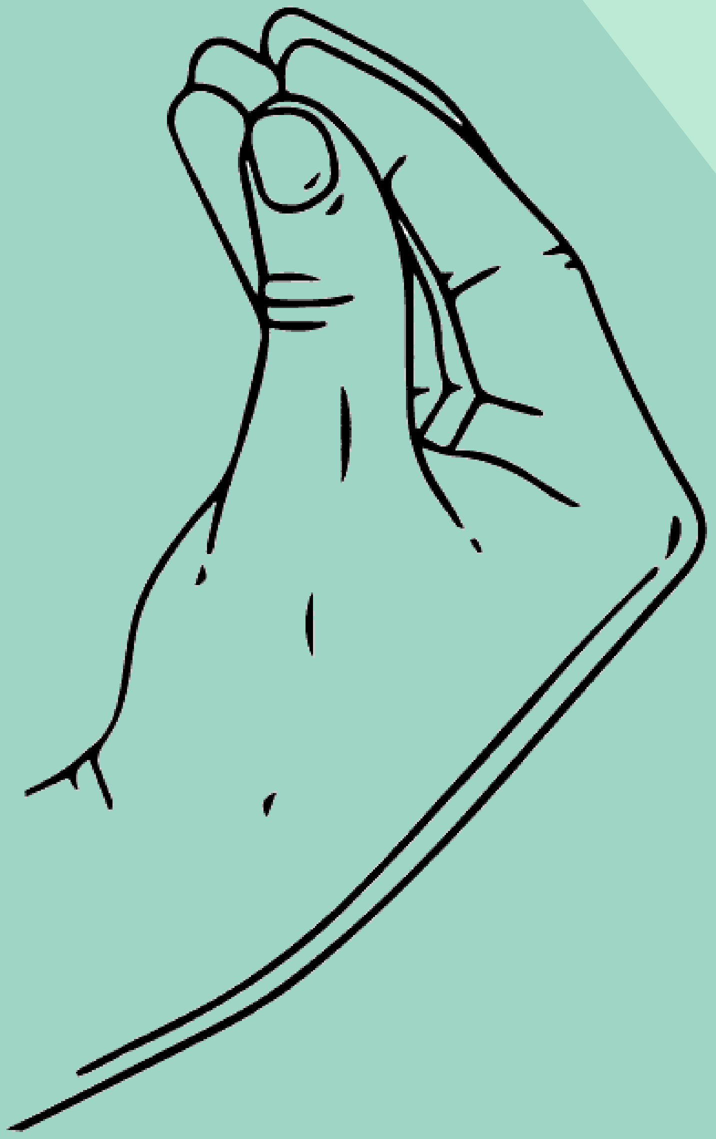
CANZONERI DANIELE
GRILLEA FRANCESCO



Speak Italian

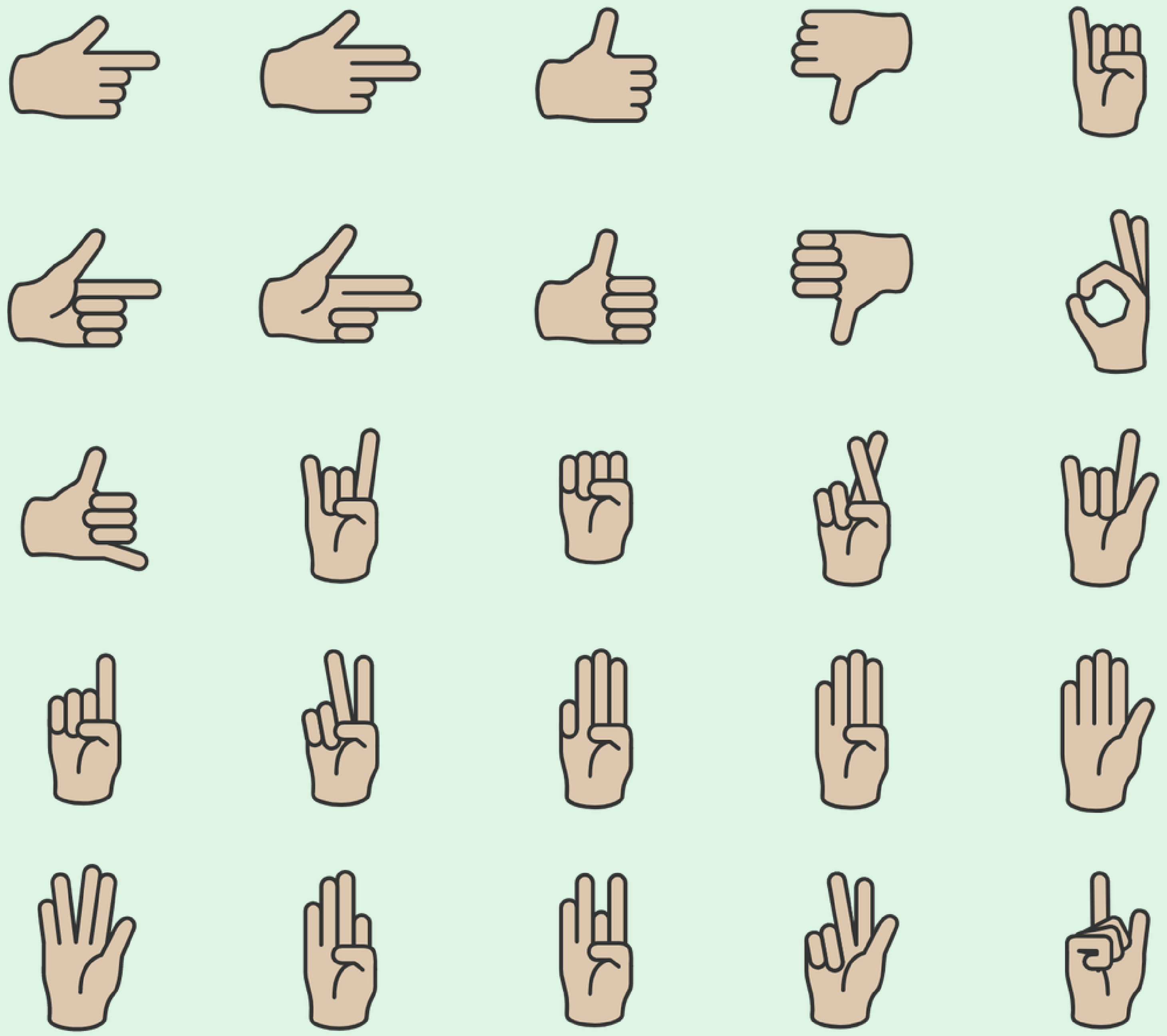
Data Mining and Machine Learning project

Context

- Hand gesture recognition task
    - Camera to record frames

- Pre-trained deep learning model to extract features from the image of a gesture (MediaPipe from Google)

