Age prediction from EEG signals

Chaïmaa Kadaoui **Pierre-Alain Langlois**

Abstract

The goal of this challenge is to predict people's ages using variables extracted from their brain activity during sleeping time. We used two different algorithms to deal with the data: A Convolutional Neural Network for the electroencephalogram and a Random Forest for the hypnogram. Furthermore, a pre-processing was applied in each case. Afterwards, a Linear Regression was applied to both predictions to combine the approaches. The error used to evaluate our performance was the mean average percentage error (MAPE). We had a final error of: 20.18% and were ranked 11^{th}.

1 Introduction

This challenge was provided by the company Rythm. Rythm is a neurotechnology company that uses neuroscience fundamentals to enhance human performance by researching, monitoring, and understanding the brain.

The brain activity has been widely studied by doctors in order to better understand human sleep. Besides, while aging, the brain activity tends to change. Many studies ([1], [2], [3], [4]) have looked into this issue using a medical approach. Hence, we would like to use mathematical tools to see if we can accurately predict the age of one subject given its brain activity.

We were given two datasets: one for the training part and another one for the testing part. We had 581 subjects in the training set and 249 in the testing set. Each dataset had the following columns:

- ID: A unique ID for each subject
- DEVICE: Indicates which device was used to record the brain activity (0 or 1)
- EEG: 75000 columns corresponding to 5 minutes of the electroencephalogram signal during the deep sleep period
- Hypnogram: A list describing the evolution of sleep stages during the night for each subject

2 The electroencephalogram (EEG)

2.1 First contact with the data

We are given an EEG taken from the subject's deep sleep period which has been recorded for 5 minutes at 250Hz. This kind of data is particularly hard to understand at first sight because it does

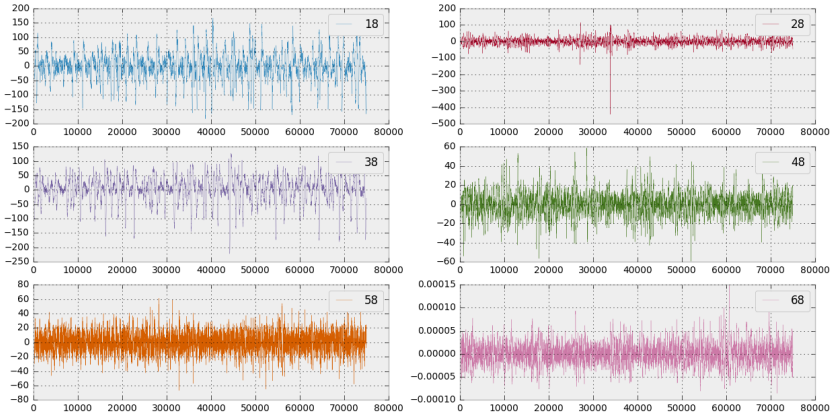

Figure 1: Representation of the raw EEG for a few subjects

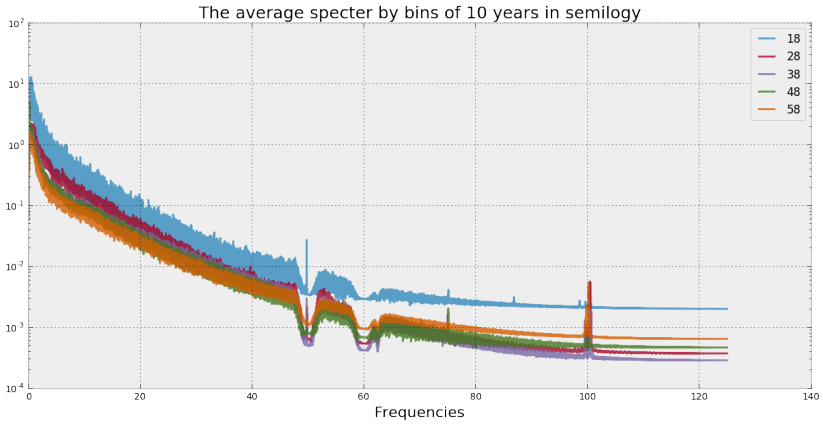

Figure 2: Representation of the fft for groups of people that span every 10 years

not show any general trend, it is very different from a subject to another, and it is also very noisy due to the difficulties of acquisition. We can observe these data for a few subjects in Figure 1.

