                                       0
Litigation by public                                          0
Penalties/settlement paid or actions imposed                  0
Imposed penalties or actions on organisation                  0
Fines issued by government or relevant body                   0
Settlement paid                                               0
Summary                                                       0
Unnamed: 43                                                1145
dtype: int64
```

Figure 8: Removed Null Values

b. Transformation of the dataset

The dataset was then split into two parts based on the associated data types. This was done to implement Label encoding to categorical attributes i.e., in the form of [YES, NO] etc. This is done because machine learning does not work on character values so using **label encoding** each value in a column is converted into an integer by assigning a numerical value to every first character in the dataset based on its categories. For example, the feature attribute **"Installed Malware"** from the target column starts with the letter **'I'** which comes first as per the alphabetical sequence. Hence the value **'0'** will be assigned to that attribute.

```
df1=dframe.select_dtypes("object") # seperating dataset into object
df2=dframe.select_dtypes("float") # seperating dataset into numeric
```

Figure 9: Splitting of the dataset

```
from sklearn.preprocessing import LabelEncoder

le =LabelEncoder() # object for label encoder
for col in df1:
    df1[col]=le.fit_transform(df1[col]) # encoding entire dataset.
```

```
<ipython-input-72-4ec78abe8fe1>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df1[col]=le.fit_transform(df1[col]) # encoding entire dataset.
```

Figure 10: Label Encoding

	Organisation	Critical Industry	Organisation size	Level of digital intensity	Sector	Country	Cyber security role	Cyber security frameworks	Education and awareness policy	Policy	...	Track 1 - Credit card details leaked/exposed	Social security number/tax number leaked/exposed	Subsequer frauduler use of dat
0	0	0	0	3	33	1	1	0	0	1	...	1	0	
1	1	1	1	2	16	23	0	0	0	1	...	0	1	
2	2	0	0	0	20	23	0	0	0	1	...	1	0	
3	9	1	2	2	16	23	0	0	0	1	...	0	1	
4	10	1	3	2	16	23	0	0	0	1	...	0	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1140	1101	0	0	2	29	23	0	0	0	1	...	0	0	
1141	1102	1	1	1	0	23	0	0	0	1	...	1	0	
1142	1103	0	0	2	29	23	0	0	0	1	...	0	0	
1143	1104	1	0	2	16	23	0	0	0	1	...	1	0	
1144	1105	0	0	0	2	23	0	0	0	1	...	0	0	

1145 rows × 38 columns

Figure 11: Encoded dataset

The datasets were then concatenated and then scaling procedures were performed. Scaling is done to remove any outliers from the dataset and bring all the values within one frame or scale range which in this case was **[-1 to 0 to +1].** As seen in figure [13], the values were too high which needed to be normalised and brought in one range which was done in figure [14].

```
from sklearn.preprocessing import StandardScaler  # scaling used to equalise

for col in profiling:
    sc=StandardScaler()
    profiling[col]=sc.fit_transform(profiling[[col]])
```

Figure 12: Scaling

Settlement paid	Number of users affected
0	0.0
0	55447.0
0	0.0
0	23000.0
0	2650000.0
...	...
1	24000000.0
0	0.0
0	0.0
0	0.0
0	17000000.0

Figure 13: Unscaled data

Without scaling

```
profiling
```

	Organisation	Critical Industry	Organisation size	Level of digital intensity	Sector	Country	Cyber security role	fr
0	-1.732942	-1.220299	-0.756048	1.243705	1.989299	-5.418687	2.895729	
1	-1.729834	0.819471	0.103609	0.330272	0.034645	0.237333	-0.345336	
2	-1.726726	-1.220299	-0.756048	-1.496594	0.494563	0.237333	-0.345336	
3	-1.704970	0.819471	0.963267	0.330272	0.034645	0.237333	-0.345336	
4	-1.701862	0.819471	1.822925	0.330272	0.034645	0.237333	-0.345336	
...	...	...	...	...	...	...	...	...
1140	1.688987	-1.220299	-0.756048	0.330272	1.529380	0.237333	-0.345336	
1141	1.692095	0.819471	0.103609	-0.583161	-1.805030	0.237333	-0.345336	

Figure 14: Scaling Output

Scaled dataset

Once this was done, the dataset was then split into the **X & Y-axis**. Here the target column ***attack type*** was dropped from the x-axis as it shall be used for prediction and the Y-axis was created by calling the target column in it.

```
: x = profiling.drop("Attack type",axis=1) # taking all columns except attack
```

Figure 15: X-axis

X-axis

```
y = df["Attack type"] # keeping attack type on y axis
```

Figure 16: Y-axis

Y-Axis

Then the training and test sets were created by importing the **sklearn.model** library.