- Machine Learning approach to classify the gesture



Roadmap

- Hand gesture recognition task
  - Camera to record frames
- Pre-trained deep learning model to extract features from the image of a gesture (MediaPipe from Google)
- Machine Learning approach to classify the gesture

Dataset - 1

Data Pre-Processing - 2

Train and Test Split - 3

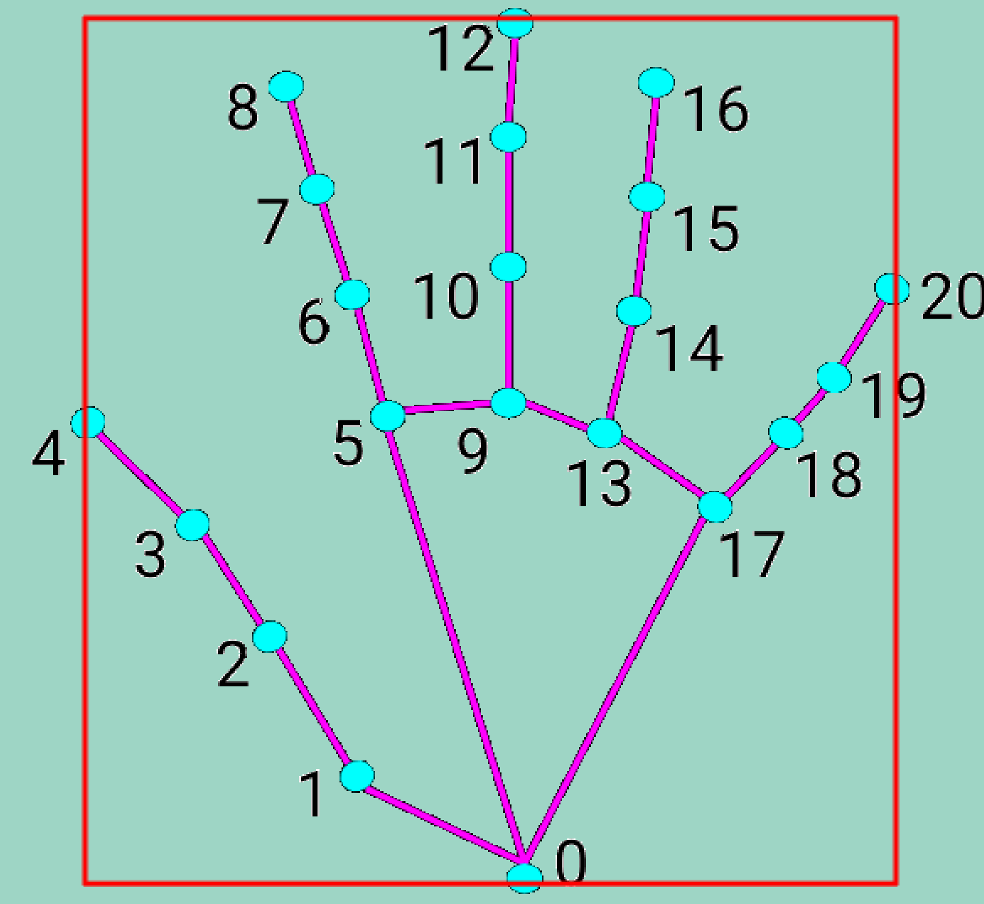
Model Comparison - 4

Real Time Demostration - 5

Dataset

- **Image pre-processing:**
  - each frame recorded is processed into a sequence of 21 hand landmarks
  - x and y are normalized to [0.0, 1.0] w.r.t. the palm width and height
  - z represents the landmark depth compared to the wrist depth

- **Data acquisition:**
  we asked 10 different people to record some frames for every gesture

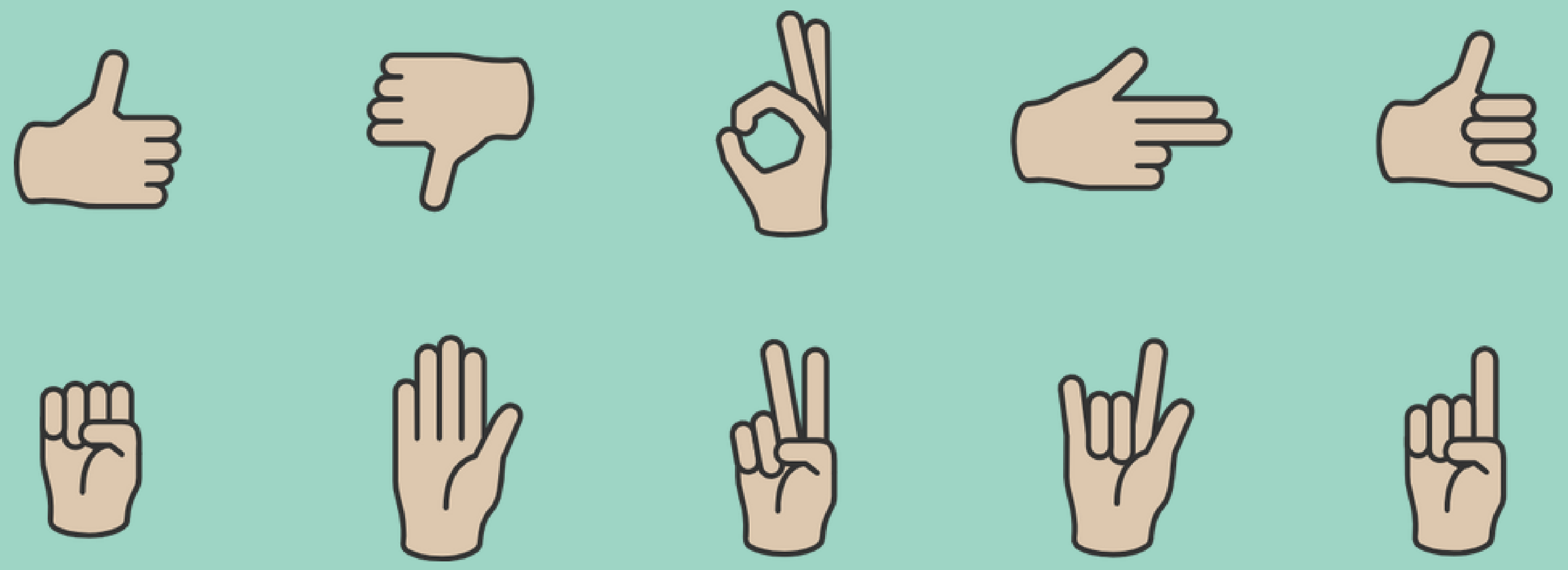


0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

Dataset

In the end we collected

- **40.000 data instances**
  - 200 data instances for each gesture made by the same person (with the same hand)
- **63 Features**: each of the 21 landmarks is represented by three points X, Y, Z
- **20 Classes**: for each gesture we considered left and right hand
- **No data cleaning needed**
  - Prerfectly balanced classes
  - Noise tolerant / No outliers



