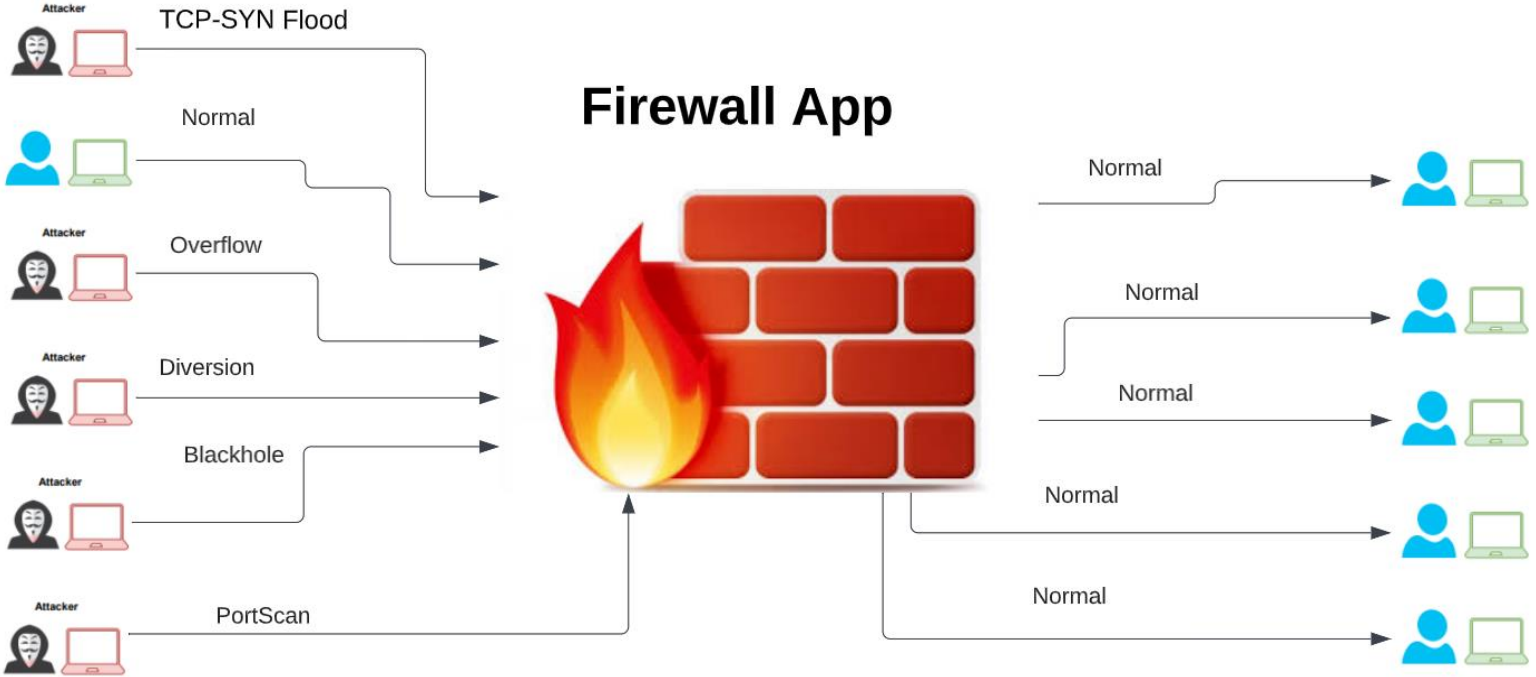
Term Project template

Group 11

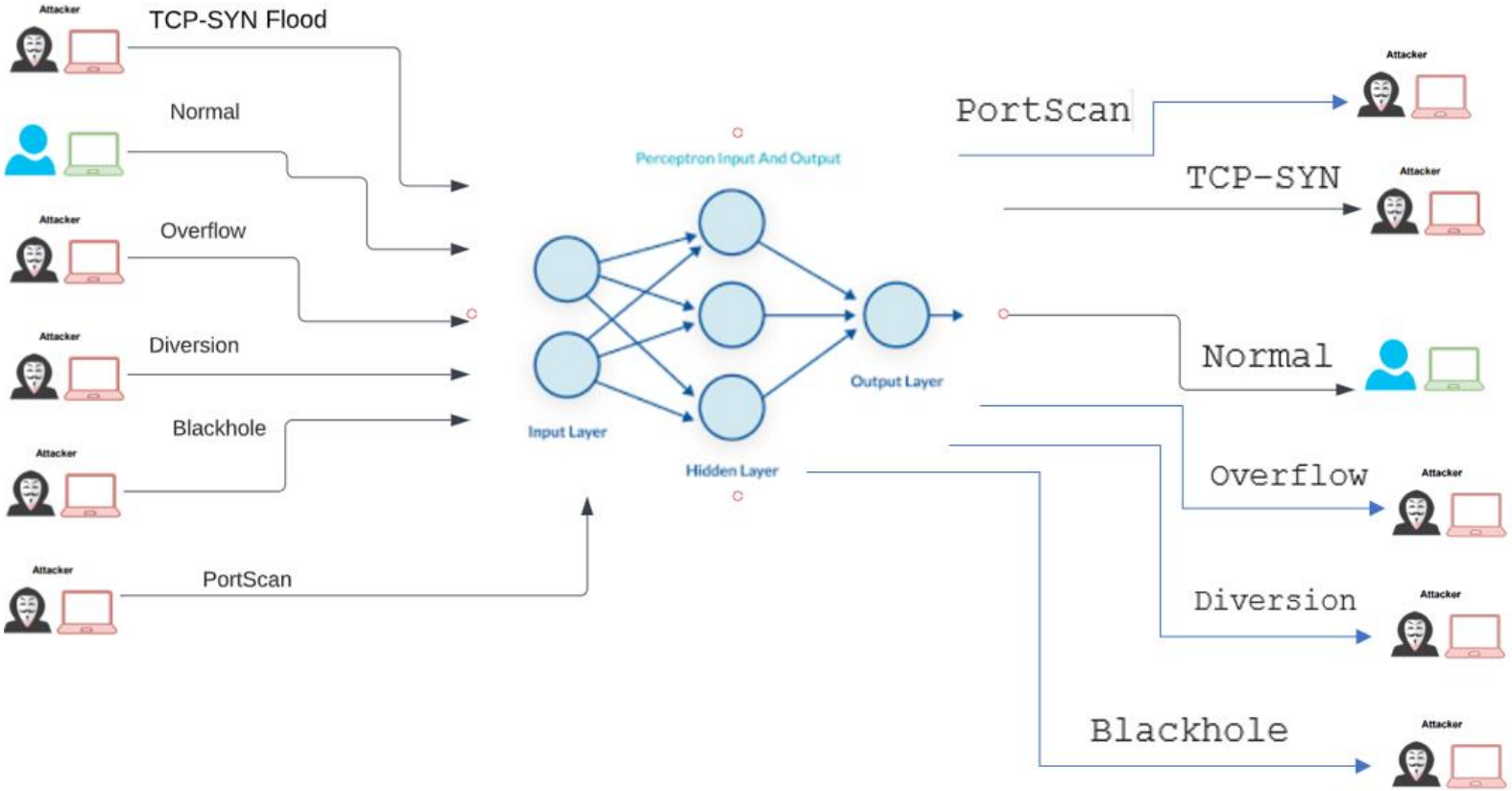
Hadeer Mamoduh Abdelfattah Mohammed	300327273
Nada Abdellatef Shaker Seddik	300327294
Mostafa Mahmoud Abdelwahab Nofal	300327286

