Reunion 26/09/2024

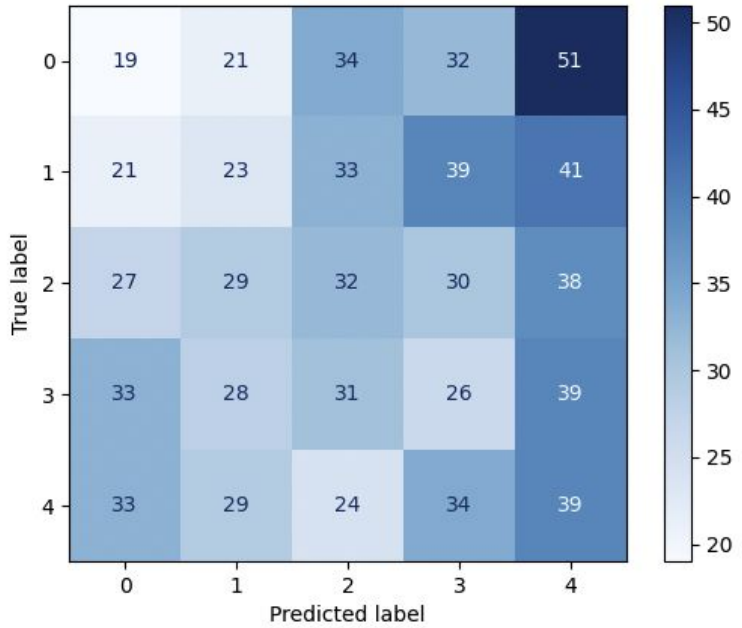
To do list from last time

- Change IQR boundaries for eliminating outliers, use:
  - lower boundary = 1s
  - upper boundary = 6s
- Add features to the simple nn see if results change
- Use Transformers:
  - Look into:
    - Camembert
    - Roberta
    - Flaubert
    - FastText
    - Treegram

Test 1

```
Bin edges for 'pauseDur':
Bin 0: 1.5099999904632568 to 1.7999999523162842
Bin 1: 1.7999999523162842 to 2.2200000286102295
Bin 2: 2.2200000286102295 to 2.859999895095825
Bin 3: 2.859999895095825 to 4.139999866485596
Bin 4: 4.139999866485596 to 5.980000019073486
```

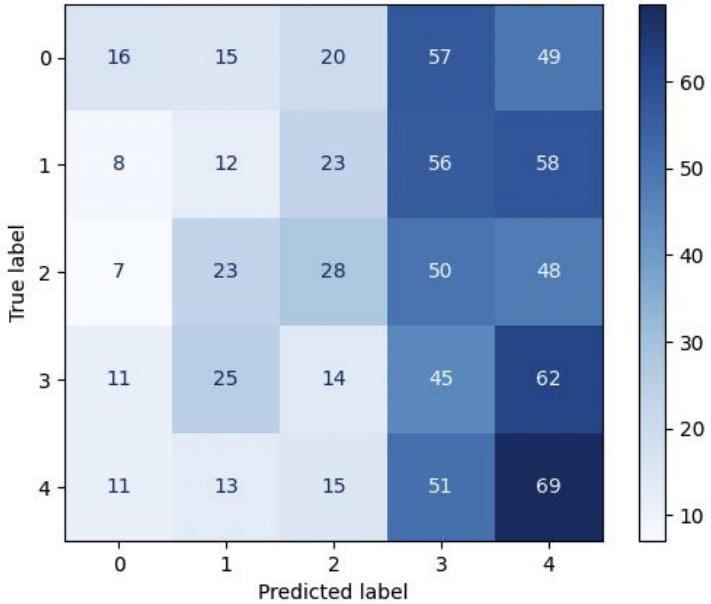
- 1-6s pause interval
- 5 categories classification
- Gradient Boosting classifier
- learning_rate = 0.1
- pauses after burst
- **Accuracy: 0.17684478371501272**



