BABELE

Carmine
Ippolito

1 - CONTEXT OF THE PROJECT

Lips and **lip movements** are biometrics that have been used for several purposes.

- improve lip reading systems
- improve applications in the forensic field

2 - GOALS OF THE PROJECT

Realize a **recognition system** that uses **videos containing only the lip area**.

Evaluate whether the **knowledge of the subject's language** (that needs to be identified) produces better results.

Do an **initial analysis** on some of the **non-functional requirements** (**fairness** and **explainability**).

3 - DATSET

The **provided dataset** contained 258 videos **unevenly split** into 6 languages (Italian, English, German, Spanish, Dutch and Russian).

For this reason, 32 videos were chosen for each language based on **frame rate**, **duration**, **size** and **thumbnail** (192 videos in total).

Language	Men (Under 30)	Men (Over 30)	Women (Under 30)	Women (Over 30)	Total
Italian	8	7	9	8	32
English	7	5	10	10	32
German	8	6	9	9	32
Spanish	7	8	10	7	32
Dutch	8	15	2	7	32
Russian	2	12	3	15	32

3 - DATSET

The videos were downloaded from YouTube by different people **without a standardized process**.

strange frame rates
(27)

very short duration
(less than a minute)

disproportionate size
for their content

3 - METHODOLOGICAL STEPS CONDUCTED TO ADDRESS THE GOALS

Old methodology

Script 1:

cut a video into **3 sub-videos** of **15 seconds** where a face is detected by dlib.

New methodology

Script 1

cut a video into **5 sub-videos** of **10 seconds** where a face is detected by dlib.

Script 2:

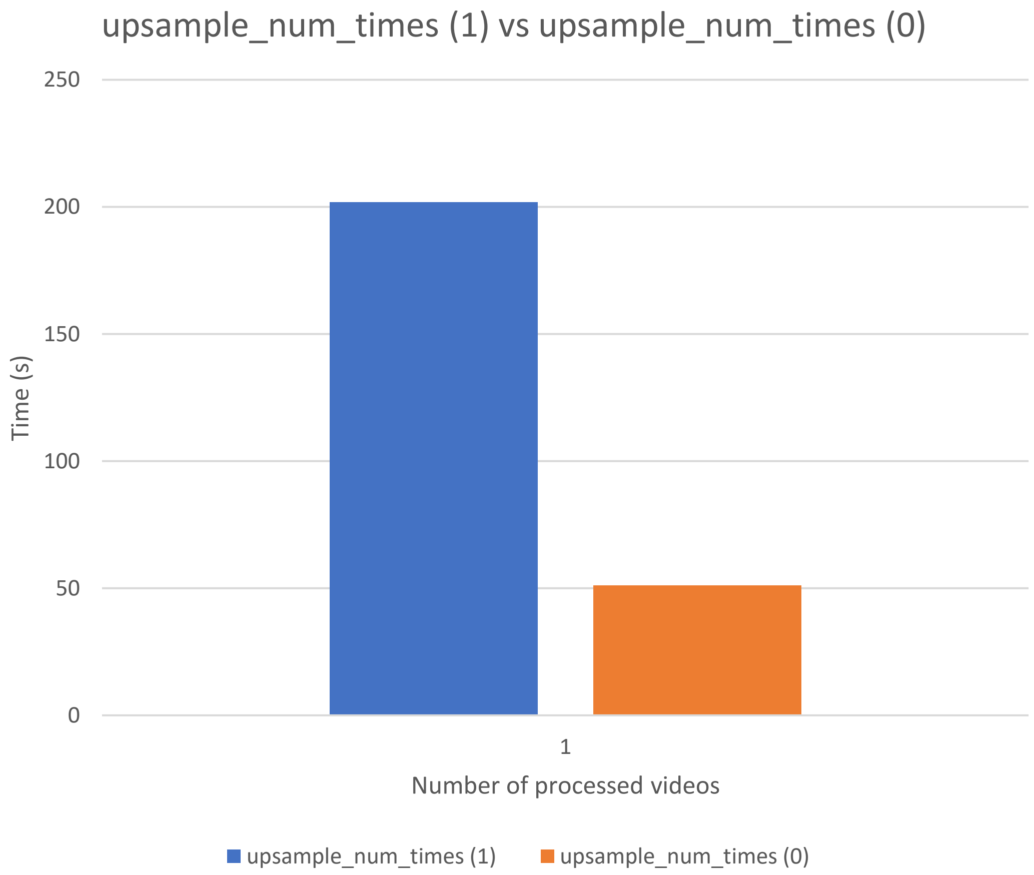
extract the **area of the labial zone** and the **Euclidean-Manhattan* distances** between the landmarks of the inner lip.

* Manhattan distances are extracted only with the new script.