Problem’s overview

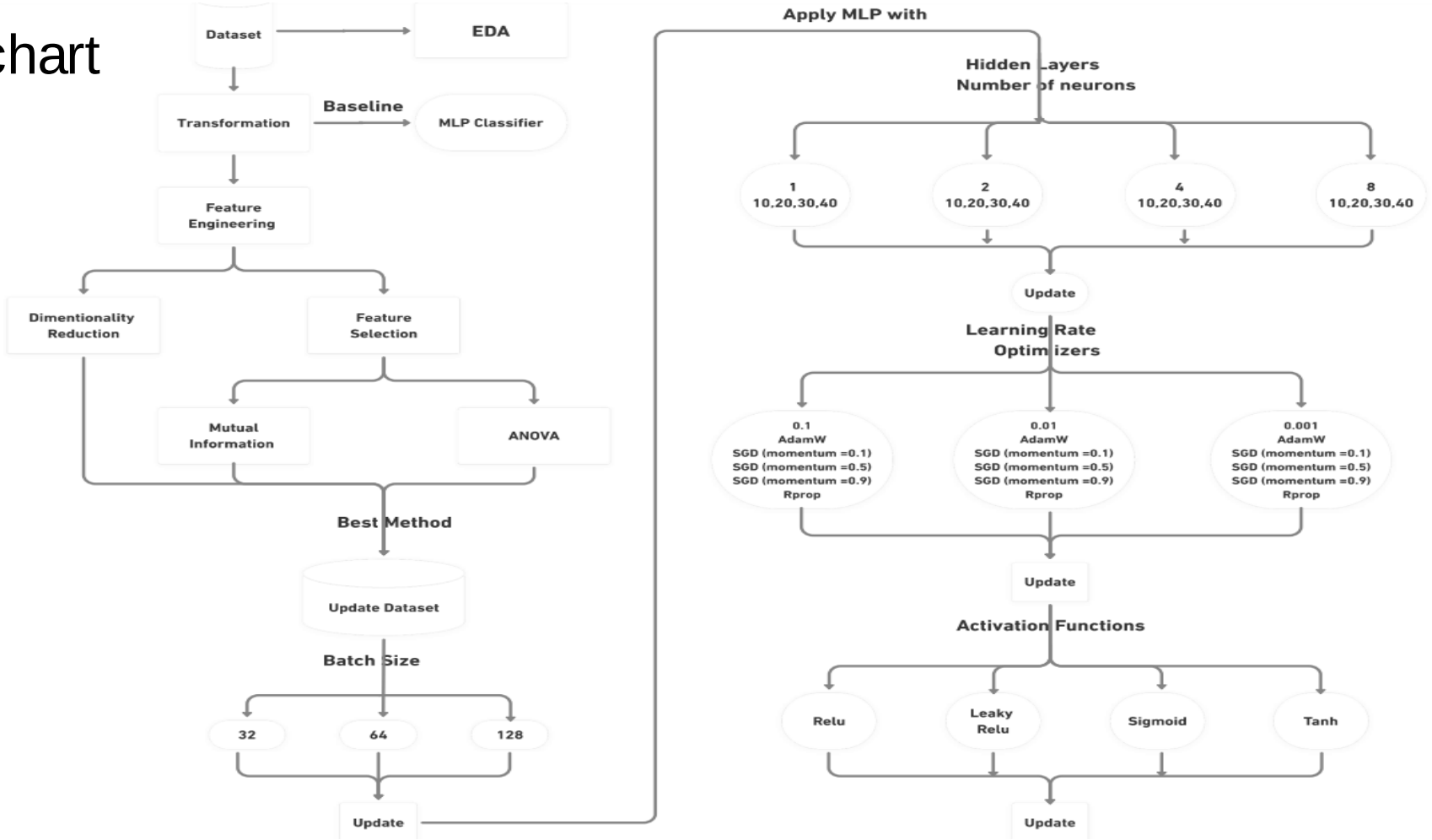
- Before using DL-Model



- After using DL Model



Flowchart


Dataset
EDA
Transformation
Baseline
MLP Classifier
Feature Engineering
Dimentionality Reduction
Feature Selection
Mutual Information
ANOVA
Best Method
Update Dataset
Batch Size
32
64
128
Update
Apply MLP with
Hidden Layers
Number of neurons
1
10,20,30,40
2
10,20,30,40
4
10,20,30,40
8
10,20,30,40
Update
Learning Rate
Optimizers
0.1
AdamW
SGD (momentum =0.1)
SGD (momentum =0.5)
SGD (momentum =0.9)
Rprop
0.01
AdamW
SGD (momentum =0.1)
SGD (momentum =0.5)
SGD (momentum =0.9)
Rprop
0.001
AdamW
SGD (momentum =0.1)
SGD (momentum =0.5)
SGD (momentum =0.9)
Rprop
Update
Activation Functions
Relu
Leaky Relu
Sigmoid
Tanh
Update

