# Post-Doctorate position: Sequential set-valued learning and application to citizen sciences


## Supervisers


- **J. Salmon** (IMAG - Université de Montpellier),
- **M. Hebiri** (LAMA, Université Gustave Eiffel)


## Keywords

Supervised learning, set-valued classification, conformal prediction, citizen science, Pl@ntNet dataset

## In brief

The general idea is to exploit conformal prediction approaches to sequentially improve plant annotations for citizen sciences, especially for the Pl@ntNet app.

## Context

Set-valued classifiers are powerful alternatives to single outputs methods in multi-class classification; they aim at producing for each input $\mathbf{X}$ a set of possible labels $\Gamma(\mathbf{X}) \subset 2^{\mathcal{Y}}$ that possibly contains the label $Y$ of $\mathbf{X}$ and that has some appealing statistical properties [4, 8]. In this project we focus on citizen sciences application (especially on Pl@ntNet, see for instance [6]) where set-valued classification has successfully been applied [2, 5, 9]. In particular, the main goal of the project is to improve set-valued classifiers accuracy based on conformal prediction arguments resulting with a procedure that builds sequentially smaller and smaller set-valued classifier to end up with unambiguous outputs for all input images. An independent task could be also to investigate improved calibration strategies that could help the overall calibration [3, 7].

From the statistical point of view, the goal is to introduce a new statistical framework that fits well to the sequential learning of set-valued classifiers and to analyze the performance of the proposed algorithm in terms of a suitable measure of error. An important aspect of the project is that we shall consider several possible visualizations of the same image – possibly rotations and translations of the initial image – and one big challenge of the internship is to understand how to merge all this information to induce a reduction of the size of the set-valued classifiers. From the application side, this means that the user could be asked to take additional pictures of the plant that he wants to classify, whenever the uncertainty is too wide.

## Objectives

Several previous works focus on building for a given $\mathbf{X}$ either a smallest set-valued classifier with $1-\alpha$ coverage for small $\alpha \in (0,1)$, or the set-valued with the highest accuracy with a given (average) size $L$ [1, 4, 10]. One big challenge of this project is to extent the proposed methodology to sequential classification, where the marginal distribution of $Y|\mathbf{X}$ may evolve with time/iterations. The objectives are three-folds

1. characterize the optimal set-valued classifier in the context of sequential learning. An important step is to define a well suited loss function;
2. deduce an algorithm for this framework and implement it in `Python`;
3. study the theoretical performance of the proposed method and evaluate its numerical properties. In particular, evaluation on the subset of Pl@ntnet dataset available here would be key: https://www.imageclef.org/lifeclef/2015/plant.

## Practical information

- Location: Univ. Montpellier
- Salary: Minimum gross monthly salary, **2700** Euros (could be higher upon experience).
- Duration: **one-year position**, with a possible additional one-year extension./
- Deadline for applications: **Feb. 1, 2024**
- Expected starting date: **April 2024**
- Required skills: PhD in statistics/machine learning/optimization. Python programming.
- Grant: ANR Chaire IA Camelot

## Collaborators

- **A. Joly** (Inria Montpellier),
- **M. Servajean** (Univ. Paul Valery)

## Application

Interested candidates should send a CV along with contact information of two references to the PI. Any inquiries regarding the position can also be addressed via the same email address.

# References


[1] R. Barber, E. Candès, A. Ramdas, and R. Tibshirani. "The limits of distribution-free conditional predictive inference". *Information and Inference: A Journal of the IMA* 10.2 (2021), pp. 455–482.

[2] E. Chzhen, C. Denis, M. Hebiri, and T. Lorieul. "Review set-values". *ArXiv* (2021).

[3] L. Clarté, B. Loureiro, F. Krzakala, and L. Zdeborová. "Expectation Consistency for Calibration of Neural Networks". *UAI*. Pittsburgh, PA, USA: JMLR.org, 2023.

[4] C. Denis and M. Hebiri. "Confidence sets with expected sizes for multiclass classification". *Journal of Machine Learning Research* 18.102 (2017), pp. 1–28.

[5] C. Garcin, M. Servajean, A. Joly, and J. Salmon. "Stochastic smoothing of the top-K calibrated hinge loss for deep imbalanced classification". *ICML*. Vol. 162. 2022, pp. 7208–7222.

[6] C. Garcin et al. "Pl@ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution". *NeurIPS Datasets and Benchmarks 2021*. 2021.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Weinberger. "On calibration of modern neural networks". *ICML*. 2017, p. 1321.

[8] J. Lei. "Classification with confidence". *Biometrika* 101.4 (2014), pp. 755–769.

[9] T. Lorieul. "Uncertainty in predictions of Deep Learning models for fine-grained classification". Theses. Université Montpellier, Dec. 2020. URL: https://theses.hal.science/tel-03040683.

[10] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.