3.1 - DATA ACQUISITION (SCRIPT 1-2)

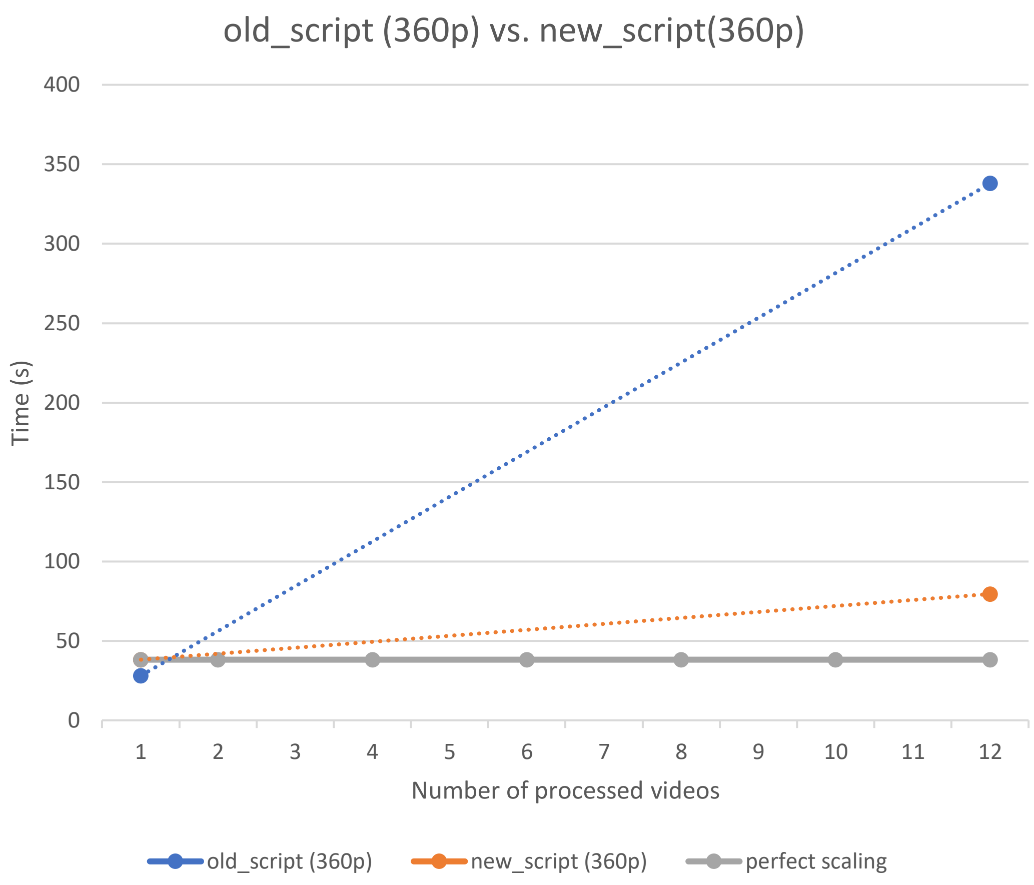
upsample_num_times (dlib) was set from 1 to 0, reducing the execution time by up to 75% with **negligible side effects**.

In general, it should be preferable to use **higher resolution videos** rather than **upsampling**.



3.1 - DATA ACQUISITION (SCRIPT 1)

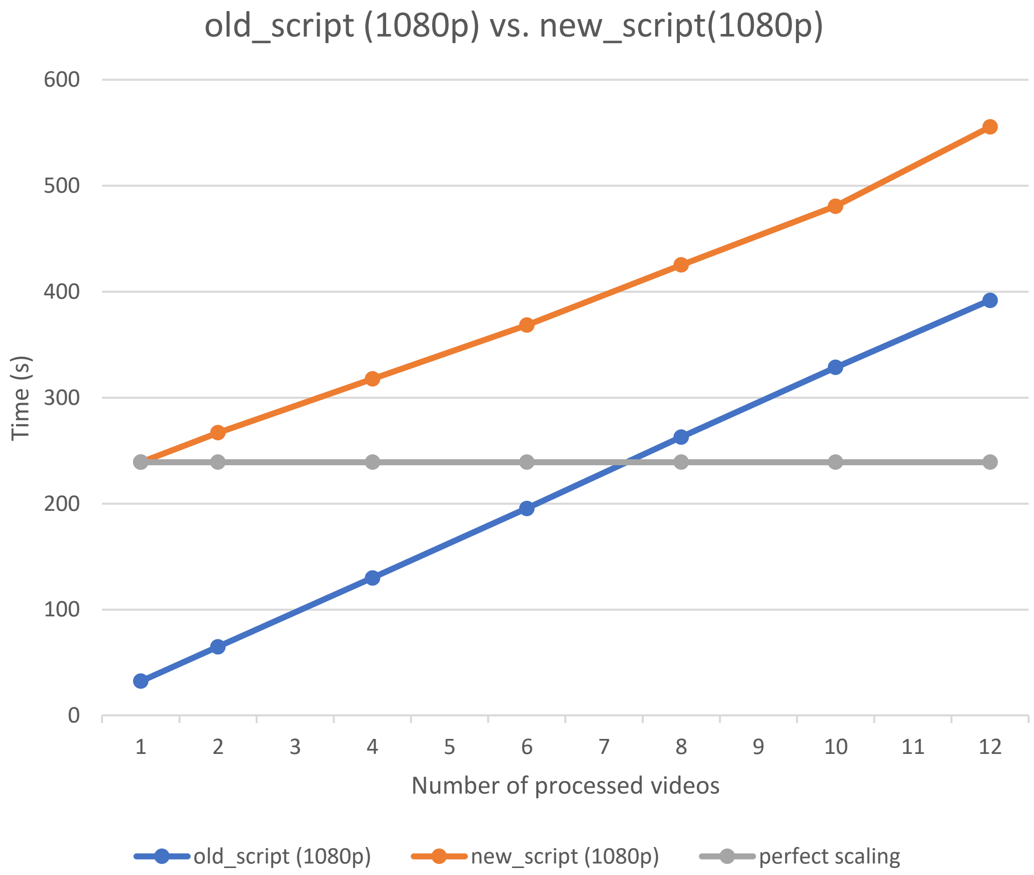
The **performance** of Script 1 varies greatly depending on the **resolution** (the old one scaled everything to 360p).