Dataset's overview (EDA)

- EDA is applied to **investigate** the data and **summarize** the key insights.

```
1 train_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2619 entries, 0 to 2618
Data columns (total 22 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Switch ID                       2619 non-null   object
 1   Port Number                     2619 non-null   object
 2   Received Packets                2619 non-null   int64
 3   Received Bytes                  2619 non-null   int64
 4   Sent Bytes                      2619 non-null   int64
 5   Sent Packets                    2619 non-null   int64
 6   Port alive Duration (S)         2619 non-null   int64
 7   Delta Received Packets          2619 non-null   int64
 8   Delta Received Bytes            2619 non-null   int64
 9   Delta Sent Bytes                2619 non-null   int64
 10  Delta Sent Packets              2619 non-null   int64
 11  Delta Port alive Duration (S)   2619 non-null   int64
 12  Connection Point                2619 non-null   int64
 13  Total Load/Rate                 2619 non-null   int64
 14  Total Load/Latest               2619 non-null   int64
 15  Unknown Load/Rate               2619 non-null   int64
 16  Unknown Load/Latest             2619 non-null   int64
 17  Latest bytes counter            2619 non-null   int64
 18  Active Flow Entries             2619 non-null   int64
 19  Packets Looked Up               2619 non-null   int64
 20  Packets Matched                 2619 non-null   int64
 21  Label                           2619 non-null   object
dtypes: int64(19), object(3)
memory usage: 450.3+ KB
```

```
1 train_data.duplicated().sum()
```

```
0
```

```
1 test_data.duplicated().sum()
```

```
0
```

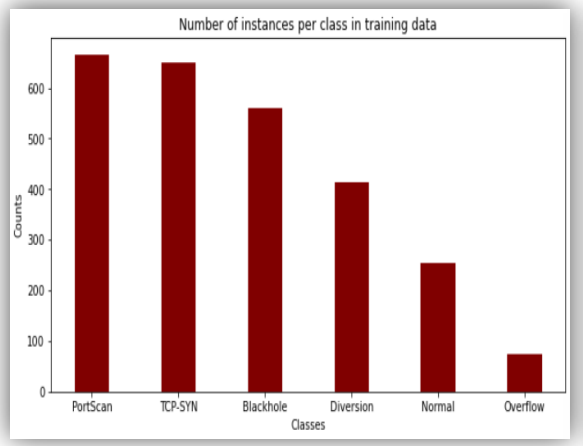
```
1 val_data.duplicated().sum()
```

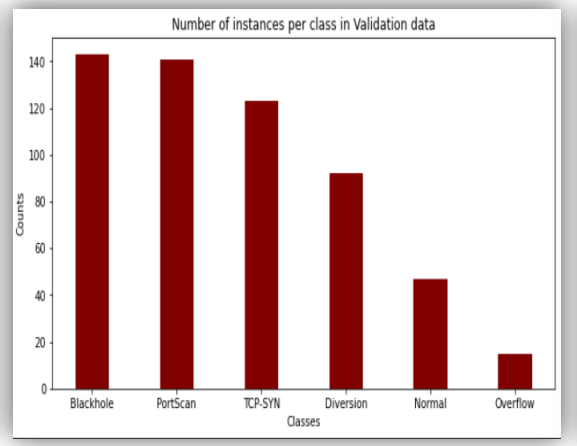
```
0
```

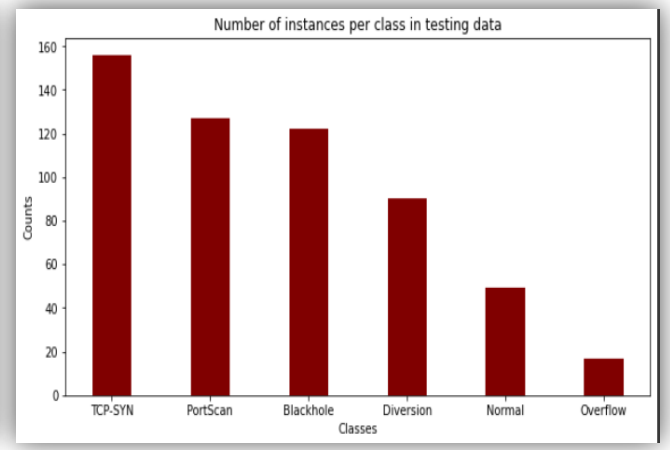
Cont.

- Number of instances per class

```
PortScan      666
TCP-SYN       651
Blackhole     561
Diversion     414
Normal        254
Overflow       73
Name: Label, dtype: int64
```







Dataset’s overview (EDA)

- Number of null values per feature

Train

```
Switch ID                       0
Port Number                     0
Received Packets                0
Received Bytes                  0
Sent Bytes                      0
Sent Packets                    0
Port alive Duration (S)         0
Delta Received Packets          0
Delta Received Bytes            0
Delta Sent Bytes                0
Delta Sent Packets              0
Delta Port alive Duration (S)   0
Connection Point                0
Total Load/Rate                 0
Total Load/Latest               0
Unknown Load/Rate               0
Unknown Load/Latest             0
Latest bytes counter            0
Active Flow Entries             0
Packets Looked Up               0
Packets Matched                 0
Label                           0
dtype: int64
```

Test

```
Switch ID                       0
Port Number                     0
Received Packets                0
Received Bytes                  0
Sent Bytes                      0
Sent Packets                    0
Port alive Duration (S)         0
Delta Received Packets          0
Delta Received Bytes            0
Delta Sent Bytes                0
Delta Sent Packets              0
Delta Port alive Duration (S)   0
Connection Point                0
Total Load/Rate                 0
Total Load/Latest               0
Unknown Load/Rate               0
Unknown Load/Latest             0
Latest bytes counter            0
Active Flow Entries             0
Packets Looked Up               0
Packets Matched                 0
Label                           0
dtype: int64
```

