Extension of thesis Project: Exploring Model Stacking Methods for Prediction in Near-Infrared Spectroscopy

Keywords: Near-Infrared Spectroscopy, Machine Learning, Model Stacking, Artificial Intelligence, Calibration, Stacking

Host laboratory

UMR AGAP Institute, Avenue Agropolis - 34398 Montpellier Cedex 5 (https://umr-agap.cirad.fr/)

- **Thesis director:** Fabien Michel (HDR), LIRMM, Équipe SMILE
- **Supervisors:**

Grégory Beurier, CIRAD, AGAP Institute

Lauriane Rouan, CIRAD, AGAP Institute

Denis Cornet, CIRAD, AGAP Institute

Context

Near-Infrared Spectroscopy (NIRS) is a fast, non-destructive, and low-cost analytical technique widely used in various fields such as health, chemistry, agri-food, and particularly agronomy. It allows for determining the chemical composition and the functional properties of samples like grains, forage, food, and tissues. The spectral data generated by NIRS are rich in information but require advanced statistical processing for accurate predictions. Historically, methods like PLS regression have been used, but advances in machine learning (neural networks, SVM, random forests, etc.) and access to extensive NIRS databases have led to the growing adoption of these artificial intelligence methods, which often demonstrate better predictive performance.

The democratization of spectrometers and the increasing number of non-specialist users in both the North and South reinforce the need to develop a generic and efficient approach to NIRS model calibration. The stacking method, which combines predictions from multiple models, has shown potential for leveraging the complementary strengths of different algorithms to improve prediction performance. However, stacking strategies remain underexplored for NIRS data analysis. In this context, the Python package Pinard (a Pipeline for NIRS Analysis ReloadeD, https://pypi.org/project/pinard/) developed by the supervising team provides an ideal base for implementing and testing stacking-based prediction approaches.

Thesis Project Objectives

The main objective of this thesis is to develop and optimize stacking strategies suitable for prediction from NIRS spectra using the Pinard package. Pinard already provides tools for NIRS data processing and analysis, including individual predictive models, but currently does not offer model assembly methodologies. This research aims to fill this gap by integrating advanced stacking techniques to significantly improve predictive performance.

The thesis work will focus on the following axes (which may evolve during the doctorate and vary in importance):

1. **Axis 1**: Study and design data standardization methods to feed the different model classes in the stack, particularly considering machine learning model constraints or different sources. This work will also include a thorough analysis and handling of the available datasets.
2. **Axis 2**: Select, integrate, and hyperparameterize predictive models (existing or new) within a "traditional" stack and study the impact of each on overall accuracy depending on datasets and assembly methods (random selection, performance-based selection, algorithm diversity, prediction dissimilarity, etc.).
3. **Axis 3**: Design and explore effective strategies to improve model stacking strategies in terms of accuracy, efficiency, and frugality:
    - Heuristics from distributed artificial intelligence (multi-agent systems) or optimization (evolutionary methods)
    - Real-time calculation of model contribution and/or explainability
    - Dynamic organization and selection of data preprocessing
    - Partial real-time hyperparameterization
    - Etc.
4. **Axis 4**: Work on disseminating the obtained results by facilitating the reuse of the stack or access to tools and methods:
    - Transfer models to new analytes/datasets/machines
    - Study the underlying explainability of the stack models and identify signal components
    - Integrate developments into the Pinard package

This work will provide innovative and efficient approaches to exploit the richness of NIRS data, improving the accuracy and robustness of NIRS analyses for issues such as rapid identification of varieties adapted to climatic challenges, detection, and quantification of biotic and abiotic contaminants in crops, optimization of the quality and nutritional value of processed foods, etc., thus contributing to themes dear to CIRAD such as food security, sustainable resource management, and health improvement in Southern countries.

Materials and Methods

This project will rely on a diverse database already built, comprising 30 datasets from various fields, ranging from starch analysis in sorghum stems to active substance content in medications, fat content in meat, and octane rating of industrial fuels. This data richness offers a unique opportunity to test and apply new methodologies in a wide range of contexts. Special attention will be paid to the energy efficiency of computational processes, aiming to minimize

the environmental impact as a beneficial consequence of algorithm optimization. Efficiently identifying model combinations leading to the best predictions, contrary to traditional exhaustive exploration methods, helps minimize required resources (time, computation, energy), aligning with sustainable development and environmental responsibility.

Expected Results

- **Axis 1**: Study and develop effective data preprocessing methods integrated into Pinard to train all algorithm classes used in stacking on all available NIRS data formats.

- **Axis 2**: Efficient use of model assembly strategies and significant improvement in predictive performance compared to single models on reference data. Preconfigured stack models will be integrated into Pinard for various use cases.

