

**THESIS REPORT**  
**ON**  
**Pattern Elicitation & Recognition of Cyber Attacks by Machine Learning**



**DAFFODIL INTERNATIONAL UNIVERSITY**  
**DHAKA, BANGLADESH**  
**JANUARY 2023**

**Pattern Elicitation & Recognition Of Cyber Attacks by Machine Learning.**

**BY**

**Md. Naeem Aziz**

**ID: 213-25-035**

This Report Presented in Partial Fulfillment of the Requirements for the Degree of  
Master of Science in Computer Science and Engineering

Supervised By

**Ms. Naznin Sultana**

Assistant Professor

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

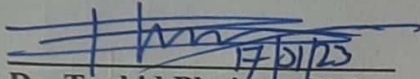
**DHAKA, BANGLADESH**

**17 January 2023**

## APPROVAL

This thesis titled “**Pattern Elicitation & Recognition Of Cyber Attacks by Machine Learning**”, submitted by Md. Naeem Aziz, ID No: 213-25-035, to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering (MSc) and approved as to its style and contents. The presentation has been held on 17-01-2023.

### BOARD OF EXAMINERS

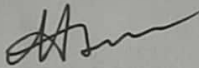


**Dr. Touhid Bhuiyan, PhD**

**Professor and Head**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Chairman**

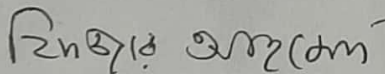


**Ms. Nazmun Nessa Moon**

**Associate Professor**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**

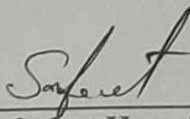


**Dr. Fizar Ahmed**

**Associate Professor**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



**Md. Safaet Hossain**

**Associate Professor & Head**

Department of Computer Science and Engineering  
City University

**External Examiner**

## DECLARATION

I hereby declare that this thesis has been done by me under the supervision of **Ms. Naznin Sultana, Assistant Professor, Department of CSE**, and Daffodil International University. I'm also notifying that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

**Supervised by:**

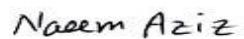


---

**Ms. Naznin Sultana**

Assistant Professor  
Department of CSE  
Daffodil International University

**Submitted by:**



---

**Md. Naeem Aziz**

ID: 213-25-035  
Department of CSE  
Daffodil International University

## ACKNOWLEDGMENT

First, I am very tons thankful to Almighty Allah for his divine blessing on me which enables me to finish this studies successfully.

I express my cordial gratitude and am obstructed to my enterprising supervisor Ms. Naznin Sultana, Assistant Professor, Department of CSE, Daffodil International University, Dhaka. Deep consciousness & piercing convenience of my supervisor in the field of "*Machine Learning*" to carry out thisthesis. Her infinite staying energy, erudite steering, chronic enkindling, perpetual and active superintendence, high-quality complaint, treasured recommendation, analyzing many deficient diagrams, and castigating them in any observance ranges have made it possible to exhaustive this assignment.

I would like to expansive my pompous cordial reception to Prof. Dr. Thouhid Bhuiyan, Department Head, Department of CSE, for his affectionate facilitation to conclude this thesis and also to other faculty members and the staff of the CSE department of Daffodil International University.

I would like to thank the faculty of Daffodil international university, who took elements in this entire painting in lots of ways.

Ultimately, I am well known with due appreciation to my mother and father who're continually assisting me and praying for me.

## **ABSTRACT**

This research is a finding and recognition type of research. That paper focuses on exploratory data analysis and YOLOv3, you only look once at version three, a real-time recognizing model. The hypothesis of this research are, a cyber-attack gives the same pattern in all the different, different IP addresses, and the research model, the YOLOv3 model can recognize the cyber-attack by seeing its pattern. First, the researcher collects more than one hundred seventy-eight thousand data from the cyber department of some companies. Then, the researcher does exploratory data analysis of that data in the jupyter notebook. Then, the researcher finds all the important information about cyber-attacks. Then, the researcher finds the patterns of some cyber-attacks from the information of the data collection. Then, the researcher collects pictures of the patterns. Then, with pictures of those patterns, the researcher labeled those pictures by labelling software and create a zip file of them. Then the researcher use Google colab and trained, "you only look once version three", the YOLOv3 model to detect the name of the pattern. The machine can detect the pattern and can tell us about what cyber-attack it is by only seeing the picture of the pattern. The researcher finds the patterns for eight IP addresses and in all the IP addresses, the attacks give the same patterns. So, the hypothesis is true in all that cases. By this research model, we can easily know about any kind of cyber-attack in detail and also can find the cyber-attacks by their patterns. So, the researcher use jupyter notebook, Google Colab, and pycharm environment. Pattern study always plays an important role to know about anything. By knowing and learning about anything's pattern, we can easily understand anything. This research does that thing greatly. The researcher first finds the patterns of some cyber-attacks, then make a model which can recognize the cyber-attack by just seeing the pattern.

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>Page</b>
Board of examiners	ii
Declaration	iii
Acknowledgment	iv
Abstract	v
Table of contents	vi-ix
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-5</b>
1.1 Prelude	1
1.2 Motivation	2
1.3 Rationale of study	3
1.4 Research Questions	3
1.5 Research Objectives	4
1.6 Expected output	4
1.7 Report Layout	5
<b>CHAPTER 2: BACKGROUND</b>	<b>6-9</b>
2.1 Introduction	6
2.2 Related works	6
2.3 Research Summary	7
2.4 Scope of the problem	8
2.5 Challenges	8

<b>CONTENTS</b>	<b>Page</b>
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>10-33</b>
3.1. Introduction	10
3.2 Research subject	10
3.3 Working procedure	13
3.3.1 Data Collection	14
3.3.2 Data processing in Jupyter Notebook	19
3.3.3 Exploratory Data Analysis	22
3.3.4 Data Supposition	22
3.3.5 Training Machine & Finding Patterns	28
3.3.6 Training Google Colab	29
3.3.7 YOLOv3 Model	29
<b>CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION</b>	<b>34-45</b>
4.1 Introduction	34
4.2 Experimental Results	36
4.3 Discussion	44
4.4 Summary	45
<b>CHAPTER 5: IMPACT ON SOCIETY, CYBER WORLD AND SUSTAINABILITY</b>	<b>46-47</b>
5.1 Impact on Society	46
5.2 Impact in Cyber World	46
5.3 Ethical Aspect	47
5.4 Sustainability Plan	47



<b>CONTENTS</b>	<b>Page</b>
<b>CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH</b>	<b>48-51</b>
6.1 Summary of study	48
6.2 Conclusion	49
6.3 Recommendation	50
6.4 Future research	50
<b>Appendix</b>	<b>51-53</b>
<b>References</b>	<b>54</b>
<b>PLAGIARISM REPORT</b>	<b>55</b>

## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE</b>
Figure 3.1: Working procedure	13
Figure 3.2: Data after cleaning	21
Figure 3.3: EDA process	22
Figure 3.4: Number of attacks	23
Figure 3.5: Top six attacks	24
Figure 3.6: Correlation matrix	25
Figure 3.7: Heat map	26
Figure 3.8: Pair plot	27
Figure 3.9: Scatterplot	28
Figure 3.10: General YOLOv3 architecture	30
Figure 3.11: Layer architecture of YOLOv3	31
Figure 3.12: YOLOv3 training model procedure	32
Figure 4.1: Patterns in eight IPv4 addresses	36
Figure 4.2: Final Patterns	37
Figure 4.3: Usher & result 1	38
Figure 4.4: Usher & result 2	39
Figure 4.5: Usher & result 3	40
Figure 4.6: Reconnaissance attack	41
Figure 4.7: Shellcode attack	41
Figure 4.8: Worms attack	41
Figure 4.9: Fuzzers attack	42
Figure 4.10: Exploits attack	42
Figure 4.11: DOS attack	42
Figure 4.12: Generic attack	43
Figure 4.13: Backdoor attack	43
Figure 4.14: Analysis attack	43

# CHAPTER 1

## INTRODUCTION

### 1.1 Prelude

In the modern world, a pattern is the maximal common and known word for problem solvers. Knowing a pattern of something is a great advantage to gaining knowledge about that. Patterns are a series of numbers, shapes, or gadgets that observe a positive rule to stay identical or change. Patterns offer an experience of order in what would possibly otherwise seem chaotic. Researchers have observed that information is capable of identifying routine patterns allowing us to make knowledgeable guesses, assumptions, and hypotheses; it enables us to expand essential talents of important thinking and good judgment. Machine Learning is a field where developer trains their machines and teach them what to do. The machine is faster than a human. So, when we teach them what and which way to do things they do those things faster than humans, and a lot of time saves from it. Python is a high-level computer language by which we can also train machines. So, ultimately python can be used for machine learning techniques. The researcher uses Jupiter notebook to do python programming and train machine to do what the researcher wants. Jupiter notebook is a web application by which we can do cryptograms and also usher the cryptogram and see the results. Here, the researcher collects the data of some software companies' cyber department's servers results and then analyzes the data and finds the patterns of some cyber-attacks. It is a groundbreaking finding in the cyber-security world. Because when we learn the pattern of anything, we can know everything about it and also gain the knowledge of how to solve that problem and what to do to stop that from happening again in the future. After collecting the data researcher finds some attacks which happened the most, then the researcher uses python cryptogram and teaches the machine so that machine can give the output. Researcher finds the pattern of some cyber-attacks. These are Reconnaissance, Fuzzers, Analysis, Backdoor, Exploit, Generic, Shellcode, and Worm. To do a reconnaissance attack, the attacker first gathers all the information of computer networks that they want to attack, then circulate security controls. In the fuzzers attack, the attacker first nourishes the computer with some massive random invalid data to block it, and then they break the security loopholes of a computer. In an analysis-gestalt attack, the attacker creates a kind of intrusion that penetrates web programs via ports, emails, and net scripts.

The backdoor attack is a stealthy technique to keep away from ordinary authentication to make sure unauthorized faraway get the right of entry to a tool. An exploit is a cryptogram that takes acquire of its penetrable or maintenance blemish. It is composed both via protection researchers as an evidence-of-idea threat or via malignant actors for use of their operations. If a thrust with a hash function occurs on an encrypted message, the message can be blocked using a generic attack. By using a shell cryptogram, the attacker can penetrate a mediocre shred of cryptogram from the shell and gain control of the compromised device regardless of the encryption settings. To spread from one computer to another, worms replicate malignant scripts. The hypothesis of this research are, a cyber-attack gives the same pattern in all the different, different IP addresses, and the research model, the YOLOv3 model can recognize the cyber-attack by seeing its pattern. The researcher finds the most targeted destination IP Address, most logical ports attacked, the most common gestalt of attack, different times of the day and most important and main subject is to find the pattern of the cyber-attacks. So, the researcher trains the machine with python to find the pattern of the attack. Jupyter notebook is used to do python cryptogram and do the solution. For data collection, the researcher collects some company's cyber department's data and then works with that. The researcher knows that, when more data are used, the more correct the result will be. So, that's how the whole management has happened. It can be a great finding for the future cause when we learn any pattern of something then we can easily understand how to solve any problem created by that thing.

## **1.2 Motivation**

The researcher has acquired motivation a lot while conducting this research. The motivations are:

- The researcher is very much interested in machine learning and python programming.
- Wanted to do something with the cyber-security sector.
- Wanted to solve the pattern of cyber-attacks.
- Motivated to see the problems that everybody faces with unknown cyber-attacks.
- Have a great mentor who motivates the researcher about this topic to solve.
- To show how to gain knowledge about cyber-attacks and that desire also works as a motivation.

### **1.3 Rationale of the study**

This research is necessary because of the needs of this study. In this modern world, we use the cyber world to do many things in our daily life. Without it, almost all of us can't even stand a chance. But there are a lot of hackers and bad people who use their knowledge to destroy or steal other people's important data. And we even can't understand what kind of attack they do to harm us. That's why pattern study comes into the field. With pattern learning, we can easily understand every kind of knowledge about cyber-attacks. Knowing a pattern of something is a great advantage to gaining knowledge about that. Patterns are a series of numbers, shapes, or gadgets which observe a positive rule to stay identical or change. Patterns offer an experience of order in what would possibly otherwise seem chaotic. Researchers have observed that information is capable of identifying routine patterns allowing us to make knowledgeable guesses, assumptions, and hypotheses; it enables us to expand essential talents of important thinking and good judgment. When we understand the attack then it becomes too easy to solve the problems. It's like a car and road. When we know what road we have to use to reach our destination then it will be very easy to reach that destination. The pattern is like the easy road map. Almost every people face cyber-attacks in their life. Even, many experts can't identify the attack and don't know how to get out of this situation. Knowing the pattern of cyber-attacks can be the solution to this problem. That's why we need this gestalt of research so we can find the pattern of any attacks. So, we can do a solution to the problem. We also can learn everything about that attack. We can learn the most targeted destination IP Address, most Logical Ports attacked most Frequent/common gestalts of attacks, and different times of the day of the attack. All can know that important information and get rid of cyber-attack problems with the help of those processes. So, those are the rationale of the story and that's why this research is necessary.

### **1.4 Research Question**

This research is based on cyber security done by machine learning. Python is the programming the researcher used for this research. To do this research we face many questions and also make the solution to some questions. Those are:

- What are the most targeted Destination IP Address?
- What are the most logical ports attacked?

- What is the most frequent/common gestalt of attack?
- What are the times of the day the attack happened?
- How to find the Pattern of the attacks?
- Are all cyber-attacks giving the same pattern in all the different, different IP addresses?
- Is the research model, the YOLOv3 model recognizing all cyber-attacks from their patterns?

## 1.5 Research Objective

### General Objective:

- To find Patterns of cyber-attacks.
- To recognize the cyber-attacks from their patterns by making a YOLOv3 model.
- To prove that all cyber-attacks give the same pattern in all the different, different IP addresses.

### Specific Objective:

- To find the most targeted destination IP address.
- To find the most common gestalt of attacks.
- To find the most logical ports which have attacked.
- To find all the times of attacks.
- To develop a YOLOv3 model for recognition the cyber-attacks.
- To find what kind of pattern the cyber-attacks are giving.

## 1.6 Expected Outcome

The expected outcome of the research is to find all hidden data and information of all the attacks that happened in different cyber departments of some companies. That means the expected outcome is the findings of patterns of many attacks. And detecting the name of the attacks from the pattern. And for that findings researcher has to do python programming in the jupyter notebook platform. The researcher used the YOLOv3 model to recognize the patterns of the attacks.

## **1.7 Report Layout**

### **Chapter 1:**

This is the chapter where the researcher writes the introduction of the research. In this chapter, we write the goal, the objective, and the motivation to do the research.

### **Chapter 2:**

In this chapter, the researcher gives literature review of the research means the previous similar gestalt works and some related gestalt of information in that segment.

### **Chapter 3:**

In this chapter, the researcher describes how the work has been done, the full methodology of the research, and also describes the proposed working process model.

### **Chapter 4:**

In this chapter, the researcher shows the results of the research. The pattern models of the cyber-attacks and description of those models.