Test 2

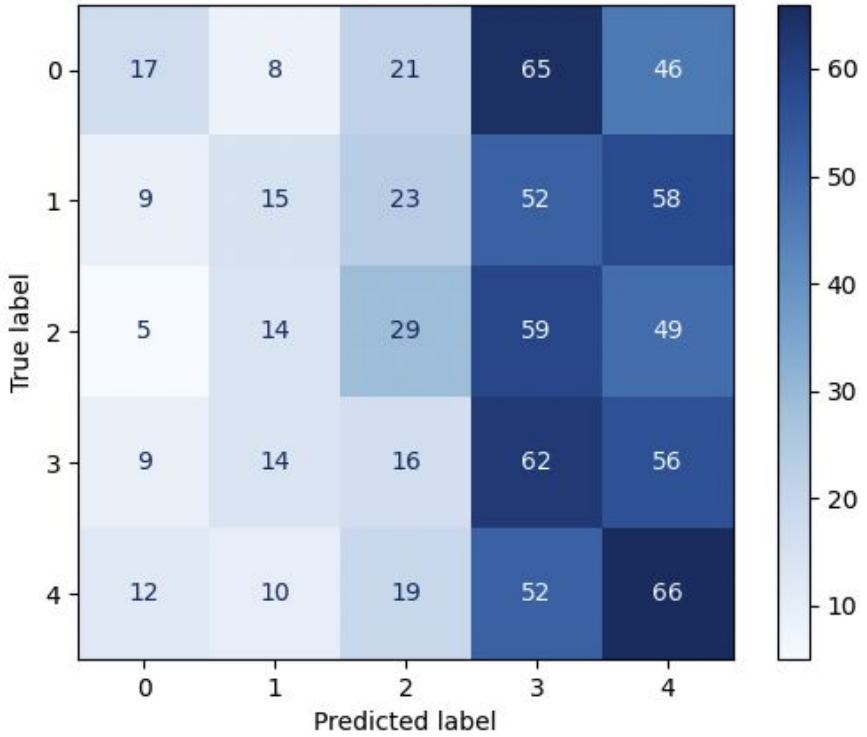
```
Bin edges for 'pauseDur':
Bin 0: 1.5099999904632568 to 1.7999999523162842
Bin 1: 1.7999999523162842 to 2.2200000286102295
Bin 2: 2.2200000286102295 to 2.859999895095825
Bin 3: 2.859999895095825 to 4.139999866485596
Bin 4: 4.139999866485596 to 5.980000019073486
```

- 1-6s pause interval
- 5 categories classification
- Random Forest classifier
- pauses after burst
- **Accuracy: 0.21628498727735368**



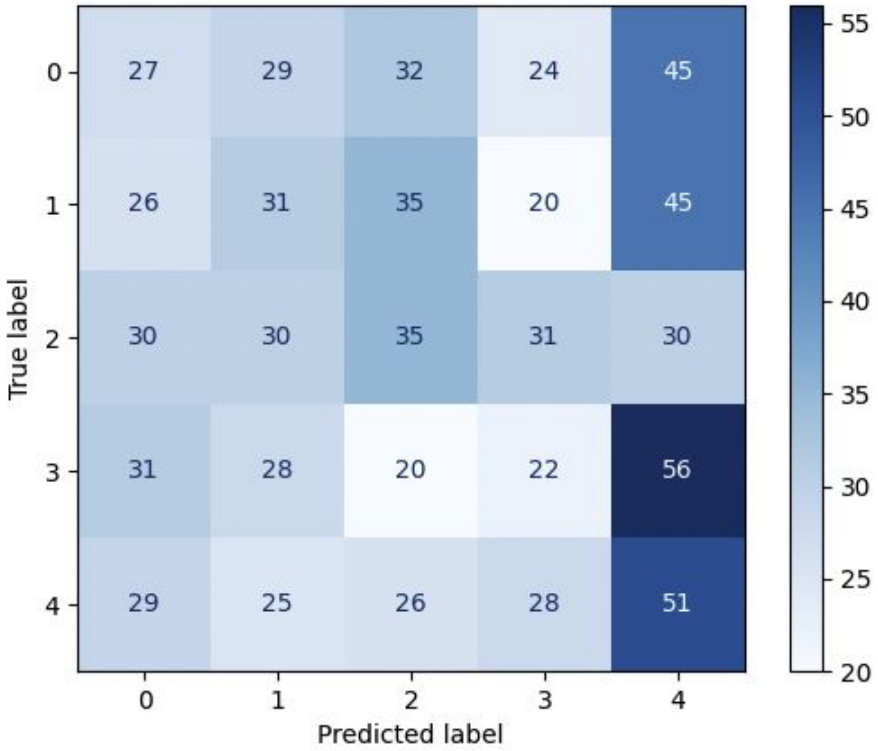
Test 3

- 1-6s pause interval
- 5 categories classification
- Random Forest classifier
- pauses before burst
- **Accuracy: 0.2404580152671755**