```
from sklearn.model_selection import train_test_split

X_train,X_test,Y_train,Y_test=train_test_split(x,y1,test_size=0.3,random_sta
```

Figure 17: Train & Test sets

The test size was set as **0.3** as the aim of prediction was to use **70%** train and **30%** test sets respectively.
The feature selection and balancing of data were done in further steps thus making the dataset ready for prediction as shown in figure [5] and figure [14] respectively.

5.2 Languages and Libraries used

a. Python:

Python is a high-level, object-oriented and interpreted programming language that was seen as feasible for this research. The main reason for selecting python was its robustness, simple lines of code syntax and availability of a vast library to be used in data classification and mining. Another reason for selecting the language was its speed and quick debugging with a majority of online resources available to help the developer overcome any issues[9].

b. Libraries used:

- Pandas

It is a python library that is used for manipulation procedures on data. The practical data analysis was done using the panda's library.

- NumPy

It is used for working with arrays of large sizes. The Numerical manipulation of the dataset was done using this library.

- Seaborn and Matplotlib

Data visualization and statistical plotting of data were done using these libraries.

- Sklearn

The machine learning library is used for classification, regression and other numerical computations. Label encoding, scaling, classification reports, confusion matrices, feature selection and prediction models were all generated using the Scikit library.

- Imblearn

The Imbalanced data was balanced using the Imblearn library which is used for generating a dataset with equal classes and ratio distribution. This is done to achieve maximum accuracy in prediction as an unbalanced dataset would be trained based on a biased class i.e., the attribute with the highest value.

6 Evaluation

6.1 The Criteria for evaluation

The degree of effectiveness and accuracy is an important aspect of evaluating classification models. It identifies the best suitable model to be used for a particular problem and is based on evaluation metrics as per Roberto Salazar[10]. This is done by evaluating the four-score metrics namely accuracy, precision, recall and F1-score which is based on the four-variable criteria namely TF- True positives (correctly predicted outcome), TN-Negatives (correctly predicted invalid outcome), FP-False positives (incorrectly predicted outcome) and FN- False Negatives (incorrectly predicted invalid outcome) respectively. The variables are computed to produce the four-score metrics with the formula stated in table [2] which are then displayed visually in the confusion matrix.

[9] https://www.python.org/doc/essays/blurb/

[10] https://towardsdatascience.com/machine-learning-classifiers-comparison-with-python-33149aecdbca

Table 2: A performance metric

Performance Metric	Formula
Accuracy	(TP+TN) / (TP + TN + FN + FN)
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F1-Score	(2 * RECALL * PRECISION) / (RECALL + PRECISION)

6.2 Experiment 1: Logistic Regression

As a basis of comparison, logistic regression was performed on unbalanced data. This gave an idea as to what needs to be done for making predictions more accurate. It was found that the accuracy and F1 score of the model improved significantly as the data was balanced. As can be seen from Table [2] below, the accuracy improved from 91% to 94% based on the collective F1 scores.

Table 3: Logistic Comparison

createm(Logreg) precision recall f1-score support Installed malware 0.97 0.90 0.93 202 Misuse of resources 0.67 0.84 0.74 50 Physical Theft 0.93 0.96 0.94 69 Web compromise 1.00 0.96 0.98 23 accuracy 0.91 344 macro avg 0.89 0.91 0.90 344 weighted avg 0.92 0.91 0.91 344 [[182 18 2 0] [6 42 2 0] [0 3 66 0] [0 0 1 22]] LogisticRegression()	predictionmodel(Logreg) precision recall f1-score support 0 0.92 0.90 0.91 202 1 0.87 0.93 0.90 202 2 0.96 0.97 0.96 202 3 1.00 0.96 0.98 202 accuracy 0.94 808 macro avg 0.94 0.94 0.94 808 weighted avg 0.94 0.94 0.94 808 [[181 21 0 0] [15 187 0 0] [0 6 196 0] [0 0 9 193]] LogisticRegression()
Unbalanced Data	Balanced Data

6.3 Experiment 2: Random Forest Classification

The model was trained using the 70% train dataset and the accuracy of prediction was verified based on F1-scores. The results produced show that there was a significant improvement seen in the prediction of attack type in the trained dataset with an accuracy of 94% having F1-scores ranging in the 91-100 range.

Table 4: Random Forest Comparison

```
from sklearn.ensemble import RandomForestClassifier

randomforestclass=RandomForestClassifier(max_depth = 4)

createm(randomforestclass)
```

```
              precision    recall  f1-score   support

           0       0.82      0.96      0.88       202
           1       0.68      0.42      0.52        50
           2       0.97      0.99      0.98        69
           3       1.00      0.22      0.36        23

    accuracy                           0.84       344
   macro avg       0.87      0.65      0.68       344
weighted avg       0.84      0.84      0.81       344

