Department of Electronic & Electrical Engineering

BEng/MEng in EEE/EDS/EES/EME: 19.496/EM401 Individual Project

INTERIM REPORT

- A draft e-copy of the report (as pdf file and entitled "name_interim-report") must be submitted online via MyPlace by 12h00 on Friday 2nd December 2022.
- The final e-copy must be submitted (and confirmed) by Friday 9th December at 12h00. (The draft submission from the 2nd Dec will be considered as the final submission after the deadline.) No submission will be permitted after the 9th December.
- Formal written feedback will be provided after the project oral in January.

Student Name: Julien Priam	Registration Number: 202243899	Supervisor: Robert Atkinson
Project Title: AI (Machine Learning) for Cyber Security using Limited Data		

Please read the following extract from the University of Strathclyde Regulations 5.4 and 5.6.

"Essays, assignments, dissertations, project reports and other forms of written material, submitted by either individual students or groups, form important components of your assessed work leading to credit awards. It is your responsibility to ensure that such material is your/your group's own work, and not that of others. Failure to comply with this standard of academic honesty will result in penalties being imposed through reduction in marks and possible disciplinary action."

Full details of the University's Academic Dishonesty Policy and definitions of Plagiarism and Collusion are given in the University Regulations and reproduced in all Undergraduate and Graduate School Handbooks.

Declaration

By submitting the work I am confirming that that I have read and accepted the University policy and hereby declare that this work has not been submitted for any other degree/course at this University or any other institution and that, except where reference is made to the work of other authors, the material presented is original.

A key part of project conduct and reporting is the requirement to compare current progress to that anticipated in any previous reporting stage and then highlight where changes have occurred, the reasons for the changes and the actions taken in response to such changes. As part of the interim report, students are required to reflect upon the progress of the project to date, reflect upon on the degree to which this progress has met expectations and intimate the impact that such changes have had on the planned project objectives and deliverables. Limited to this current page. Comments should be simple and clear. Not duplicate content from main body of interim report.

A. Project Objectives

Discuss below how your project's objectives have changed or been revised in comparison to those detailed in your statement of intent. Describe the nature of these changes and explain the motivations for these changes.

Project global objectives have not changed since the SoI. However, the first objective (create the dataset from pcap files) proved to be more time consuming than expected.

It appears that, even if I already had knowledges about networks, a lot of time was spent to read about how packets network work, how are pcap files constructed…

B. Project Progress:

How does current project progress compare with that anticipated in the Statement of Intent? (Better, Worse or As Expected?) Please reflect upon the appropriateness of your progress and the degree to which you are satisfied with your current progress.

Regarding what have been said in the previous section, the current project progress is slower than expected as I am still working on creating the dataset and adding label to it during week 9.

Despite these changes, I would still say that I am satisfied with my progress. The delay is due to a lack of understanding of the project at the beginning of the year, but I now have a better perception of what I must do, and I understand that some part would take longer than I first expected.

C. Project Deliverables:

Have the project deliverables changed significantly from those listed in the Statement of Intent? If so, discuss why such changes were necessary. Comment upon how any further significant changes to the schedule of project deliverables can be avoided.

The project progress so far does not require a change in deliverables.

Project Context & Background

Interim Report;

The maximum length of the main report body is 4 pages and **must** follow the structure given below:

Page1-2:	Project Context and Background material
Page 3:	Project plan – full-page Gantt chart - landscape. Indicating tasks completed as well as tasks planned.
Page 4:	Discussion of Project Plan and Technical Risks; include any ethics/sustainability issues appropriate to the project

Additional allowed content:

- Bibliography & Reference List

Font should be between 10-12 point. Around 400-500 words per page.

What and Why:

Cybersecurity aims to protect the integrity of technologies such as networks and informatic devices to keep data safe and private. Attackers can proceed in a lot of different ways to achieve their goals. Indeed, they can corrupt people, use pishing campaigns, use malwares… This project aims to prevent attacks coming from the network through IP packets. To do so, we will use an Artificial Intelligence (AI) that will classify in real time the packets of a network into different classes representing different types of attacks.