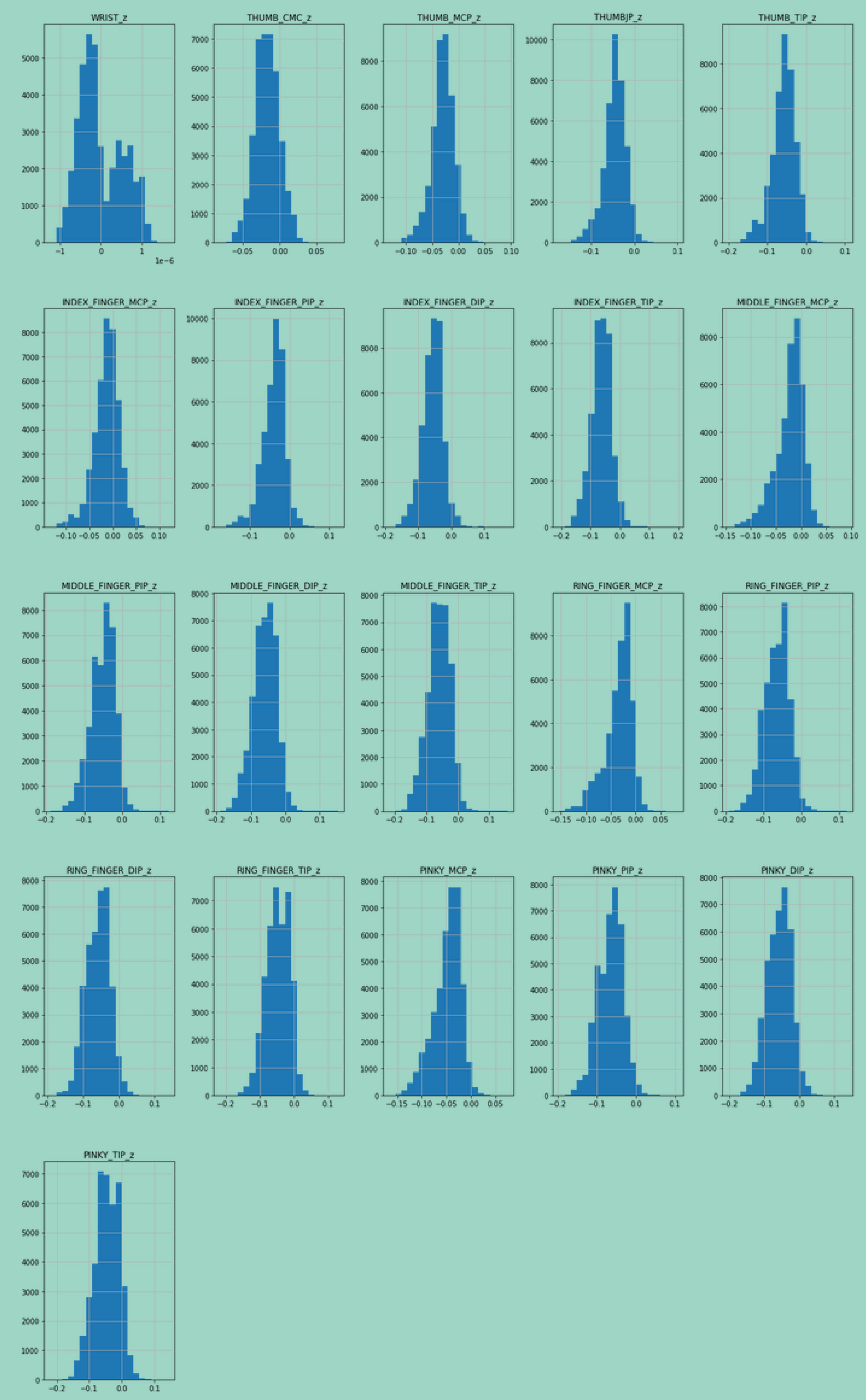
10 Different Gestures

- ThumbUp
- ThumbDown
- Okay
- Gun
- Call
- Fist
- Stop
- Peace
- Rock
- Index

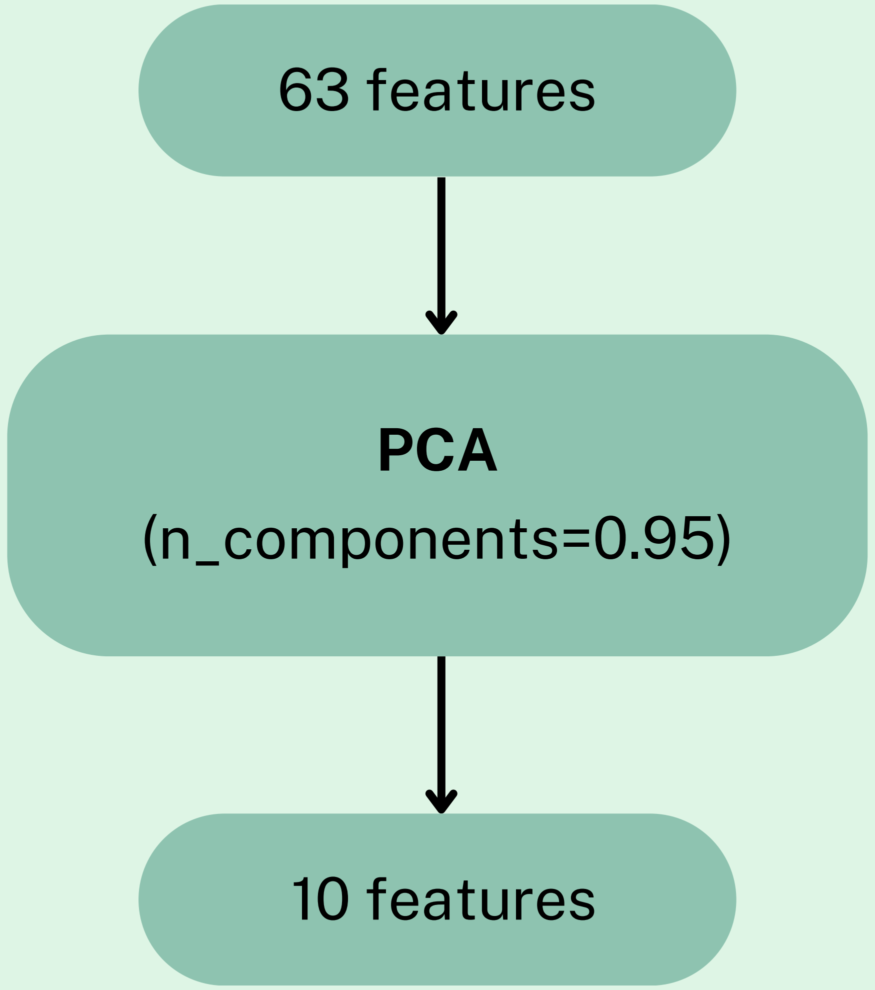
Pre–Processing

Attributes that represent the z point of a landmark have similar behaviour among all gestures.

