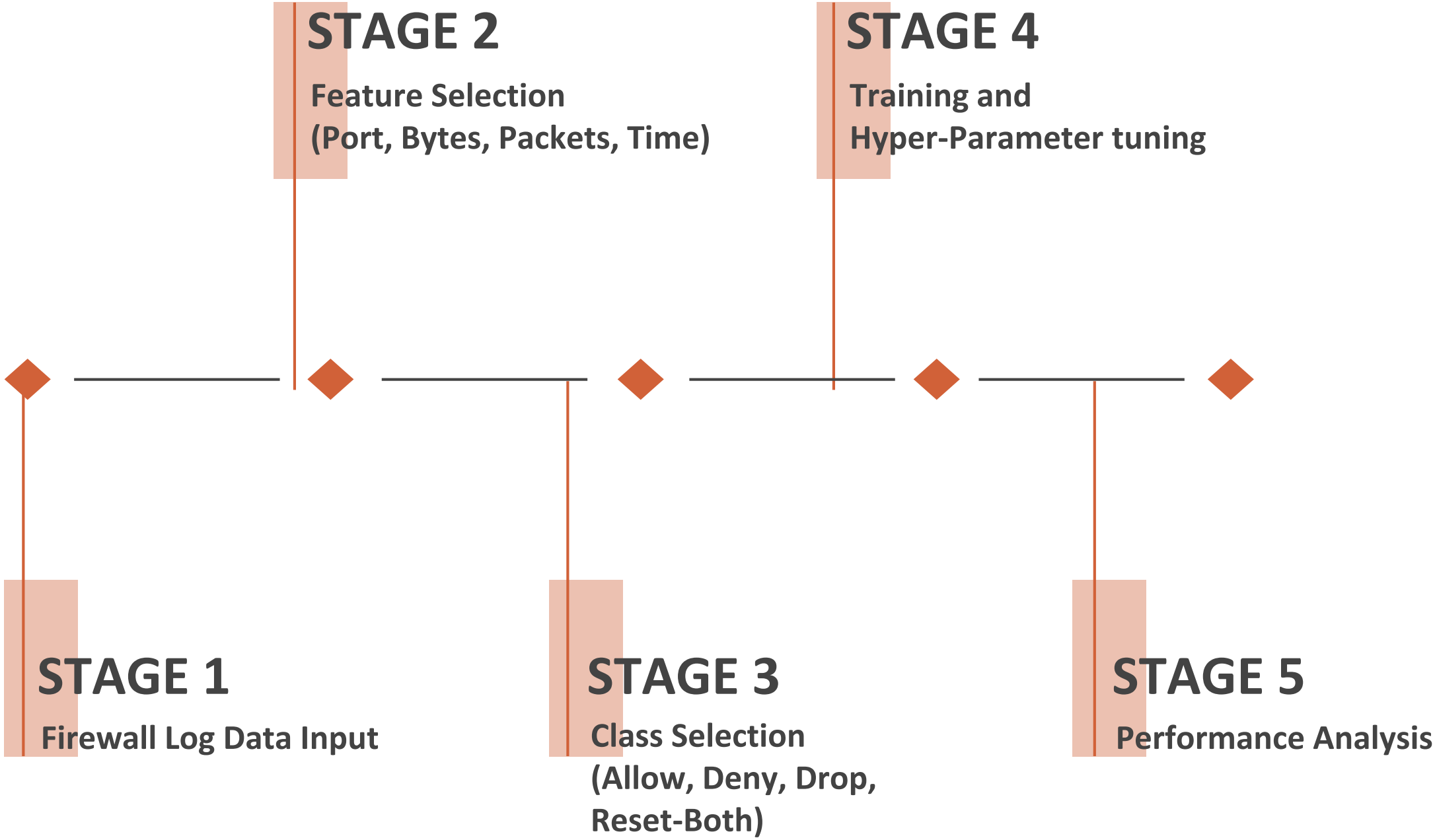
數位影像處理
期末報告
PART1

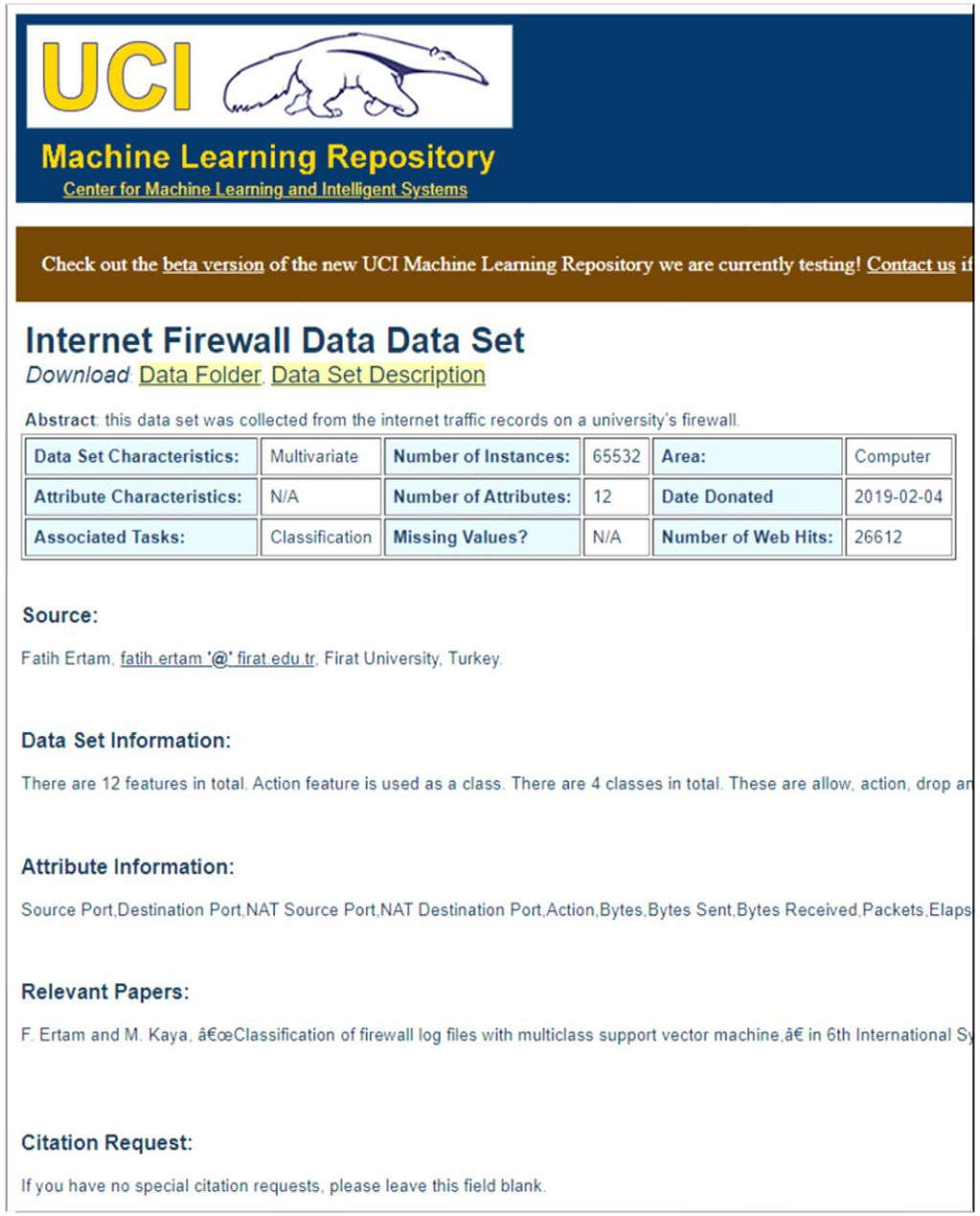
109321041 資工三 劉彥汝

目錄


STAGE 2
Feature Selection
(Port, Bytes, Packets, Time)
STAGE 4
Training and
Hyper-Parameter tuning
STAGE 1
Firewall Log Data Input
STAGE 3
Class Selection
(Allow, Deny, Drop,
Reset-Both)
STAGE 5
Performance Analysis

資料集的敘述

01



UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Check out the beta version of the new UCI Machine Learning Repository we are currently testing! Contact us if

Internet Firewall Data Data Set

Download: Data Folder, Data Set Description

Abstract: this data set was collected from the internet traffic records on a university's firewall.

Data Set Characteristics:	Multivariate	**Number of Instances:**	65532	**Area:**	Computer
Attribute Characteristics:	N/A	**Number of Attributes:**	12	**Date Donated**	2019-02-04
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	26612

Source:

Fatih Ertam, fatih.ertam '@' firat.edu.tr, Firat University, Turkey.

Data Set Information:

There are 12 features in total. Action feature is used as a class. There are 4 classes in total. These are allow, action, drop ar

Attribute Information:

Source Port,Destination Port,NAT Source Port,NAT Destination Port,Action,Bytes,Bytes Sent,Bytes Received,Packets,Elaps

Relevant Papers:

F. Ertam and M. Kaya, â€œClassification of firewall log files with multiclass support vector machine,â€ in 6th International Sy

Citation Request:

If you have no special citation requests, please leave this field blank.