### **Chapter 5:**

The complete short description of the research, findings, impact on society, and future plan and future development with the research.

## **CHAPTER 2**

### **BACKGROUND**

#### **2.1 Introduction**

The background of the research is something where we see the literature review and the previous works similar to the current work. A literature evaluation is a bit of educational writing demonstrating expertise and knowledge of the educational literature on a specific subject matter positioned in context. A literature overview additionally consists of a critical evaluation of the fabric; this is why it is called a literature overview rather than a literature record. Here, the researcher talks about the related gestalt of work like this research done previously by others. The researcher also describes the research summary, the scope of the problem, and the challenges of the problem.

#### **2.2 Related works**

Now, we are going to see some works related to this research.

##### **Finding patterns by EDA**

To do this research, they use Google analysis tools and exploratory data analysis to do this research. This analysis is done by Pamela Fox. It's sometimes helpful to discern that records in a catalog, like an epoch dilution, overstep sketch, or scatter pare lot, based on the statistics and styles. Statistics and styles can sometimes be viewed in an unswerving tabular format. There are other times when catalogs, such as time series, epistles, and scatter plots, help discern the records. In some cases, one can discern the sample in a catalog, such as a time collection, epistle, or scatter plot, depending on the statistics and styles. Depending on the statistics and the styles, they sometimes discern that sample through an unswerving tabular presentation. Other times, cataloging bits of help, whether it's a time collection, an epistle, or a scatter plot. The statistics and styles allow us to present the sample occasionally in a tabular format. Occasionally, an epistle, epistle, or scatter plot helps us discern the data. There are times when an unswerving tabular presentation of the statistics is enough. Other times, we can discern the records in a catalog, as with a time collection, epistle, or scatter plot [1].



### **Fruit recognition using YOLOV3 algorithm**

Computers can learn from deep learning, which is a gestalt of machine learning by a specimen, just as humans do. Driverless cars use deep learning to recognize stop signs and distinguish pedestrians and lampposts. Deep learning enables machines to do what humans do naturally: learn from a specimen. This is the key to driverless cars, which can recognize a stop sign and distinguish pedestrians from lampposts. Here, in that research, they do deep learning to recognize specific fruits' names by seeing their picture [2], [3].

### **Object Recognition Using YOLOv3 algorithm**

Machine Learning is a field where developer trains their machines and teach them what to do. The machine is faster than a human. So, when we teach them what and which way to do things they do those things faster than humans, and a lot of time saves from it. They use the YOLOv3 model and OpenCV means open-source computer vision, a library of machine learning. It also uses deep learning for this research. Computers can learn from deep learning, which is a gestalt of machine learning, just as humans do. Deep learning enables machines to do what humans do naturally: learn from a specimen. This is the key to driverless cars, which can recognize a stop sign and distinguish pedestrians from lampposts. In this case, they use data with information about every animal and its details information. So, the machine can recognize them by the similarity which is provided. That's how the machine is working for this one. Now, we'll see how the machine work, a visual representation of this research [4].

## **2.3 Research Summary**

At first, the researcher collect the related data from some companies' cyber departments. Then, the researcher makes the proper environment for python and jupyter notebook. Then, have to import some engines and libraries. Those libraries are pandas, seaborn, NumPy, missingo, etc. Then, have to clean the data. For cleaning the data, missingo is the library that worked. By this, we find the missing data and delete those missing or unavailable data from rows or columns. After cleaning the data we do python programming and find out some arrays, matrix, and other visual representations of data. And at last, researchers find the pattern of cyber-attacks. Then, the researcher creates more environments on the computer. Created a tensor-flow platform. It's a

python friendly open source library. Then created an open-source vision library environment. After that, the information about all attacks is given so, the machine can identify the targeted destination. Then researcher loads the image by python programming. Then, the researcher commands the machine to detect the pattern and identify them. And the machine detects all the attack patterns correctly. So, now we can say that the machine is working perfectly.

## **2.4 Scope of the problem**

The scope of the research is to give the cyber world the proper services. To solve the problem of cyber-attacks first, we have to learn about everything about a cyber-attack. This research helps us to learn everything about cyber-attacks and detect the cyber-attacks. Some scopes of the research are:

- Can identify the attack gestalt.
- Can identify the name of the cyber-attacks.
- Finding the pattern of the attack.
- Can learn every detail of the attack so, anyone can get rid of the problems.
- Can detect the name of the attack from the pattern.

## **2.5 Challenges**

The IT industry may find it extremely difficult to acquire Machine Learning knowledge of initiatives. The difficulty of these tasks can be an epithet of a variety of factors. Several factors contribute propriety and reliability of the systems, including the amount of information to be processed, the complexity of the algorithms to be used, and the volume of information to be processed. Similarly, a system gaining knowledge of tasks may be time-eating and costly to expand and install. While system-mastering tasks in the IT industry can be challenging, they are also very rewarding [5]. Python is a high-level computer language by which we can also train machines. So, ultimately python can be used for machine learning techniques. The researcher uses Jupiter notebook to do python programming and train machine to do what the researcher wants. Jupyter notebook is a web application by which we can do cryptograms, usher the cryptogram, and see the results. Here, the researcher collects the data of some software companies' cyber department's servers results and then analyzes the data and finds the patterns of some cyber-attacks. It is a groundbreaking finding in the cyber-security world. Because when we learn the pattern of anything, we can know everything about it and also gain the knowledge of how to solve that

problem and what to do to stop that from happening again in the future. After collecting the data, the researcher finds some attacks which happened the most, then the researcher uses python cryptogram and teaches the machine so that machine can give the output. Researcher finds the pattern of some cyber-attacks. Patterns are a series of numbers, shapes, or gadgets which observe a positive rule to stay identical or change. Patterns offer an experience of order in what would possibly otherwise seem chaotic. Researchers have observed that information is capable of identifying routine patterns allowing us to make knowledgeable guesses, assumptions, and hypotheses; it enables us to expand essential talents of important thinking and good judgment. Then created an open-source vision library environment. After that, the information about all attacks is given so, the machine can identify the targeted destination. Then researcher loads the image by python programming. Then, the researcher commands the machine to detect the pattern and identify them. And the machine detects all the attack patterns correctly. So, now we can say that the machine is working perfectly. Then, after doing the detection process done by python, the machine now can detect the name of the attacks by seeing their patterns.

The challenges that are faced are:

- Collecting all the data related to the research.
- The collected data was disorganized and unclean. Had to clean them too.
- Resolved plenty of issues to find information about cyber-attacks because the data was not organized. We had to organize them first.
- The best challenging part was, finding the pattern of cyber-attacks unerringly.
- Teaching the machine to recognize the cyber-attack name from its patterns was also a big challenge in this research.
- Overall, the creation of the proper environment for all the libraries and engines to do this research was also challenging.

## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Introduction

In this part, will know the methodology of this research, and the processes I do to complete this research. At first, the researcher collect the related data from some companies' cyber departments. Then, the researcher makes the proper environment for python and jupyter notebook. Then, have to import some engines and libraries. Those libraries are pandas, seaborn, NumPy, missingo, etc. Then, have to clean the data. For cleaning the data, missingo is the library that worked. By this, we find the missing data and delete those missing or unavailable data from rows or columns. After cleaning the data we do python programming and find out some arrays, matrix, and other visual representations of data. And at last, researchers find the pattern of cyber-attacks. Then, the researcher creates more environments on the computer. Created a tensor-flow platform. It's a python friendly open source library. Then created an open-source vision library environment. After that, the information about all attacks is given so, the machine can identify the targeted destination. Then researcher loads the image by python programming. Then, the researcher commands the machine to detect the pattern and identify them. And the machine detects all the attack patterns correctly. So, now we can say that the machine is working perfectly. This research is all about finding and recognizing patterns of cyber-attacks. And this research proves that in every IP address the pattern of a cyber-attack is the same.

#### 3.2 Research Subject

The research subject is to find the pattern of the cyber-attacks and also recognize the name of the cyber-attacks. The researcher collect the related data from some companies' cyber departments. Researchers have observed that information is capable of identifying routine patterns allowing us to make knowledgeable guesses, assumptions, and hypotheses; it enables us to expand essential talents of important thinking and good judgment. Machine Learning is a field where developer trains their machines and teach them what to do. The machine is faster than a human. So, when we teach them what and which way to do things they do those things faster than humans, and a lot of time saves from it. Python is a high-level computer language by which we can also train machines.

So, ultimately python can be used for machine learning techniques. The researcher uses Jupiter notebook to do python programming and train machine to do what the researcher wants. Jupyter notebook is a web application by which we can do cryptograms and also usher the cryptogram and see the results. Here, the researcher collects the data of some software companies' cyber department's servers results and then analyzes the data and finds the patterns of some cyber-attacks. It is a groundbreaking finding in the cyber-security world. Because when we learn the pattern of anything, we can know everything about it and also gain the knowledge of how to solve that problem and what to do to stop that from happening again in the future. After collecting the data researcher finds some attacks which happened the most, then the researcher uses python cryptogram and teaches the machine so that machine can give the output. Researcher finds the pattern of some cyber-attacks. These are Reconnaissance, fuzzers, Analysis, Backdoor, Exploit, Generic, Shellcryptogram, and worms. To do a reconnaissance attack, the attacker first gathers all the information of computer networks that they want to attack, then circulate security controls. In the fuzzers attack, the attacker first nourishes the computer with some massive random invalid data to block it, and then they break the security loopholes of a computer. In an analysis-gestalt attack, the attacker creates a kind of intrusion that penetrates web programs via ports, emails, and net scripts. The backdoor attack is a stealthy technique to keep away from ordinary authentication to make sure unauthorized faraway get the right of entry to a tool. An exploit is a cryptogram that takes gain of software or safety fissures. It is written both via protection researchers as an evidence-of-idea threat or via malignant actors for use in their operations. A generic attack is a technique that attempts to block encryption using a hash function for thrust regardless of encryption settings. In the shellcryptogram technique, the attacker penetrates a mediocre shred of cryptogram from the shell to control the compromised regardless of encryption settings. A worm attack replicates a malignant script to spread it to other computers. Often, it uses a computer network to spread depending on security fissures in the destination computer. Those are the gestalts of attacks the researcher works with. The researcher finds the most targeted destination IP Address, most logical ports attacked, the most common gestalt of attack, different times of the day and most important and main subject is to find the pattern of the cyber-attacks. So, the researcher trains the machine with python to find the pattern of the attack. Jupyter notebook is used to do python cryptogram and do the solution. Then, the researcher makes the proper environment for python and jupyter notebook. Then, have to import some engines and libraries. Those libraries are pandas, seaborn,

NumPy, missingo, etc. Then, have to clean the data. For cleaning the data, missingo is the library that worked. By this, we find the missing data and delete those missing or unavailable data from rows or columns. After cleaning the data we do python programming and find out some arrays, matrix, and other visual representations of data. And at last, researchers find the pattern of cyber-attacks. Then, the researcher creates more environments on the computer. Created a tensor-flow platform. It's a python friendly open source library. Then created an open-source vision library environment. After that, the information about all attacks is given so, the machine can identify the targeted destination. Then researcher loads the image by python programming. Then, the researcher commands the machine to detect the pattern and identify them. And the machine detects all the attack patterns correctly. So, now we can say that the machine is working perfectly. Then, the researcher use Google Colab to use the YOLOv3 machine learning algorithm. YOLOv3 is an actual-time item detection set of jus that identifies precise gadgets in motion pictures, live nourishes, or snapshots [5]. The YOLO system mastering algorithm ultra-modern functions discovered by way of a deep convolucional neural community to hit upon an object. The 1/3 version of today's YOLO device's modern-day set of jus is a more accurate model modern-day the unique ML algorithm. The first version of modern-day YOLO became created in 2016, and version 3, that's discussed considerably in this newsletter, become made two years later in 2018. YOLOv3 is an improved model of modern-day than the quondam others. YOLO has been applied with the use of the Keras or OpenCV machine learning libraries for this gestalt of work. The researcher use OpenCV libraries for this research. YOLO is a Convolutional Neural community (CNN) for appearing object detection in real-time. CNN's are classifier-based systems that could manner enter umbrages as established arrays of information and recognize styles between them (view image underneath). YOLO has the gain of being a great deal quicker than different networks and nevertheless continues accurate. It permits the version to examine the entire photo at test time, so its predictions are knowledgeable employing the global context inside the photo. Excessive-scoring areas are referred to as effective detections of something class they most carefully perceive with. For specimen, in a stay nourish of visitors, YOLO may be used to stumble on one-of-a-kind kinds of vehicles depending on which regions of the hieroglyphic score extraordinarily in evaluation to predefined training of cars. The researcher does YOLO to detect the exact cyber-attack name by seeing the patterns of some cyber-attacks. So, that's how the process was done in this research.

### 3.3 Working procedure

This is huge research with a huge amount of work. First, we see the working procedure of this research and then we describe the procedure. First, collect the data and then finds the patterns of the attacks. Then, create a model which can recognize the attacks by only seeing the pattern of the attacks.

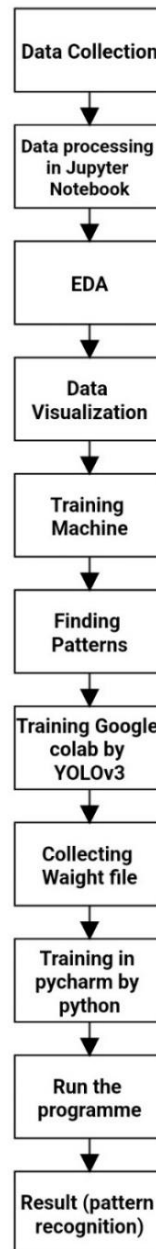


Figure 3.1: Working Procedure

### **3.3.1 Data Collection**

The researcher worked with two gestalts of data. The first is text data and then the researcher finds the pattern's picture with the text data. Then the researcher works with the picture data of the patterns and makes a recognition machine on the picture data. The amount of text data is one lakh seventy-eight thousand plus. For data collection, the researcher takes four companies' cyber department's data. The process of collecting and studying the right facts from various sources to find solutions to explore problems, features, options, etc. to evaluate viable implications is called fact gathering [6]. Keep scrolling for a better understanding. Information is power, statistics is understanding, and facts are facts in digitized form, at least as described in IT. For this reason, data is electricity. But before you can use these facts directly to a successful method of your business or commercial enterprise, you need to acquire them. To help you get started with this technique, we highlight a collection of boards. Moreover, what are the special kinds of statistics collection? And what gestalts of streak equipment and streak strategies are there? If you need to accelerate approximately what is a record collection method, you're in the right place.

### **Challenges in data collection**

#### **Data Quality Issues**

The primary chance for a wide and successful machine mastering software is terrible facts satisfactory. Exceptional statistics must be your top priority if you want technology like device recognition to deliver the results you want. In this weblog article let's tell approximately the number of maximum widespread information problems and the way to restore them.