- **Axis 3**: Explore and design methods to improve stacking efficiency (training time reduction, model number reduction, result improvement, etc.). Results will depend on the success of various developed and tested strategies, and these results will be integrated into the Pinard library.

- **Axis 4**: Beyond expected publications, dissemination levers for this work concern the usability (especially in the South) of developed methods and their transferability. Methods should be easily usable with Pinard and able to address diverse existing NIRS issues (lab data, field data, absorbance, reflectance, etc.). Special attention will be paid to the interpretability of stacking results.

Scientific and Material Conditions

This thesis project is funded through co-financing by AAP Region Emergence (50%) and Cirad (50%). Thesis operating costs are covered by ongoing projects. Calculations will be performed on local computers (Geforce(s) 4090) and the Jean Zay and Adastra clusters.

International Outreach

Participation in several international conferences will allow the PhD student to develop an international network. Collaborations with Southern partners involved in data acquisition for tropical species will also be considered.

Envisaged Collaborations

Within the UMR AGAP Institute, collaborations will be established with the Phenomen and GSP teams currently mobilizing deep learning approaches in genomic prediction. Internationally, many partnerships are possible to adapt Pinard developments to the needs of

Northern and Southern users (e.g., CNRS, INRAe, University of Potsdam, Boyce Thompson Institute, IITA, CNRA, NRCRI).

Valorization Objectives

Results will be valorized through publications in impact factor journals and presentations at international conferences. Optimizing a recent methodology will allow relatively easy valorization of research results highly anticipated in this field. A paper on the database used and its analysis is also envisaged, along with any other technical or methodological advances made during this thesis.

Desired Profile

- **Required Degrees**: Master's in computer science, bioinformatics, applied mathematics, statistics, or agricultural sciences with a data science specialization.
- **Required Skills**:
  - Python development
  - Data science and/or statistics
  - English (reading, writing, speaking)
  - French (basics)
  - Knowledge of R (optional)
  - Signal processing (optional)
  - Interest in interdisciplinarity

Continuous Training of the PhD Student

Necessary training will be chosen from catalogs offered by the University of Montpellier (Doctoral School I2S), AGAP Institute, Cirad, and INRAE according to the project's specific needs.

Thesis Project Progress Monitoring

An annual thesis monitoring committee will be organized to ensure the thesis proceeds smoothly and redirect certain parts if necessary. This committee will consist of members whose expertise areas relate to the subject matter (Deep Learning, signal processing, algorithms and computation, chemometrics, etc.). A monthly meeting with the thesis director and supervisors will discuss overall progress, obstacles encountered, and upcoming objectives. Weekly meetings with the supervisors will closely monitor ongoing work, provide technical support, and adjust work plans. A shared GitHub repository will be used to centralize and track work, facilitating collaboration, traceability of changes, and secure progress sharing.

Administrative Information

- **First-year doctoral enrollment academic year**: 2024
- **Thesis start date**: 01/12/2024
- **Application deadline**: 23:59, 22/09/2024
- **Required documents**: CV, cover letter, github/gitlab repository if available
- **Submit all documents to**: denis.cornet@cirad.fr

Contacts

• **Grégory Beurier** – beurier@cirad.fr
• **Lauriane Rouan –** lauriane.rouan@cirad.fr
• **Denis Cornet –** denis.cornet@cirad.fr
CIRAD Agricultural Research for Development
AGAP Mixed Research Unit "Genetic Improvement and Adaptation of Mediterranean and Tropical Plants".
Avenue Agropolis - 34398 Montpellier Cedex 5
France

• **Fabien Michel** – fmichel@lirmm.fr
LIRMM -Université de Montpellier - CNRS, Équipe SMILE
Montpellier, France
http://www.lirmm.fr/~fmichel

Relevant Publications of the Supervising Team

1. Vasseur F. et al. (2022). A Perspective on Plant Phenomics: Coupling Deep Learning and Near-Infrared Spectroscopy. https://doi.org/10.3389/fpls.2022.836488

2. Houngbo M. E. et al. (2023). Convolutional neural network allows amylose content prediction in yam (Dioscorea alata L.) flour using near infrared spectroscopy. https://doi.org/10.1002/jsfa.12825

3. Alamu E. O. et al. (2020). Near-infrared spectroscopy applications for high-throughput phenotyping for cassava and yam: A review. https://doi.org/10.1111/ijfs.14773

4. Sambakhé D. et al. (2019). Conditional optimization of a noisy function using a kriging metamodel. https://doi.org/10.1007/s10898-018-0716-0

5. Bonnici I. et al. (2022). Input addition and deletion in reinforcement: towards protean learning. https://doi.org/10.1007/s10458-021-09534-6