ABOUT THE PROJECT

蒐集防火牆設備生成的數據，檢查生成的數據所使用的策略來允許或阻止流量，希望能夠透過機器學習來找出有問題的地方。

Dataset 網址

這是一個 Classification 的任務

FEATURES AND DESCRIPTION 1

01

Source Port

Client Source Port

02

Destination Port

Client Destination Port

03

NAT Source Port

Network Address Translation Source Port

04

NAT Destination Port

Network Address Translation

Destination Port

05

Elapsed Time (sec)

Elapsed Time for flow

06

Bytes

Total Bytes

FEATURES AND DESCRIPTION 2

07

Bytes Sent

Bytes Sent

08

Bytes Received

Bytes Received

09

Packets

Total Packets

10

pkts_sent

Packets Sent

11

pkts_received

Packets Received

12

Action

Class
(allow, deny, drop, reset-both)

資料觀察

我們可以發現到當 Action 不是 allow 的時候，NAT Source Port、NAT Destination Port、Bytes Received、Packets、Elapsed Time(sec)、pkts_sent、pkts_received 有明顯的改變；而 Source Port 和 Destination Port 並沒有什麼影響。

Source Port	Destination Port	NAT Source Port	NAT Destination Port	Action	Bytes	Bytes Sent	Bytes Received	Packets	Elapsed Time (sec)	pkts_sent	pkts_received
50816	443	50681	443	allow	134	60	74	3	6	2	1
50057	80	26007	80	allow	1570	754	816	10	16	6	4
52146	80	44068	80	allow	366	240	126	7	33	5	2
54139	53	64934	53	allow	168	78	90	2	31	1	1
49410	53	29049	53	allow	177	94	83	2	31	1	1
53994	16605	15809	16605	allow	70	70	0	2	8	2	0
35242	8635	35242	8635	allow	138	78	60	2	6	1	1
52148	80	27900	80	allow	366	240	126	7	33	5	2
49900	443	28761	443	allow	1820	1101	719	18	73	10	8
43931	53	36161	53	allow	168	78	90	2	31	1	1
51048	445	0	0	drop	70	70	0	1	0	1	0
51045	445	0	0	drop	70	70	0	1	0	1	0
13394	23	0	0	deny	60	60	0	1	0	1	0
61078	57470	0	0	deny	62	62	0	1	0	1	0
55725	445	0	0	drop	70	70	0	1	0	1	0
55723	445	0	0	drop	70	70	0	1	0	1	0
55724	445	0	0	drop	70	70	0	1	0	1	0
51125	445	0	0	drop	66	66	0	1	0	1	0
51123	445	0	0	drop	66	66	0	1	0	1	0
51122	445	0	0	drop	66	66	0	1	0	1	0
1024	21854	0	0	reset-both	157	157	0	1	0	1	0
11317	53563	0	0	reset-both	143	143	0	1	0	1	0

防火牆安全政策

(Action)

Allow

允許網路通過。

Deny

阻止網路通過並強制拒絕應用程序。

Drop

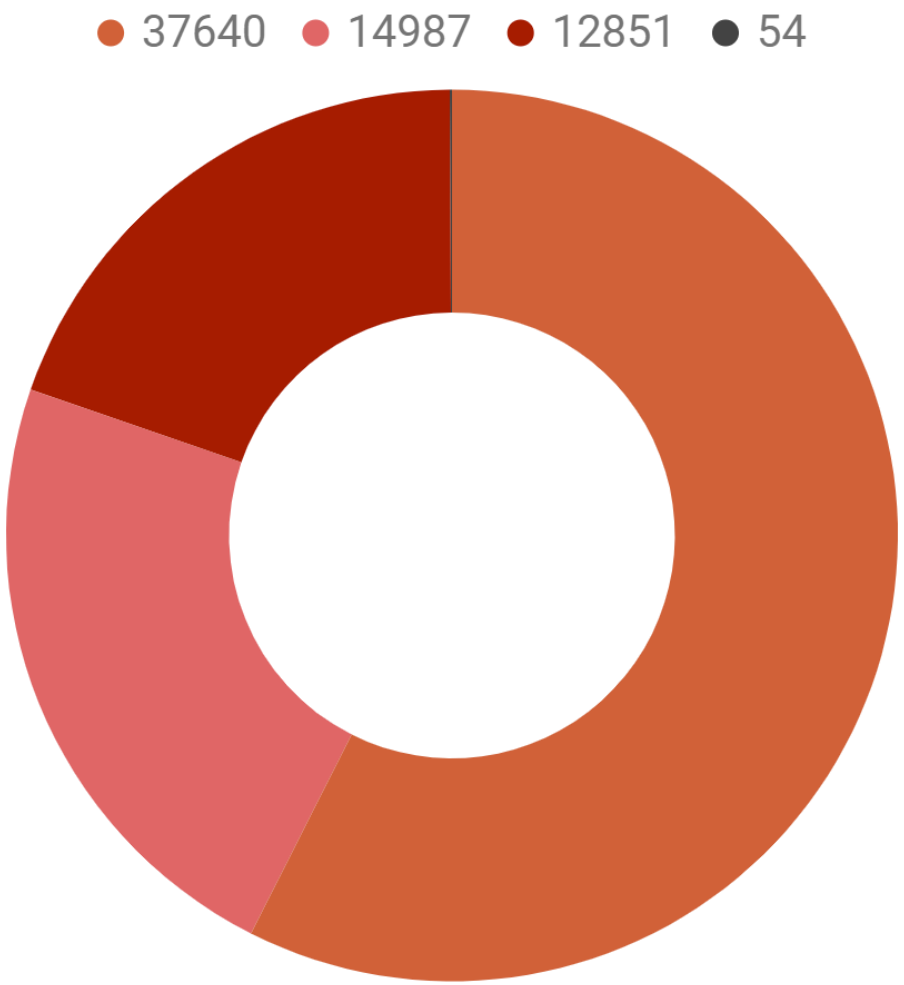
降低網路流量，會漸漸覆蓋掉預設的拒絕指令，但並不會送出 TCP Reset 指令至Host / Application。