Moreover, we found that all attributes relative to z-axis have a variance below a threshold fixed to 0.002.



Pre–Processing

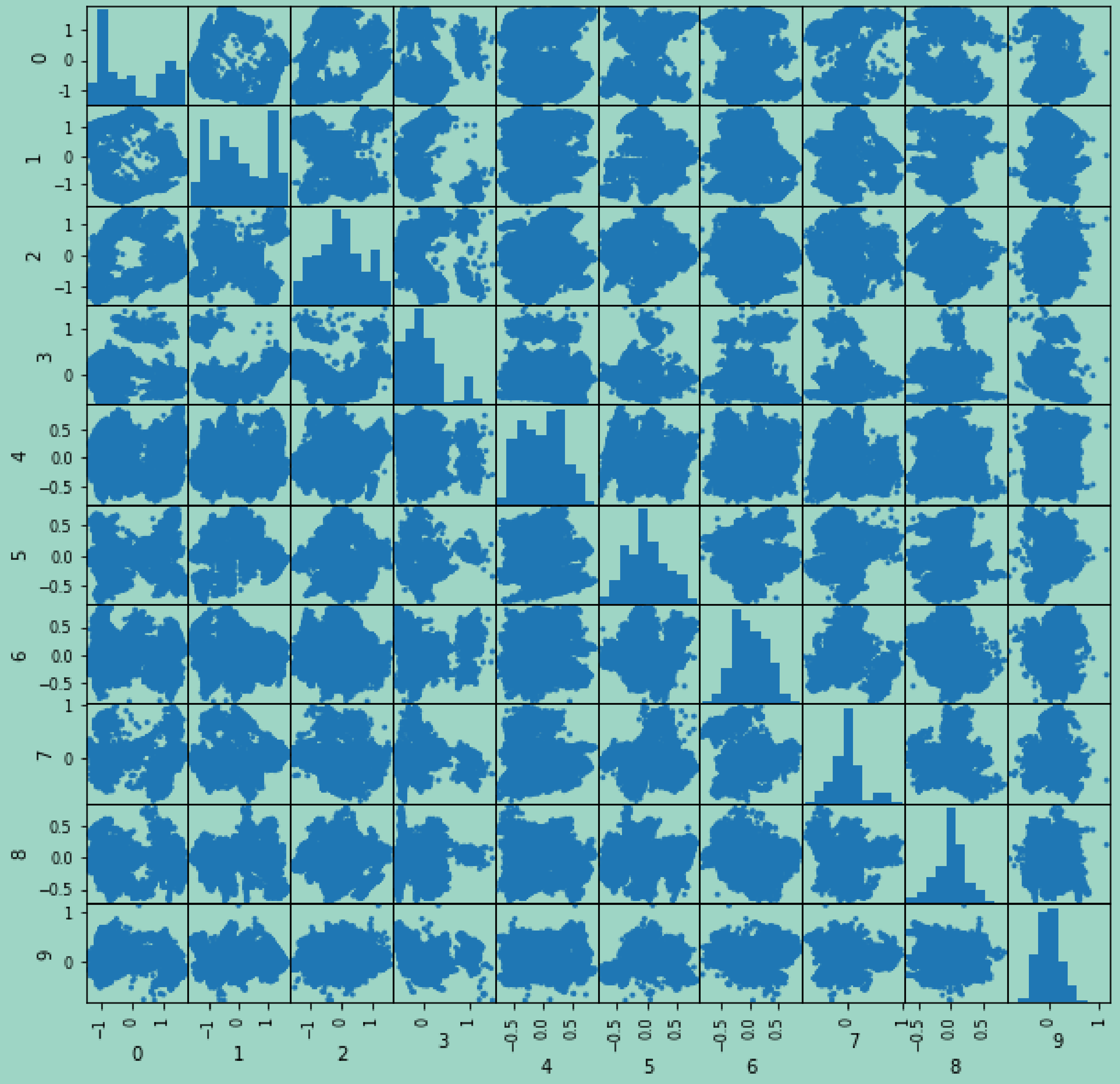


63 features

PCA

(n_components=0.95)