Validation

```
Switch ID                       0
Port Number                     0
Received Packets                0
Received Bytes                  0
Sent Bytes                      0
Sent Packets                    0
Port alive Duration (S)         0
Delta Received Packets          0
Delta Received Bytes            0
Delta Sent Bytes                0
Delta Sent Packets              0
Delta Port alive Duration (S)   0
Connection Point                0
Total Load/Rate                 0
Total Load/Latest               0
Unknown Load/Rate               0
Unknown Load/Latest             0
Latest bytes counter            0
Active Flow Entries             0
Packets Looked Up               0
Packets Matched                 0
Label                           0
dtype: int64
```

Dataset's overview (EDA)

- Number of possible outliers

Train

Active Flow Entries	152
Connection Point	0
Delta Port alive Duration (S)	441
Delta Received Bytes	648
Delta Received Packets	649
Delta Sent Bytes	604
Delta Sent Packets	604
Label	0
Latest bytes counter	396
Packets Looked Up	536
Packets Matched	536
Port Number	0
Port alive Duration (S)	0
Received Bytes	186
Received Packets	312
Sent Bytes	222
Sent Packets	379
Switch ID	0
Total Load/Latest	501
Total Load/Rate	396
Unknown Load/Latest	501
Unknown Load/Rate	396

dtype: int64

Test

Active Flow Entries	31
Connection Point	0
Delta Port alive Duration (S)	89
Delta Received Bytes	124
Delta Received Packets	130
Delta Sent Bytes	127
Delta Sent Packets	127
Label	0
Latest bytes counter	76
Packets Looked Up	108
Packets Matched	108
Port Number	0
Port alive Duration (S)	0
Received Bytes	38
Received Packets	69
Sent Bytes	54
Sent Packets	78
Switch ID	0
Total Load/Latest	99
Total Load/Rate	76
Unknown Load/Latest	99
Unknown Load/Rate	76

dtype: int64

Validation

Active Flow Entries	30
Connection Point	0
Delta Port alive Duration (S)	86
Delta Received Bytes	105
Delta Received Packets	112
Delta Sent Bytes	119
Delta Sent Packets	120
Label	0
Latest bytes counter	60
Packets Looked Up	118
Packets Matched	119
Port Number	0
Port alive Duration (S)	0
Received Bytes	32
Received Packets	74
Sent Bytes	54
Sent Packets	75
Switch ID	0
Total Load/Latest	90
Total Load/Rate	60
Unknown Load/Latest	90
Unknown Load/Rate	60

dtype: int64

Dataset’s overview (EDA)

- Basic statistical analysis for every feature (mean, std, min, max)

Test

	Received Packets	Received Bytes	Sent Bytes	Sent Packets	Port alive Duration (S)	Delta Received Packets	Delta Received Bytes	Delta Sent Bytes	Delta Sent Packets	Delta Port alive Duration (S)	Connection Point	Total Load/Rate
count	561.000000	5.610000e+02	5.610000e+02	561.000000	561.000000	561.000000	5.610000e+02	5.610000e+02	561.000000	561.000000	561.00000	5.610000e+02
mean	23141.636364	2.726677e+07	2.438093e+07	30155.998217	910.654189	149.048128	4.546660e+05	3.171120e+05	169.782531	4.841355	2.44385	1.992058e+04
std	69278.589077	3.627827e+07	3.439084e+07	82370.302150	981.303212	861.268543	1.381151e+06	1.207563e+06	1030.484551	0.365671	1.20006	1.105112e+05
min	10.000000	8.560000e+02	6.854000e+03	49.000000	26.000000	0.000000	0.000000e+00	2.780000e+02	2.000000	4.000000	1.00000	-4.042080e+05
25%	353.000000	1.042010e+05	4.480100e+04	322.000000	136.000000	2.000000	2.780000e+02	2.800000e+02	2.000000	5.000000	1.00000	0.000000e+00
50%	1376.000000	1.267023e+07	1.262299e+07	1215.000000	256.000000	4.000000	5.560000e+02	5.560000e+02	4.000000	5.000000	2.00000	0.000000e+00
75%	3562.000000	3.809469e+07	3.176783e+07	4054.000000	1742.000000	19.000000	1.823000e+03	7.590000e+02	5.000000	5.000000	3.00000	0.000000e+00
max	352572.000000	2.589394e+08	2.138743e+08	420932.000000	3307.000000	11130.000000	6.323770e+06	6.647966e+06	13840.000000	5.000000	5.00000	1.194922e+06