Reset-Both

送出 TCP Reset 指令至 client 端與 server 端。

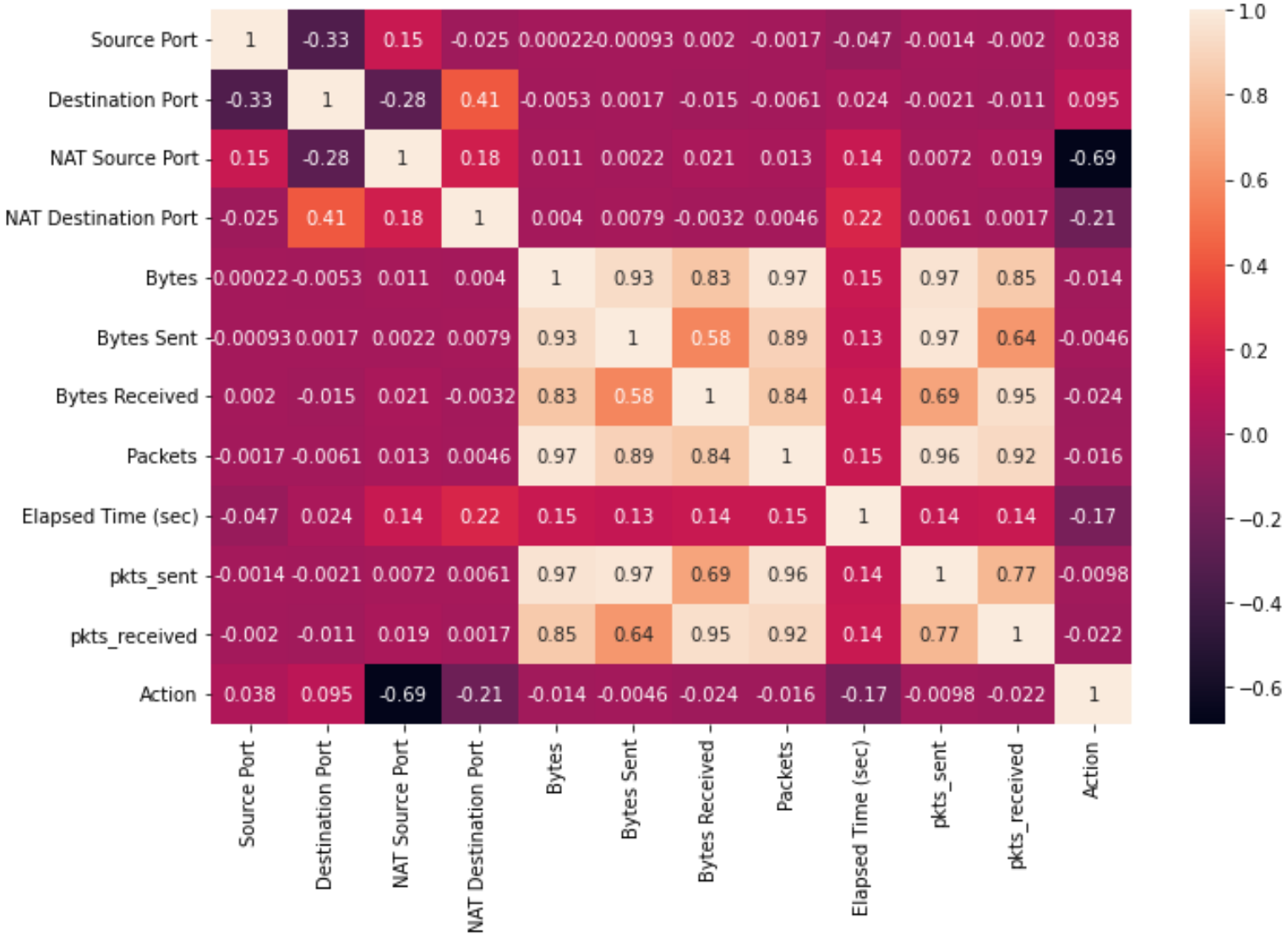
資料數量分析

整體數據極度不平衡，尤其是 **Reset-Both** 只佔據了 **0.2%** 而已，數量小到可以忽略其存在，而又因為這份資料集是從防火牆蒐集而來的，不像 image 可以旋轉縮放等等，可以做 Data Augmentation。