3.1 - DATA ACQUISITION (SCRIPT 1)

It may also take longer as it:

- encodes with a **better codec**;
- looks for a **single 10 second sequence** instead of merging three 5 second sequences;
- **skips some frames** to avoid possible transactions, special effects and occlusions.

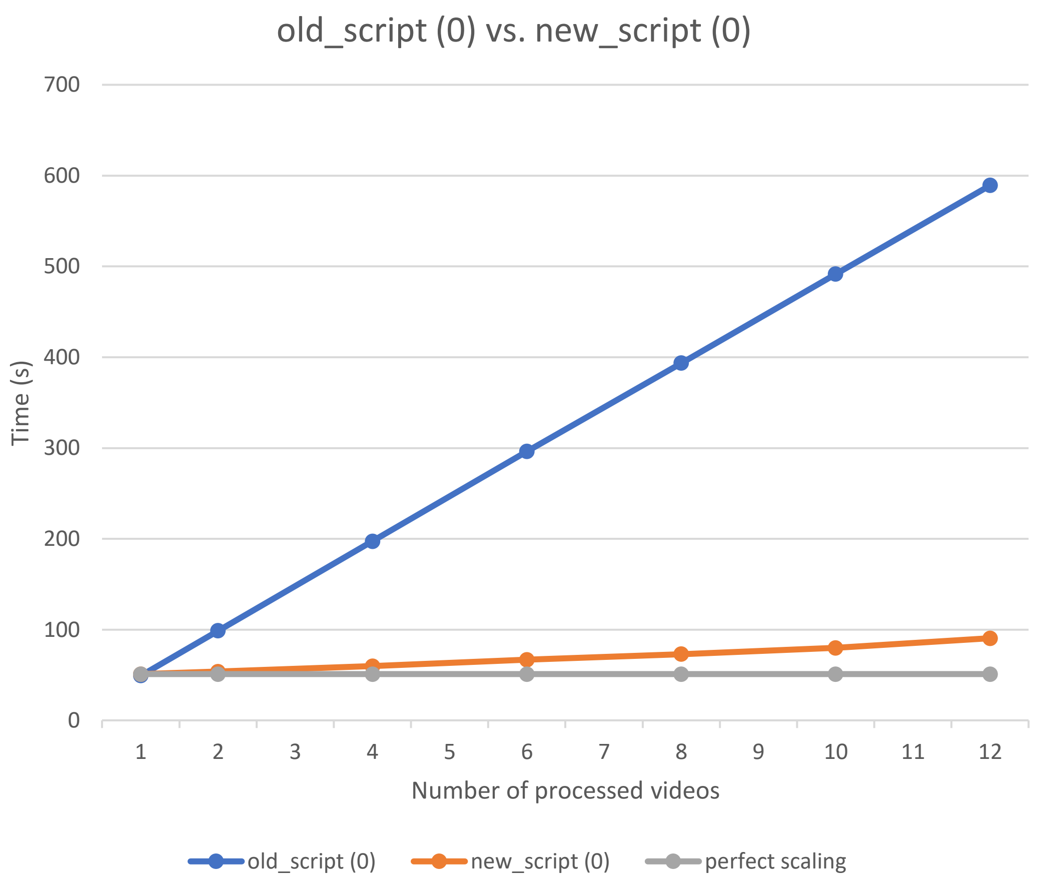


3.1 - DATA ACQUISITION (SCRIPT 2)

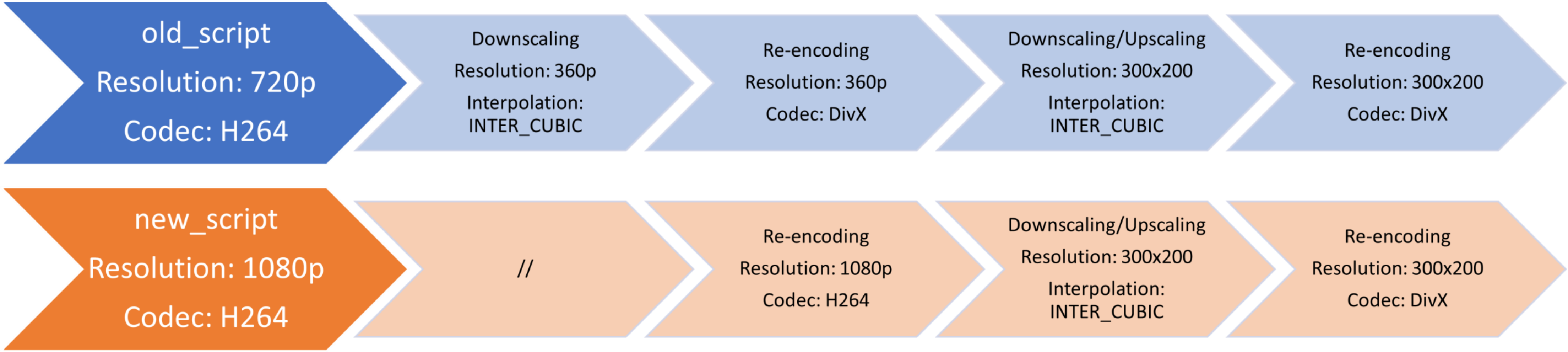
Script 2 reduces the execution time by up to 85%.

Experiments with **Oliver-Ricci curvature** have had little success and require more in-depth analysis.

Some sub-videos could not be extracted due to **dlib** (with the new methodology the **missing sub-videos** dropped from 15.3% to 1.8%).