10 features

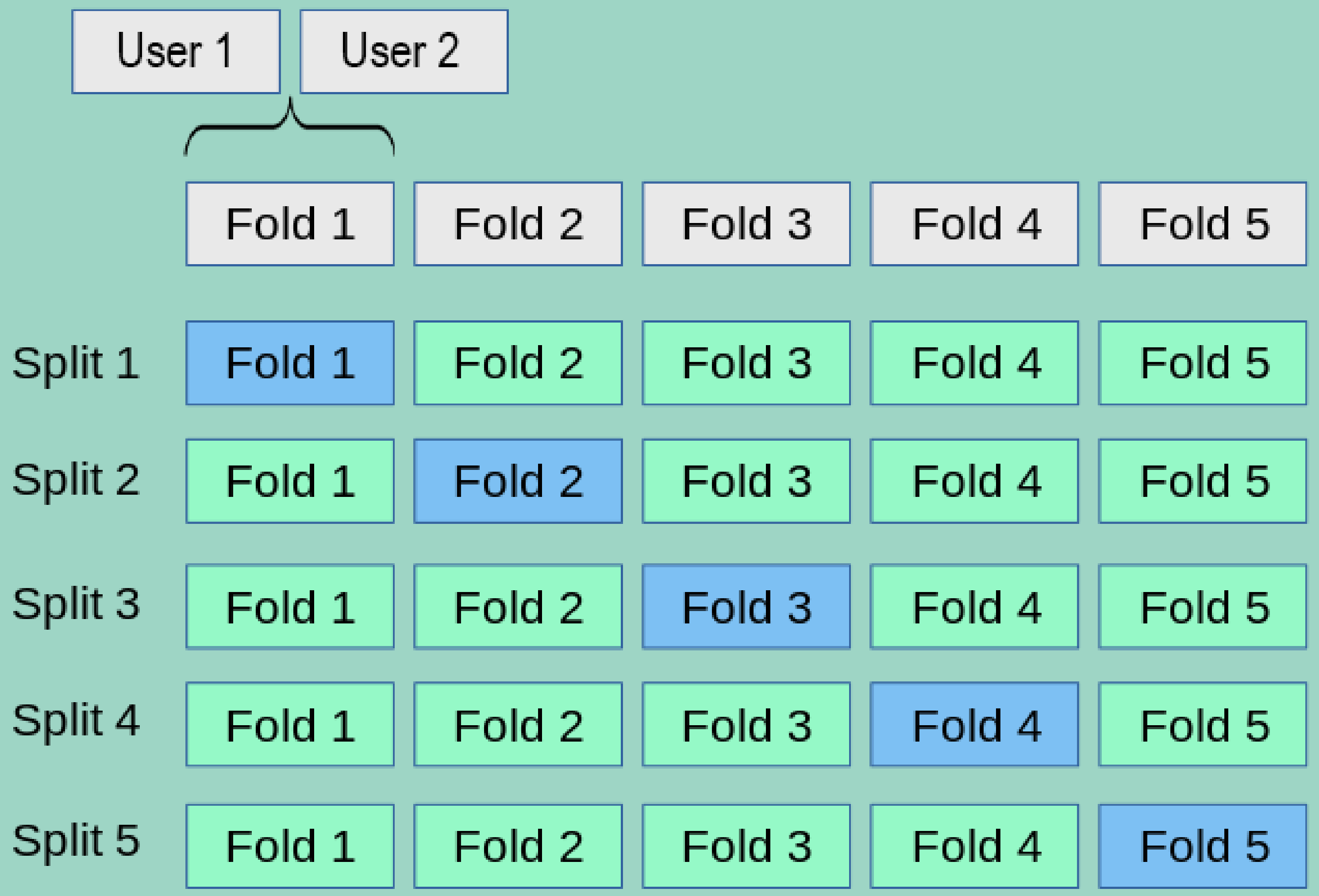


Train–Test Split

- **Ordered dataset**:
  - *user* as primary key
  - *gesture* as secondary key

- Stratified **5-Fold Cross Validation**
  - 8000 instances for each fold

- Each fold is made by all gestures made by **two different users**
  - the test set is made of landmarks belonging to *unseen* users

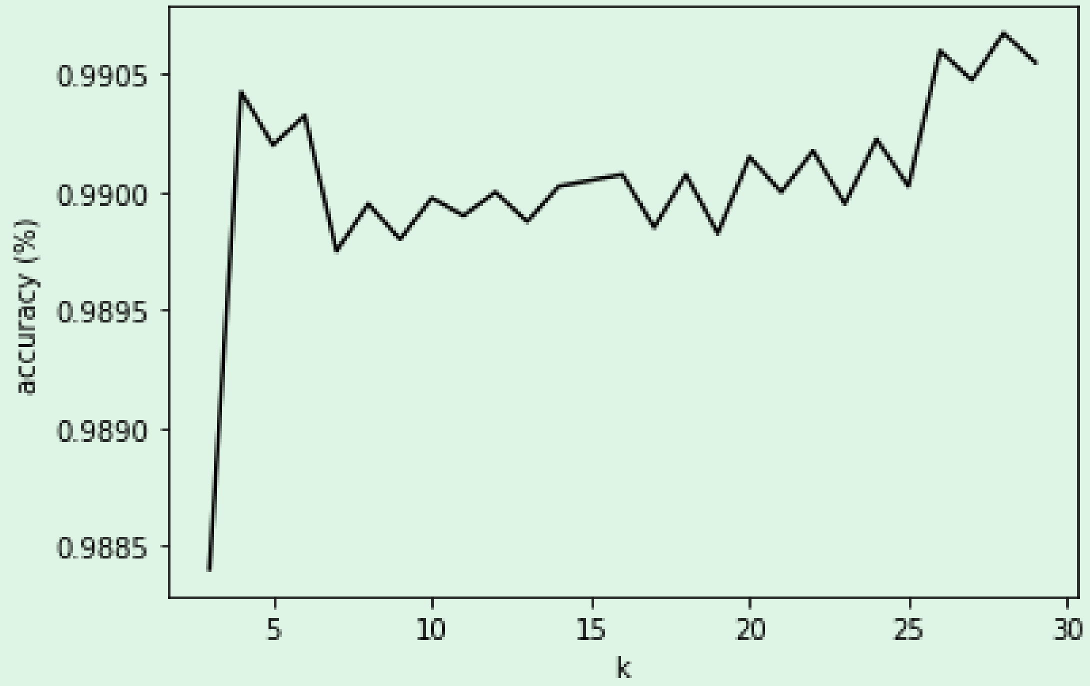


User 1 User 2

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Hyperparameters

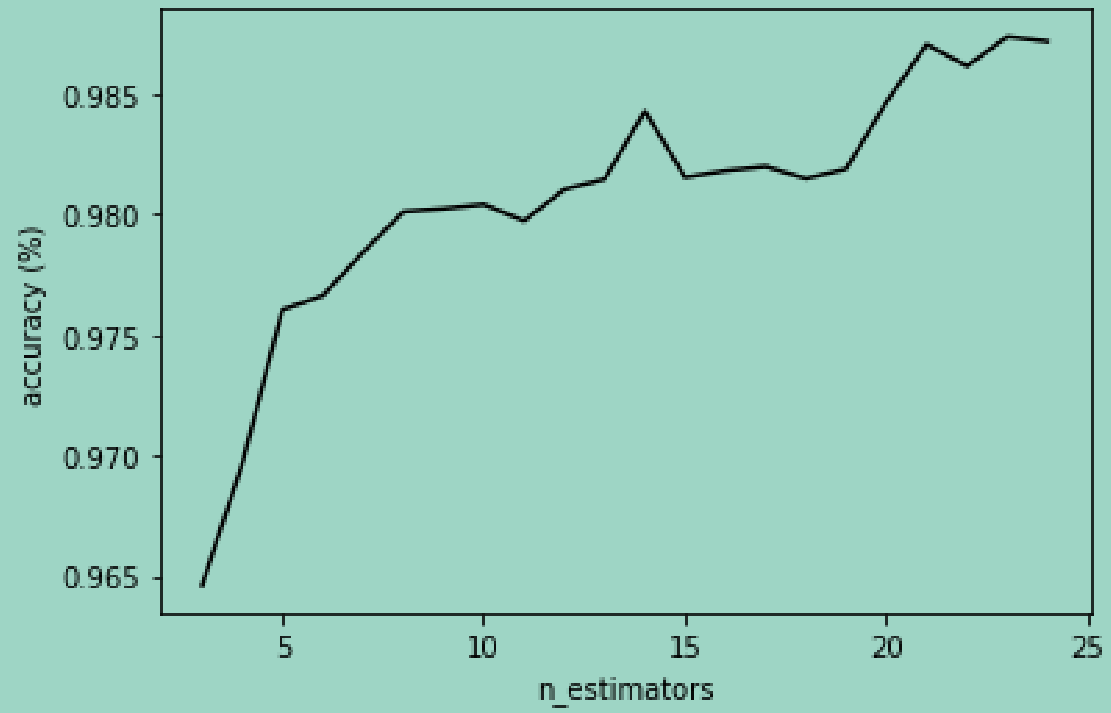
kNN



Chosen value = 4

- not the highest accuracy but the difference is very low, so we choose the simplest