Allow	57.4%
Deny	22.8%
Drop	19.6%
Reset-Both	0.2%

各參數之間的 Pearson相關係數



從 Pearson 相關係數的熱力圖裡，我們可以發現與 Action 最有相關的居然是 NAT Source Port，其他幾乎都低於 0.1。

資料集的
前處理程式碼

02

資料清理

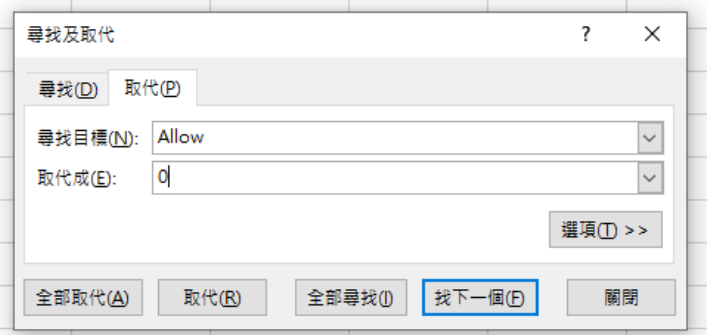
我們在其中一項資料發現極端值，它可能會影響到訓練的成果，因此需要刪除此筆異常值資料，且我們沒有發現有值缺失，所以不需要補空值。

	Source Port	Destination Port	NAT Source Port	NAT Destination Port	Bytes	Bytes Sent	Bytes Received	Packets	Elapsed Time (sec)	pkts_sent	pkts_received	Action
10221	52125	80	43801	80	2178	1003	1175	10	15	6	4	0
10222	57235	15187	23276	15187	1.269E+09	948477220	320881795	1036116	9283	747520	288596	0
10223	50410	443	7615	443	14049	602	13447	10	35	6	13	0

資料正規化

我們使用 Excel 的取代功能，把 Action 裡的 allow、deny、drop 和 Reset-Both 分別替換成 0、1、2、3。

	Source Por	Destination	NAT Sourc	NAT Desti	Action	Bytes	Bytes Sent	Bytes Rece	Packets	Elapsed Tir	pkts_sent	pkts_received
2	57222	53	54587	53	allow	177	94	83	2	30	1	1
3	56258	3389	56258	3389	allow	4768	1600	3168	19	17	10	9
4	6881	50321	43265	50321	allow	238	118	120	2	1199	1	1
5	50553	3389	50553	3389	allow	3327	1438	1889	15	17	8	7
6	50002	443	45848	443	allow	25358	6778	18580	31	16	13	18
7	51465	443	39975	443	allow	3961	1595	2366	21	16	12	9
8	60513	47094	45469	47094	allow	320	140	180	6	7	3	3
9	50049	443	21285	443	allow	7912	3269	4643	23	96	12	11
10	52244	58774	2211	58774	allow	70	70	0	1	5	1	0
11	50627	443	16215	443	allow	8256	1674	6582	31	75	15	16
12	43676	80	45378	80	allow	696	378	318	12	35	7	5
13	52190	443	16680	443	allow	7942	870	7072	22	15	10	12
14	50690	80	20479	80	allow	4805	3639	1166	16	31	9	7
15	55597	53	45448	53	allow	168	86	82	2	30	1	1
16	49164	443	45916	443	allow	7292	950	6342	19	75	9	10
17	36887	443	63451	443	allow	10922	2532	8390	27	28	13	14
18	1939	53	33288	53	allow	210	78	132	2	30	1	1
19	50281	53	33175	53	allow	195	102	93	2	30	1	1
20	57222	53	51448	53	allow	177	94	83	2	30	1	1
21	56710	53	57885	53	allow	177	94	83	2	30	1	1



資料標準化

在大數據資料中，是用不同資料欄位與資料值所組成，他們可能分佈狀況可能都不盡相同，因此必須將特徵資料按比例縮放，讓資料落在某一特定的區間。

我們使用 Matlab 裡的normalize，使得資料分佈區間在 [0, 1] 之間。

```
predicator = normalize([port, Byte, other])
```