Train

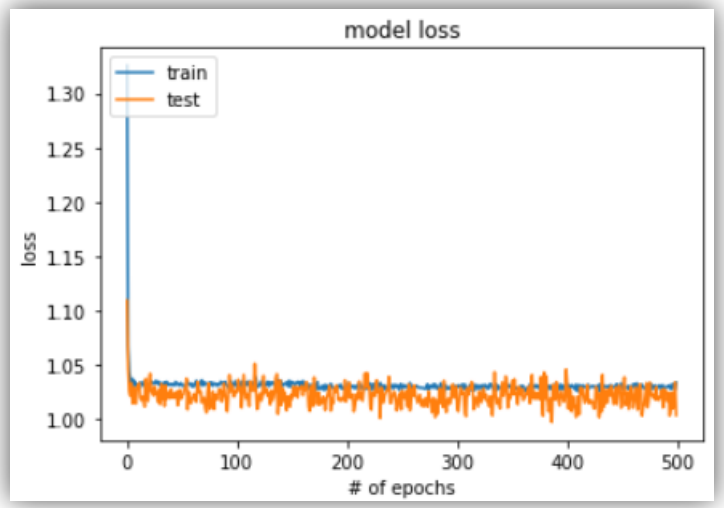
	Received Packets	Received Bytes	Sent Bytes	Sent Packets	Port alive Duration (S)	Delta Received Packets	Delta Received Bytes	Delta Sent Bytes	Delta Sent Packets	Delta Port alive Duration (S)	Connection Point	Tota Load/Rat
count	2619.000000	2.619000e+03	2.619000e+03	2619.000000	2619.000000	2619.000000	2.619000e+03	2.619000e+03	2619.000000	2619.000000	2619.000000	2.619000e+0
mean	19929.095074	2.606639e+07	2.372645e+07	33676.567774	890.539137	174.948835	3.254661e+05	2.957171e+05	160.607866	4.831615	2.450935	2.044276e+0
std	61817.860816	3.670801e+07	3.309908e+07	90565.150583	970.916789	1028.907797	1.166624e+06	1.107859e+06	972.484039	0.374279	1.159286	1.141013e+0
min	10.000000	8.560000e+02	6.025000e+03	44.000000	26.000000	0.000000	0.000000e+00	2.780000e+02	2.000000	4.000000	1.000000	-6.446240e+0
25%	321.500000	8.464900e+04	5.360900e+04	333.500000	136.000000	2.000000	2.780000e+02	2.800000e+02	2.000000	5.000000	1.000000	0.000000e+0
50%	1108.000000	1.262083e+07	1.262176e+07	1243.000000	254.000000	4.000000	5.560000e+02	5.560000e+02	4.000000	5.000000	2.000000	0.000000e+0
75%	3268.500000	3.736785e+07	3.170219e+07	3727.000000	1721.000000	6.000000	8.310000e+02	7.590000e+02	5.000000	5.000000	3.000000	0.000000e+0
max	352584.000000	2.589422e+08	2.130728e+08	420806.000000	3317.000000	15588.000000	6.171714e+06	6.302910e+06	15593.000000	5.000000	5.000000	1.260657e+0

Validation

	Received Packets	Received Bytes	Sent Bytes	Sent Packets	Port alive Duration (S)	Delta Received Packets	Delta Received Bytes	Delta Sent Bytes	Delta Sent Packets	Delta Port alive Duration (S)	Connection Point	Total Load/Rate
count	561.000000	5.610000e+02	5.610000e+02	561.000000	561.000000	561.000000	5.610000e+02	5.610000e+02	561.000000	561.000000	561.000000	5.610000e+02
mean	20265.901961	2.651052e+07	2.462073e+07	26309.124777	949.360071	238.839572	3.268137e+05	2.927781e+05	162.711230	4.846702	2.422460	1.349392e+04
std	59032.911054	3.505715e+07	3.420225e+07	75202.041514	977.609275	1149.069110	1.181636e+06	1.106077e+06	909.905479	0.360596	1.150215	9.124726e+04
min	10.000000	8.560000e+02	7.202000e+03	50.000000	36.000000	0.000000	0.000000e+00	2.780000e+02	2.000000	4.000000	1.000000	0.000000e+00
25%	379.000000	1.962210e+06	4.856800e+04	377.000000	141.000000	2.000000	2.780000e+02	2.800000e+02	2.000000	5.000000	1.000000	0.000000e+00
50%	1385.000000	1.266395e+07	1.263028e+07	1239.000000	317.000000	4.000000	5.560000e+02	5.560000e+02	4.000000	5.000000	2.000000	0.000000e+00
75%	3642.000000	3.785762e+07	3.176313e+07	3889.000000	1750.000000	5.000000	6.260000e+02	7.590000e+02	5.000000	5.000000	3.000000	0.000000e+00
max	350280.000000	2.652568e+08	1.837435e+08	419567.000000	3287.000000	11273.000000	6.249706e+06	6.302708e+06	11273.000000	5.000000	5.000000	1.260664e+06

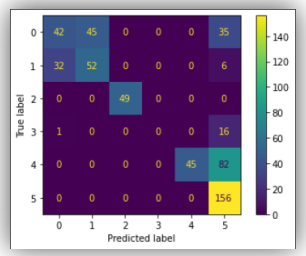
Q1) Baseline Performance

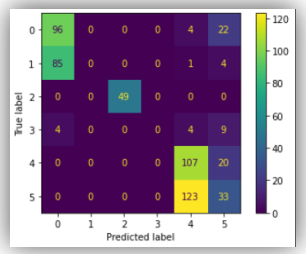
Training and testing losses vs. the number of epochs

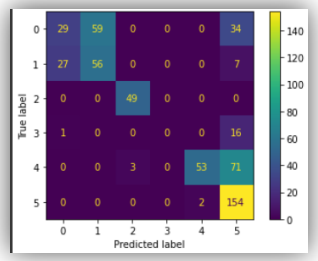


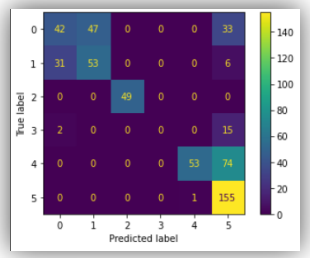
- Confusion matrix for 5 runs

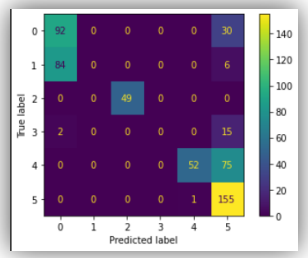
Max training acc	Max test acc	Min training acc	Min test acc	Avg training acc	Avg test acc
62.12%	67.37%	50.47%	48.84%	59.59%	60.28%