At first, we wanted to try the most simple models in order to have a hint about how much the data tell us about the subjects' age. We wanted to perform dimensionality reduction because we did not know how to deal with 75000 dimensions. However, methods such as Principal Component Analysis are not immediately possible because we have less data than the number of dimensions. As a consequence, we decided to try a Partial Least Squares regression which is a model in which we try to find a relation between the label Y and the data X by projecting both of them in spaces in which the covariances between them is maximized. More precisely, we look for :

$$X = TP^T + E$$
$$Y = UQ^T + F$$

T and U are the projections of X and Y ; P and Q are the projections matrices and E and F are error terms.
We fitted this model thanks to the implementation available in scikit learn and we observed that unfortunately, the model always return the most represented value of the dataset Y which is 41 years old. As a matter of fact, this value gives an error around 31%. With this results raises the main difficulty of the problem : the trivial average predictor gives a not so bad error and therefore, a lot of algorithms we will try will tend to converge to this solution if we are not careful.

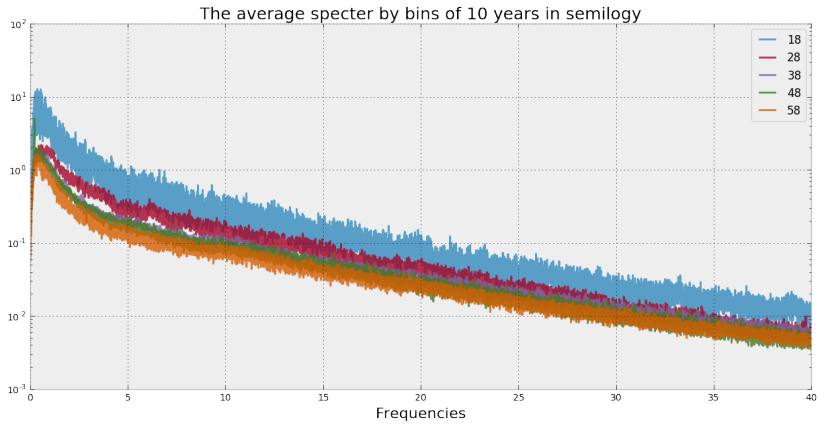

Figure 3: Representation of the fft for groups of people that span every 10 years - 0-40Hz

2.2 Fourrier approach

As seen in the previous subsection, a trivial solution would not solve the problem. That's why we decided to go deeper in the analysis of the model. The first thing that came to mind was that the temporal domain shows only what looks like an indescribable noise. As a consequence, we decided to observe the data in frequency thanks to a Fast Fourrier Transform. In order to have an insight about the relation between the frequency representation and the age of the subjects, we decided to average FFT for groups of people that spans 10 years old (18-28, 28-38, 38-48, 48-58 and 58-68). The result of that can be observed on Figure 2. We need to be very careful with this figure. At first sight, it seems that the very high frequencies (80-120Hz) differ a lot when the age changes. However, academic papers [3] and [5] tells us that these bands of frequency are not relevant in EEG. What's more, the standard deviation which is not represented in this figure overlaps a lot between the groups of people. However, the aforesaid papers focus on bands of frequencies that are located around 0.5-40 Hz and therefore, we decided to focus on this band of frequency. As a matter of fact, we can observe these frequencies more precisely in Figure 3. We see that some interesting phenomena are observed around 5Hz. It seems that the frequency representation increases with the age in this area. This is the main observation that led to our solution.