[[194   8   0   0]
 [ 28  21   1   0]
 [  0   1  68   0]
 [ 16   1   1   5]]
```

```
RandomForestClassifier(max_depth=4)
```

Unbalanced Data

```
predictionmodel(randomforestclass)
```

```
              precision    recall  f1-score   support

           0       0.98      0.89      0.94       202
           1       0.90      0.96      0.93       202
           2       0.93      1.00      0.96       202
           3       1.00      0.96      0.98       202

    accuracy                           0.95       808
   macro avg       0.95      0.95      0.95       808
weighted avg       0.95      0.95      0.95       808

[[180  21   1   0]
 [  3 194   5   0]
 [  0   0 202   0]
 [  0   0   9 193]]

RandomForestClassifier(max_depth=4)
```

Balanced Data

Following a similar approach, the other two models namely SVC and KNN were tested. The results obtained from these tests are mentioned in the table [4]. As results show, the accuracy increased in the case of SVC with an F1-Score of 90% but dropped to 60% from 74% significantly while classifying prediction for a balanced dataset in KNN. This may be because KNN works on nearest neighbours and unbalanced data has imbalanced classes i.e., unequal number of entries for each attribute in the dataset.

Table 5: Classifier Comparison

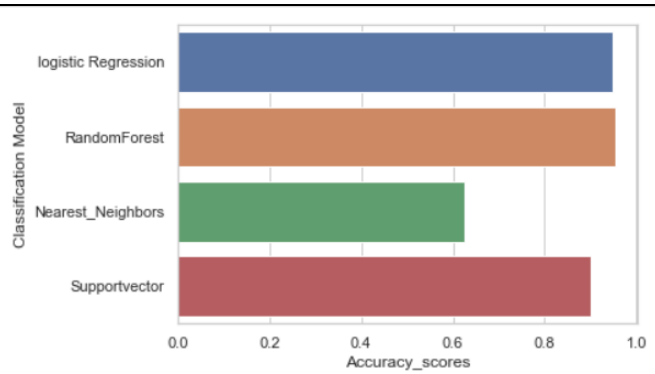
Model name	Unbalanced score	Balanced Score
SVC	0.88	0.90
KNN	0.74	0.60

6.4 Comparison of Classification Models.

Table 6: Comparison of accuracy

	Classification Model	Accuracy_scores
0	logistic Regression	0.946782
1	RandomForest	0.952970
2	Nearest_Neighbors	0.626238
3	Supportvector	0.900990

Accuracy scores of Classification Models



Accuracy Plot

Table [5] depicts the comparison of models which shows the average accuracy of 94% and 95% respectively. Random forest classifier outperformed Logistic Regression with an F1-Score of 95%.

6.5 Discussion and Final Prediction

Table 7: Final Prediction

Input parameters	Output produced

`test = pd.Series(attacktype).values.reshape(1,-1)`
 	
`randomforestclass.predict(test)`
	`randomforestclass.predict(test)`
	`array([0])`

Based on comparisons made on the dataset with structured and unstructured data, it was noted that the two classification models from the test scenario in Table [7] out of the four names, Logistic Regression and Random Forest Classifier were able to produce an accuracy score in the 90-100 range. Taking this into consideration Random Forest was taken as the final prediction classifier and a new attack type for a given input problem was predicted.
For better understanding, the Output: **array([0])** has been converted into categorical value in table [7] to signify what attack type has been predicted. This is done by changing the label encoded Y-axis to a non-encoded one while creating the train and test sets.

Table 8: Label encoding comparison

Without Label Encoding	With Label Encoding
`X_train,X_test,Y_train,Y_test=train_test_split(x,y,test_size=0.3,random_`	`y1=le.fit_transform(y1)` `X_train,X_test,Y_train,Y_test=train_test_split(x,y1,test_size=0.3,random_sta`

The result is the same output as in Table [6] but in a character format.