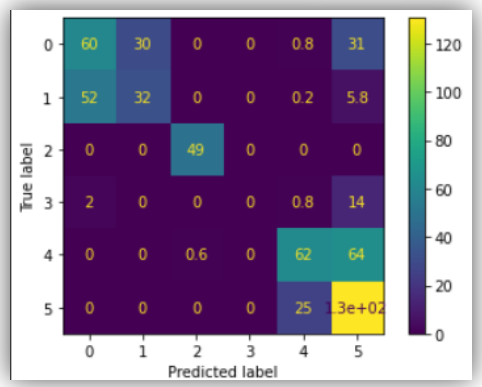




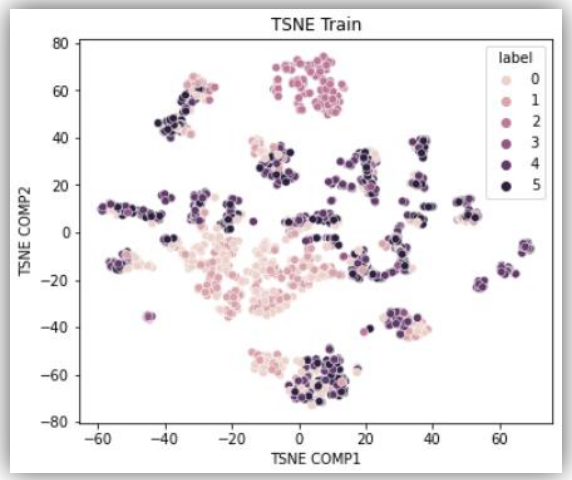


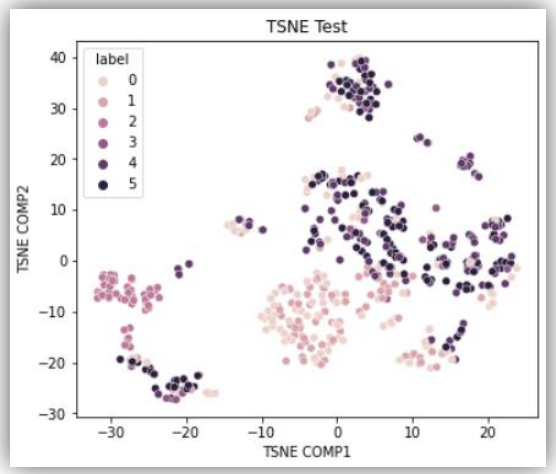


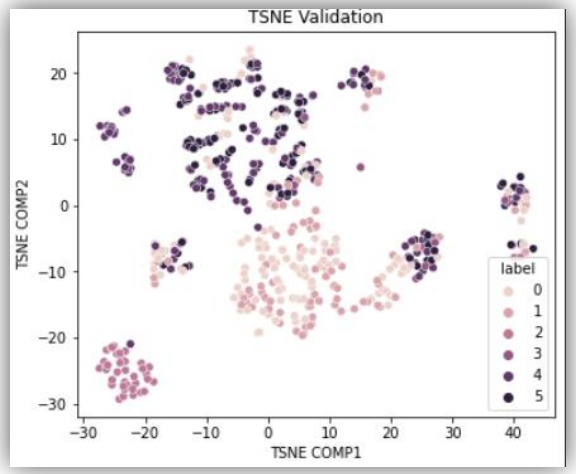
Average of the 5 runs



- TSNE Plots

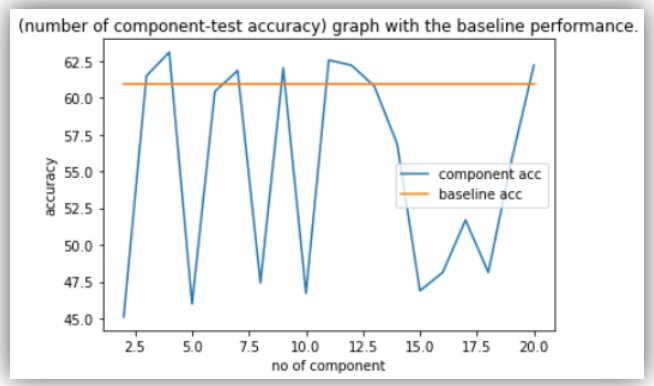

TSNE Train
label
0
1
2
3
4
5
TSNE COMP2
TSNE COMP1

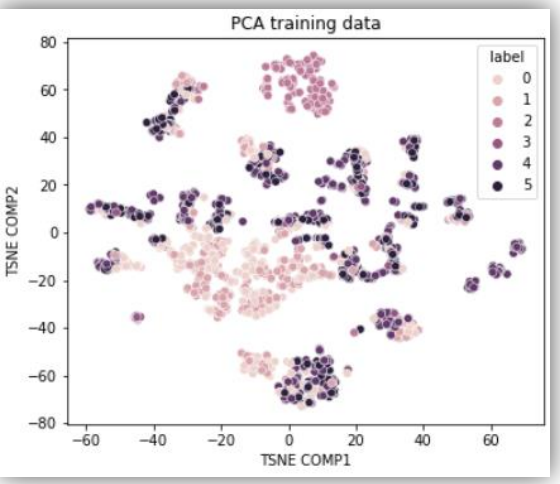

TSNE Test
label
0
1
2
3
4
5
TSNE COMP2
TSNE COMP1

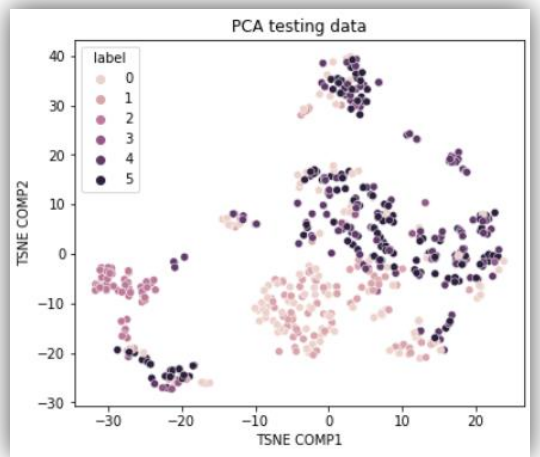

TSNE Validation
label
0
1
2
3
4
5
TSNE COMP2
TSNE COMP1

Q2) Compare dimensionality reduction to feature selection

Q2.1) Dimensionality reduction







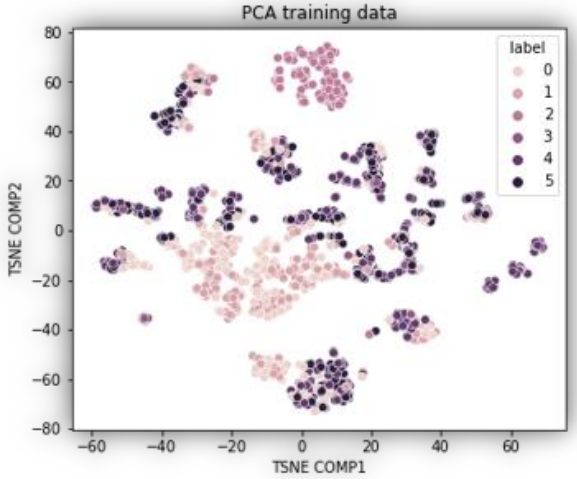
- **Q2.2) Feature selection**
  - **ANOVA**

