

Mini-Project (ML for Time Series) - MVA 2025/2026

Firstname Lastname youremail1@mail.com

Firstname Lastname youremail2@mail.com

November 30, 2025

What is expected for these mini-projects? The goal of the exercise is to read (and understand) a research article, implement it (or find an implementation), test it on real data and comment on the results obtained. Depending on the articles, the task will not always be the same: some articles are more theoretical or complex, others are in the direct line of the course, etc... It is therefore important to balance the exercise according to the article. For example, if you have reused an existing implementation, it is obvious that you will have to develop in a more detailed way the analysis of the results, the influence of the parameters etc... Do not hesitate to contact us by email if you wish to be guided.

The report The report must be at most FIVE pages and use this template (excluding references). If needed, additional images and tables can be put in Appendix, but must be discussed in the main document. The report must contain a precise description of the work done, a description of the method, and the results of your tests. Please do not include source code! The report must clearly show the elements that you have done yourself and those that you have reused only, as well as the distribution of tasks within the team (see detailed plan below.)

The source code In addition to this report, you will have to send us a Python notebook allowing to launch the code and to test it on data. For the data, you can find it on standard sites like Kaggle, or the site <https://timeseriesclassification.com/> which contains a lot of signals!

The oral presentations They will last 10 minutes followed by 5 minutes of questions. The plan of the defense is the same as the one of the report: presentation of the work done, description of the method and analysis of the results.

Deadlines Two sessions will be available :

- **Session 1**

- Deadline for report: December 14th (23:59)
- Oral presentations: December 15th and 17th (precise times TBA)

- **Session 2**

- Deadline for report: January 4th (23:59)
- Oral presentations: January, 5th and 7th (precise times TBA)

1 Introduction and contributions

The Introduction section (indicative length : less than 1 page) should detail the scientific context of the article you chose, as well as the task that you want to solve (especially if you apply it on novel data). **The last paragraph of the introduction must contain the following information:**

- Repartition of work between the two students
- Use of available source code or not, percentage of the source code that has been reused, etc.
- Use of existing experiments or new experiments (e.g. test of the influence of parameter that was not conducted in the original article, application of the method on a novel task/data set etc.)
- Improvement on the original method (e.g. new pre/post processing steps, grid search for optimal parameters etc.)

2 Methodology

2.1 Formation Change-Point Detection (FormCPD)

2.2 Role Change-Point Detection (RoleCPD)

The goal of RoleCPD [?] is to detect long-term tactical changes in player roles (e.g., a winger swapping sides with another winger permanently) while ignoring temporary switches (e.g., overlapping runs or covering defensive duties). This process operates within a single Formation Period (T_i) identified in the previous step.

2.2.1 Formal Representation

We define the inputs and mathematical framework based on the SoccerCPD protocol:

- **Input:** A sequence of "Temporary Role Permutations" $\{\pi_t\}_{t=1}^{|T_i|}$.
- **Role Permutation (π_t):** At every frame t , the Role Representation step assigns a role X_p to every player p . Since roles are distinct, this assignement is a permutation of the initial canonical role:

$$\beta_t(p) = \pi_t(X_p)$$

Where β_t is the player-to-temporary-role mapping at time t .

2.2.2 The Distance Metric

To detect a change, we must quantify the difference between the team's configuration at time t and time t' . Since the data consists of permutations (non-Euclidean), we cannot use standard Euclidean distance. We use the Hamming Distance normalized by the number of roles ($N = 10$ outfield players):

$$d(\pi_t, \pi_{t'}) = \frac{1}{N} \sum_{p \in P} \mathbb{1}_{\pi_t(X_p) \neq \pi_{t'}(X_p)}$$

This metric represents the "Switch Rate" or the proportion of players whose roles differ between two frames.

2.2.3 Change-Point Detection Algorithm

The paper utilizes Discrete g-segmentation, a graph-based change-point detection method effective for repeated observations in non-Euclidean space. The procedure is as follows:

1. **Preprocessing:** Calculate the Switch Rate relative to the dominant permutation. Exclude frames with a switch rate > 0.7 (likely temporary switches during set-pieces or abnormal situations).
2. **Segmentation:** Apply the detection algorithm recursively on the sequence of permutations using the Hamming distance.
3. **Significance Test:** A change point τ is significant if:
 - The p-value of the scan statistic is < 0.01 .
 - The segmentation duration is sufficient (robustness against noise).
 - The most frequent permutations (Instructed Roles) in the segments before and after τ must be distinct.

3 Data

The Data section (indicative length : 1 page) should provide a deep analysis of the data used for experiment. In particular, we are interested here in your capacity to provide relevant and thoughtful feedbacks on the data and to demonstrate that you master some "data diagnosis" tools that have been dealt with in the lectures/tutorials.

4 Results

The Result section (indicative length : 1 to 2 pages) **should display numerical simulations on real time series** (even if the original article was focused on images). If you re-used some existing implementations, it is expected that this section develops new experiments that were not present in the original article. Results should be discussed not only based on quantitative scores but also on qualitative aspects. In particular (especially if your article focuses on black box methods), please provide some feedbacks whether the method was adapted to the data or not and whether the hypothesis behind the approach you used were validated or not.