Random Forest

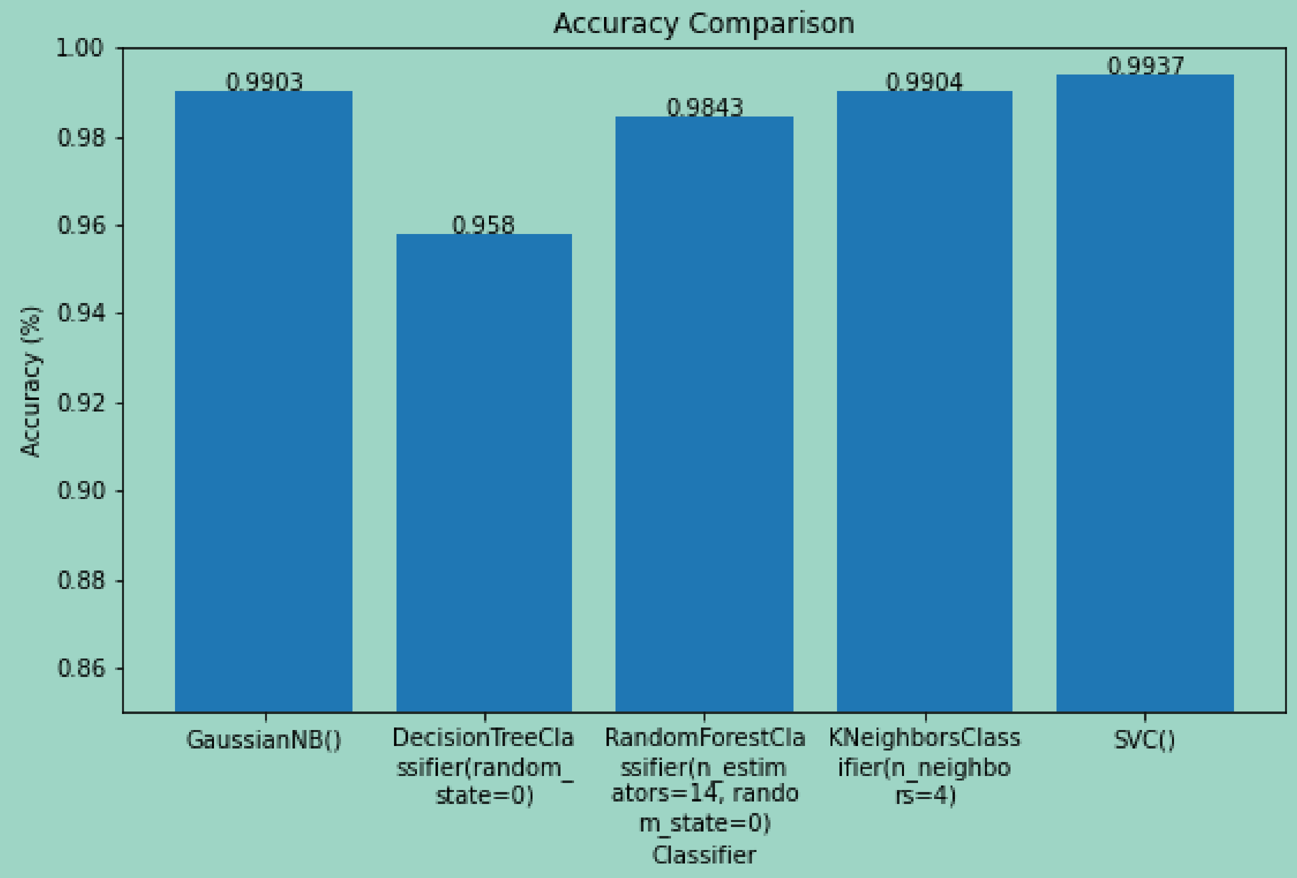


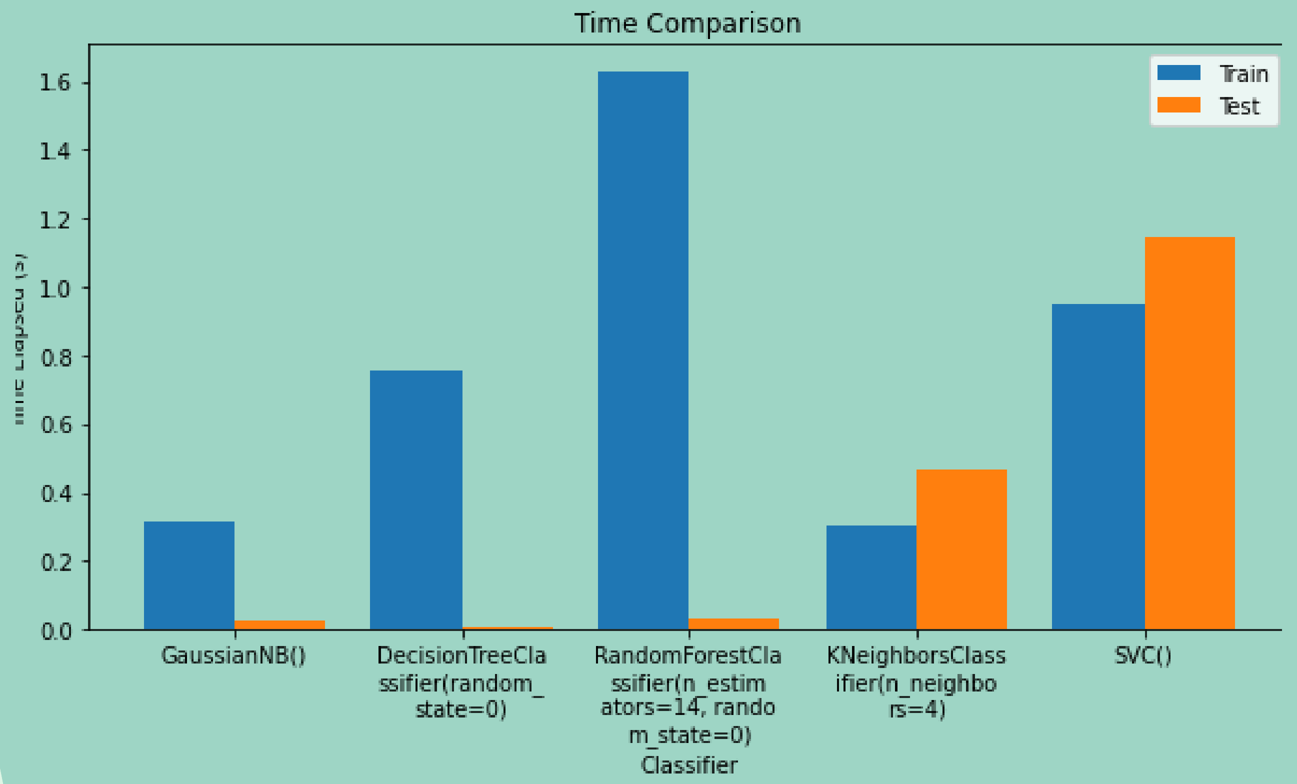
Chosen value = 14

- highest accuracy

Comparison

	Accuracy	Precision	Recall	F1_score
Gaussian Naive Bayesian	0.99	0.99	0.99	0.99
DecisionTree Classifier	0.958	0.964	0.958	0.956
Random Forest (n=14)	0.984	0.986	0.984	0.984
kNN (k=4)	0.99	0.991	0.99	0.99
SVC	0.9936	0.9939	0.9936	0.9936