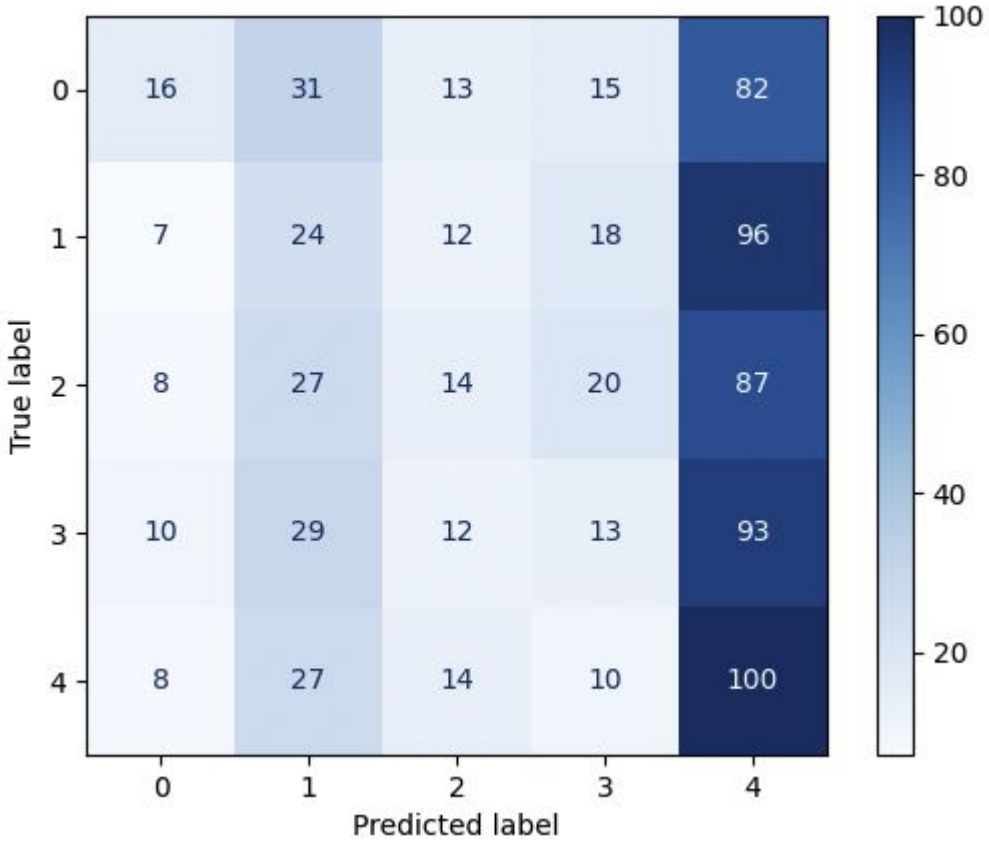
Test 4

- 1-6s pause interval
- 5 categories classification
- Gradient Boost
- pauses before burst
- **Accuracy: 0.21119592875318066**