#### **Inconsistent Data**

When working with many information assets, it is achievable that the same facts will have differences between sources. Variations can be in cryptograms, widgets, or often spelling. In addition, the emergence of inconsistent information may occur during mergers or relocations of companies. Inconsistencies in facts will be prone to build up and depreciate the information if usually not resolved. Companies that have narrowly focused on record consistency will achieve this because they simply want reliable data to support their analysis.



## **Data Downtime**

Information is the use of pressure in the back of selections and operations of information-pushed agencies. However, there may be brief periods when their data is unreliable or not ready. Client management and subpar analytical effects are the most adept methods of how this unavailability of facts will have a broad impact on companies. A statistical engineer spends approximately eighty percent of his time updating, maintaining, and ensuring the integrity of the information nourish. When it comes to asking the following question about commercial enterprises, there is excessive marginal dread due to the extended operational time from information capture to understanding. Schema modifications and migration issues are just two specimens of reasons for downtime. Statistics can be difficult due to their size and complexity. Data downtime must be constantly monitored and should be shortened by automation.

## **Ambiguous & Duplicate Data**

Regardless of thorough oversight, some errors can still occur in huge databases or data lakes. For recordings streamed at high speeds, the problem will be even greater. Spelling errors can go unnoticed, formatting problems can occur, and column headings can be misleading. This unclear information can cause some problems when creating reports and analyses. Streaming statistics, neighborhood databases, and cloud recording lakes are just a few of the sources of statistics that today's companies should contend with. They could additionally have application and system silos. These resources are likely to replicate and overlap each other in nice shreds. For specimen, replica contact information has a huge impact on consumer entertainment. If certain options are ignored, even when others engage repeatedly, advertising and marketing campaigns fail. The likelihood of biased analytical results increases, while reproductive information is a gift. It is also able to bring ML mods with biased facts about education.

## **Too Much Data**

While we emphasize record-based analytics and its benefits, there is a data-satisfying problem with excessive statistics. There is a threat of getting lost in the mass of information when searching for statistics related to your analytics efforts. Records scientists, information analysts, and enterprise customers spend 80% of their work finding and organizing the right data. As the range of facts increases, various information issues become more important, especially when dealing with streaming facts and large files or databases.

### **Inaccurate & Hidden Data**

For a fantastically regulated group like healthcare, factual propriety is essential. Given the current enjoyment, it is more important than ever that the data for COVID-19 and subsequent pandemics is first class. Bad records now don't give you a true picture of the scenario and can't be used to plan a quality course of action. Personalized consumer stories and advertising and marketing strategies don't work if your consumer information is incorrect. Statistical inaccuracies can be an epithet of several things, including degradation of records, human error, and drift of facts. The breakdown of international facts occurs at a fee of about three% following the month, which is quite alarming. The integrity of information may be compromised as it is transferred between disparate systems, and good information may deteriorate over time. Most corporate people make the best use of some component in their facts, with the rest lost here and there in the silos of facts or discarded in record graveyards. For specimen, a customer support group may not obtain customer data from revenue and cannot create specific and complete customer profiles. Missing out on opportunities to develop new merchandise, embellish offerings, and streamfootsie strategies is because of hidden information.

### **Finding Relevant Data**

Finding usable statistics is not so straightforward. There are several elements that we want to keep in mind when searching for relevant statistics, including relevant domain, relevant demoepistleics, relevant terms, and many other factors that we want to keep in mind even when trying to find relevant data. Statistics that do not apply to our view of any of the elements make it obsolete and we cannot effectively continue to analyze it. This will result in incomplete research or analysis, re-gathering the facts over and over again, or turning off the look feature.

### **Deciding the Data Collect**

Determining what information to collect is one of the most critical factors in gathering statistics and should be one of the first elements in gathering data. We need to choose the topics the statistics will cover, the sources we will use to collect them, and the number of facts we can require. Our answers to these questions will depend on our goals or what we expect to achieve by using your facts. As a specimen, we can also choose to gather facts about categories of articles that often get between 20 and 50 website traffic at most. We may also choose to collect data on the daily age of all clients who have purchased your business within the previous month. If you don't already address it, it can cause double work and irrelevant statistics, or ruin your examination whole.

## **Dealing with Big Data**

Huge statistics refer to extremely large statistical sets with extra complicated and varied structures. These homes generally make for better problems in saving, parsing, and using other remediation strategies. In particular, large records refer to information units that are relatively large or complicated, and traditional information processing facilities are inadequate. The substantial amount of data, both unstructured and structured, that businesses deal with every day. The amount of records produced by medical packages, the internet, social networks, social networks, sensor networks, and many different businesses is growing rapidly due to current technological advancements. Huge information refers to large amounts of facts compiled from many sources in various formats at an incredibly fast pace. Handling this kind of information is one of the many challenges of fact-gathering and is a critical step toward effective statistical series.

## **Low Response and Other Research Issues**

Negative design and low response fees are two problems with information gathering, specifically in fitness surveys that have used questionnaires. This could result in insufficient or insufficient statistics being provided for the study. Creating a stimulating statistical series software is probably useful in this situation to get more answers.

## **Data collection process**

### **Decide the working data gestalt**

The first factor we want to do is decide what information we want to collect. We need to choose the topics that will hide the information, the sources that we can use to get it, and the number of facts that we will need. As a specimen, we can also choose to get facts about the kinds of products that the average e-commerce website tourist between the ages of 30 and 45 most searched for.

### **Fixed time for collection**

The technique of creating access for information series can now begin. We must set a deadfootsie for information collection at the beginning of our planning phase. Several gestalts of facts that we would like to obtain continuously. For specimen, we would have to create a way to track transaction statistics and website traffic facts over the long term. But we tune the data at a certain stage in a certain time frame if we are tracking it for a specific campaign. In these conditions, we can have an agenda for when we can start and stop the collection of statistics.

## **Selecting a certain technique**

We can choose the technique of a series of facts to act as inspiration for our statistics gathering plan in this degree. We have to remember the kind of statistics we want to collect, the period in a certain phase in which we can receive them, and other factors that we directly determine to choose a good collection approach.

## **Collecting information**

Once our plan is complete, we can put our data collection plan into action and start collecting data. In our DMP, we can store and organize our data. We have to be careful to stick to our plan and see how things are going. Especially if we collect data regularly, it can be useful to set up a schedule to check how our data collection is going. As circumstances change and we learn new details, we may need to adjust our plan.

## **Getting ready for data processing**

It's time to track our stats and organize our findings after we've gathered all our information. The degree of analysis is important because it transforms raw facts into insightful understanding that can be executed to better our advertising plans, items, and commercial enterprise judgments. The analytics tools included in our DMP can be used to help with this segment. Once we find styles and insights in our information, we can use those discoveries to beautify our business. There may be a lack of freely available facts. There are stats here and there, but we don't have to access them. As a specimen, unless we have a compelling motive, we can't shamelessly look at another man or woman's medical statistics. It can be difficult to rate the numerous record styles. Don't forget how time-consuming and difficult it will be to collect individual data when figuring out what records to take. Identifiers or information describing the context and offer of a survey response are simply as important as records approximately the challenge or program we are discovering. At the strip, adding more identifiers will allow us to more accurately determine the successes and failures of our application, but moderation is the key. Even though pleasant motion management (detection/tracking and hitting) occupies the area always after and at some stage of the statistical series, the specifics must be carefully special in the process guide. Organizing surveillance structures requires a special form of communication, which is a necessary prerequisite. After devising the issue of fact-gathering, there may be no ambiguity about the flow of information

between the number one investigator and the personnel staff. A poorly designed verbal exchange system promotes a lack of oversight and reduces the ability to detect errors. As a part of detection or monitoring, direct calls can be used with conventions to observe workers at some stage of site visits, or frequent or routine checks of data reports for discrepancies, disproportionate numbers, or invalid cryptograms. Website visits will not be suitable for all disciplines. However, without a routine audit of records, whether qualitative or quantitative, it will be difficult for investigators to confirm that record collection is being conducted according to the methods described in the manual. Now, the researcher will describe the data processing part work, and how the data is processed in jupyter notebook will be described next.

### **3.3.2 Data processing in Jupyter Notebook**

Furnish is processed individually and in batches in transaction processing structures. Batch processing involves grouping similar furnish together and processing this company as a batch. Salary assessments, for specimens, can be processed in batches. All time cards for the pay period are gathered and the resulting payrolls are processed and published in a set or batch. It is common for receivables and money owed to be processed in batches. For instance, invoices to providers are often processed in batches. Actual-time processing takes place even as furnish are processed without delay. This processing is interactive as the transaction is processed while the miles are entered. While determining whether or not or now not batch processing or real-time processing is suitable, system specialists have to bear in mind reaction time, performance, complexity, dealing with, and garage media. Batch systems have sluggish responses due to the fact furnish are not processed till the whole group is ready for the method. Real-time systems are responsive due to the fact furnish are processed as they're entered. Batch processing is greater inexperienced for a huge extent of similar furnish. That is proper for plenty of motives. First, people whose attention is on gathering and processing comparable varieties of furnish are getting parsimonious in processing the one's furnish. That means they end up extra specialized.

### **Data Cleaning**

Data cleansing is the procedure of fixing or casting off wrong, corrupted, incorrectly formatted, duplicated, or incomplete statistics inside a dataset. Whilst combining multiple information sources, there are many opportunities for records to be duplicated or mislabeled. For data cleaning, had to check some shreds of information.

## **Propriety**

How close is the price of the statistics to the genuine fee? In other words, how as it should be does the cost of the information describe the object or event being described?

## **Syntactic propriety**

In this situation, the cost is probably accurate, but it doesn't belong to the appropriate area of the variable. As a specimen: A bad dread for duration or age or a percentage better than one hundred.

## **Semantic propriety**

In this case, the cost is in an appropriate area, but it isn't accurate.

## **Consistency**

Do all of the dreads of 1 variable represent an identical definition? As a specimen: Distance is recorded in the same unit for the duration of the dataset.

## **Completeness**

How entire is the dataset concerning variable dreads and/or facts? Variable dreads: Are there dreads lacking for positive variables? Facts: Is the dataset entire for the analysis at hand? For instance, you got down to survey 1,000 households but most effective have 900 finished. The incidence of missing dreads will have one-of-a-kind causes. Humans may have refused or forgotten to reply to a question in a questionnaire, or a variable might not apply to a sure object. As a specimen, the variable "pregnant" with the two possible dreads, sure and no, does not make sense for men. Of route, one may want to constantly input the dread no for the characteristic pregnant. However, this could result in a grouping of men with now not-pregnant women.

## Time footsiess

How well-timed are the statistics? For instance: records must be accumulated within a described term.

## Distinctiveness

Are there any replica facts? Checking the individuality in a dataset consists of identifying and correcting duplicated rows (observations). For the specimen, the water point has been mapped two times.

## Validity

Does the information agree with the described regulations? For instance: The challenge identification needs to usually be 3 characters between A-Z. A dread that includes and is consequently invalid.

## Data analysis after cleaning

Attack subcategory	Protocol	Source IP	Source Port	Destination IP	Destination Port	Attack Name	Attack Reference	Start time	Last time	Destination Port Service
HTTP	TCP	175.45.176.0	13284	149.171.126.16	80	Domino Web Server Database Access: /doladmin.n...	-	1421927414	1421927416	HTTP
Unix 'r' Service	UDP	175.45.176.3	21223	149.171.126.18	32780	Solaris rwallid Format String Vulnerability (ht...	CVE 2002-0573 ( <a href="http://cve.mitre.org/cgi-bin/cv...">http://cve.mitre.org/cgi-bin/cv...</a> )	1421927415	1421927415	NaN
Browser	TCP	175.45.176.2	23357	149.171.126.16	80	Windows Metafile (WMF) SetAbortProc() Code Exe...	CVE 2005-4560 ( <a href="http://cve.mitre.org/cgi-bin/cv...">http://cve.mitre.org/cgi-bin/cv...</a> )	1421927416	1421927416	HTTP
Miscellaneous Batch	TCP	175.45.176.2	13792	149.171.126.16	5555	HP Data Protector Backup (https://strikecenter...	CVE 2011-1729 ( <a href="http://cve.mitre.org/cgi-bin/cv...">http://cve.mitre.org/cgi-bin/cv...</a> )	1421927417	1421927417	PERSONAL-AGENT
Cisco IOS	TCP	175.45.176.2	26939	149.171.126.10	80	Cisco IOS HTTP Authentication Bypass Level 64 ...	CVE 2001-0537 ( <a href="http://cve.mitre.org/cgi-bin/cv...">http://cve.mitre.org/cgi-bin/cv...</a> )	1421927418	1421927418	HTTP

Figure 3.2: Data after cleaning

### 3.3.3 Exploratory Data Analysis (EDA)

After the data processing period, the real exploratory data analysis begins. It's a concept of data science. EDA is all about, how to test a dataset. By EDA we can check data and make a picturesque form.

	Attack category	Attack subcategory	Protocol	Source IP	Source Port	Destination IP	Destination Port	Attack Name
174347	Generic	IXIA	udp	175.45.176.1	67520	149.171.126.18	53	Microsoft_DNS_Server_ANY_Query_Cache_Weakness_...
174348	Exploits	Browser	tcp	175.45.176.3	78573	149.171.126.18	110	Microsoft Internet Explorer 6.0 Png pngfilt.dl...
174349	Reconnaissance	HTTP	tcp	175.45.176.1	71804	149.171.126.10	80	Domino Web Server Database Access: /internet.n...
174350	DoS	Ethernet	pnni	175.45.176.3	0	149.171.126.19	-753	Cisco IPS Jumbo Frame System Crash (https://st...
174351	Fuzzers	OSPF	trunk-1	175.45.176.0	73338	149.171.126.13	0	Fuzzer: OSPF Hello Packet: Long Neighbor Lists...
...	...	...	...	...	...	...	...	...
178026	Generic	IXIA	udp	175.45.176.0	72349	149.171.126.12	53	Microsoft_DNS_Server_ANY_Query_Cache_Weakness_...
178027	Exploits	Browser	sep	175.45.176.3	67647	149.171.126.18	0	Persits XUpload ActiveX Method MakeHttpRequest...
178028	Exploits	Office Document	tcp	175.45.176.0	78359	149.171.126.13	110	Microsoft Excel SxView Memory Corruption (POP3...
178029	Exploits	Browser	tcp	175.45.176.2	68488	149.171.126.19	80	Internet Explorer createTextRange() Code Execu...
178030	Reconnaissance	ICMP	unas	175.45.176.3	77929	149.171.126.19	0	IP Options: Loose Source Route (IP Option 3) (...)