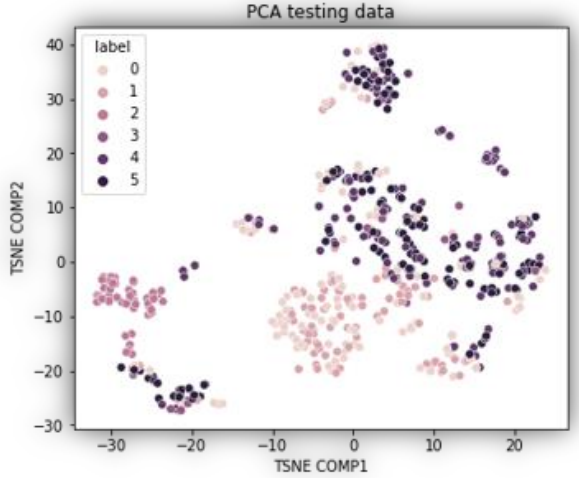
```
max ANOVA 0.6737967729568481
Best value of n components:  14 from ANOVA filter method
```

  - **Mutual Information**

```
max mutal 0.6595365405082703
Best value of n components:  20 from Mutual information for a discrete target filter method
```

  - **provide 2D TSNE plots for ANOVA testing data and training data**





Update the dataset based on highest validation accuracy

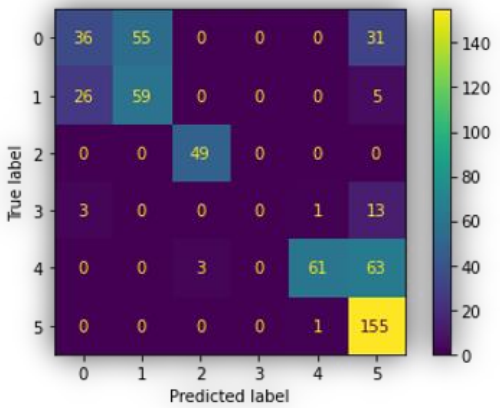
PCA Validation acc	Anova validation acc	Mutual validation acc
65.41%	**66.66%**	**65.77%**

- Update the dataset

	1	2	3	4	5	6	7	10	12	14	16	18	19	20
0	-1.179390	-0.318563	-0.709785	-0.335537	-0.366763	-0.787591	-0.170066	-0.163127	-1.251816	-0.280256	-0.280256	-0.109868	-0.416408	-0.416164
1	-0.181386	-0.319889	-0.709677	-0.715896	-0.367768	-0.741234	-0.166178	-0.120959	-0.389051	-0.280256	-0.280256	-0.108660	-0.415565	-0.415482
2	0.816618	-0.291219	0.321926	1.333552	4.203925	0.418716	-0.170066	-0.161070	-0.389051	-0.280256	-0.280256	-0.105037	3.759201	3.759206
3	1.814622	-0.307835	-0.194000	-0.176054	0.711352	-0.736083	-0.167150	4.979328	0.473713	-0.280256	-0.280256	-0.105037	0.409165	0.409257
4	-1.179390	-0.315602	-0.709446	-0.138653	-0.354670	0.577359	-0.170066	-0.161070	-1.251816	-0.280256	-0.280256	-0.109868	-0.400727	-0.400698
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2614	-1.179390	-0.311994	-0.194503	-0.716178	-0.369126	-0.761837	-0.168122	-0.163127	1.336478	-0.280256	-0.280256	-0.107452	-0.413259	-0.413150
2615	-0.181386	-0.271496	-0.015737	0.243498	-0.336524	1.478742	-0.166178	-0.161070	0.473713	-0.280256	-0.280256	-0.109868	-0.357344	-0.357375
2616	0.816618	-0.316200	-0.537908	-0.716567	-0.370849	-0.813345	-0.143820	-0.161070	-0.389051	-0.280256	-0.280256	-0.105037	-0.408497	-0.408406
2617	-0.181386	-0.309389	-0.365867	0.046601	-0.358248	-0.540355	-0.124378	-0.057192	-0.389051	3.229734	3.229734	-0.107452	-0.406804	-0.406772
2618	-0.181386	-0.313628	-0.537534	-0.335429	-0.366431	-0.751536	-0.168122	-0.163127	1.336478	-0.280256	-0.280256	-0.107452	-0.413170	-0.413070

2619 rows × 14 columns

- provide the confusion matrix



Q3) Vary the MLP parameters [1/5]

Q3.1) Batch size

Batch size= 32						Batch size= 64						Batch size= 128					
Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	**Avg test acc**
71.21	**72.9**	**23.17**	**69.87**	**68.76**	**71.08**	**72.47**	**72.01**	**17.64**	**70.58**	**69.56**	**71.47**	**71.7**	**74.33**	**27.97**	**71.56**	**68.01**	**72.72**

- Maximum average testing accuracy: **`72.72% from batch size 128`**
- Maximum average training accuracy: **`69.56% from batch size 64`**
- The combination that achieves the highest average test accuracy is **batch size 128**