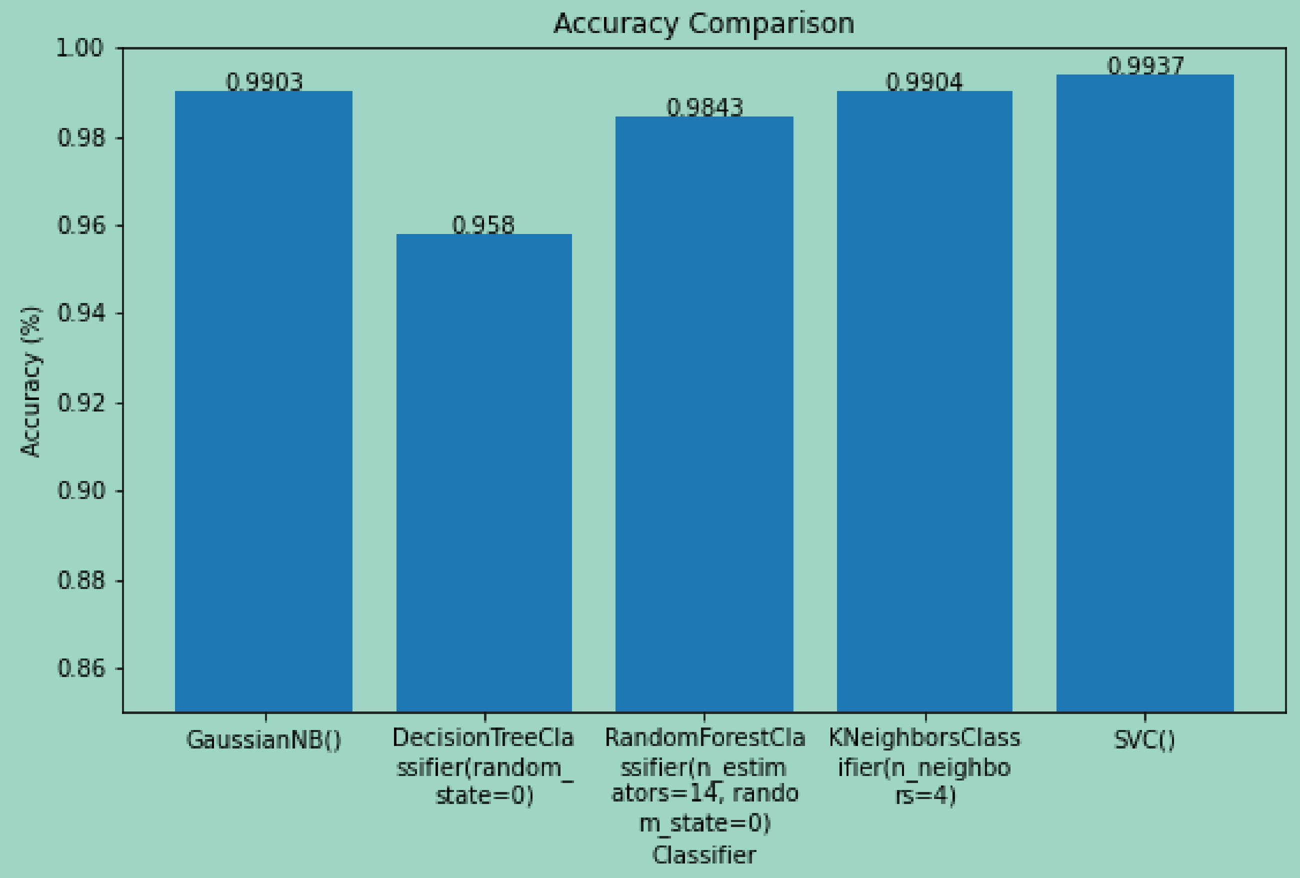


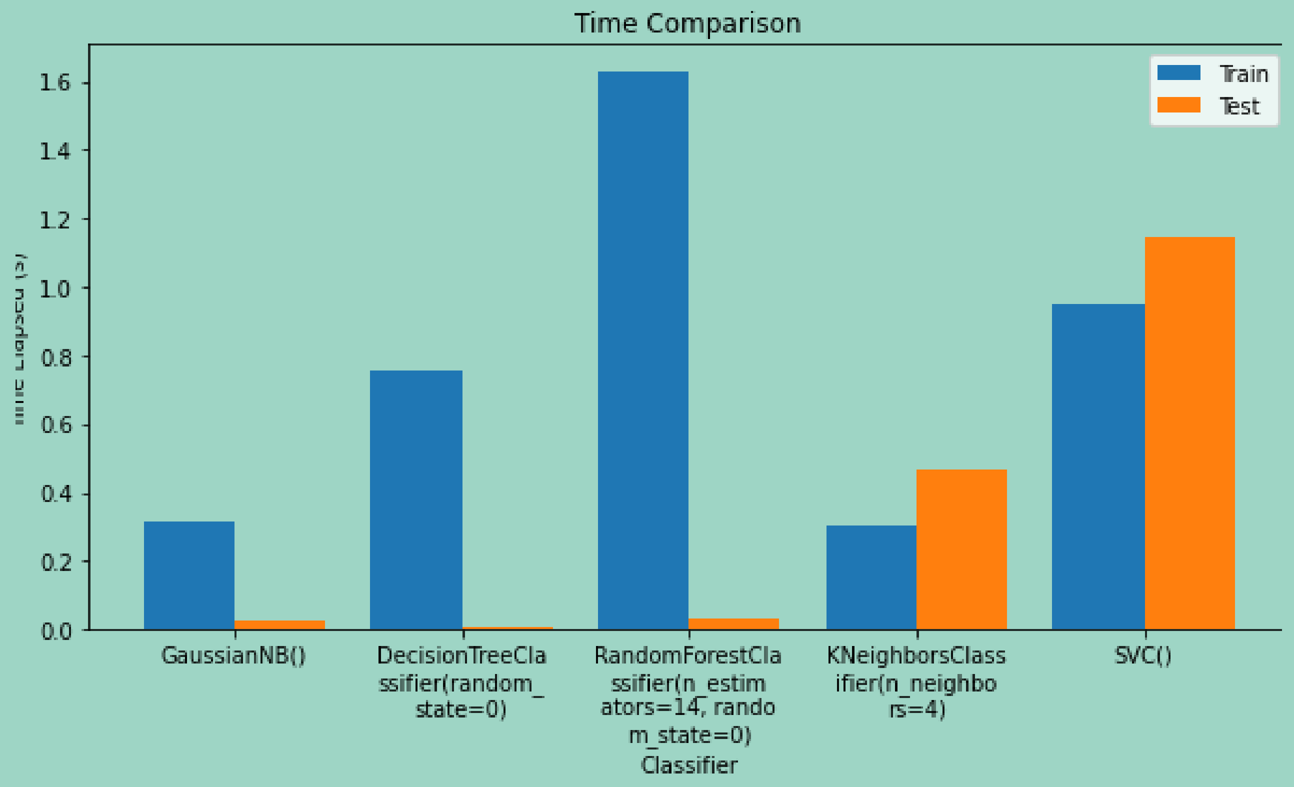
Comparison

The chosen model is

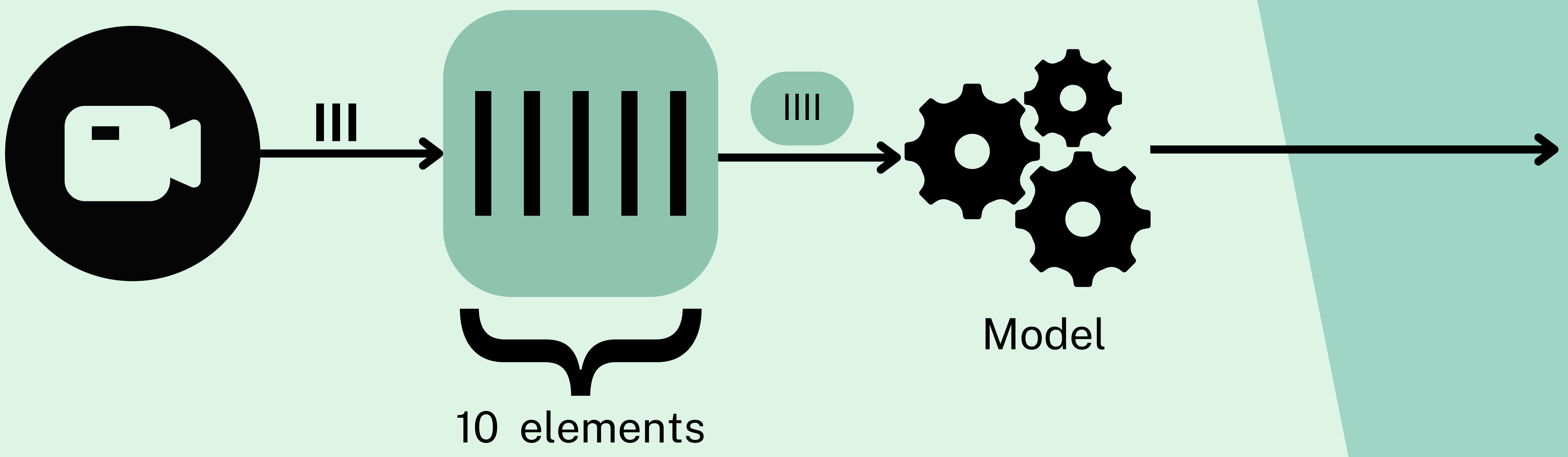
Gaussian Naive Bayesian Classifier

- For real-time application we need low latency:
  - eager learners are preferred to lazy learners.