2.3 Neural Network

In order to take advantage of this last observation, we tried to find in the bibliography approaches that had already been tested for this kind of data. We found in [5] that an Artificial Neural Network had already been used in order to determine sleep stages out of EEG. Even if this is not the problem we are trying to tackle, we found the connection with our data interesting and we decided to try a similar approach in our work.
As specified in [5], we decided to create a fully connected artificial neural network with only a few intermediate layers (2 or 3 in practice). In the same reference, it is also specified to preprocess the data with a Butterworth band-pass filter in order to select only the relevant frequency bands.
Despite our effort, training this kind of network with the usual method (stochastic gradient descent with dropout), we kept on ending with the trivial result. As a consequence, we understood that our analysis needed to go one step further.
Here is the assumption we made : since the sleep is commonly divided in stages ; it might be reasonable that what makes these stages unique is there stationarity. Since we know that all the EEG we have are taken during one single sleep stage (the deep sleep), we make the hypothesis that each observed signal is stationary.
With this hypothesis, we can easily make data augmentation : for a given individual, we can divide the 5 minutes observation into several parts that we consider to be new independent observations. The only connection between these new observations is that they represent the same age. What's more, we saw in [5] that the neural network are fed with 5 seconds EEG samples. Therefore, we had a clue. The final modification we made in order to have things working was to feed the neural network with

only the frequencies wanted instead of feeding it with the bandpass filtered FFT. Finally, we obtained something interesting with the following parameters :

- Lowest frequency : 5Hz
- Highest frequency : 20Hz
- Order of the Butterworth filter : 2
- Dimension of the output : 90 (ability to predict age between 0 and 89 years)
- Size of the hidden layers : 200
- Number of layers : 3
- Size of an observation : 5000 (which means 20s instead of 5 minutes)
- Batch size : 4000
- Non linearity : sigmoid
- Learning rate : 10^{-5}
- Number of iterations : 100000
- Dropout probability : 0.5

Our ages are represented by the one-hot representation (which means that we try to find a probability distribution over the possible ages). Since evaluating the problem's loss for the Rythm challenge would require a argmax on this distribution, it is not possible to use this loss as a training loss (it is non-differentiable). As a consequence, we call accuracy the loss used by Rythm challenge, and in order to train the model, we use the classical softmax cross entropy with logits, which is very often used for this kind of problem and will be referred to as loss. We implemented the neural network in Google's Tensorflow framework and we monitored the training with Tensorboard. The accuracy and loss are represented for both the training set in Figure 4 and the validation set in Figure 5. We see that the evaluation accuracy dropped under 25% ; however, the evaluation loss started to increase which expresses a potential risk of overfitting. This is actually what we observed : after 100k iterations, the performances started to decrease when we submitted the results.
Finally, in order to test the algorithm, we need to perform a pulling of the result. As a matter of fact, we made a data augmentation and as a consequence, we have a few numbers for a given person. Our strategy here was to make a mean pulling, because we want to evaluate the expectation of the age. The method presented here gave a result of 21.95% when submitted.

3 The hypnogram

We can define two different periods during the sleep: Rapid Eye Movement sleep (REM) and Non Rapid Eye Movement sleep (N-REM). The American Academy of Sleep Medicine has divided the NREM in 3 stages. Each one of them is characterized by different properties. During a night, we have multiple cycles of sleep which consist of a succession of these stages.

An hypnogram is a graph which represents the sleep stages (plus the wake periods) during a night's sleep. For this challenge, we were provided with hypnograms for each subject. The hypnograms were stored in strings that represent lists of different lengths. Each element in one of these lists characterizes the sleep stage during a period of 30 seconds. The values are 0 (for Wake), 1 (for N1), 2 (N2), 3 (Deep sleep - N3), 4 (REM), -1 (when scoring was not possible).

For example, for one subject, the hypnogram had this shape: ['0', '0', '0',. . ., '4', '4']. We can see a visualization of an hypnogram in the figure 6. This feature couldn't be used for our prediction because of the following reasons:

- The hypnograms have different lengths.
- The numerical values for each hypnogram are categorical. Thus, we can't feed them directly to an estimator. This is because an order will be induced (4 > 1 for example) which is misleading.
- There are missing values (-1) in the hypnogram.

Hence, a pre-processing was necessary to extract meaningful informations from this feature.