3684 rows x 11 columns

Figure 3.3: EDA process

### 3.3.4 Data Supposition

Data supposition is the discerned presentation of EDA results. We will see data epistemically in this part. Records supposition is a way to symbolize information epistemically, highlighting patterns and tendencies in statistics and helping the reader to achieve quick insights. Also called “interactive visible exploration,” it enables the exploration of facts via the manipulation of catalog umbrages, with the shade, brightness, size, shape, and motion of visible gadgets representing factors of the dataset being analyzed. It consists of an array of supposition options that go beyond those of pie, bar, and footsie catalogs, together with warmness and tree maps, geoepistleic maps, scatter plots, and different unique-reason visuals. That equipment enables users to investigate the data with the aid of interacting at once with a visible representation of it. The supposition of facts is one of the steps of the information science method which says that once the statistics have been collected, processed, and modeled, they need to be discerned to conclude. Statistics supposition is also a discipfootsie detail of the wider record presentation structure (DPA), which aims to discover,



discover, manipulate and deliver records as efficiently as possible. Visualizing facts is essential to almost any career. It could be used by instructors to demonstrate the effects of checking scientists, using laptop scientists exploring advances in artificial intelligence (AI), or executives trying to share records with stakeholders. It also plays a vital role in huge statistical projects. As agencies amassed large collections of information during the early years of the Big Statistics trend, they wanted a way to quickly and easily gain insight into their information. For comparable reasons, the supposition is essential for advanced analysis. While a photoepistleer writes advanced predictive analytics or machine Mastering (ML) algorithms, it becomes critical to discern the outputs to display the results and make sure the models are working as intended. This is because suppositions of complex algorithms are normally easier to interpret than numerical outputs.

### Number per attack Category

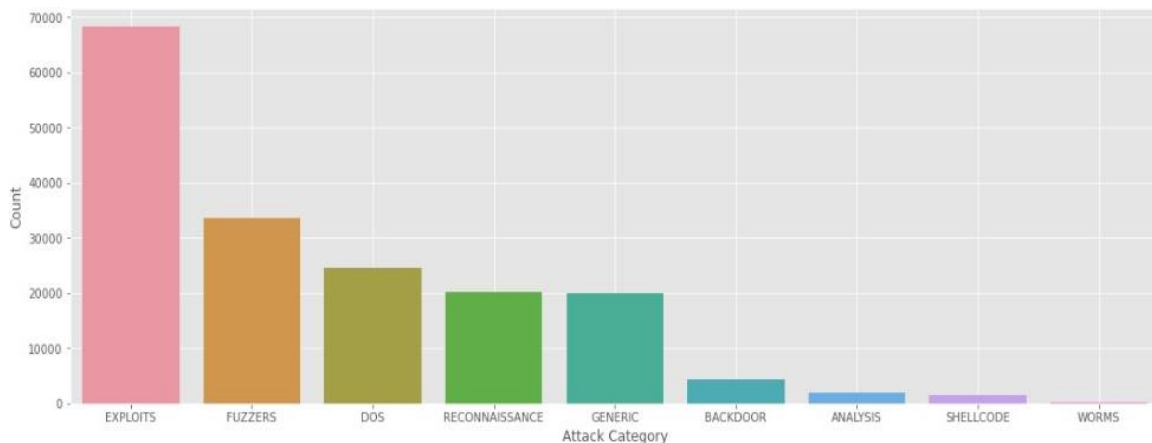


Figure 3.4: Number of attacks

This epistle shows the attack number and how many cyber-attack happened. We can see that exploits attack happens the most. The next most happened attack is fuzzers. Then DOS, then reconnaissance, then generic, then backdoor, then analysis, then shellcode, and the last one is worms. To do a reconnaissance attack, the attacker first gathers all the information of computer networks that they want to attack, then circulate security controls. In the fuzzers attack, the attacker first nourishes the computer with some massive random invalid data to block it, and then they break the security loopholes of a computer. In an analysis-gestalt attack, the attacker creates a kind of intrusion that penetrates web programs via ports, emails, and net scripts. The backdoor attack is a stealthy technique to keep away from ordinary authentication to make sure unauthorized faraway get the right of entry to a tool. An exploit is a cryptogram that takes gain of a software vulnerability.

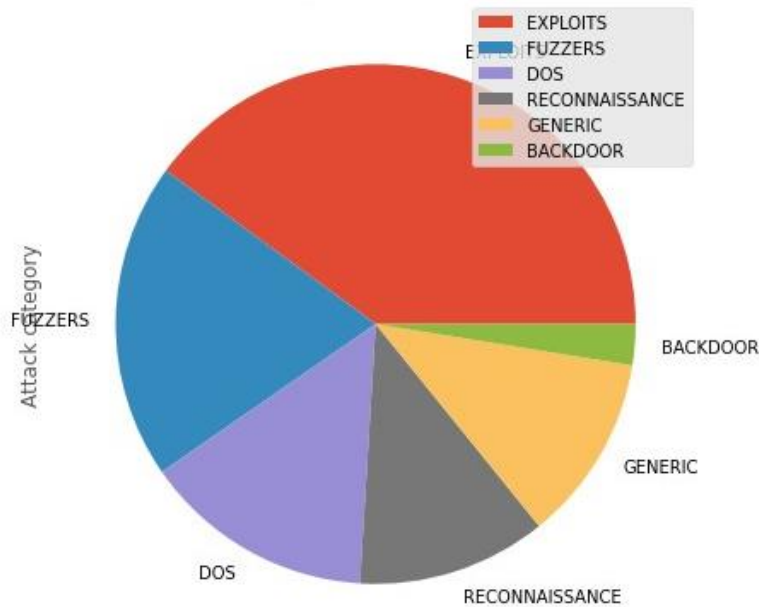


Figure 3.5: Top six attacks

### Correlation Matrix

A correlation matrix is a desk that shows the correlation coefficients for one-of-a-kind variables. Similarly, this type of one has regularly utilized the side of other varieties of statistical analysis. As an instance, it may be beneficial in the evaluation of multiple footwear regression fashions. Remember the fact that the fashions include several impartial variables. Now, we will see the correlation matrix of this research. A correlation matrix is genuinely a desk that presents the correlation coefficients for exclusive variables. It's for a powerful device to summarize a big dataset and to perceive and discern patterns inside the given facts. The correlation matrix is in reality a table of correlations. The maximum commonplace correlation coefficient is Pearson's correlation coefficient, which compares two gestalts of program language period variables or ratio variables. But there are many others, relying upon the gestalt of information you want to correlate.



Figure 3.6: Correlation matrix

### Heat map

A heat map is a statistics supposition method that shows the significance of a phenomenon as color in dimensions. The version in color can be via hue or intensity, giving visible cues to the reader about how the phenomenon is clustered or varies over an area. Working with mediocre and huge facts sets, information scientists and information analysts have a look at and decide essential relationships and characteristics among different points in a data set in addition to the capabilities of those records factors. Records scientists and analysts work with a team of others in distinct professions. Using warmth maps make for a visually clean manner to summarize findings and primary components. There are other ways to symbolize records, however, warmth maps can discern those facts factors and their relationships in an excessive dimensional space without turning too compact and visually unappealing. Heat maps in statistics analysis, permit particular variables of rows and/or columns on the axes or even at the diagonal. Heat maps represent exceptional densities of statistics points on a geopistleic map to help customers see the intensities of sure phenomena and show objects of the finest or least importance. Normally, warmth maps utilized in geopistleic supposition are wrong for heat maps, but the difference comes in how certain statistics are supplied that differentiate them.

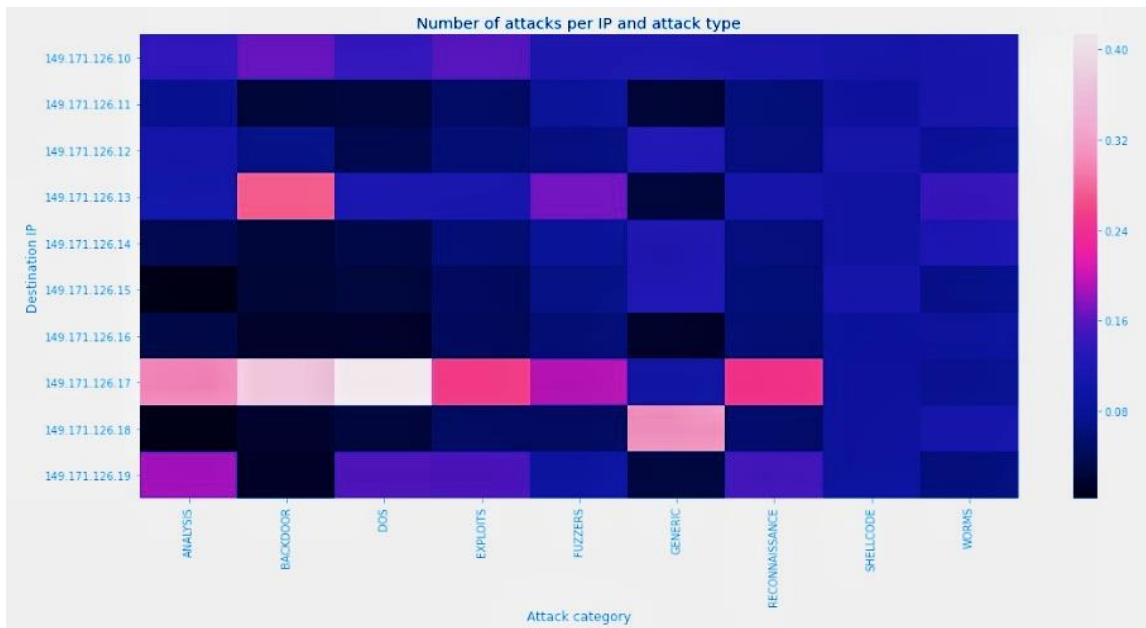


Figure 3.7: Heat map

### Pair-plot

Pair-plot supposition comes on hand when you want to go for exploratory facts evaluation. Plot pairwise relationships in an information set. Pair-plot is a module of the seaborn library which offers an excessive-level interface for drawing attractive and informative statistical pics. A pair plot is used to apprehend the best set of capabilities to explain a dating among variables or to form the maximum separated clusters. It also helps to form some unswerving category models employing drawing a few easy traces or making a footwear separation in our statistics set.

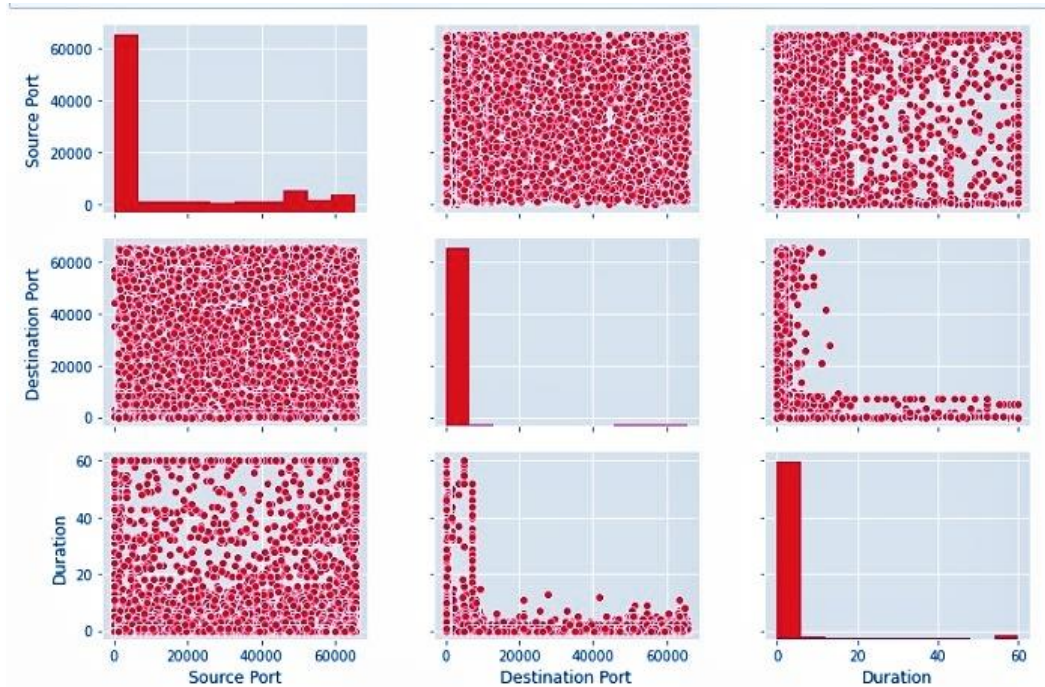


Figure 3.8: Pair plot

## Scatterplot

A scatter plot is a gestalt of the epistle or mathematical diagram that makes use of Cartesian coordinates to display trends for common variables for difficult and rapid information. If the elements are cryptograms, one extra variable may be displayed. If no based variable exists, both kinds of variables can be plotted on either axis and a scatter plot will illustrate best the diploma of correlation (no longer causation) between two variables. A scatterplot can advocate different kinds of correlations between variables with a certain confidence period c programming language. For specimens, weight and height might be on the y-axis and top on the x-axis. Correlations can be subtle (increasing), negative (decreasing), or zero (uncorrelated). If the pattern of dots slopes from a decrease from the left to the upper right corner, it shows a nice correlation between the variables under study. If the pattern of dots slopes from the top left to the bottom right, this indicates a negative correlation. A series of first-class healthy may be interested in observing the relationship between variables. An equation for the correlation between variables can be determined using fitted goodness-of-fit approaches. For a footsie correlation, the excellent form procedure is known as footsie regression and is guaranteed to generate the correct answer in a finite time. No recognized high-quality matching procedure is guaranteed to generates the exact

answers for any relationship. A scatterplot is also very useful as we want to see how similar statistical units agree to reveal non-footsiear relevances between reducible.

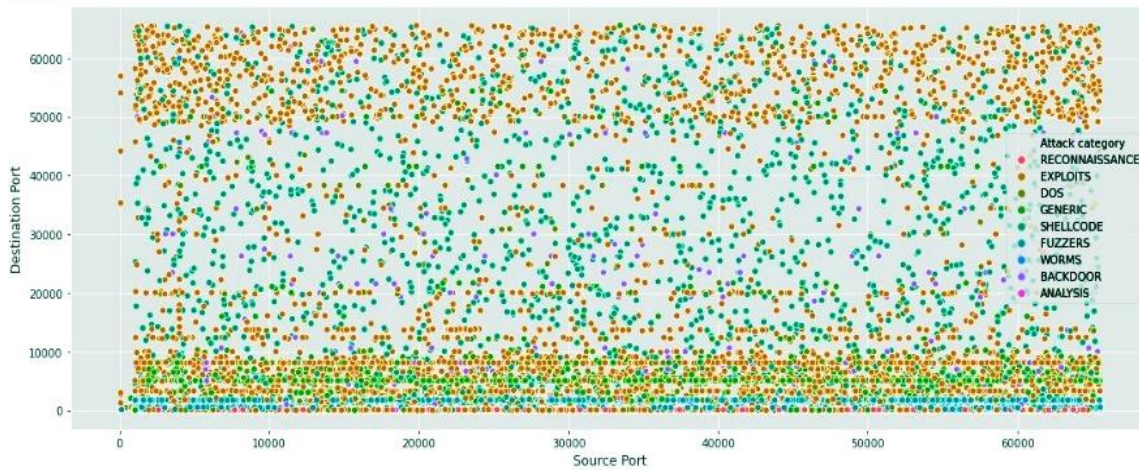


Figure 3.9: Scatterplot

### 3.3.5 Training Machine & Finding Patterns

Machine training is a process wherein a system studying (ML) set of jus is fed with sufficient training records to research from. ML fashions may be skilled to advantage production techniques in several ways. In this part, the researcher does some python cryptograms to train the machine in jupyter notebook to find the patterns of the cyber-attacks. For this, the researcher uses the dread of attacks in the destination port and categorizes them into different source IP addresses means the researcher finds the patterns for different IP addresses. The researcher finds the most targeted destination IP Address, most logical ports attacked, the most common gestalt of attack, different times of the day and most important and main subject is to find the pattern of the cyber-attacks. So, the researcher trains the machine with python to find the pattern of the attack. Jupyter notebook is used to do python cryptogram and do the solution. For data collection, the researcher collects some company's cyber department's data and then works with that. The researcher knows that, when more data are used, the more correct the result will be. So, that's how the whole management has happened. It can be a great finding for the future cause when we learn any pattern of something then we can easily understand how to solve any problem created by that thing. And that is how the researcher finds the pattern of all attacks.