- No need to store structures in memory unlike Decision Tree Classifiers or Random Forest.



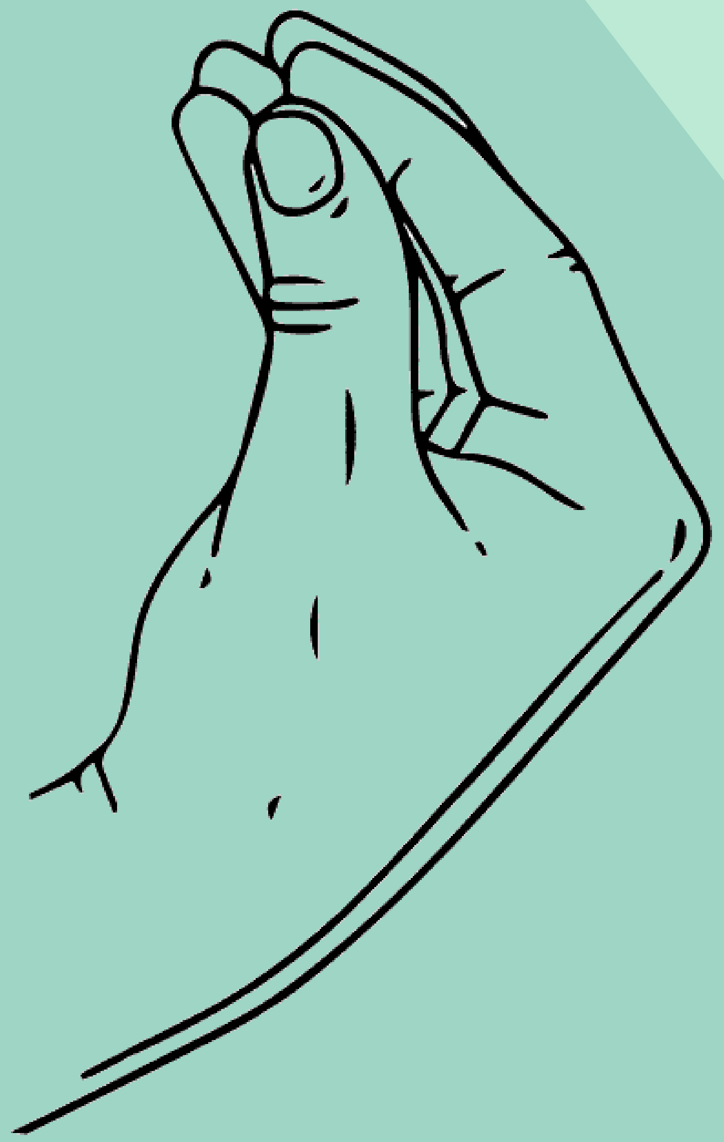


Real Time Usage



MOST COMMON PREDICTED VALUE

CANZONERI DANIELE
GRILLEA FRANCESCO



Live Demo

model chosen: Gaussian Naive Bayes