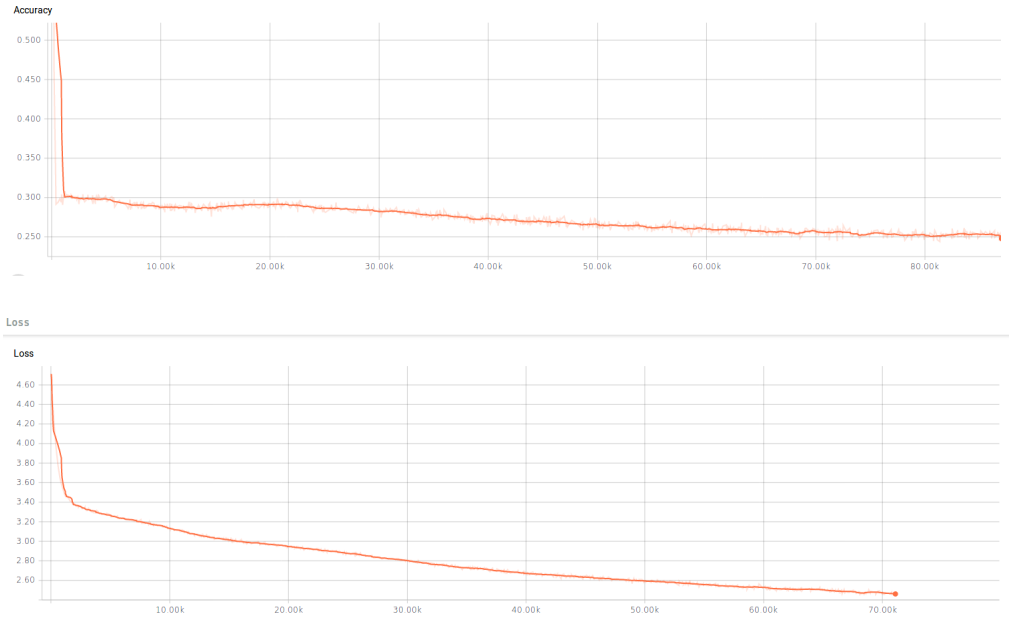

Figure 4: Accuracy and Loss along the iterations for the training set

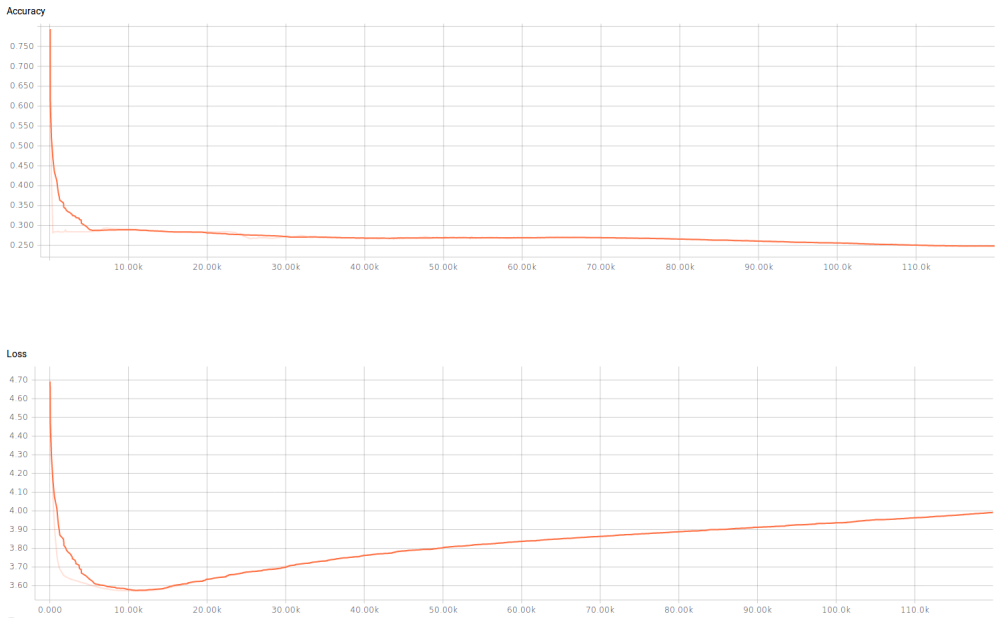

Figure 5: Accuracy and Loss along the iterations for the validation set

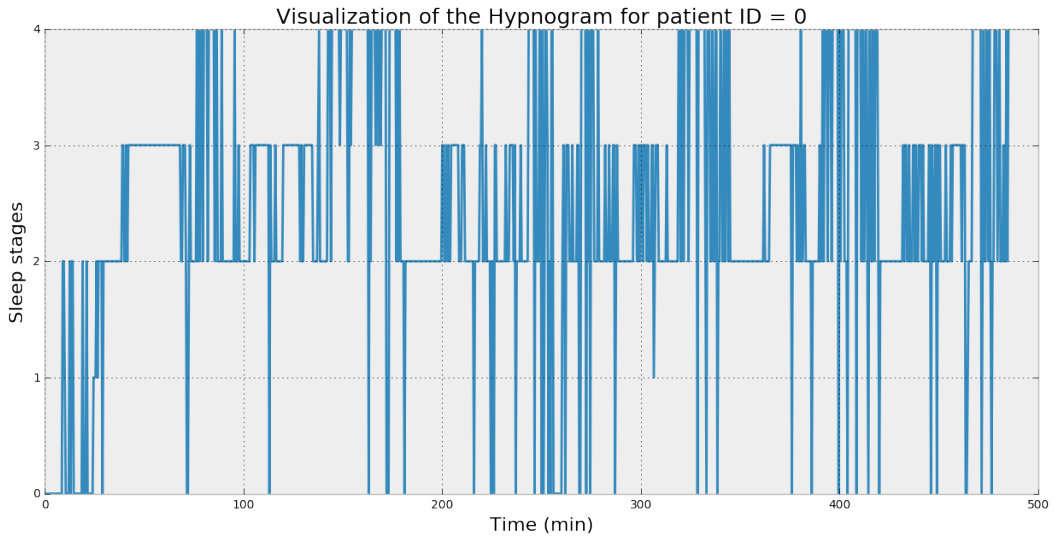

Figure 6: An example of hypnogram

3.1 Pre-processing

We can see hypnograms as time-series. We supposed then that the missing stages were most likely to be the same as the stages which happened before. Using this hypothesis, we filled the -1 values by propagating the last valid observation forward.

Many articles in the literature ([1], [2], [3], [4]) have studied the relationship between sleep stages and aging. To do so, they introduced different features and analyzed their evolution with age. Inspired by these studies, we extracted similar features:

- Total sleep time (TST)
- Sleep efficiency (SE)
- Sleep Latency (SL)
- Percentage of each stage including wake time (Si_PERC for $i \in \{0, \dots, 4\}$)
- The average duration of each stage (Si_MEAN) and its maximum (Si_MAX) because sleep becomes more fragmented with age.

In Figure 7 we can see the evolution of the deep sleep time as a function of age (we gathered people by bins of 5 years). We can see that the curve is slowly decreasing.

3.2 Random Forest

After the pre-processing was done, we decided to use a Random Forest regressor in order to predict ages using these features. We chose the Random Forests algorithm because it allows the variables to be of different scales (in our case: percentages vs. seconds) while being robust to irrelevant features. Furthermore, Random Forests can be used to estimate features importances.

Random Forests work by averaging multiple deep decisions trees trained on different parts of the training set. The `scikit-learn` implementation of this algorithm allows us to tune many parameters. Mainly, we focused on the number of trees in the forest and chose the highest number possible given our CPUs (500 estimators). Increasing the number of trees makes the algorithm more robust and helps avoiding overfitting. Another interesting parameter is the Out-of-bag (OOB) error: it is the mean prediction error for each sample x_i using only the trees that didn't fit to this sample. Hence, the OOB error was used to evaluate the performance of the algorithm.

After running the algorithm on the train set, we had an error of 24.35% on the train set. To check which one of the features extracted was most relevant, we plotted the importance for each variable. This is possible with Random Forest by modifying the values for one feature and then calculating