```
randomforestclass.predict(test)
array(['Installed malware'], dtype=object)
```

Figure 18: Categorical result

The prediction of attack type for a cybercrime improved significantly when cased based reasoning (CBR) methodology was followed. (Rasmi, 2013) proposed an (SAI) model based on Attack intentions algorithm that achieved a prediction score of (68%) which is less than 95% which was computed in this research using the case-based reasoning methodology. The result of the prediction can be interpreted as an event where an attacker has tried to inject malware into the system, thus indicating that this attacker is a malware hacker. The investigation thus can proceed towards identifying similar attack vectors by narrowing down the search of attack types and providing an efficient approach in apprehending the perpetrator.

The conclusion drawn from these experiments is that while a majority of the Classification algorithm work efficiently on categorically numeric data, it was seen that the KNN algorithm performed poorly on categorially equal dataset.

7 Conclusion and Future Work

The Research was conducted to evaluate and detect cybercriminal attack types using machine learning. Multiple classification models were selected to identify the highest accuracy measure and thus predict the attack type of a newly fed problem to the system. Based on the literature review conducted, the importance of criminal profiling was discussed with emphasis on cyber-crime or computer-based crime events. Furthermore, the types of cybercrimes were discussed to form a baseline for current research with gradually progressing towards the contributions of machine learning in this field.

Based on research conducted, the Case-based reasoning methodology was selected as the foundation of a classification model. Sequentially, four classification models were finalised based on their applicability and availability of resources needed for implementation and their relevance in the application of CBR methodology. Since the data was in an unorganised format, a data cleansing procedure was carried out to rid the data of null values and prepare the data for classification and prediction. Other activities performed on data included sampling, label encoding and scaling. Upon completion of the said activities, training and testing models were computed and classification was performed. The evaluation of these classification models was then done using the confusion matrix generated which determined that Random Forest and Logistic Regression models produced an accuracy score of 95% and 94% respectively. Thus, using one of the suitable models a prediction for a new attack type was made.

Discussing limitations, it was noted that due to lack of resources available regarding cyber-criminal profiling and time constraints relating to the creation of a new dataset from scratch with customised features, the dataset consisting of a small amount of data was selected for prediction. Apart from that, the classification models such as XGB, Naïve Bayes weren't included in the research due to a lack of research in the said models.

For the work in this research, a dataset with a relatively large amount of data and more categorical features such as gender, employment, age, etc. of the attacker shall be used to predict the attack type of an attacker with specific physical traits with emphasis on implementing more classification models and improving the accuracy of the models that performed at a low success rate in this research.

8 References

Aminnezhad, A., 2012. A Survey on Privacy Issues in Digital Forensics. *international journal of Cyber-Security and Digital Forensics,* 2012(2305-0012), pp. 183-199.

Baumgartner, K., 2008. Constructing Bayesian networks for criminal profiling from limited data. *Knowledge-based systems,* 21(7), pp. 563-572.
Bernstein, C., 2019. *case-based reasoning (CBR).* [Online]
Available at: https://searchenterpriseai.techtarget.com/definition/case-based-reasoning-CBR
[Accessed 01 December 2021].
Bilen, Abdulkadir & Özer, A. B., 2021. *Cyber-attack method and perpetrator prediction using machine learning algorithms.,* Turkey: PeerJ. Computer science.
Chanjin, L., 2015. Digital Forensic for Location Information using Hierarchical Clustering and k-means Algorithm. *Journal of Korea Multimedia Society,* 19(1), pp. 30-40.
Ferrante, A., 2018. *The impact of GDPR on WHOIS: Implications for businesses facing cybercrime.* [Online]
Available at: https://www.fticonsulting.com/~/media/Files/us-files/insights/articles/impact-gdpr-whois-implications-businesses-facing-cybercrime.pdf
[Accessed 05 December 2021].
Garcia, N., 2018. *THE USE OF CRIMINAL PROFILING IN CYBERCRIME INVESTIGATIONS.* [Online]
Available at:
https://www.researchgate.net/publication/327187114_The_use_of_criminal_profiling_in_cybercrime_investigations
[Accessed 05 December 2021].
I Watson, 1999. Case-based reasoning is a methodology and not a technology. *Journal of Cleaner Production,* 12 (5), pp. 303-308.
Jantan, M. R. &. A., 2011. AIA: Attack Intention Analysis Algorithm Based on D-S Theory with Causal Technique for Network Forensics - A Case Study. *International Journal of Digital Content Technology and its Applications,* 5(9), pp. 230-237.
Ko, R., Tsen, E. and Slapnicar, S., 2020. *Dataset of data breaches and ransomware attacks over 15 years from 2004.* Australia: The University of Queensland.
Mantaras, R. L. D., 2005. Retrieval, reuse, revision, and retention in casebased reasoning. *The Knowledge Engineering Review,* 0(1), pp. 1-2.
Mee Lan Han, 2019. CBR-Based Decision Support Methodology for Cybercrime Investigation: Focused on the Data-Driven Website Defacement Analysis. *Security and Communication Networks,* 2019(1939-0122), p. 21.
Nakid, S., 2021. *An analysis on the importance of criminal profiling in cybercrimes and how digital evidence acquired during forensic investigations can help further improve it. ,* Dublin: National College of Ireland.
Rasmi, m., 2013. *A new algorithm to estimate the similarity between the intentions of the cyber crimes for network forensics.* Malaysia, Science Direct.
Rogers, M., 2003. The role of criminal profiling in computer forensics process. *Computers & Security,* 22(04), pp. 292-298.
W. Peng, Z. Wang and J. Chen, 2009. Research on Attack Intention Recognition Based on Graphical Model. *Fifth International Conference on Information Assurance and Security,* 1(10.1109), pp. 360-363.