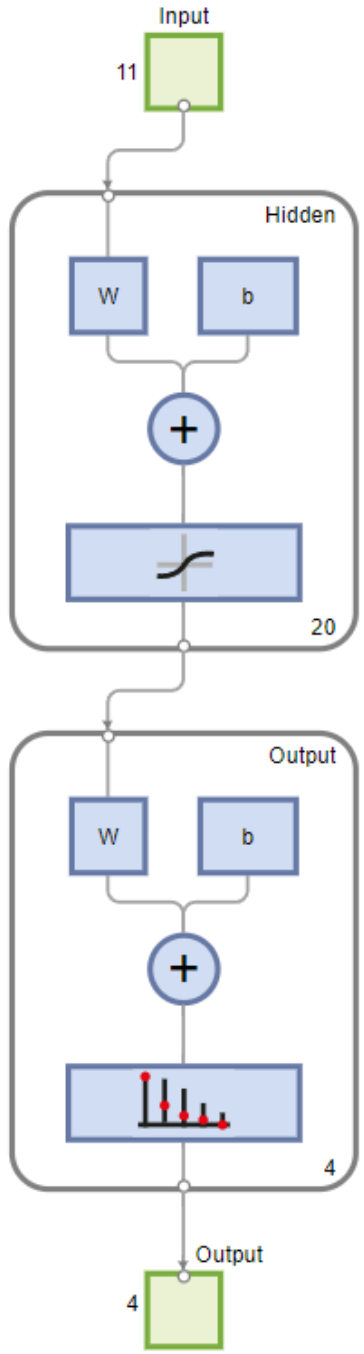
predicator = 65531×11

	1	2	3	4	5	6	7	8	9	10	11
8	0.0431	-0.5488	0.0911	-0.2288	-0.0264	-0.0048	-0.0307	-0.0203	0.1009	-0.0133	-0.0240
9	0.1870	2.6100	-0.7770	5.7606	-0.0294	-0.0081	-0.0329	-0.0273	-0.2021	-0.0214	-0.0298
10	0.0810	-0.5488	-0.1396	-0.2288	-0.0263	-0.0065	-0.0298	-0.0178	0.0310	-0.0111	-0.0214
11	-0.3747	-0.5685	1.1877	-0.2660	-0.0292	-0.0078	-0.0328	-0.0238	-0.1022	-0.0170	-0.0272
12	0.1834	-0.5488	-0.1185	-0.2288	-0.0264	-0.0073	-0.0296	-0.0206	-0.1688	-0.0148	-0.0235
13	0.0851	-0.5685	0.0544	-0.2660	-0.0276	-0.0044	-0.0324	-0.0225	-0.1155	-0.0155	-0.0261
14	0.4067	-0.5699	1.1909	-0.2688	-0.0294	-0.0081	-0.0329	-0.0269	-0.1189	-0.0214	-0.0293
15	-0.0149	-0.5488	1.2122	-0.2288	-0.0267	-0.0072	-0.0299	-0.0216	0.0310	-0.0155	-0.0246

