ORL&IA **LittleBigCode**

Presentation of the team

We are a team of 3 data Scientist consultants from LittleBigCode, an AI solutions creator. We contribute to this challenge on our off-mission time.

LittleBigCode
AI Solution Creator

Our approach

The objective we have set is to predict the class of overall survival (short, medium or long survival) of a patient with ORL cancer due to human papillomavirus (HPV).
To deal with this multiclass classification problem, two approaches were tested :

- A machine learning approach considering the quantity of antibodies
- An approach using Deep Learning on the images to get further insights.

N.B: We have also submitted predictions on a Naïve DecisionTree predicting OS from Alcool, Tabacco and OMS score.

Useful links

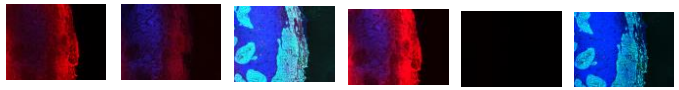
- Medium : Medium Description des données du challenge
- Github: Github folder of Epidemium Season 3

Data

Data

For predicting the overall survival of patients, we have **clinical data** about them (Age, OMS Score, information of consumption of tobacco, alcohol) and **image data** (immune labeling revealing the antibodies in red color)

Images on the right represent markers for a patient for a specific zone. There may be between 1 and 6 different stamps for a patient.



Feature Engineering

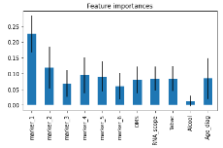
- Because of the extremely limited amount of data, we **simplify the problem of predicting the overall survival**. We defined 3 classes which can be interpreted as *short survival* (between 0 and 2 years), *medium survival* (between 2 and 6 years) and *long survival* (more than 6 years)
- For clinical data that are tabular data, most variables were **categorical or binary variables,** so no processing has been done on them. We simplify the age of diagnosis only considering the decade.
- For image data :
  - For machine learning approach, we calculate per image the **percentage of red pixels**.
  - For Deep Learning approach, we use **a CNN using only the red channel** of each marker image**.** A **BagNet** model is created **for each marker individually using the full RGB image.**
  - Creating combination of markers : For the same patient and the same marker, there can be between 1 and 6 stamps representing different areas. We do a preprocessing which consists of **considering a single image per marker and creating all possible combinations of marker images**. This Data Augmentation technique has the advantage of creating more samples **as if they were different patients** and the models will be to generalize its understandings.

Modeling and features importance

We are aware of the need for medical teams to have information on the variables that most influence prediction. Therefore, we were interested in **interpretable models.** Therefore, we have decided to not have the best performance on the leaderboard but to analyze statistics and correlations outlined by our models. Nevertheless, to appear on the leaderboard, we have submitted predictions made by a default DecisionTree trained on OMS, Alcool and Tabacco data reaching almost 7 MAPE.
To analyze the results, we split our training dataset in two parts : **70% dedicated to model training** and 30% to model validation (this represents 12 patients).

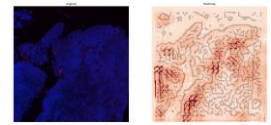
Machine Learning with Random Forest and XGBoost Models

The advantage of these models is that both clinical data and images can be easily used. Hyperparameter optimization has been done to optimize performance even though it is not very stable given the small amount of data.



Deep Learning with CNN and BagNet models

We have decided to use these models knowing that we could use their weights associated with input images to understand their attention mechanism.
For the BagNet models we did one model for each marker using the full RGB image.
We researched the best models for each marker using a random and Bayesian optimization. Some marker gave better results than others. To simplify our findings, based on the very limited research we had time to do, we have the following results :
Marker 2 > Markers 1, 3, 4 > Markers 5, 6.



Heatmap of images of incorrect predictions

Our best model could predict the correct category on 8 out of the 12 validation patients : 66.7% accuracy.

Conclusion

During this challenge, we have faced two main issues :

- The first one was obviously **the limited amount of data**. There are questions about the **representativity of these 47 samples compared to the real data**, i.e., are these patient good representatives of the overall patients of this disease? Furthermore, it is very hard and scientifically not valide to conclude anything with such a short sample of examples. Even so, our Machine Learning models have found a causality of the quantity of red pixels over the total number of pixels to predict the overall survival score. It might suggest that the quantity of bio markers on images might have a correlation on the survival of a patient.

- The other challenge was that the competition was an open issue, and we were not sure that we could find anything useful for users. So, we have not tried to achieve the best prediction but rather to interpret attention heatmaps of Deep Learning. As you can observe on the part Modelling of the poster, we have visualized some interesting values. The model seems to get rid of artifacts induced by the shape of cells and tend to focus on part of images which are light blue. In association with the proportion of red marker, this might be a lead for further research.