### **3.3.6 Training Google Colab**

After collecting pictures of the patterns of cyber-attacks, the researcher labeled those pictures with labeling software. Then, make a zip of all those data and put it in Google Drive, and then the researcher does a cryptogram in Google colab and connects Google colab with Google Drive, then commands in colab to read that specific file and unzip the file. Then train the machine more than a thousand times to collect the training weight file. Google colab is Google hosted Jupyter notebook product that provides an unfastened compute environment, which includes GPU and TPU. Colab comes batteries included with many famous Python applications installed, making it the desired device for clean version experimentation. Because of this, the Robo-flow version Library consists of many loose, open-source laptop vision fashions available on Google Colab. Colab does come with boundaries. The compute assets allocated are limited to 12 hours. Saving a model's weights way saving the shape of a version after schooling. Reloading the version weights means the usage of those stored weights in a future test even though that could be a new session in colab. To keep model weights, we need to first have weights we need to store and a vacation spot in which we are searching to keep those weights. As soon as we have got the document direction of our weights file, we can shop this file domestically or to our Google power. We advise saving weights for your Google force. Colab is a loose Jupyter paperback environment that ushers entirely in the cloud. Most importantly, your teammates can edit the notebooks you create at the same time, with no setup required. It's like editing a document with Google Doctors. Colab supports many popular device learning libraries that can be easily loaded into your pocketbook. So, after collecting the weight file, the researcher can use them with the model (in the YOLOv3 project, this is the description of work before ushering the `yolo_object_detection.py` cell).

### **3.3.7 YOLOv3 Model**

YOLOv3 is an actual-time item detection set of jus that ascertains undistorted annihilation in motion pictures, live hieroglyphics, or snapshots. The YOLO system mastering algorithm ultra-modern functions discovered by way of intricate convolutional neural assemblages to hit upon an object. The 1/3 version of today's YOLO device's modern-day set of jus is a more accurate model modern-day the unique ML algorithm. The first version of modern-day YOLO became created in 2016, and version 3, that's discussed considerably in this newsletter, become made two years later

in 2018. YOLOv3 is an improved model of modern-day than the previous versions [7]. YOLO has been applied with the use of OpenCV machine learning libraries for this gestalt of work. The researcher use OpenCV libraries for this research. YOLO is a Convolutional Neural community (CNN) for appearing recognition.

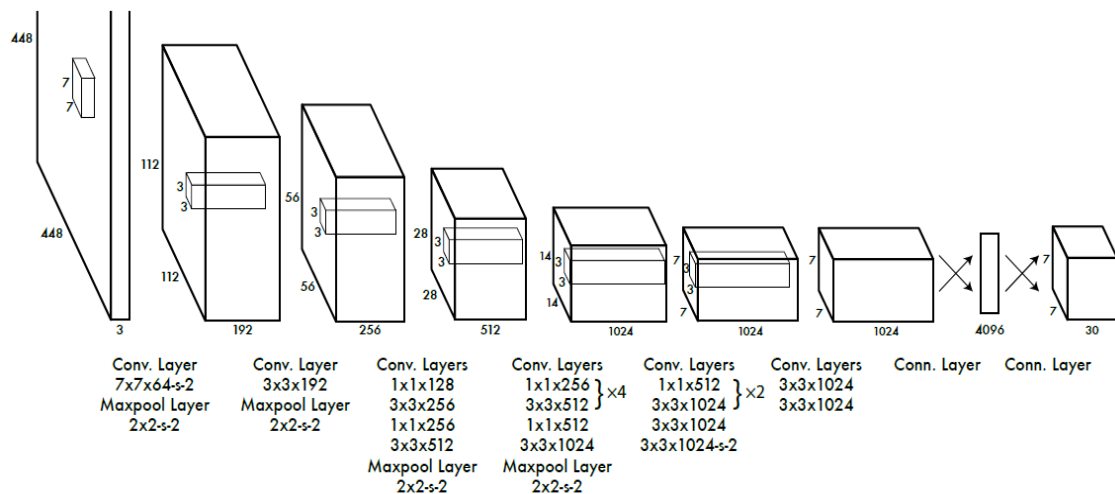


Figure 3.10: General YOLOv3 architecture [8].

With YOLOv3, a single CNN concurrently predicts multiple demarcated bins and class possibilities for those bins. YOLOv3 backsides on whole photos and at once optimistically unearth overall substantiation. This version has convenience regarding other item detection techniques:

- YOLOv3 is quicker than others.
- YOLO sees the exhausted simulacrum during formulation and takes a look at time so it consistently enciphers applicable compassions approximately training in addition to their look.
- YOLO learns generalizable representations of items so that after skilled on herbal pix and tested on paintings, the set of jus outperforms other pinnacle detection methods.





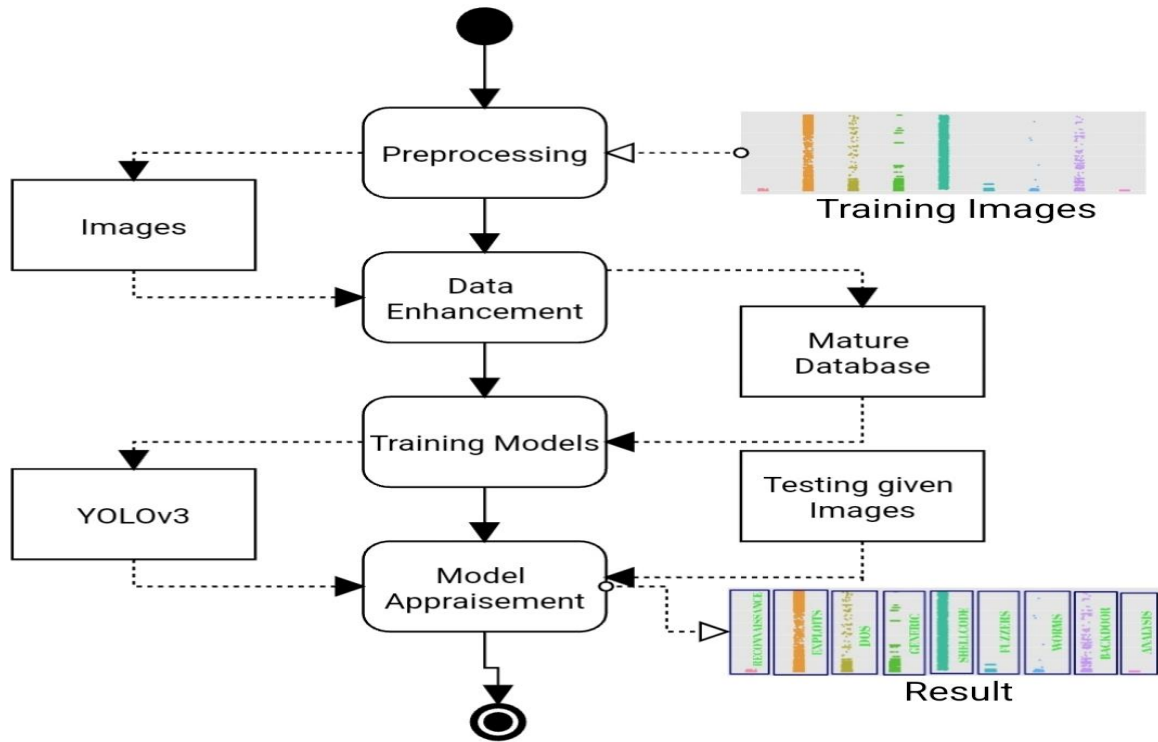


Figure 3.12: YOLOv3 training model procedure.

This is the working UML diagram of this research. First, the researcher train the umbrages. Then, the researcher preprocesses them and then the researcher enhances the data, and then the researcher train the models according to YOLOv3. It's the mature database that the researcher is working with. After testing the given umbrages, the model appraisalment is happened by the researcher. Then, we see the result of the work. That's how the working procedure is completed.

### YOLOv3 Loss Function

The YOLO loss function has three parts. Those are demarcated boxes, confidence, and classification. Now, we will see and understand the equation of the YOLO loss function. The equation of the YOLO loss function is given below:

$$\begin{aligned}
 & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] \\
 & + \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} [(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2] \rightarrow \text{Bounding Box} \\
 & + \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} (C_i - \hat{C}_i)^2 + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{noobj} (C_i - \hat{C}_i)^2 \rightarrow \text{Confidence} \\
 & + \sum_{i=0}^{S^2} 1_i^{obj} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \rightarrow \text{Classification}
 \end{aligned}$$

In this loss feature (Equation: four),  $(1_{obj})$  refers to the presence of an item in cellular  $(i)$ , and  $(1_{ij})$  refers to  $(j)$ th the item in mobile is anticipated using the demarcate box. The regularisation parameters  $(\lambda_{coord})$  and  $(\lambda_{noobj})$  are essential for the loss feature to be balanced. The loss corresponding with predicted demarcated container vicinity coordinates  $(x, y)$  is computed within the first component and the ground fact information within the education setting has demarcated container coordinates of  $(\hat{x}, \hat{y})$ . Inside the Yolo algorithm  $(\lambda_{coord})$  the fee is taken to be five.0 and whether or not a mistake happens, it suggests a constant that will increase the penalty. The quantity demarcated ate bins within the sieve is given by using  $B$ , whilst the range of cells within the sieve is given via  $S^2$ . Inside the 2nd element,  $(C)$  represents the extent of self-assurance and the expected demarcate field with the ground truth box's IOU is  $(C)$ . on this model  $(\lambda_{noobj})$  the cost is taken to be zero.5 and while there's no object, it's miles applied to make the loser less concerned approximately self-belief. Within the closing component (Equation: 4), for the category, this loss is the sum of squared blunders loss. Inside the period  $(1_{obj})$ , whilst there's an item on a cell then it is 1, and when there is not, it is 0. (Zafar et al., 2018, #).

## CHAPTER 4

### EXPERIMENTAL RESULTS AND DISCUSSION

#### 4.1 Introduction

This research is finding and detecting the gestalt of research. At first, the researcher finds the patterns of some cyber-attacks from the information of the data collection. Then, with the picture of those patterns, the researcher trained only looks once version three, the YOLOv3 model to detect the name of the pattern. The machine can detect the pattern name by only seeing the picture of the pattern. The researcher collect the related data from some companies' cyber departments. Researchers have observed that information is capable of identifying routine patterns allowing us to make knowledgeable guesses, assumptions, and hypotheses; it enables us to expand essential talents of important thinking and good judgment. Machine Learning is a field where developer trains their machines and teach them what to do. The machine is faster than a human. So, when we teach them what and which way to do things they do those things faster than humans, and a lot of time saves from it. Python is a high-level computer language by which we can also train machines. So, ultimately python can be used for machine learning techniques. The researcher uses Jupiter notebook to do python programming and train machine to do what the researcher wants. Jupyter notebook is a web application by which we can do cryptograms and also usher the cryptogram and see the results. Here, the researcher collects the data of some software companies' cyber department's servers results and then analyzes the data and finds the patterns of some cyber-attacks. It is a groundbreaking finding in the cyber-security world. Because when we learn the pattern of anything, we can know everything about it and also gain the knowledge of how to solve that problem and what to do to stop that from happening again in the future. After collecting the data researcher finds some attacks which happened the most, then the researcher uses python cryptogram and teaches the machine so that machine can give the output. Researcher finds the pattern of some cyber-attacks. These are Reconnaissance, Fuzzers, Analysis, Backdoor, Exploit, Generic, Shellcode, and Worm. Then researcher loads the image by python programming. Then, the researcher commands the machine to detect the pattern and identify them. And the machine detects all the attack patterns correctly. So, now we can say that the machine is working perfectly. Then, the researcher use Google colab to use the YOLOv3 machine learning algorithm. YOLOv3

is an actual-time section recognition set of jus that ascertain undistorted targets in motion pictures, live hieroglyphics, or snapshots. The YOLO system mastering algorithm ultra-modern functions discovered by way of a profound flexional neural assemblage to hit upon an object. The 1/3 version of today's YOLO device's modern-day set of jus is a more accurate model modern-day the unique ML algorithm. By you only look once algorithm one can easily make a detection model which can detect any object easily. Modern day model is upgradable than older versions. But version 3 is better and popular in its own way. The first version of modern-day YOLO became created in 2016, and version 3, that's discussed considerably in this newsletter, become made two years later in 2018 [9], [10]. YOLOv3 is an improved model of modern-day than the previous versions. YOLO has been applied with the use of the OpenCV machine learning libraries for this gestalt of work. The researcher use OpenCV libraries for this research. CNN's are classifier-based systems that could manner enter umbrages as established arrays of information and recognize styles between them (view image underneath). YOLO has the gain of being a great deal quicker than different networks and nevertheless continues accurate [11]. It permits the version to examine the entire photo at test time, so its enumeration is knowledgeable employing the yearlong connection inside the photo. Excessive-scoring areas are referred to as effective detections of something class they most carefully perceive [12].