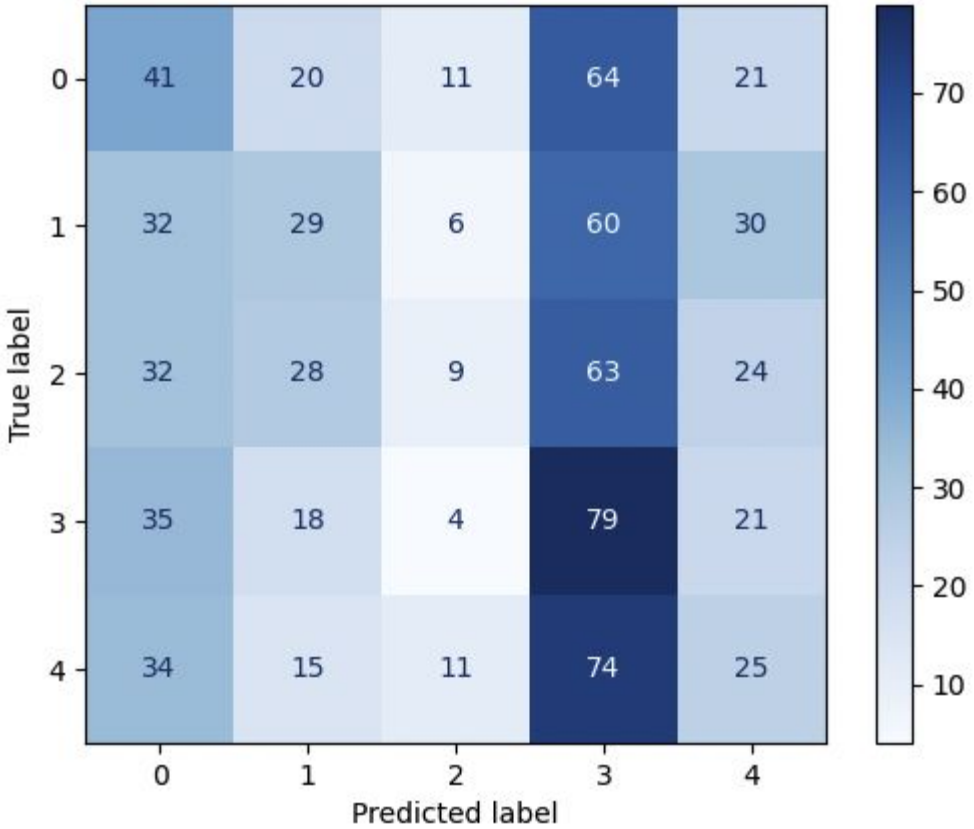
Test 5

- 1-6s pause interval
- 5 categories classification
- Sequential - dropout rate, early stopping, batch_size=32, 100 epochs
- pauses before burst
- **Accuracy: 0.2124681919813156**



Test 6

- 1-6s pause interval
- 5 categories classification
- Sequential - dropout rate, early stopping, batch_size=256, 100 epochs
- pauses before burst
- **stopped after 6 epochs to avoid overfitting**
- **Accuracy: 0.2328244298696518**



Transformers

- Essayé de suivre des tutoriels en ligne pour utiliser BERT, mais ajusté pour CamemBERT.
- Difficultés avec la structure des données - idées?
- Finalement, on a decidé d'utiliser CamembertForTokenClassification, et on a restructuré les données dans ce format:

  cat_2 L'intention de l cat_5 'aéroport de biard cat_5 de cat_4 diminuer cat_1 la poussé des gaz sur le décollage de ses avions cat_1 au dessus cat_5 des zones cat_5 habité cat_3 à ses avantage et ses inconvéniant OUTLIER.

- Problemes:
  - tokenization un peu bizarre - a cause de cat?
  - explicabilité - on ne sait pas ce qu'influence le resultat
  - resultats un peu bizarre - on n'arrive pas a en trouver la cause

cat_2 L'intention de l cat_5 'aéroport de biard cat_5 de cat_4 diminuer cat_1 la poussé des gaz sur le décollage de ses avions cat_1 au dessus cat_5 des zones cat_5 habité cat_3 à ses avantage et ses inconvéniant OUTLIER.

```
Token: _ses, Predicted Label: 0, Actual Label: 0
Token: _avantage, Predicted Label: 0, Actual Label: cat_5
Token: _et, Predicted Label: 0, Actual Label: 0
Token: _ses, Predicted Label: 0, Actual Label: 0
Token: _incon, Predicted Label: 0, Actual Label: 0
Token: vé, Predicted Label: cat_5, Actual Label: cat_5
Token: ni, Predicted Label: cat_5, Actual Label: cat_5
Token: ant, Predicted Label: cat_5, Actual Label: cat_5
```