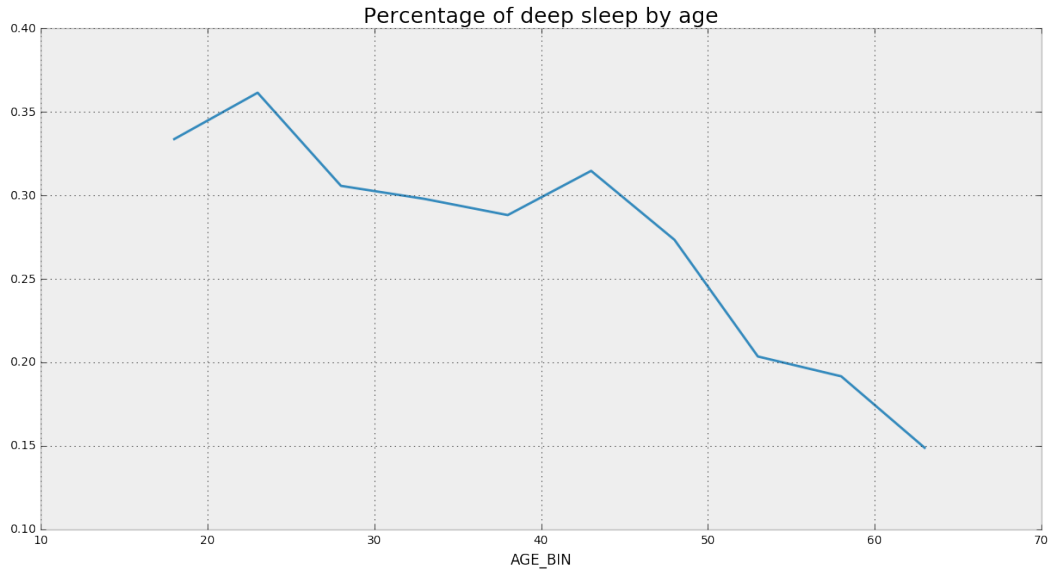

Figure 7: Percentage of deep sleep (stage 3)

the difference between the OOB error after and before the perturbation. Afterwards, the features are ranked based on this score. In `scikit-learn`, we can have this result easily after fitting the algorithm.

In the figure 8, we can see that the most relevant variable is the average duration of deep sleep (which represents how fragmented this stage is). Moreover, the other attributes related to the deep sleep stage (S3) had good scores in comparison with other features. Hence, we can conclude that this stage is the most affected with aging.

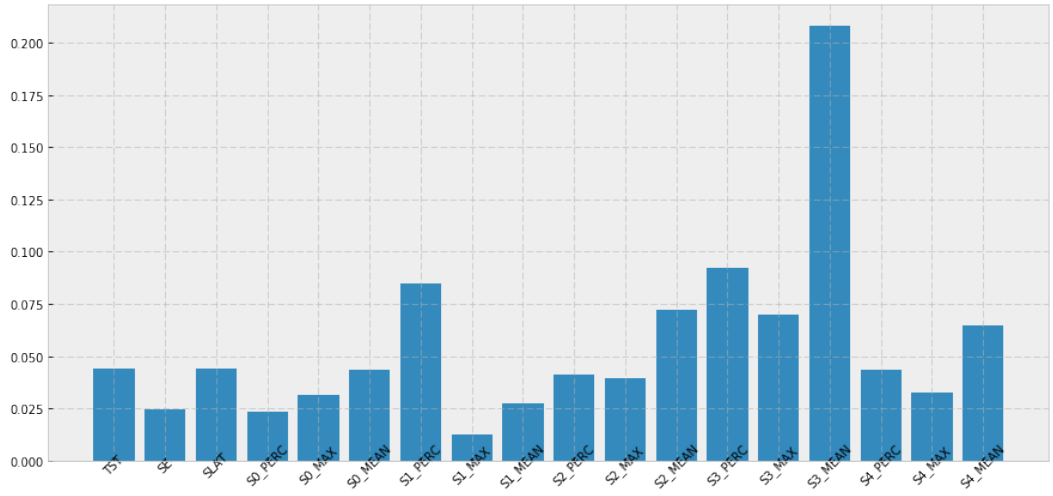

Figure 8: Features importance after fitting the Random Forests

4 Results

We uploaded our results for each one of our approaches and we had the following error values on the test dataset:

- ANN on the EEG: 21.95%

- Random Forest on the Hypnogram: 21.69%

The errors are better than with the train set, we conclude that our models didn't overfit.