3.1 - PIPELINE FOR THE EXTRACTION OF THE LABIAL ZONE



old_script
Resolution: 720p
Codec: H264

Downscaling
Resolution: 360p
Interpolation: INTER_CUBIC

Re-encoding
Resolution: 360p
Codec: DivX

Downscaling/Upscaling
Resolution: 300x200
Interpolation: INTER_CUBIC

Re-encoding
Resolution: 300x200
Codec: DivX

new_script
Resolution: 1080p
Codec: H264

//

Re-encoding
Resolution: 1080p
Codec: H264

Downscaling/Upscaling
Resolution: 300x200
Interpolation: INTER_CUBIC

Re-encoding
Resolution: 300x200
Codec: DivX

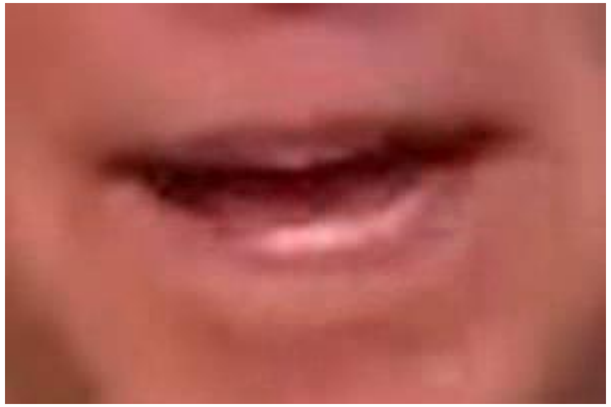
The **H264** codec requires **ffmpeg**.

3.1 - QUALITATIVE RESULTS

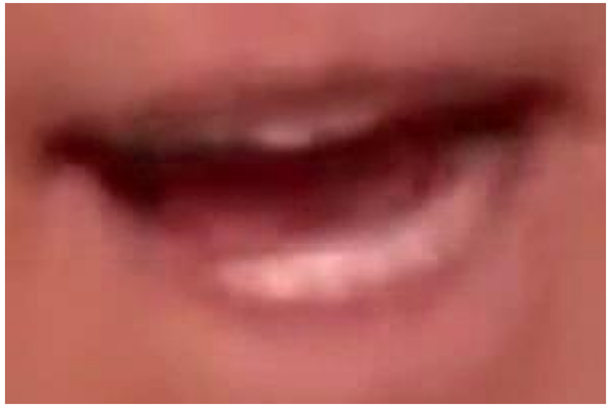
The increase in quality does not entirely pay off the decrease in performance.

The difference between the results is caused by the **downscaling**.

It remains desirable to experiment with more advanced codecs (**VP9**) and use the downscaling to improve performance.