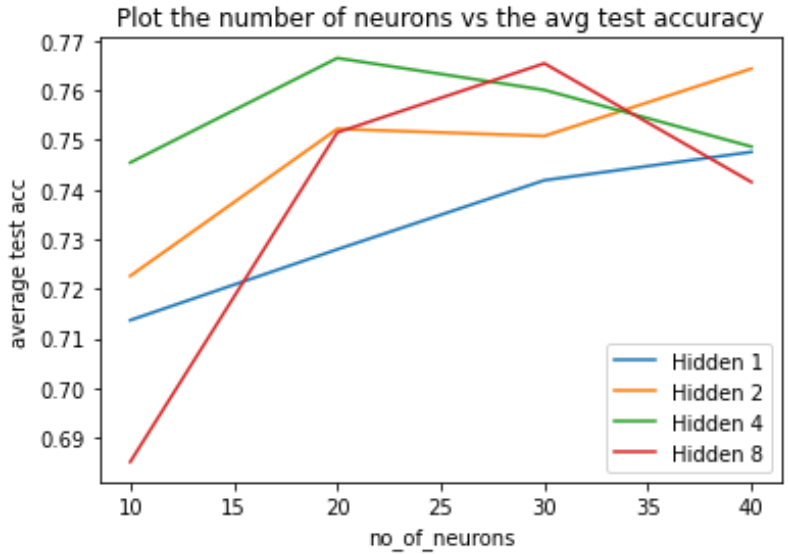
Q3) Vary the MLP parameters [2/5]

Q3.2) Hidden layers vs. neurons/layer

	1 hidden layer						2 hidden layer						4 hidden layer						8 hidden layer					
# of neurons for each layer	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc
10	**72.2**	**73.8**	**27.18**	**69.87**	**69.14**	**71.44**	**71.55**	**73.61**	**21.95**	**71.65**	**67.69**	**72.26**	**73.5**	**76.47**	**27.94**	**72.9**	**69.04**	**74.54**	**66.97**	**73.44**	**25.23**	**64.52**	**63.72**	**68.52**
20	**73.53**	**74.15**	**24.62**	**71.65**	**70.79**	**72.79**	**75.06**	**75.75**	**20.23**	**74.15**	**72.16**	**75.22**	**77.2**	**77.89**	**32.95**	**75.4**	**74.3**	**76.64**	**78.57**	**77.36**	**25.27**	**72.54**	**93.79**	**72.54**
30	**75.75**	**75.22**	**35.85**	**71.65**	**73.29**	**74.18**	**76.97**	**77**	**28.94**	**73.79**	**74.74**	**75.8**	**78.46**	**78.78**	**33.44**	**70.23**	**75.36**	**76**	**78.23**	**79.67**	**25.08**	**72.72**	**74.64**	**76.54**
40	**75.71**	**75.75**	**32.45**	**73.08**	**73.26**	**74.75**	**78.12**	**77.71**	**41.77**	**74.68**	**75.65**	**76.43**	**78.84**	**77.36**	**37.91**	**72.54**	**76.14**	**74.86**	**80.22**	**77.71**	**33.52**	**70.4**	**76.74**	**74.15**

- Highest validation accuracy from Q3.1 is at batch size **64**

Plot the number of neurons vs the avg test accuracy.



Q3) Vary the MLP parameters [3/5]

Q3.3) Learning rate and different optimizers

	Learning rate=0.1						Learning rate=0.01						Learning rate=0.001					
optimizer	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc
AdamW	67.04	69.16	39.21	60.6	55.7	65.27	84.34	79.67	53.57	74.86	80.5	**78.11**	79.3	78.25	34.51	70.76	76.84	74.65
SGD (momentum =0.1)	95.34	78.25	46.08	70.94	87.93	75.57	84.72	77.36	25.62	70.4	78.39	74.29	74.6	73.61	25.65	71.47	64.49	72.4
SGD (momentum =0.5)	97.13	79.32	51.35	72.37	90	75.61	89.61	79.32	21	69.87	82.9	74.11	77.54	75.57	18.21	72.72	70.25	74.33
SGD (momentum =0.9)	79.45	78.96	23.29	78.96	56.47	78.96	94.76	78.25	37.68	73.08	87.55	75.86	84.84	79.67	15.84	71.12	78.04	74.83
Rprop	79.45	78.96	23.29	78.96	56.47	78.96	94.27	77.36	45.32	69.69	88.3	73.54	94.95	79.85	42.38	72.72	89.19	76

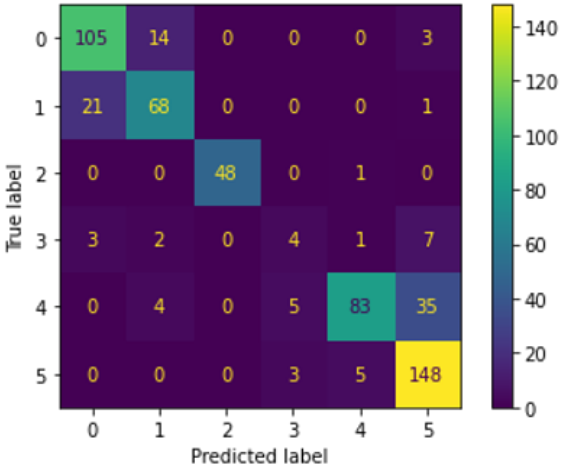
The highest validation accuracy from Q3.2 is at 4 hidden layers and 40 neurons.

Q3) Vary the MLP parameters [4/5]

Q3.4) Activation functions

Relu						Leaky Relu						Sigmoid						Tanh					
Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc	Max train acc	Max test acc	Min train acc	Min test acc	Avg train acc	Avg test acc
83.27	81.63	59.37	75.04	80.56	77.57	80.48	80.21	60.63	73.79	77.38	77.21	77.47	78.96	30.39	69.87	74.43	74.4	86.36	80.3	62.04	74.33	81.99	77.86

-the best-obtained combination from Q3.3
AdamW optimizer with learning rate 0.01



Conclusion

- EDA was performed to get insights from the data set and describe it.
- Baseline performance was applied on MLP model with testing acc=61%
- We updated the data set based on ANOVA which was the best of our 3 feature engineering methods.
- Build MLP with the best hyperparameters:
  - batch size=64
  - Hidden layers= 4, number of neurons=40
  - Learning rate= 0.01, optimizers= AdamW
  - Activation function=tanh