The next step was to find a way to aggregate both our approaches. We decided to run a Linear Regression on a new dataset formed by concatenating the data with the two predictions. With this final model, we had a score of: 20.18% which is better than our models used separately. Our final ranking was 11^{th}.

The figure 9 shows the different predictions obtained for the first 10 subjects. We see that the Linear Regression tends to average the results of the Random Forest and the ANN. This doesn't work in some cases (see the subject number 2 or 4); however in general it improves the predictions (for example the subject number 5).

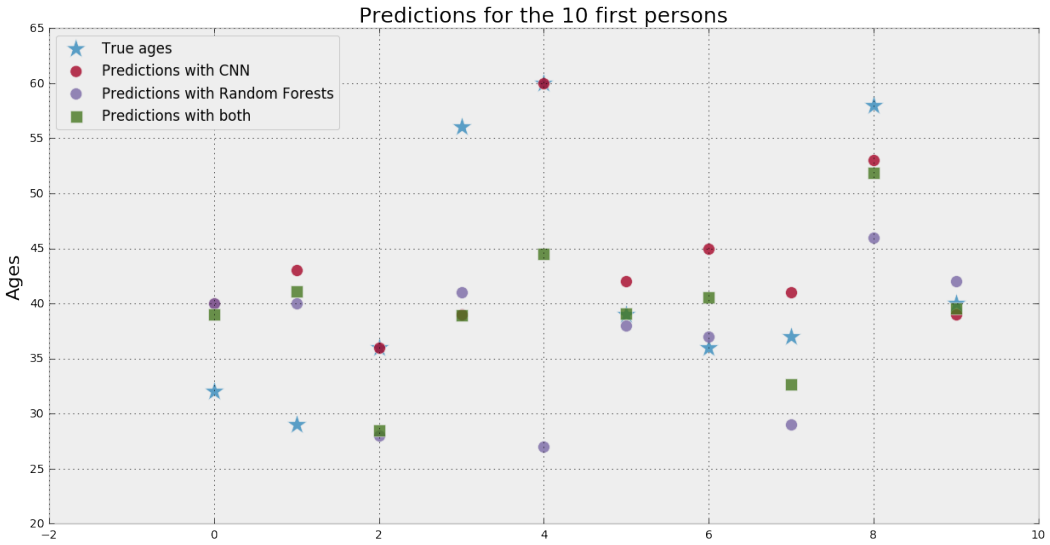

Figure 9: Results

5 Discussion

This project was the occasion for us to make intensive experiment with the TensorFlow framework. It was also the occasion for us to learn how to deal with labels that easily yield a trivial non interesting solution. In the future, we may try some techniques that penalize the trivial solution, but our feeling about this kind of problem is that it is better to focus on preprocessing in such condition, because we need to feed the machine learning algorithms with data that contain as much relevant information as possible.
On the other hand, we have noticed in the articles that we read that external parameters can influence the EEG : period of the year, alcohol, smoking, non smoking... In general, the EEG presented in the articles were selected by professionals that pre selected the examples used. What's more, we know that the EEG were acquired with two different devices, but this information did not appeared to be useful in our analysis. However, we would have preferred having technical information about the devices : what is the unit and the precision of the measured data for example. With such information, we could have improved the preprocessing part by adding a denoising module for instance.

Acknowledgments

We would like to thank Stéphane Mallat and all the team for organizing the Challenge Data. Also, our thanks go to the Rythm team for providing this interesting project.

References

[1] D. Dijk, J. Groeger, N. Stanley, and S. Deacon. Age-related reduction in daytime sleep propensity and nocturnal slow wave sleep. 2010.

[2] H. Landolt, D. Dijk, P. Achermann, and A. Borbély. Effect of age on the sleep eeg: slow-wave activity and spindle frequency activity in young and middle-aged men. 1996.

[3] L. Novelli, F. Raffaele, and B. Oliviero. Sleep classification according to aasm and rechtschaffen and kales: effects on sleep scoring parameters of children and adolescents. 2009.

[4] M. Ohayon, M. Carskadon, C. Guilleminault, and M. Vitiello. Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan. 2004.

[5] E. Tagluk, N. Sezgin, and M. Akin. Estimation of sleep stages by an artificial neural network employing eeg, emg and eog. 2009.