old_sript

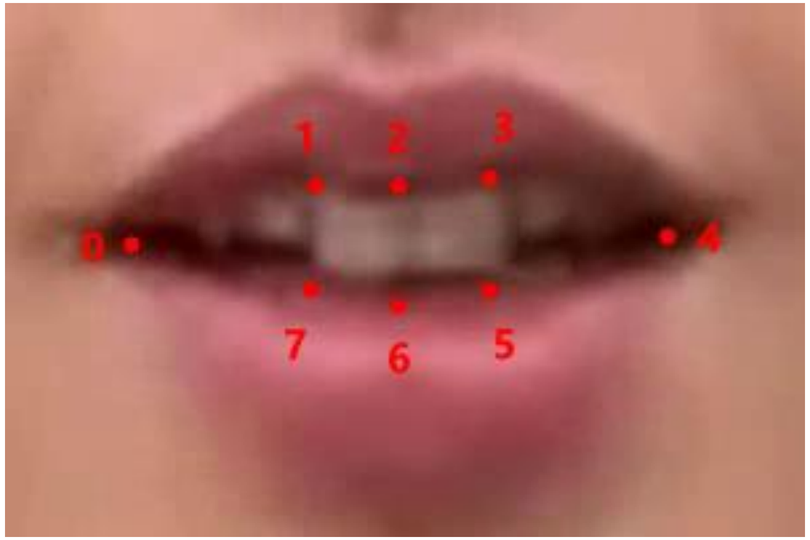


new_sript

3.2 - DATA UNDERSTANDING AND EXPLAINABILITY

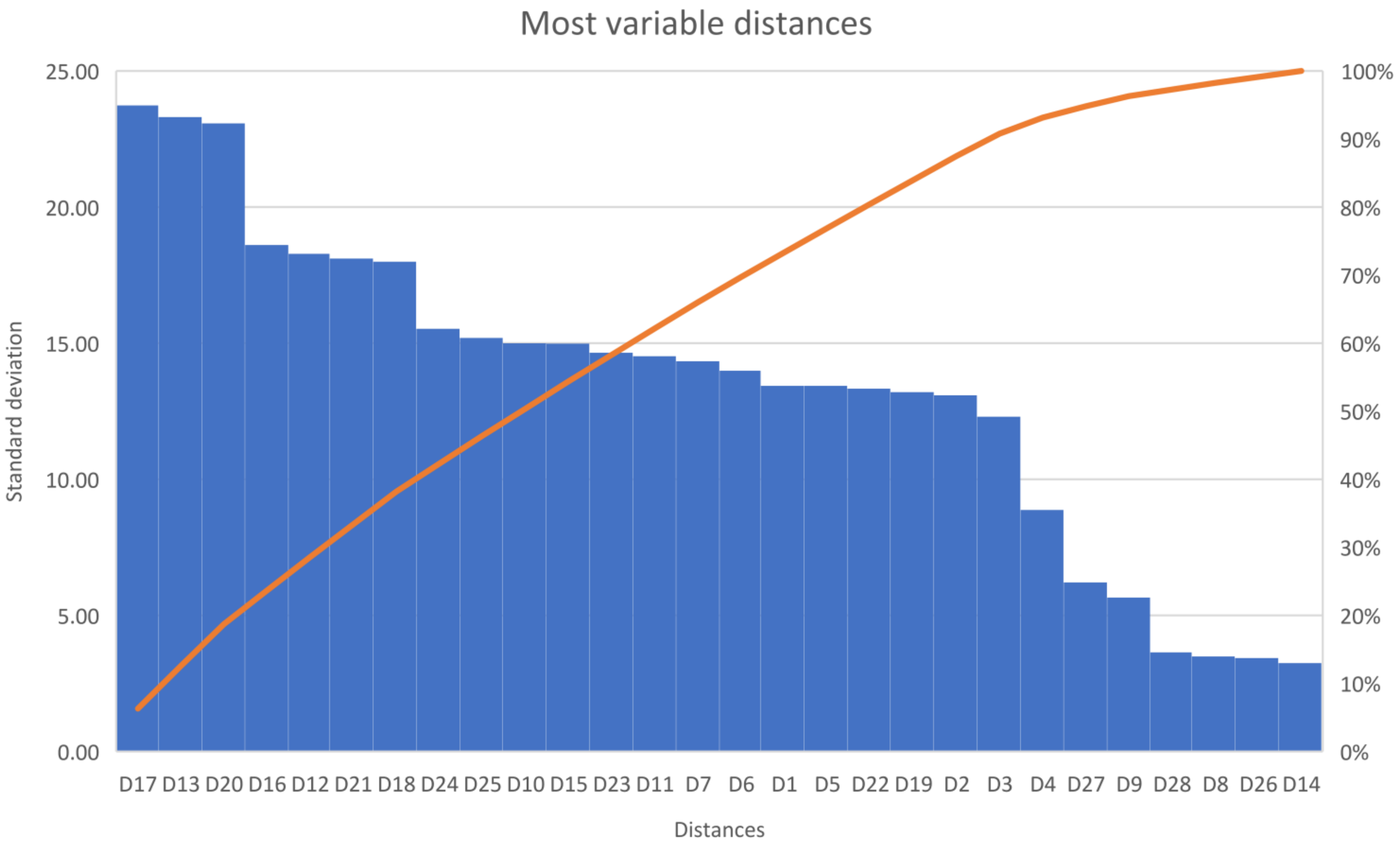
Two different methods were used to understand which distances had the least impact on the results.

The first method used the **standard deviation** to find the 4 **least variable distances**: D8 (1, 2), D14 (2, 3), D26 (5, 6), and D28 (6, 7).



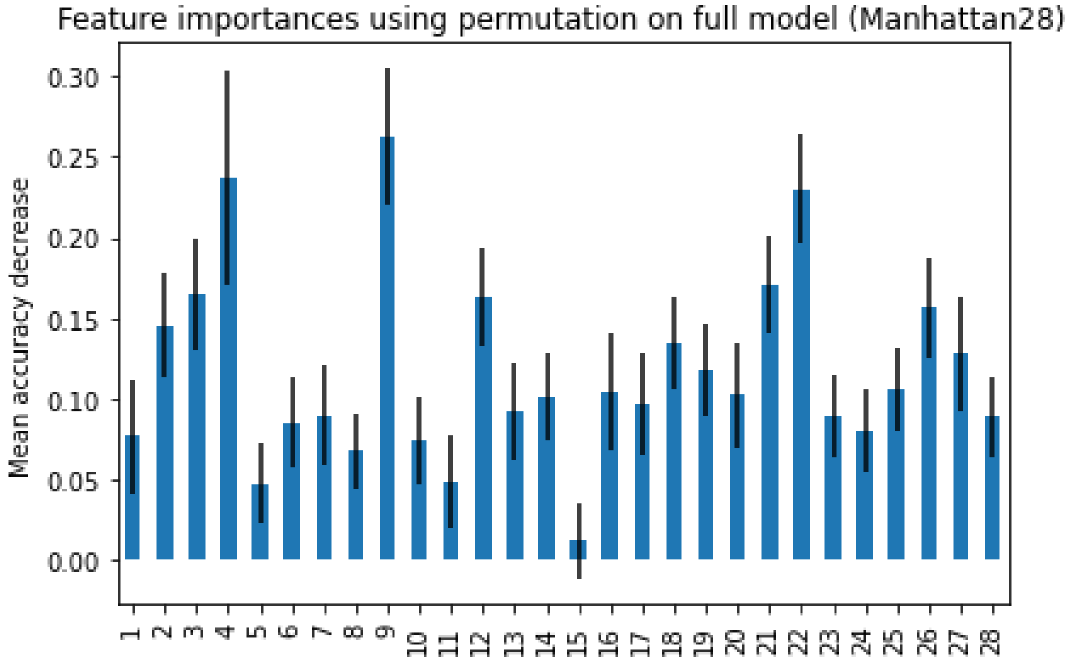
3.2 - DATA UNDERSTANDING AND EXPLAINABILITY

The first method used the **standard deviation** to find the 4 **least variable distances**: D8 (1, 2), D14 (2, 3), D26 (5, 6), and D28 (6, 7).