## 4.2 Experimental result

### Finding Patterns

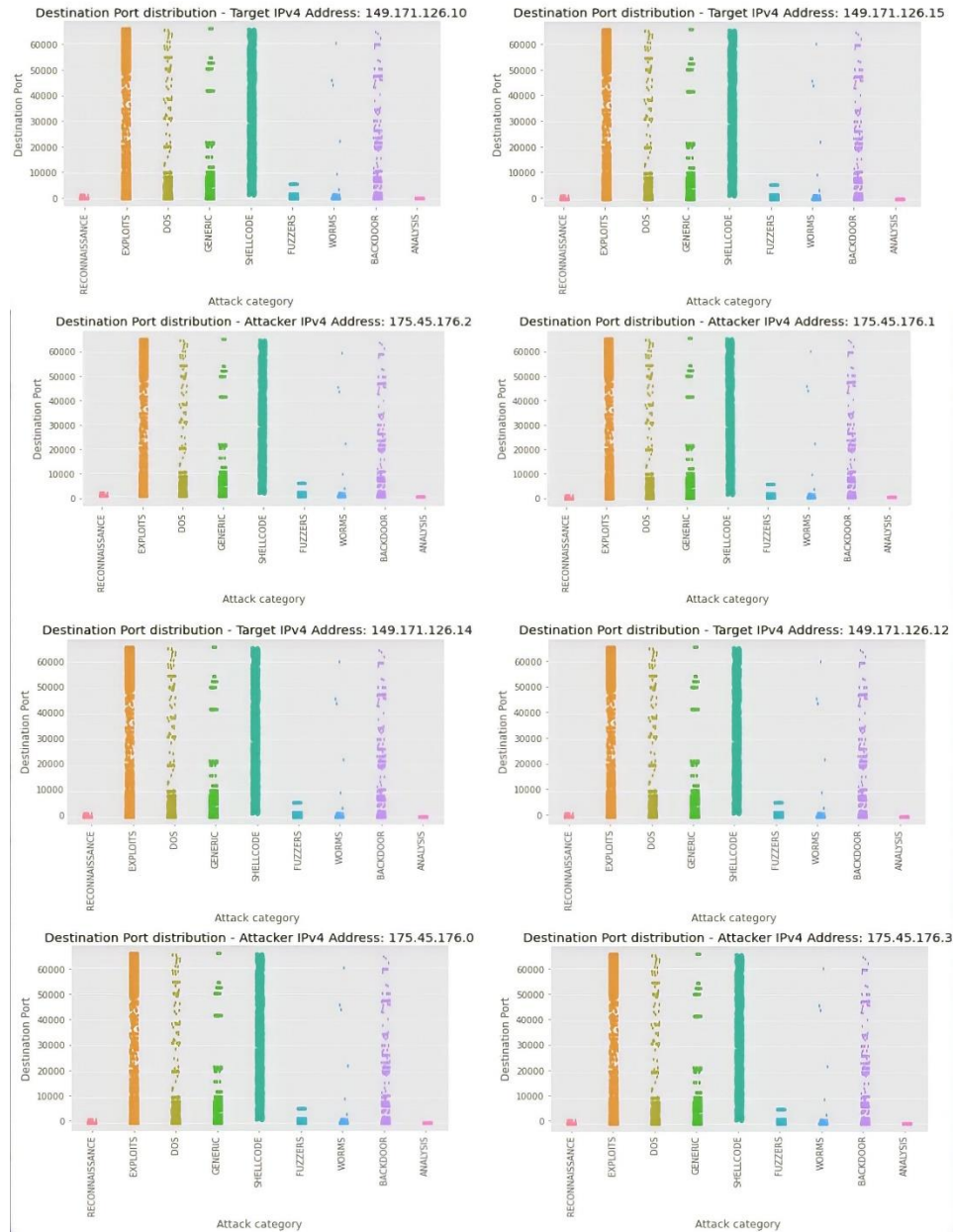


Figure 4.1: Patterns in eight IPv4 addresses

This is the pattern of nine cyber-attacks in eight IPv4 addresses. The hypothesis is, the patterns are the same in all the IP addresses. And if we look at the picture of the patterns in that eight IP addresses, we can see that all attacks are giving the same pattern in all the IP addresses.

## Final Patterns

Here is the final pattern of the cyber-attack.

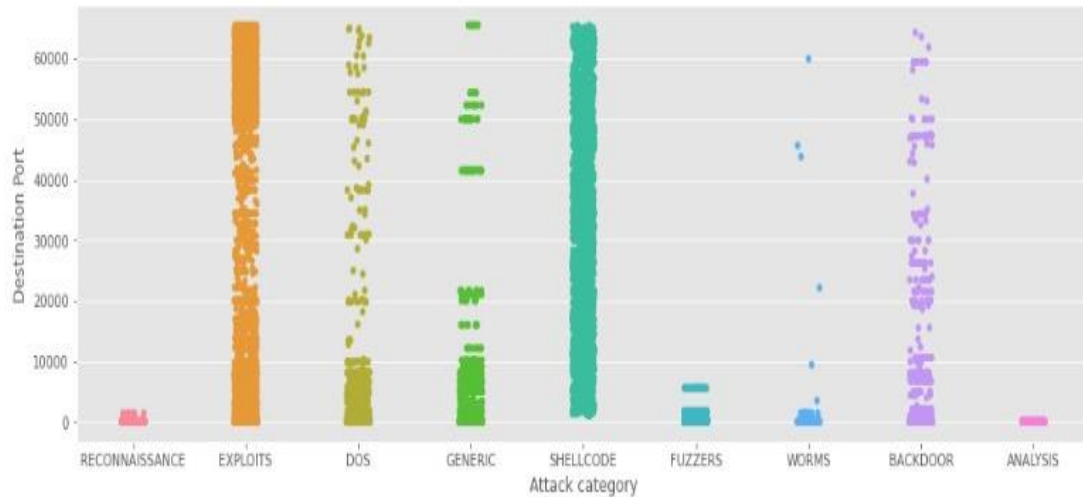


Figure 4.2: Final Patterns

This is the final pattern for all IP addresses. After finding the picture of the pattern, the researcher works with the pattern recognition model which can recognize all the patterns and can tell us the name of the patterns. The researcher finds the most targeted destination IP Address, most logical ports attacked, the most common gestalt of attack, different times of the day and most important and main subject is to find the pattern of the cyber-attacks. So, the researcher trains the machine with python to find the pattern of the attack. Jupyter notebook is used to do python cryptogram and do the solution. For data collection, the researcher collects some company's cyber department's data and then works with that. The researcher knows that, when more data are used, the more correct the result will be. So, that's how the whole management has happened. It can be a great finding for the future cause when we learn any pattern of something then we can easily understand how to solve any problem created by that thing. Now, we will see the final model for recognizing the patterns of the attacks.

## Ushering the YOLOv3 model for the result

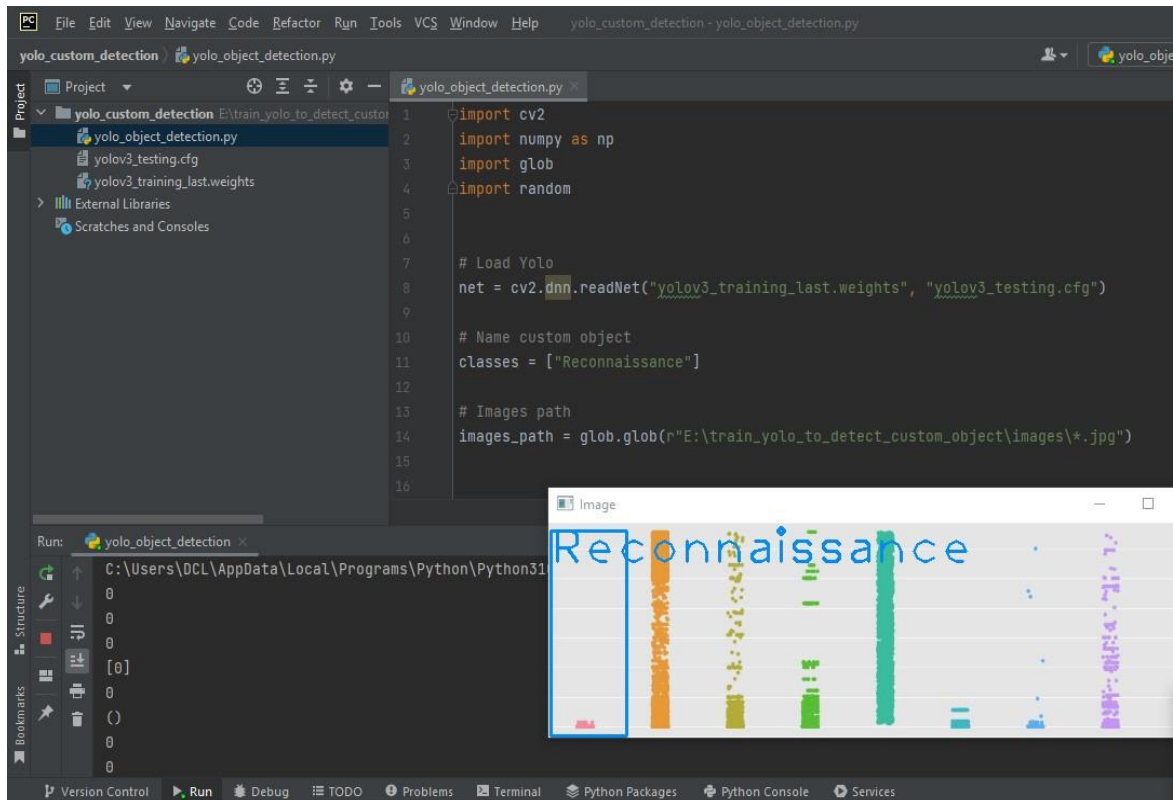


Figure 4.3: Usher & result 1

We can see that when the researcher ushers the program for detecting the reconnaissance attack pattern by typing reconnaissance in the classes, the program recognizes the pattern of reconnaissance from all the patterns of the attack.



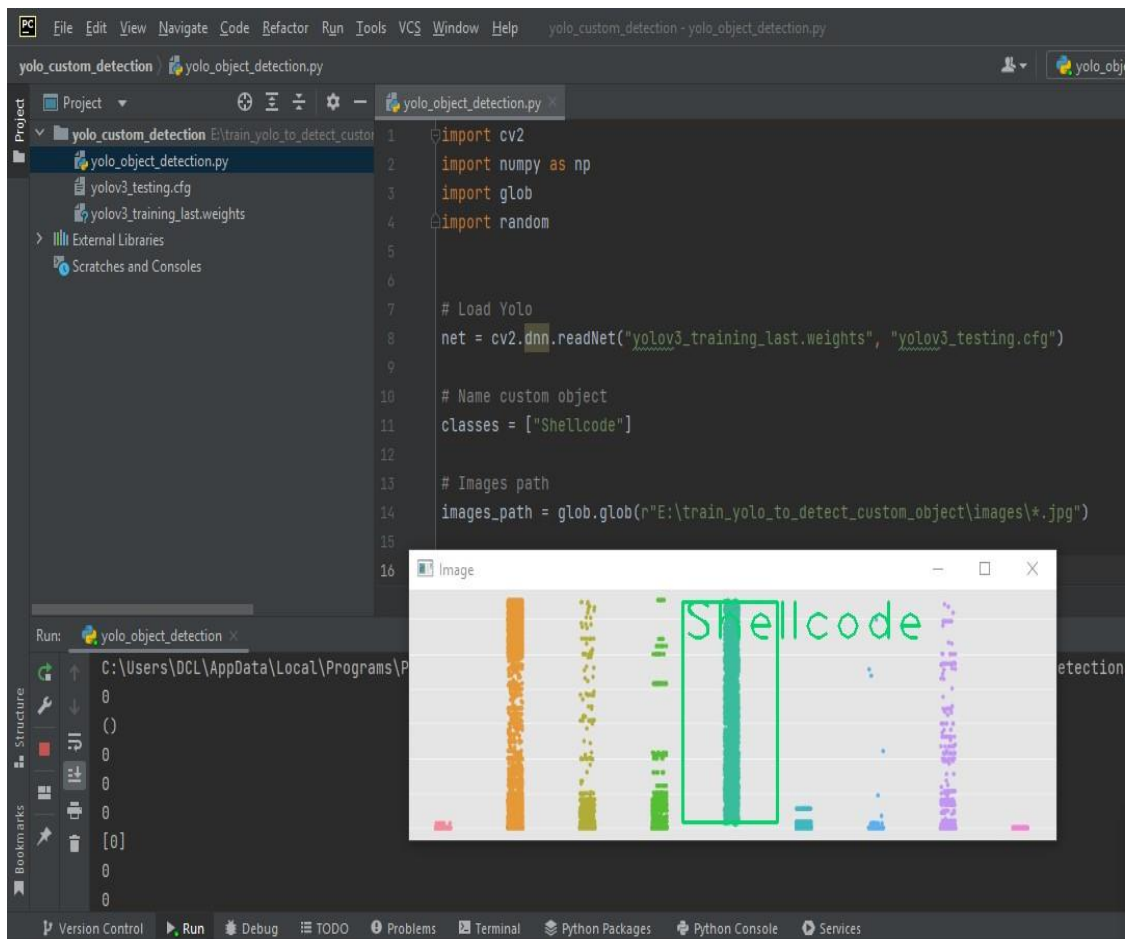


Figure 4.4: Usher & result 2

We can see that when the researcher ushers the program for detecting the shellcode attack pattern by typing shellcode in the classes, the program recognizes the pattern of shellcode from all the patterns of the attack.

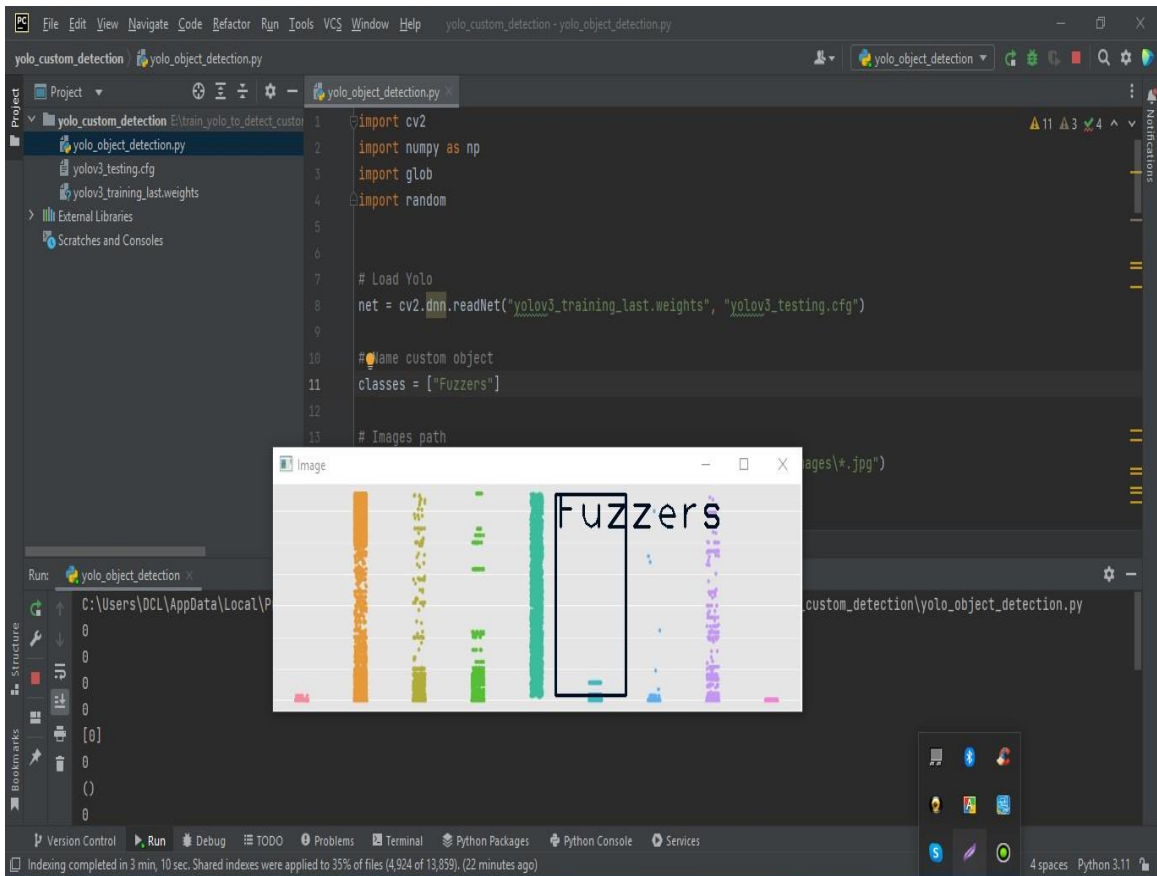


Figure 4.5: Usher & result 3

We can see that when the researcher ushers the program for detecting the fuzzers attack pattern by typing fuzzers in the classes, the program recognizes the pattern of fuzzers from all the patterns of the attack.