神經網路架構與參數調整

參數

- 20層隱藏層
- 使用 patternnet
- 80 / 10 / 10，Train / Validation / Test
- 4層輸出層

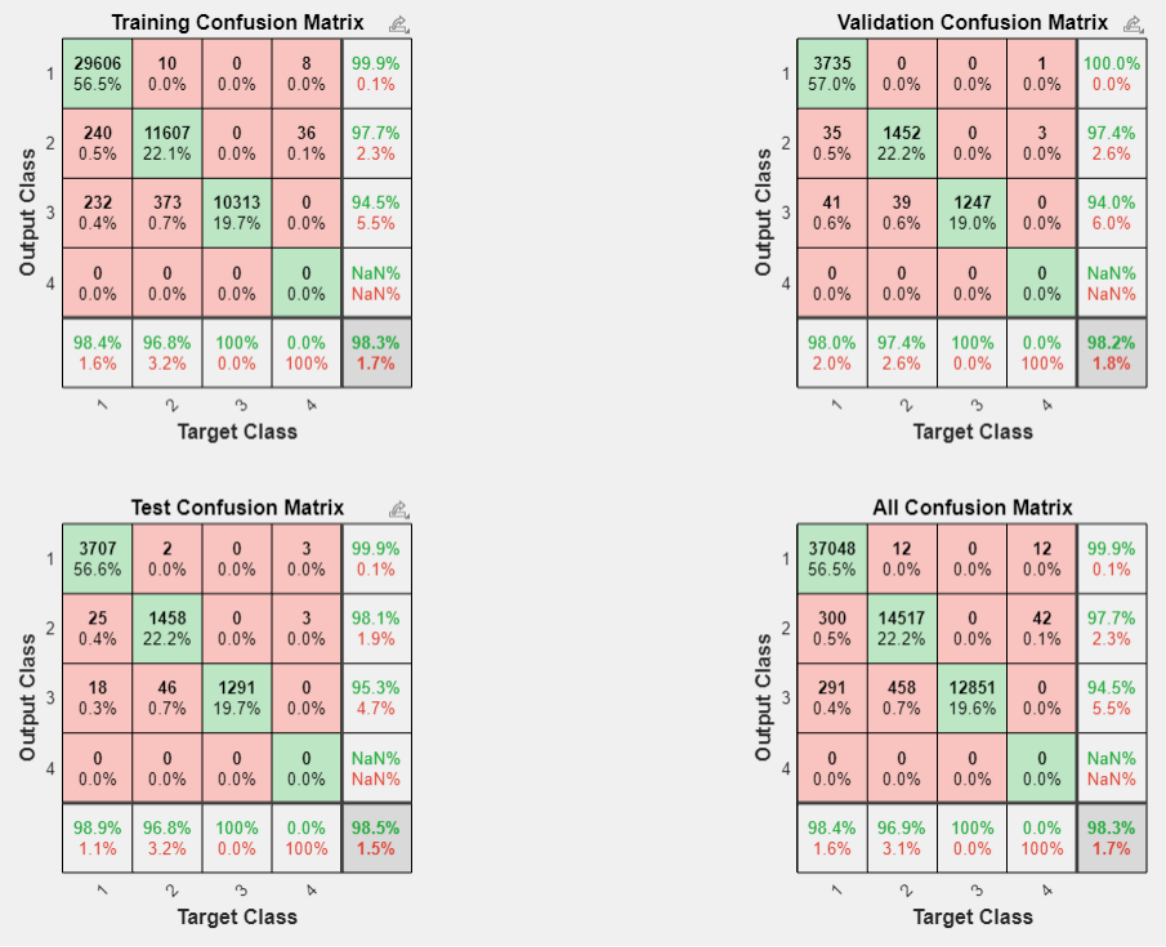


03

學習成果評估

04

Confusion Matrix



Accuracy ：98.3%

Precision ：73.025%

Recall ：73.825%

F1 Score ：73.423%

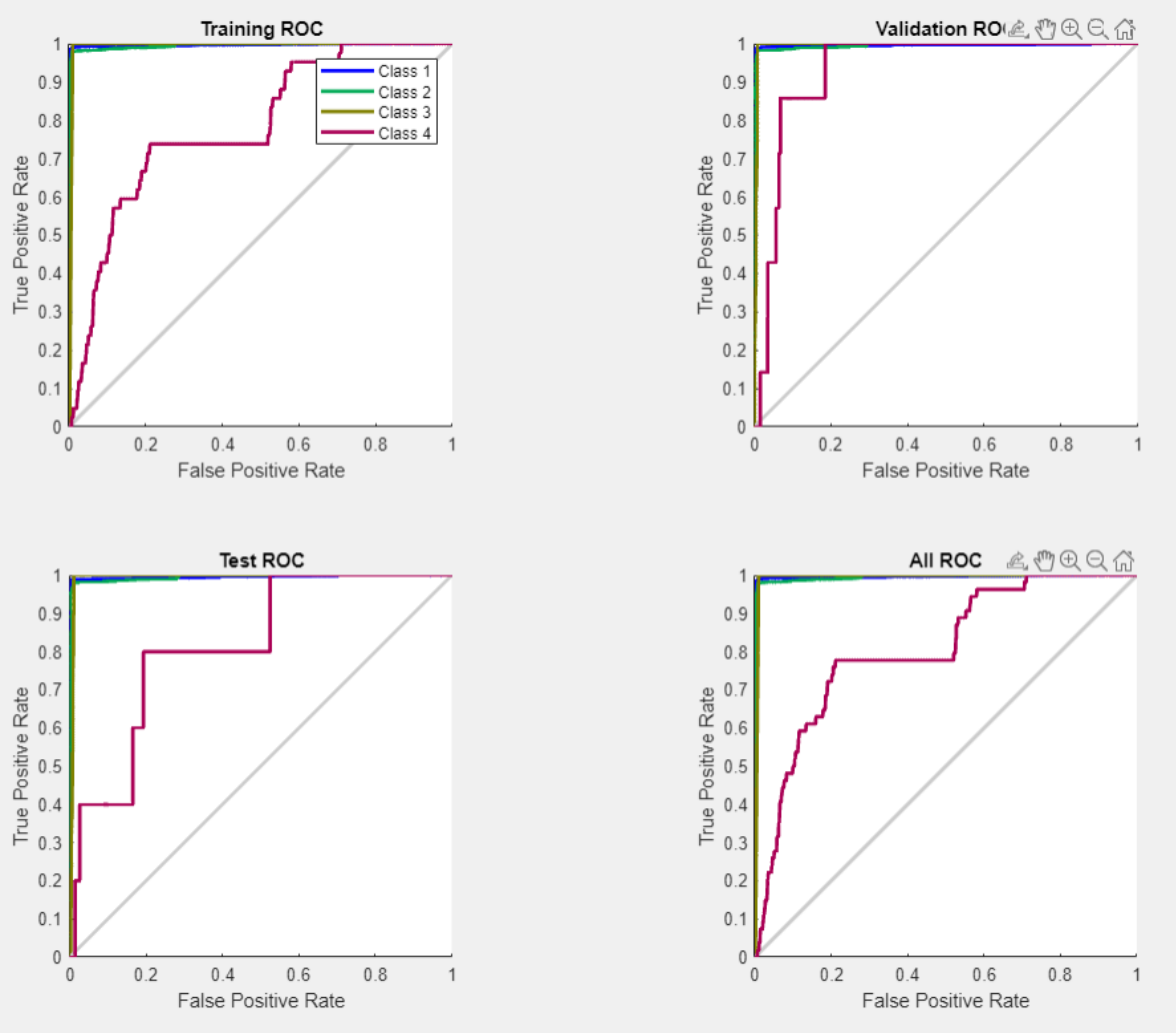
***我的模型的 Precision 比論文的模型都高**

TABLE III. EVALUATION RESULTS

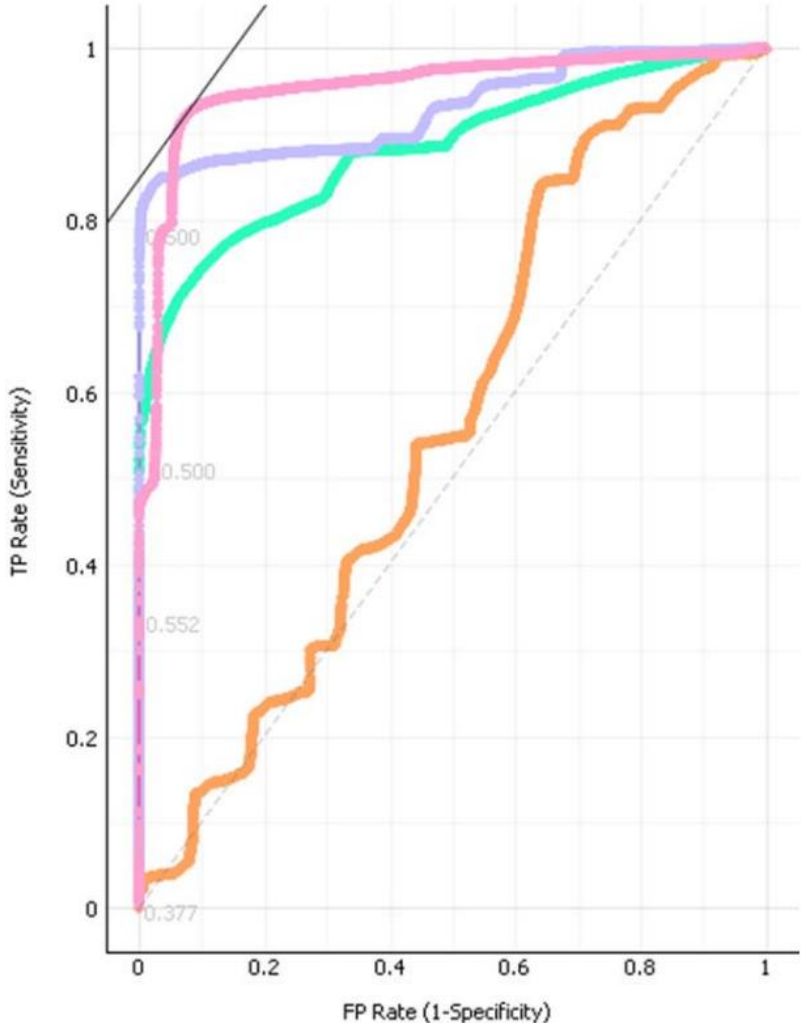
Method	F_1 Score	Precision	Recall
SVM Linear	75.4	67.5	85.3
SVM Polynomial	53.6	61.8	47.4
SVM RBF	76.4	63.0	97.1
SVM Sigmoid	74.8	60.3	98.5

***論文只有計算 F1 Score、Precision、Recall 而已，沒有給 Confusion Matrix，所以無法得知其 Accuracy**

ROC Curve



我的model



論文的model

結論與心得

我的模型的 Precision 比論文的模型都還要高，但 Precision 和 Recall 本身就是相輔相成的數值，這代表我的模型較能預測為 Positive 的結果。

回到這份 Dataset 的主軸－分析防火牆設備生成的數據，來產生指令(Allow、Deny、Drop、Reset-Both)

為了安全著想，我們應該更在乎準確率，「寧可錯殺一萬，不可放過萬一」。

所以我認為我的模型比論文的模型還要更適合，儘管兩者的 F1 score 並沒有太大的差距。

05

THANKS