3.2 - DATA UNDERSTANDING AND EXPLAINABILITY

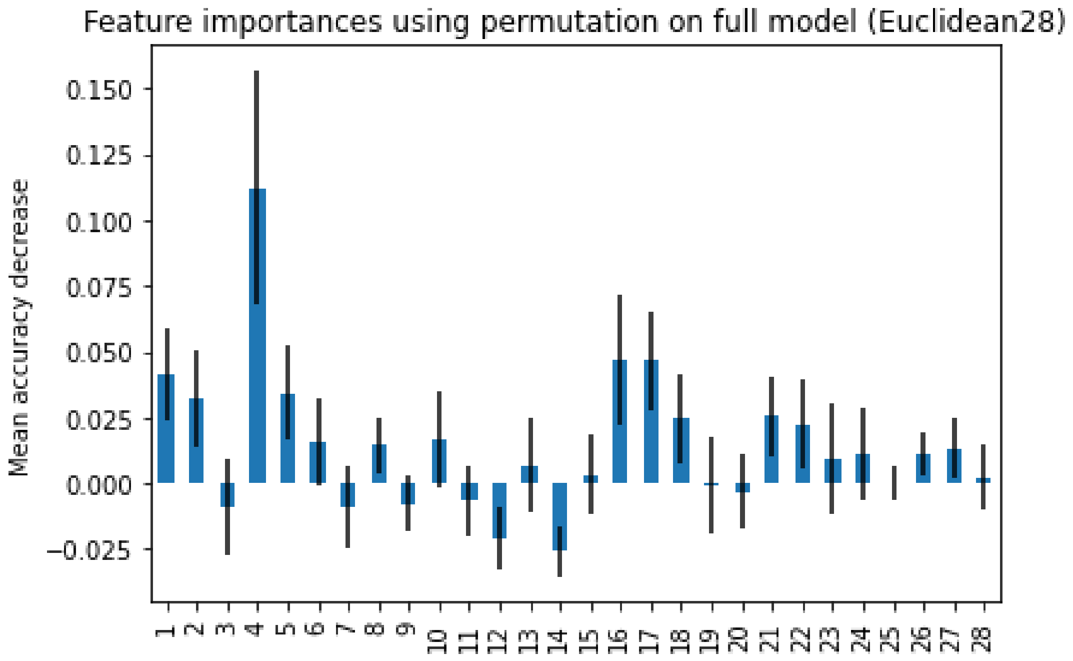
The second method used one of the trained classifiers (random forest) to find the **importance of the features based on their permutation**.



3.2 - DATA UNDERSTANDING AND EXPLAINABILITY

While the results for the **Manhattan distances** were fine, the results for the **Euclidean distances** could not be considered reliable (probably due to the lower accuracy of the model).

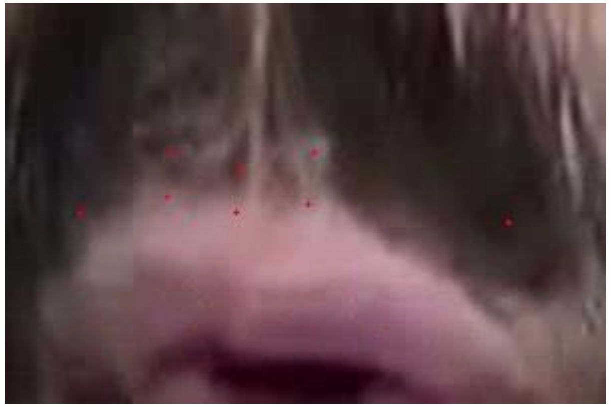
Because of this problem, only the first method was used.