However, training a classifier requires large datasets which are not easily available in this field. The project then aims to test different Machine Learning (ML) models that can be efficient with a small dataset.
This project also aims to detect and extract good features to train the ML model in an efficient way. The features should provide high accuracy but avoid too much correlation as it would only slower the classifier without improving performances.

State of art:

The use of artificial intelligence in cybersecurity is a growing field as it often provides an automatic way to secure networks and allows the treatment of huge amount of data. In the field of Intrusion Detection System (IDS), the CICIDS-2017 is known to be a good dataset, available on internet (Canadian Institution for Cybersecurity CIC). This dataset is available in a csv format gathering 70 different features and 14 different types of attacks, some of them which never appeared in previous dataset.

However, researchers show that this dataset presents some shortcomings that can bias the results of the classifiers (see A detailed analysis of CICIDS2017).
As the CIC also provides the true real-world data (pcap files), the project will be based on these files. It will then be necessary to create a parser program (using Python) to extract useful features from the packets available, in order to re-create a dataset that can be used for training ML models.

Technical:

The creation of a useful dataset based on pcap files has proved to be a big part of the project in my case.
Pcap is a file format often generated by Wireshark that contains the capture of network analysis. The data is basically formed by a series of network packets.

In order to get usable data, we need to create a parser that works in 3 major steps:

1) Loop on each packet of the pcap file to extract relevant information (Timestamp, IPs, Ports, protocol, flags…) and to store it in a python dictionary.

2) Groups the packets extracted depending on the IPs, Ports and protocols to re-create the flows. Then compute new features for each flow (such as length of flow, total number of flags, time between packets…). The result is once again stored in a python dictionary.

3) Using the dictionary of unidirectional flows, groups them 2-by-2 to re-create bidirectional flows and computes the new features. The result is stored in a python dictionary, and it can be used to create a python dataframe which is a good data format for training AI.

After this first process, we recreated a database close to the one provided by CIC in .csv format, but we are now able to add or change the features depending on what we need and what will proved to be relevant in the next parts of the project.

However, starting from the raw data, our dataset does not contain a label column. As a result, we cannot directly test the efficiency of the dataset on a simple classifier model. A second python script will be used to add labels, based on the information provided by CIC. They provide a list of all the attacks performed during the time packets were captured. This basically contains the IP of the attackers/victims, the time when the attacks were performed and the types of attacks.
The python script will then loop on all the flows, looking for a match between the information provided and the ones extracted from the pcap files. If a match is found then the appropriate label is added, otherwise a benign label is added.

At the end of this script, the dataset is constructed, contains labels, and therefor can be used to train a classifier model.

The next step is to train a simple classifier that will be used to test and select the features of the dataset. To do this we implement a simple Random Forest Classifier based on sklearn module.
However, some of the features are non-categorical, so it makes no sense to compare them directly. To avoid this problem it is necessary to implement one-hot encoding.

Training the Random Forest Classifier on 80% of the dataset provided by the Canadian Institution for Cybersecurity gave really good results when testing it on the 20% left:

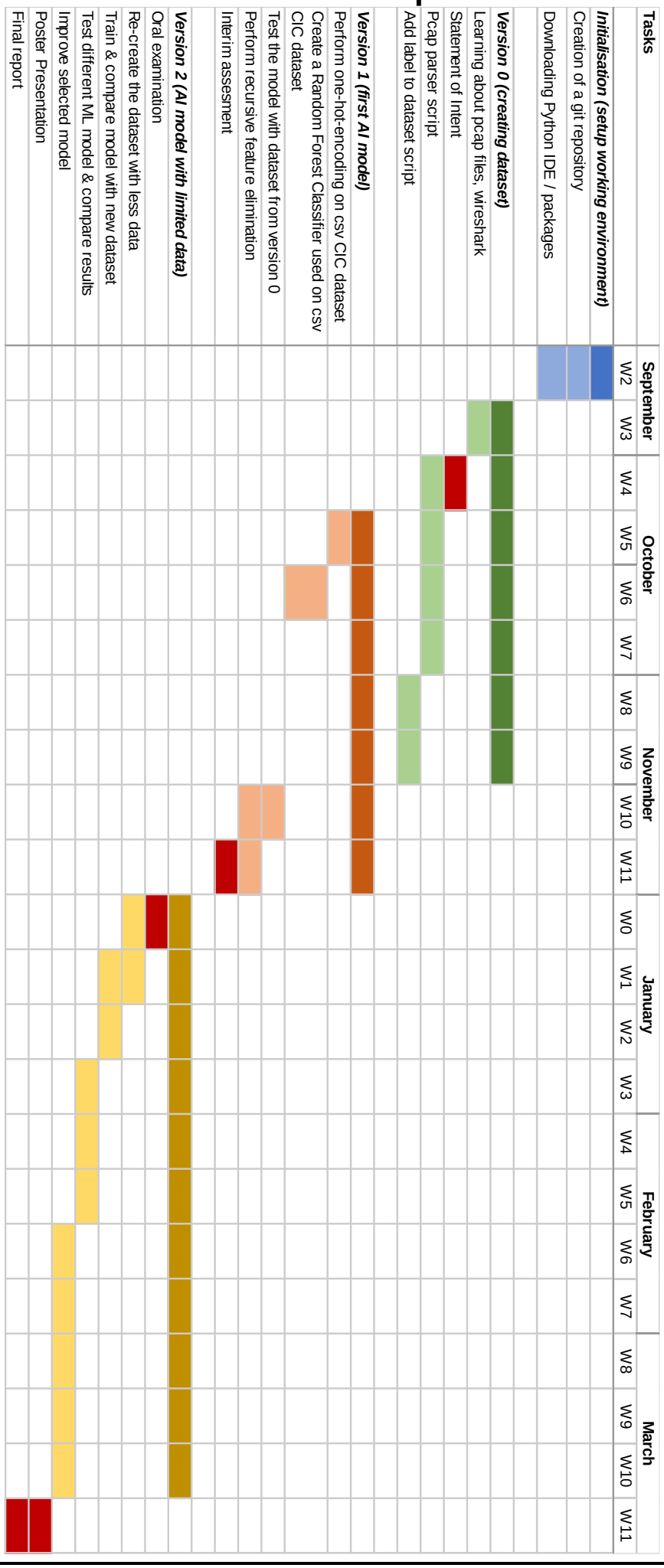
- Accuracy on training sample is 0.99998
- Accuracy on testing sample is 0.99993

At the point of the Interim Report the ML model has not yet been tested on the dataset created from the pcap files.

When this first part of the work is working, we will have to apply the python scripts on smaller data files. The aim is to compare the results with big datasets, then find a way to improve the AI to be still efficient with less data. To do so, it will be necessary to try other ML models and compare the outputs.

.

Project Plan – Gantt Chart – Landscape



Discussion of Project Plan and Technical Risks

The plan above shows what have been down until week 10, and what is planned for the rest of the year. This Gantt is different from the one stated in the SoI for first semester.

The difference is due to the time it took to create an efficient pcap parser, capable to extract the required features in a reasonable amount of time (pcap files are 50Gb which takes a long time to read).
Moreover, adding label to the dataset also took longer than expected due to an issue relative to timestamp in packets. Indeed, the packets capture is known to start at 9am but reading the data in pcap files, it was starting at 12pm. I also lost some time during the label process because I was not aware of Network Address Translation (NAT), so I could not find any correspondence between the data I had and what was stated in the CIC documentation.

During the month of November, I have been given access to a server with more computing power. Once set up, this server allows me to save a lot of time when running python scripts on huge files, but it also took me some time to get everything working on the server.

However, this delay should not have an impact on second semester plans because I started working on ML model during weeks 5 and 6 using the dataset given by CIC. Thanks to this, the ML model is already working, and I will just have to apply the model to the new dataset created from pcap files. That is why the end of version 1 should only take the 2 last weeks of December to be completed.

The main task of semester 2 will be to develop an efficient ML model having limited data available. I plan to test different ML model such as decision Trees, Naïve Bayes, Nearest Neighbours… and compare the results to determine which is the best. As scikit learn library implements these different algorithms and a lot of documentation is available, this step should not be too much time consuming. Finally, I plan to work on the selected classifier to adapt it to the current problem and so improve the results.
For this second parts, the risk specified in the SoI about the lack of computing power still apply, but thanks to the access to a dedicated server, the risk should be minimised.

The Gantt chart also includes the important deadlines (in red) and before each of these, some time will be spent to write and prepare the documents or orals.