## Recognition Result

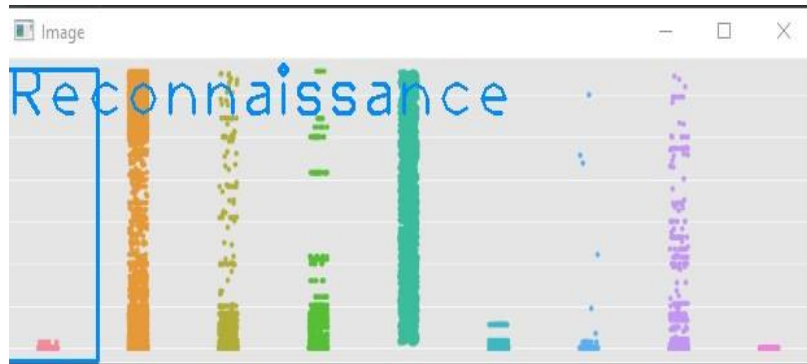


Figure 4.6: Reconnaissance attack

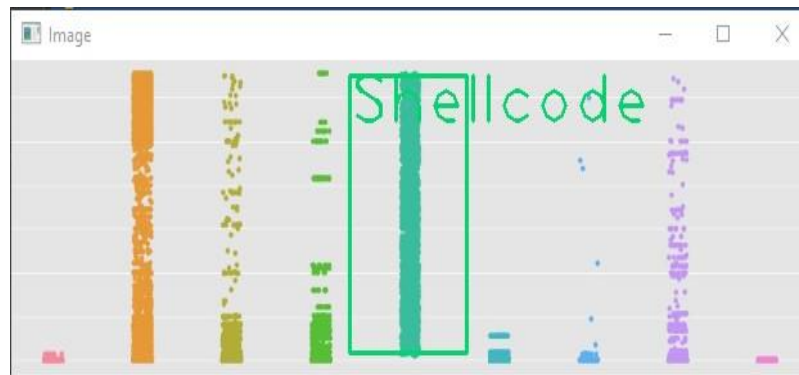


Figure 4.7: Shellcode attack

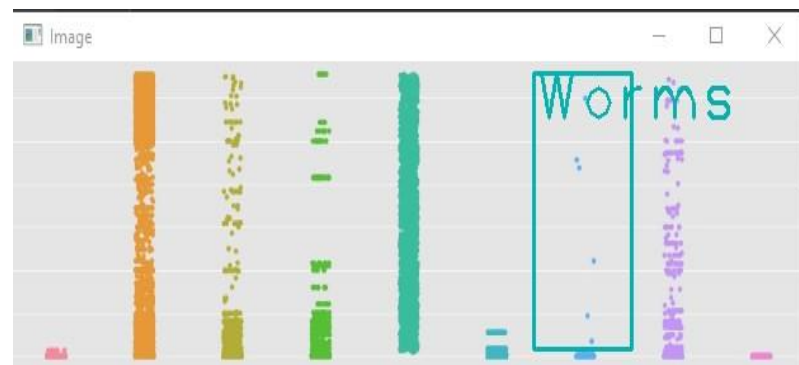


Figure 4.8: Worms attack

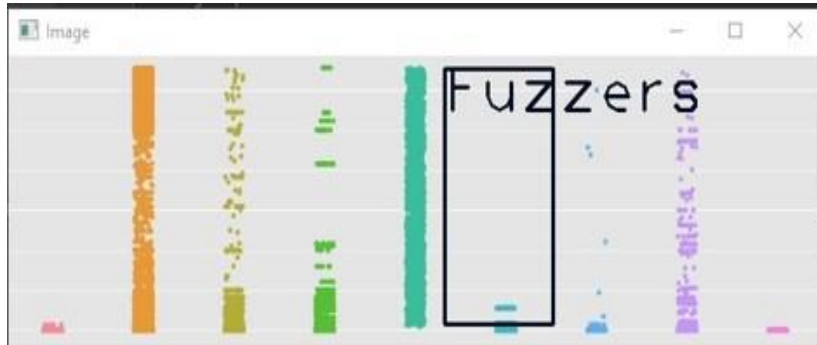


Figure 4.9: Fuzzers attack

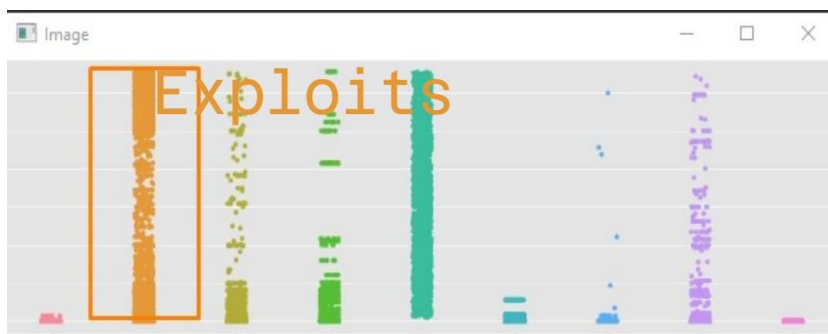


Figure 4.10: Exploits attack

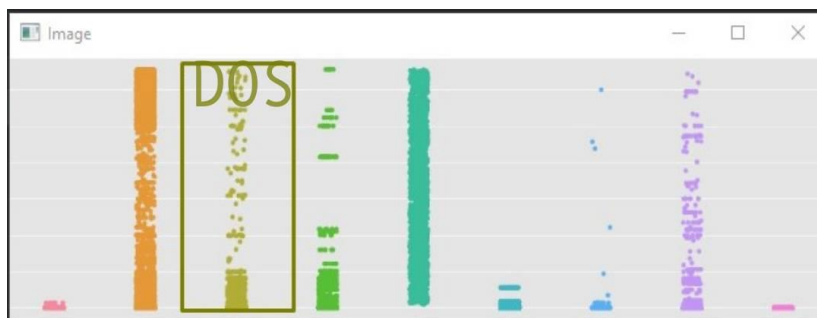


Figure 4.11: DOS attack

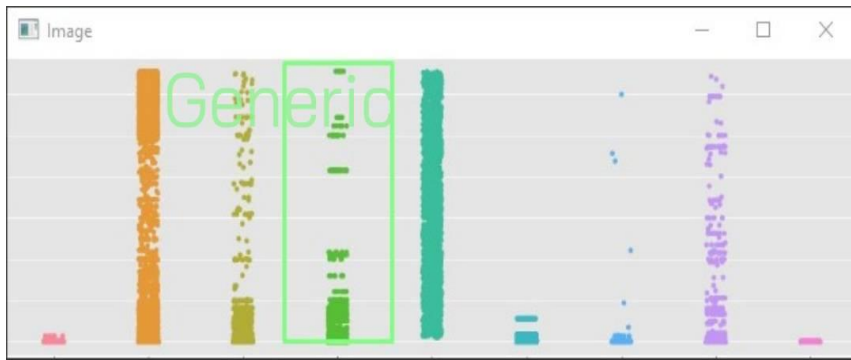


Figure 4.12: Generic attack

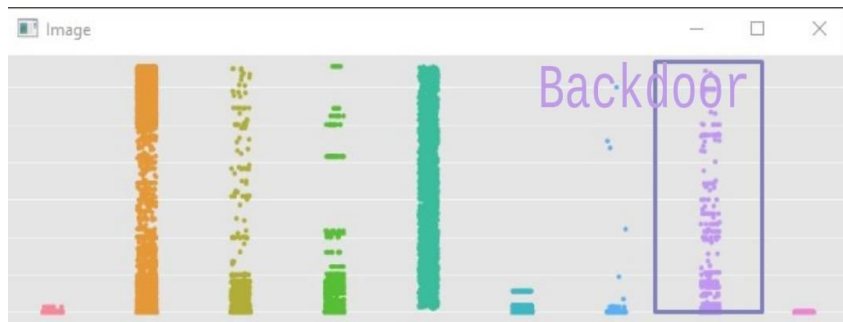


Figure 4.13: Backdoor attack

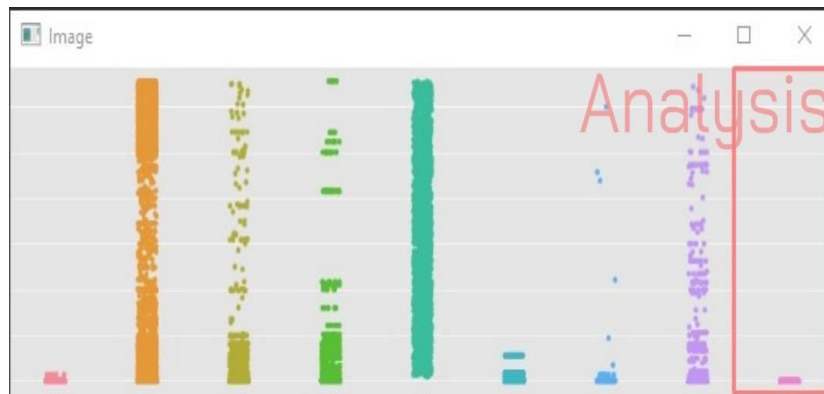


Figure 4.14: Analysis attack

### 4.3 Discussion

At first, the researcher finds the patterns of the cyber-attacks, and then the researcher makes a model of recognizing the name of the cyber-attacks by seeing the patterns by the YOLOv3 algorithm. Understanding a pattern of something is a notable advantage to gaining information approximately that. Styles are a chain of numbers, shapes, or gadgets which look at a positive rule to live the same or alternate. Styles offer to reveal a level of order in what could in all likelihood otherwise seem chaotic. Researchers have discovered that facts are able to figure out ordinary styles allowing us to make knowledgeable guesses, assumptions, and hypotheses; it enables us to amplify essential abilities of important thinking and correct judgment. Machine learning is a subject in which developer trains their machines and teach them what to do. The machine is quicker than a human. So, while we train them on what and in which manner to do matters they do the ones that matter faster than human beings, and lots of time save from it. Python is a high-degree computer language through which we can also train machines. So, in the long usher, python can be used for system learning techniques. The researcher makes use of Jupiter notebook to do python programming and teaches the device to do what the researcher wishes. Jupyter notebook is a web utility by which we can do cryptograms and additionally usher the cryptogram and notice the consequences. Right here, the researcher collects the facts of some software program companies' cyber branch's servers results and then analyzes the records and finds the patterns of a few cyber-attacks. It's far a groundbreaking location within the cyber-security world. Due to the fact while we examine the sample of anything, we are able to realize the whole thing approximately and additionally gain the knowledge of the way to clear up that hassle and what to do to stop that from going on once more inside destiny. After collecting the information researcher unearths some attacks which passed off the most, then the researcher makes use of python cryptogram and teaches the gadget so that machine can deliver the output. The Researcher uncovers the sample of some cyber-attacks. Those are reconnaissance, fuzzers, analysis, backdoor, shellcode, and worm. To do a reconnaissance assault, the attacker first gathers all the records of laptop networks that they want to attack, then circulate security controls. Inside the fuzzers' assault, the attacker first nourishes the pc some large random invalid records to block it, after which they ruin the security loopholes of a laptop. In an evaluation-kind assault, the attacker creates a form of encroachment that pervade web programs through harbors, emails, and internet scripts. The backdoor assault is an unperceived method to supersede ordinary documentation to ascertain unwarranted far-off get the proper entry

to a device. An advantage is a cryptogram that takes gain of a software enterable situation or safety deficiency. It's far more epistleical each through protection researchers as a proof-of-concept denunciation or thru malevolent agents for use in their dynamism. A popular attack is a way that tries to dam encryption through the use of a carve operation for percussion irrespective of association arrangements. Within the shellcode approach, the attacker makes way for a handy shred of cryptogram from the cartridge to regimen the compromise irrespective of encryption settings. A bug assault replicates a hypocritical inscription to unfold it to isolated computer systems. Frequently, it makes use of a laptop network to expand overhanging indemnity fissure inside the vacation spot computer. Those are the gestalts of attacks the researcher works with. The researcher unearths the maximum targeted destination IP address, most logical ports attacked, the maximum common form of assault, one-of-a-kind instances of the day, and the maximum critical and most important concern is to find the sample of the cyber-assaults. So, the researcher trains the machine with python to discover the pattern of the attack. Jupyter notebook is used to do python cryptograms and do the answer. For information collection, the researcher collects a few companies' cyber departments' information after which works with that. The researcher knows that, when more information is used, the greater correct the result can be. So, that's how the entire control has befallen. it can be a wonderful locating for the destiny reason when we study any sample of something then we are able to without difficulty apprehend a way to remedy any problem created by way of that issue. YOLOv3 model can recognize the name of the patterns of cyber-attacks. The working procedure and mechanism of the model are already described.

#### **4.4 Summary**

In this chapter, we learn a lot of things. We learn about how the researcher works in this research. We know how the model works and detects. We learn the mechanism and results of the model. We learn how the program ushers and how it shows the results. We saw the answers here. The researcher works with nine cyber-attacks. Then the researcher finds the patterns of those cyber-attacks. Then make a YOLOv3 model which can detect the name of the cyber-attack by seeing its pattern. That's how you only look once version 3, YOLOv3 is working.

## CHAPTER 5

### IMPACT ON SOCIETY, CYBER WORLD AND SUSTAINABILITY

#### 5.1 Impact on Society

The impact on society is given below:

- Will give safety to the cyber user.
- People will feel safer and tension free.
- Can easily detect and solve problems.
- People will do more business on this platform and can change their situation in life.
- People will feel safe using social platforms and can make more good relations with one another.
- People will be more confident to do business and other things on this platform.
- People will trust this platform more.
- When people will feel safe, it will impact society by itself.
- From research, we know that 60.5% of students agreed with online classes during the corona situation. Other didn't agree for many reasons. One of the reasons is cyber-security issues. So, this research can also give hope to students to do online classes in any pandemic situation [13].