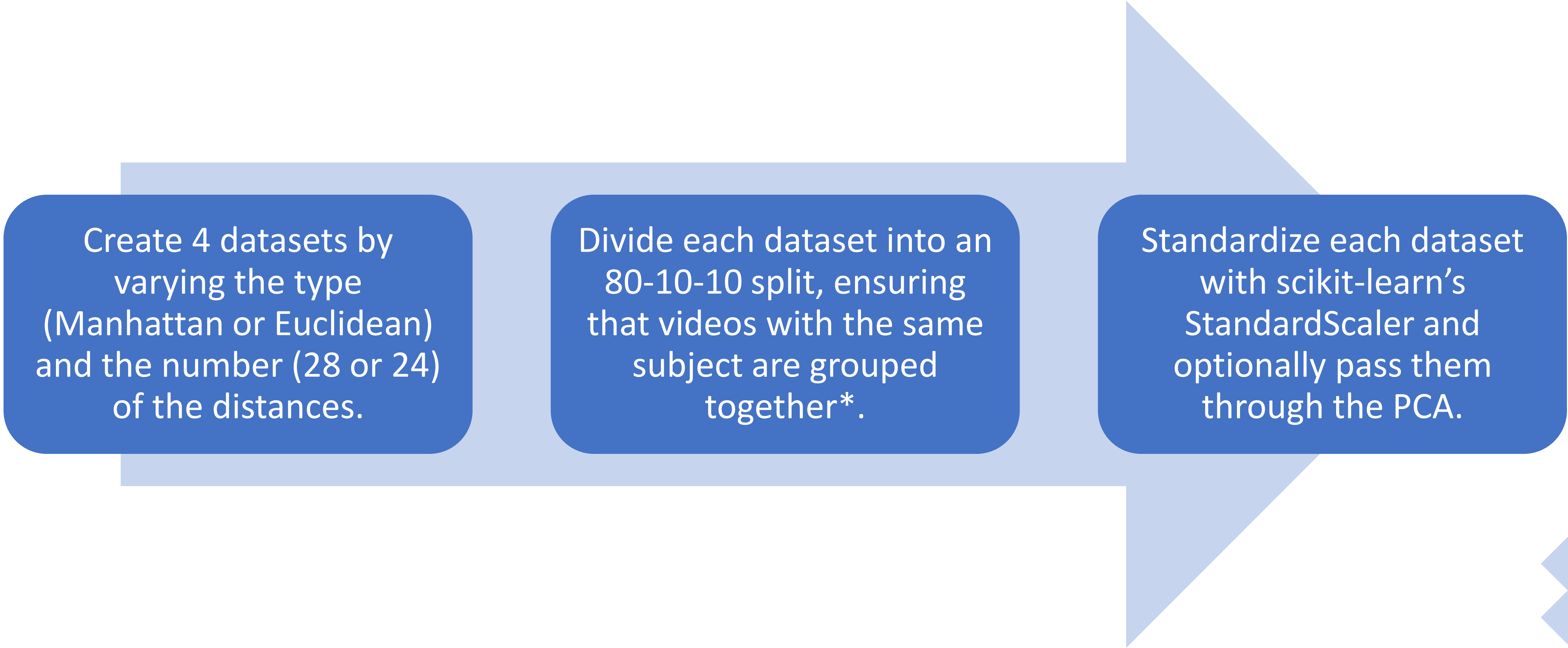
3.2 - DATA UNDERSTANDING AND FAIRNESS

Some of the videos featuring men with **mustaches** are characterized by an **incorrect localization of the landmarks**.

A better detector than **dlib** should be used to fix the problem.



3.3 – MODELING (DATASET FOR THE IDENTIFICATION OF THE LANGUAGE)



Create 4 datasets by varying the type (Manhattan or Euclidean) and the number (28 or 24) of the distances.

Divide each dataset into an 80-10-10 split, ensuring that videos with the same subject are grouped together*.

Standardize each dataset with scikit-learn's StandardScaler and optionally pass them through the PCA.

* The models need to learn the differences between the languages and not between the subjects.

3.3 – MODELING (IDENTIFICATION OF THE LANGUAGE)

The classifiers tested were the **k-nearest neighbors**, the **support vector machine** and the **random forest** (due to a problem with the **predict_proba** function the SVM was not used).

The **grid search** and the **MLFlow Tracking API** were used to select and save the **parameters**, respectively.

The **most accurate classifier** was the random forest trained with 300 estimators on the Manhattan28 dataset passed through the PCA with 0.95 components (42.2% accuracy).

3.3 – MODELING (RECOGNITION)

Two different strategies were used to create the models for the recognition: one based on machine learning (k-nearest neighbors and random forest) and one based on deep learning (CNN).

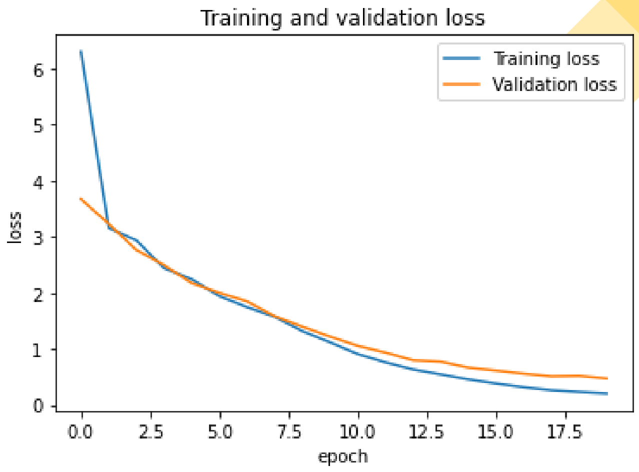
For the one based on **machine learning**, the dataset was created by merging the validation and test sets of the datasets with 28 distances and dividing them into a 60-20-20 split.

The **best classifier** was the random forest trained with 300 estimators on the Manhattan28 dataset (61.1% accuracy).

3.3 – MODELING (RECOGNITION)

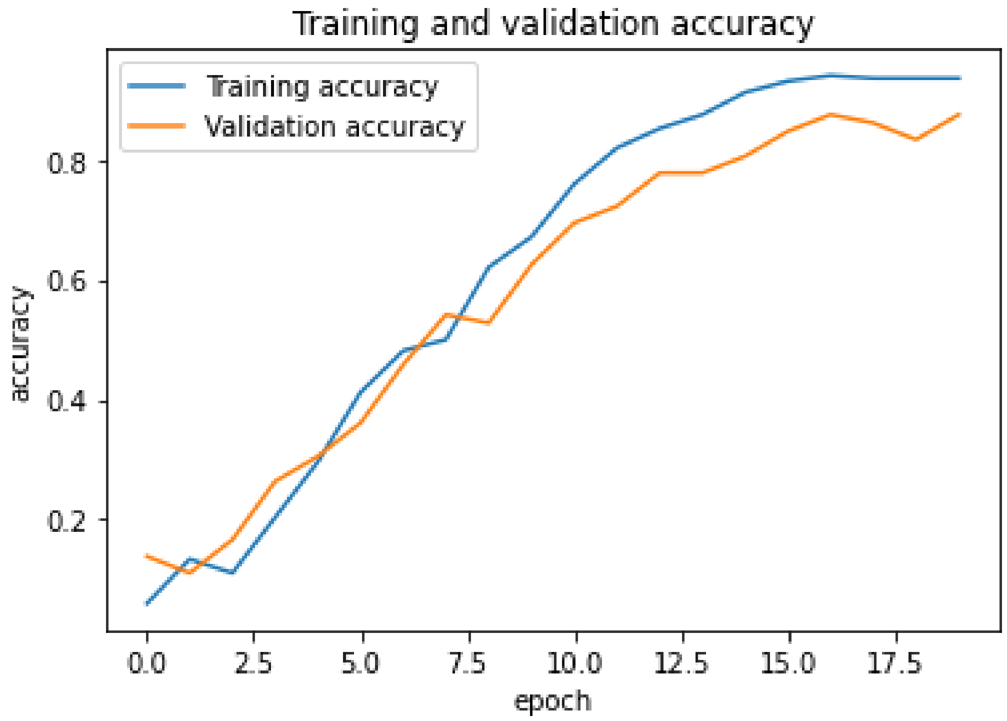
For the one based on **deep learning**, the dataset was created by extracting 10 frames for each subject present in the previously mentioned dataset (the data was preprocessed so that each pixel had zero mean and unit standard deviation).

