Final Project: Last Milestone

TEAM 3

Ella Jewison:
Model testing, medium article writing, testing advance dimensionality reduction techniques and feature extraction.

Ben Sivan:
Model testing, structuring the project for robust model testing and automatic metric logging.

Samuel Lederman:
Researching advance data augmentation techniques.

Ethan Ben-Attar:

Part 1:

Project name : odor classification

Summary:
Classification of odor's signals that where recorded using the desert locust antenna.
Labels - 8 different odors
Features - 150 time points of electical potential measured from the antenna's neurons submitted to various odors.

The dataset is mostly balanced, so our metric of choice for high level evaluation of the model was accuracy. Since we have 8 labels, base model of random sampler would have accuracy of about 12%. In the original paper published, the researcher used random forest model and achived accuracy of 69.3%.

Our achievements :
We have implemented a script that manages to test various preprocessing steps and different model to classify the odors. After testing multiple strategies for preprocessing and multiple models we also found the random forest gives the best results, but we were able to increase the accuracy to 80%.

Part 2:

The pipeline that we use to analyse and preprocess the data can be seen on our git repository :
https://github.com/EllaJewison/Final_project

the different option that you can use to preprocess the signal are :
- scale
- derivative
- umap
- clipped
- reduce

The different model option that you can use are:
- random forest
- MLP
- RNN
- dense neural net
- adaboost
- KNN
- logistic regression

The configuration that yield the best accuracy score is : clipped, scale, derivative and random forest. The best accuracy is 80 %. This is significantly better than the baseline model that is 12 % (choosing randomly across 8 odors)

Across the 8 odors we obtain the following F1-score :

label	F1-score
Benz	0.857
Hex	0.898
Ethyl	0.909
Rose	0.734
Lem	0.705
Ger	0.721
Cit	0.627
Van	0.930

We can see that some odors are more easily identified than other. The most difficult to predict is beta citronellol and the easiest is Vanilla.

The medium article can be found here: https://medium.com/p/d24525f0f9d8/edit

Part 3: challenges

1) Data augmentation was a challenge since it only yield worse performance on the models.
2) Working together was a challenge as we had difficult time to decide what to do
3) The main difficulty that we encountered was that everything that we tried made our model worse

Part 4:

The next steps would be to try on the mixed odors or find a better preprocessing method.
And spend a bit a time on hyperparameter tunning.