cat_2 L'intention de l cat_5 'aéroport de biard cat_5 de cat_4 diminuer cat_1 la poussé des gaz sur le décollage de ses avions cat_1 au dessus cat_5 des zones cat_5 habité cat_3 à ses avantage et ses inconvéniant OUTLIER.

```
{'train_runtime': 2.6689, 'train_samples_per_second': 1.124, 'train_steps_per_second': 1.124, 'train_loss': 1.66189988454
100%|                                                                                                | 3/3 [00:02<00:00,  1.12it/s]
Token: _L, Predicted Label: 0, Actual Label: 0
Token: ', Predicted Label: 0, Actual Label: cat_2
Token: intention, Predicted Label: 0, Actual Label: cat_2
Token: _de, Predicted Label: 0, Actual Label: cat_2
Token: _l, Predicted Label: 0, Actual Label: 0
Token: _, Predicted Label: 0, Actual Label: 0
Token: ', Predicted Label: 0, Actual Label: 0
Token: aéroport, Predicted Label: 0, Actual Label: 0
Token: _de, Predicted Label: 0, Actual Label: 0
Token: _bi, Predicted Label: 0, Actual Label: cat_5
Token: ard, Predicted Label: 0, Actual Label: 0
Token: _de, Predicted Label: 0, Actual Label: 0
Token: _diminuer, Predicted Label: 0, Actual Label: 0
Token: _la, Predicted Label: 0, Actual Label: cat_5
Token: _poussé, Predicted Label: 0, Actual Label: 0
Token: _des, Predicted Label: 0, Actual Label: cat_4
Token: _gaz, Predicted Label: 0, Actual Label: 0
```

```
{'train_runtime': 2.6689, 'train_samples_per_second': 1.124, 'train_steps_per_second': 1.124, 'train_loss': 1.66189988454
100%|                                                                                                  | 3/3 [00:02<00:00,  1.12it/s]
Token: _L, Predicted Label: O, Actual Label: O
Token: ', Predicted Label: O, Actual Label: cat_2
Token: intention, Predicted Label: O, Actual Label: cat_2
Token: _de, Predicted Label: O, Actual Label: cat_2
Token: _l, Predicted Label: O, Actual Label: O
Token: _, Predicted Label: O, Actual Label: O
Token: ', Predicted Label: O, Actual Label: O
Token: aéroport, Predicted Label: O, Actual Label: O
Token: _de, Predicted Label: O, Actual Label: O
Token: _bi, Predicted Label: O, Actual Label: cat_5
Token: ard, Predicted Label: O, Actual Label: O
Token: _de, Predicted Label: O, Actual Label: O
Token: _diminuer, Predicted Label: O, Actual Label: O
Token: _la, Predicted Label: O, Actual Label: cat_5
Token: _poussé, Predicted Label: O, Actual Label: O
Token: _des, Predicted Label: O, Actual Label: cat_4
Token: _gaz, Predicted Label: O, Actual Label: O
Token: _sur, Predicted Label: O, Actual Label: O
Token: _le, Predicted Label: O, Actual Label: cat_1
Token: _décollage, Predicted Label: O, Actual Label: O
Token: _de, Predicted Label: O, Actual Label: O
Token: _ses, Predicted Label: O, Actual Label: O
Token: _avions, Predicted Label: O, Actual Label: O
Token: _au, Predicted Label: O, Actual Label: O
Token: _dessus, Predicted Label: O, Actual Label: O
Token: _des, Predicted Label: O, Actual Label: O
Token: _zones, Predicted Label: O, Actual Label: O
Token: _habité, Predicted Label: O, Actual Label: O
Token: _à, Predicted Label: O, Actual Label: O
Token: _ses, Predicted Label: O, Actual Label: O
Token: _avantage, Predicted Label: O, Actual Label: cat_5
Token: _et, Predicted Label: O, Actual Label: O
Token: _ses, Predicted Label: O, Actual Label: O
Token: _incon, Predicted Label: O, Actual Label: O
Token: vé, Predicted Label: cat_5, Actual Label: cat_5
Token: ni, Predicted Label: cat_5, Actual Label: cat_5
Token: ant, Predicted Label: cat_5, Actual Label: cat_5
```