3.3 – MODELING (RECOGNITION)



Because of the **small dataset**, only a **simple CNN** was used (accuracy 81.9%). The network structure is composed of:

- 1x convolutional layer;
- 1x maxpooling layer;
- 1x dropout layer*;
- 1x flatten layer;
- 2x fully connected layers.

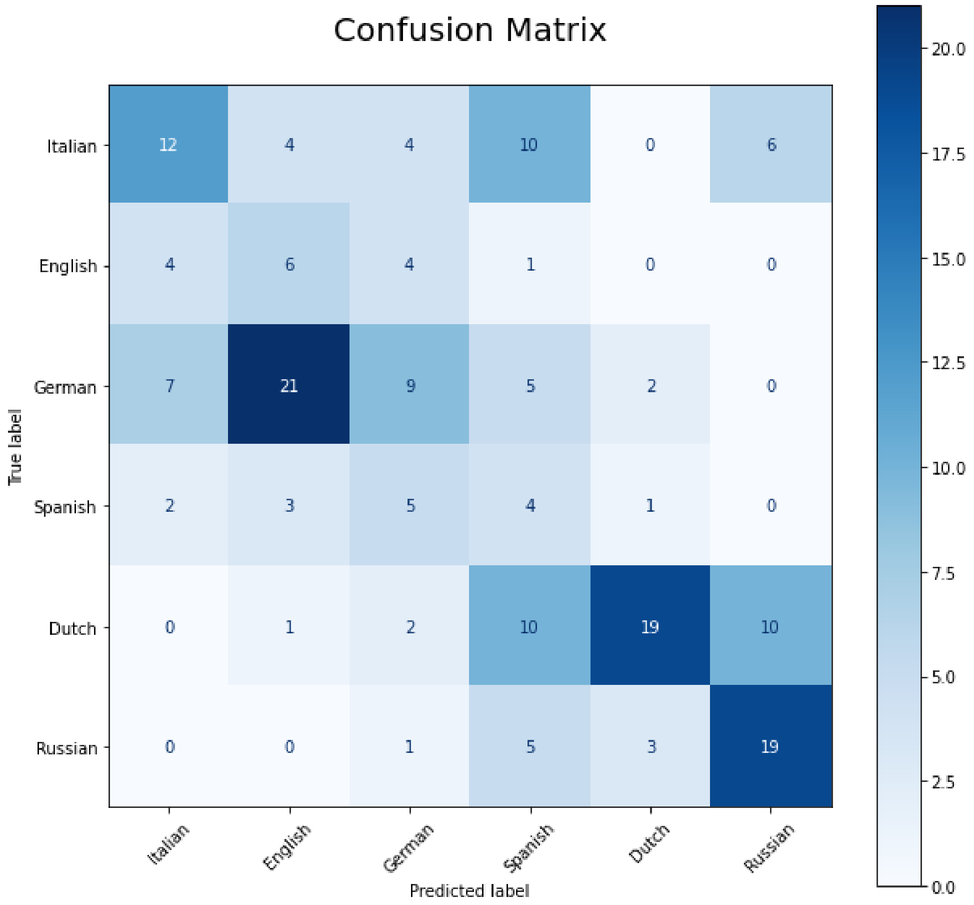


* The dropout layer was added to improve generalization.

4 - PRELIMINARY RESULTS AND FINDINGS (IDENTIFICATION OF THE LANGUAGE)

Dutch and Russian are the **most correctly classified** languages, while English and Spanish are the worst.

English, Spanish and Russian are also the languages with the **greatest number of predictions**.



4 - PRELIMINARY RESULTS AND FINDINGS

The **deep learning model** was not influenced by the language of the subject (its results were already very decisive).

On the other hand, the **machine learning model** performed better due to its uncertainty and lower accuracy.
It is safe to assume that with a **larger dataset** even the deep learning model may perform better (only 36 of the 192 subjects were included).

%-% Weights	ML accuracy	DL accuracy
100%-0%	61.1%	81.9%
80%-20%	66.7%	81.9%
60%-40%	55.6%	81.9%
40%-60%	41.7%	81.9%
20%-80%	30.6%	76.4%

5/6 - IMPLICATIONS OF THE RESULTS AND CONCLUSION

Recognition can be improved by knowing the **subject's language**.

There is still room for improvement due to **dataset limitations** (the **mustache** problem in particular).

Classifiers that use **temporal information** could also be considered.

The dataset with 24 distances always performed the same or better than the dataset with 28 distances, except for the random forest trained on the Manhattan28 dataset. Therefore, it is safe to assume that further analysis on **explainability** could lead to better accuracy.