#### 5.2 Impact in Cyber World

The impact in the cyber world is given below:

- Can easily find the pattern and details of cyber-attack.
- Can easily detect the name of the cyber-attack.
- When we know about the details of an attack, it will be easy to solve problems that are created by the cyber-attack.
- Cyber-attack will reduce day by day.
- The cyber-world will be safer than before.



### **5.3 Ethical Aspects**

To do this research, the researcher maintains all the ethical aspects. The researcher gives credit to all the authors from whom the researcher collects important information by citing their papers and also gives credit to the websites from which the researcher took the information.

### **5.4 Sustainability Plan**

The sustainability plans are given below:

- Have to advertise the research and give people the definition of what is it.
- Improve the areas of the research.
- Have to find opportunities with the research.
- Have to create an imaginative viewpoint about the research.
- Have to change some implementation changes.
- Have to learn more about the development laws and work with them.
- Have to review and change policies from time to time.
- Have to define the research model terms.
- Have to solve some problems with that model so people will understand the dread of that model perfectly.
- Have to advertise the works and benefits of the model.
- Have to improve with the demand of the cyber world's environment.

## CHAPTER 6

# SUMMARY, CONCLUSION, RECOMMENDATION, AND FUTURE RESEARCH

### 6.1 Summary of the Study

At first, the researcher collect the related data from some companies' cyber departments. Then, the researcher makes the proper environment for python and jupyter notebook. Then, have to import some engines and libraries. Those libraries are pandas, seaborn, NumPy, missingo, etc. Then, have to clean the data. For cleaning the data, missingo is the library that worked. By this, we find the missing data and delete those missing or unavailable data from rows or columns. After cleaning the data we do python programming and find out some arrays, matrix, and other visual representations of data. And at last, researchers find the pattern of cyber-attacks. Then, the researcher creates more environments on the computer. Created a tensor-flow platform. It's a python friendly open source library. Then created an open-source vision library environment. After that, the information about all attacks is given so, the machine can identify the targeted destination. Then, the researcher trained YOLOv3 in Google colab. After collecting pictures of the patterns of cyber-attacks, the researcher labeled those pictures with labeling software. Then, make a zip of all those data and put it in Google Drive, and then the researcher does a cryptogram in Google colab and connects Google colab with Google Drive, then commands in colab to read that specific file and unzip the file. Then train the machine more than a thousand times to collect the training weight file. Google colab is Google hosted Jupyter notebook product that provides an unfastened compute environment, which includes GPU and TPU. Colab comes batteries included with many famous Python applications installed, making it the desired device for clean version experimentation. Because of this, the version Library consists of many loose, open-source laptop vision fashions available on Google colab. Colab does come with boundaries. The compute assets allocated are limited to 12 hours. Saving a model's weights way saving the shape of a version after schooling. Reloading the version weights means the usage of those stored weights in a future test even though that could be a new session in colab. To keep model weights, we need to first have weights we need to store and a vacation spot in which we are searching to keep those weights. As

soon as we have got the document direction of our weights file, we can shop this file domestically or to our Google power. We advise saving weights for your Google force. Colab is a loose Jupiter notebook environment that ushers entirely in the cloud. Most importantly, your teammates can edit the notebooks you create at the same time, with no setup required. It's like editing a document with Google Colab supports many popular device learning libraries that can be easily loaded into your pocketbook. So, after collecting the weight file, the researcher can use them with the model (in the YOLOv3 project, this is the description of work before ushering the `yolo_object_detection.py` cell). Then the researcher makes a model which can recognize cyber-attacks by seeing their patterns.

## 6.2 Conclusions

The research is about pattern findings and recognition of cyber-attacks by their patterns. The sample popularity is an energetic region of studies that includes numerous packages. It is a branch of synthetic intelligence that appeals to the techniques of the gadget getting to know data. But, the goals of pattern recognition are to design and broaden wise systems which can be capable of learning and reasoning. So we can define the sample reputation because of the set of strategies that allow replicating the human perception. The researcher use the YOLOv3 algorithm to detect the patterns of the cyber-attack. YOLOv3 is rapid and has at-par precision with picked echelon identifier (on 0. five IOU) and this makes it a very herculean item identifier sampling. Encampment of object Detection in fields like media, retail, production, robotics, and many others want the models to be very fast (a little compromise on propriety is k) but YOLOv3 is likewise very correct. This makes it the pleasant version to choose in those forms of programs where the pace is essential either due to the fact the goods want to be real-time or the data is simply too big. In this paper, the researcher implemented and proposed to use the YOLOv3 set of jus for detection because of its benefits. This algorithm may be implemented in diverse fields to remedy some real-life problems like protection, monitoring site visitors' lanes, or even helping visually impaired people with help of nourish back. In this, we've created a model to hit upon only a few cyber-attacks employing its styles, which may be stripped in addition to discovering multiple many numbers of items. So, that's all for the research.

### **6.3 Recommendation**

Recommend using more models like faster R-CNN and SSD models for that detection part so we can find more accurate results. And those models are faster too. The quicker R-CNN model changed into advanced via a group of researchers at Microsoft. This is a deep convolutional community used for item detection that appears to the person as a single, cease-to-give-up, unified network. The community can accurately and speedy predict the locations of different items. With the intention to truly apprehend faster R-CNN, we ought to additionally be acquainted with the networks that it advanced from, specifically R-CNN and fast R-CNN. Quicker R-CNN is an extension of rapid R-CNN. As its name shows, quicker R-CNN is quicker than fast R-CNN way to the vicinity concept community. SSD is a single-shot detector. It has no delegated vicinity suggestion community and predicts the boundary packing containers and the instructions at once from function maps in a single unmarried bypass. It may be educated on cease-to-give-up for better propriety. SSD makes more predictions and has higher insurance on location, strip, and thing ratios. By using removing the delegated place suggestion and the use of decrease resolution pictures, the model can usher at an actual-time pace and still beats the propriety of the Faster R-CNN [14]. That's why the researcher recommends those algorithms too, to get more accurate results. And also can use other gestalts of machine learning techniques to get that result for the research.

### **6.4 Future Research**

The researcher does a lot of work in that research. This research is very big by itself. But, the researcher can make it a more interesting and vast project to do more work in the future. The future plan with that research is given below:

- This will make that research more accurate.
- Will do more algorithms.
- Will use more machine learning techniques.
- Will update more according to time.
- Will work with more data.
- Will work with more cyber-attacks.
- Will find the patterns for more IP addresses.
- Will check the similarity for more IP addresses.

- Will work with the SSD algorithm for recognition too.
- Will work with faster R\_CNN for recognition too.
- Will use the same process in other subjects too.
- Will use this for other object recognition too.
- Will work to make, a more good and advanced recognition model.

## APPENDIX

### Appendix A

```

{"cells": [
  {
    "cell_gestalt": "cryptogram",
    "execution_count": 22,
    "metadata": {},
    "outputs": [],
    "source": [
      "import pandas as pd\n",
      "import seaborn as sns\n",
      "import matplotlib.pyplot as plt\n",
      "import ipaddress\n",
      "import numpy as np\n",
      "from scipy import stats\n",
      "from scipy.stats import chi2_contingency\n",
      "from datetime import datetime, timedelta\n",
      "import math\n",
      "import missingno as msno\n",
      "plt.style.use('ggplot')\n",
      "import warnings\n",
      "warnings.filterwarnings('ignore')
    ]
    "text/plain": [
      "(178031, 11)"
    ]
  },
  "execution_count": 23,

  "metadata": {},
  "output_gestalt": "execute_result"

```

```

}
],
" <thead>\n",
" <tr style=\"text-align: right;\">\n",
" <th></th>\n",
" <th>Attack category</th>\n",
" <th>Attack subcategory</th>\n",
" <th>Protocol</th>\n",
" <th>Source IP</th>\n",
" <th>Source Port</th>\n",
" <td>1421927415-1421927415</td>\n",
" </tr>\n",
" <tr>\n",
" <th>2</th>\n",
" <td>Exploits</td>\n",
" <td>Browser</td>\n",
" <td>tcp</td>\n",
" <td>CVE 2005-4560 (http://cve.mitre.org/cgi-bin/cv...</td>\n",
" <td>.</td>\n",
" <td>1421927416-1421927416</td>\n",
" </tr>\n",
" <tr>\n",
" <th>3</th>\n",
" <td>Exploits</td>\n",
" <td>Miscellaneous Batch</td>\n",
" <td>tcp</td>\n",
" <td>175.45.176.2</td>\n",
" <td>13792</td>\n",
" <td>149.171.126.16</td>\n", }
],
" <tr style=\"text-align: right;\">\n",
" <th></th>\n",
" <th>Attack category</th>\n",
" <th>Attack subcategory</th>\n",
" <th>Protocol</th>\n",
" <th>Source IP</th>\n",
" <th>Source Port</th>\n",
" <th>Destination IP</th>\n",
" <th>Destination Port</th>\n",
" <th>Attack Name</th>\n",
" <th>Attack Reference</th>\n",
" <th>Start time</th>\n",
" <th>Last time</th>\n",

" </tr>\n", "name": "python",
"nbconvert_exporter": "python",

```

```

    "pygments_lexer": "ipython3",
    "version": "3.7.0"
}
},
"nbformat": 4,
"nbformat_minor": 4

```

## Appendix B

```

import cv2
import numpy as np
import glob
import random
net = cv2.dnn.readNet("yolov3_training_last.weights", "yolov3_testing.cfg")
classes = ["Shellcode"]
umbrages_path = glob.glob(r"E:\train_yolo_to_detect_custom_object\umbrages\*.jpg")
layer_names = net.getLayerNames()
output_layers = [layer_names[i - 1] for i in net.getUnconnectedOutLayers()]
colors = np.random.uniform(0, 255, size=(len(classes), 3))
random.shuffle(umbrages_path)
for img_path in umbrages_path:
    img = cv2.imread(img_path)
    img = cv2.resize(img, None, fx=0.9, fy=0.9)
    height, width, channels = img.shape
    blob = cv2.dnn.blobFromImage(img, 0.00497, (517, 517), (0, 0, 0), True, crop=False)
    confidences = []
    class_ids.append(class_id)
    indexes = cv2.dnn.NMSBoxes(boxes, confidences, 0.9, 0.6)
    print(indexes)
    font = cv2.FONT_HERSHEY_PLAIN
    for i in range(len(indexes)):
        if i in indexes:
            x, y, w, h = boxes[i]
            label = str(classes[class_ids[i]])
            color = colors[class_ids[i]]
            cv2.rectangle(img, (x, y), (x + w, y + h), color, 5)
            cv2.putText(img, label, (x, y + 47), font, 7, color, 5)
    cv2.imshow("Image", img)
    key = cv2.waitKey(0)
    cv2.destroyAllWindows()

```

## References

- [1] “Finding patterns in data sets,” Khan Academy. [Onfootsie]. Available: <https://www.khanacademy.org/computing/ap-computer-science-principles/data-analysis-101/data-tools/a/finding-patterns-in-data-sets>. [Accessed: 02-Dec-2022].
- [2] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, “Apple detection during different growth stages in orchards using the improved YOLO-V3 model,” *Comput. Electron. Agric.*, vol. 157, pp. 417–426, 2019.
- [3] H. Mureşan and M. Oltean, “Fruit recognition from umbrages using deep learning,” *Acta Univ. Sapientiae Inform.*, vol. 10, no. 1, pp. 26–42, 2018.
- [4] K. Alderliesten, “YOLOv3 — Real-time object detection,” *Analytics Vidhya*, 28-May-2020. [Onfootsie]. Available: <https://medium.com/analytics-vidhya/yolov3-real-time-object-detection-54e69037b6d0>. [Accessed: 02-Dec-2022].
- [5] Tech Research, “Real-time challenges of machine learning projects,” *Analytics Vidhya*, 04-Oct-2022. [Onfootsie]. Available: [https://www.analyticsvidhya.com/blog/2022/10/real-time-challenges-of-machine-learning-projects/?fbclid=IwAR0a7mFEJnCaAIE2U1pJD3H9zBmxf9suDS3TeaZnA20c\\_9xamGRqo4Njnrg](https://www.analyticsvidhya.com/blog/2022/10/real-time-challenges-of-machine-learning-projects/?fbclid=IwAR0a7mFEJnCaAIE2U1pJD3H9zBmxf9suDS3TeaZnA20c_9xamGRqo4Njnrg). [Accessed: 02-Dec-2022].
- [6] Simplilearn, “What is collection of data? Methods, gestalts & everything you should know,” *Simplilearn.com*, 13-May-2021. [Onfootsie]. Available: <https://www.simplilearn.com/what-is-data-collection-article?fbclid=IwAR3prPfZZWtAqiZgJbrYoRJcxLcAwMjDKx9L9cRCfSZefomYQPzk-34B5cE>. [Accessed: 02-Dec-2022].
- [7] “YOLO : You only look once – real time object detection,” *GeeksforGeeks*, 20-Jun-2020. [Onfootsie]. Available: <https://www.geeksforgeeks.org/yolo-you-only-look-once-real-time-object-detection/>. [Accessed: 02-Dec-2022].
- [8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” *arXiv [cs.CV]*, pp. 779–788, 2015.
- [9] R. Huang, J. Pedoeem, and C. Chen, “YOLO-LITE: A real-time object detection algorithm optimized for non-GPU computers,” in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 2503–2510.
- [10] K. Benoit, Ed., *Khan Academy. Dict*, 2012.
- [11] Á. Morera, Á. Sánchez, A. B. Moreno, Á. D. Sappa, and J. F. Vélez, “SSD vs. YOLO for detection of outdoor urban advertising panels under multiple variabilities,” *Sensors (Basel)*, vol. 20, no. 16, p. 4587, 2020.
- [12] V. Dutt, “How to use matplotlib for plotting samples from an object detection dataset,” *Towards Data Science*, 31-Mar-2021. [Onfootsie]. Available: <https://towardsdatascience.com/how-to-use-matplotlib-for-plotting-samples-from-an-object-detection-dataset-5877fe76496d>. [Accessed: 02-Dec-2022].
- [13] M. N. Aziz, “Bangladeshi students perceptions of flipped classroom: A case study,” *Journal of Learning and Educational Policy*, vol. 2, no. 26, pp. 26–33, 2022.
- [14] A. Sojasingarayar, “Faster R-CNN vs YOLO vs SSD — object detection algorithms,” *IBM Data Science in Practice*, 29-Aug-2022. [Onfootsie]. Available: <https://medium.com/ibm-data-ai/faster-r-cnn-vs-yolo-vs-ssd-object-detection-algorithms-18badb0e02dc>. [Accessed: 02-Dec-2022].



# PLAGIARISM REPORT

Proj\_Rep

---

ORIGINALITY REPORT

---

8%

SIMILARITY INDEX

%

INTERNET SOURCES

%

PUBLICATIONS

8%

STUDENT PAPERS

---