Use of convolutional neural network to predict yam (*D. alata*) tuber amylose content from near infrared spectra

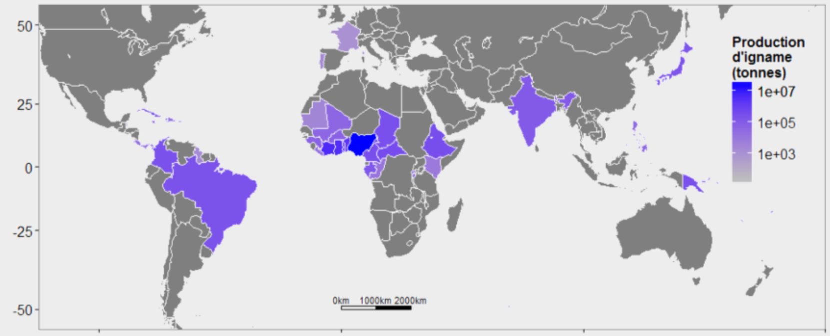
Houngbo M. E.[1], Desfontaines L.[2], Mestres C.[3], Davrieux F.[3], Meghar K.[3], Arnau G.[1], Irep JL.[4], Marie-Magdeleine C.[5], Rouan L.[1], Beurier G.[1], Cornet D.[1]

1. CIRAD UMR AGAP
2. INRAE UR ASTRO
3. CIRAD UMR QUALISUD
4. INRAE UE PEYI
5. INRAE URZ

June 2021

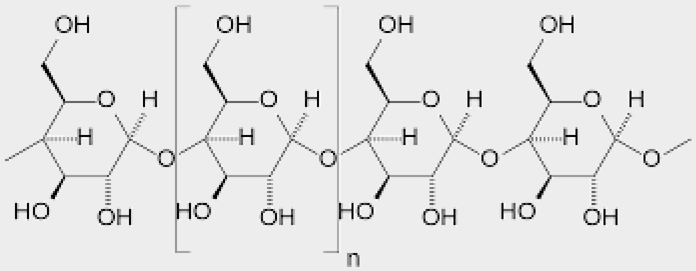
Context

- Yam importance
  - 4th most cultivated root tuber
  - Cultivate in intertropical zones
  - 60 million people's staple food
- Consumption mode
  - Boiled
  - Pounded
- Varieties from breeding programmes not widely adopted because quality not acceptable



Context

- Yam composition: starchy (80% of dry matter)
  - Amylose & Amylopectin
  - Affects starch viscosity and friability of yam products
- NIRS can help to predict tuber quality
  - Amylose is difficult to predict by NIRS for RTB
  - Mostly C-H bonds $(C_6H_{10}O_5)_n$
  - Multiple wavelengths involved and not well known
- Two unsuccessful attempts to predict amylose content using NIRS with PLS for yam
  - R^2=0,27 (Alamu et *al.*, 2019)
  - R^2=0,18 (Lebot and Malapa, 2009)



Context

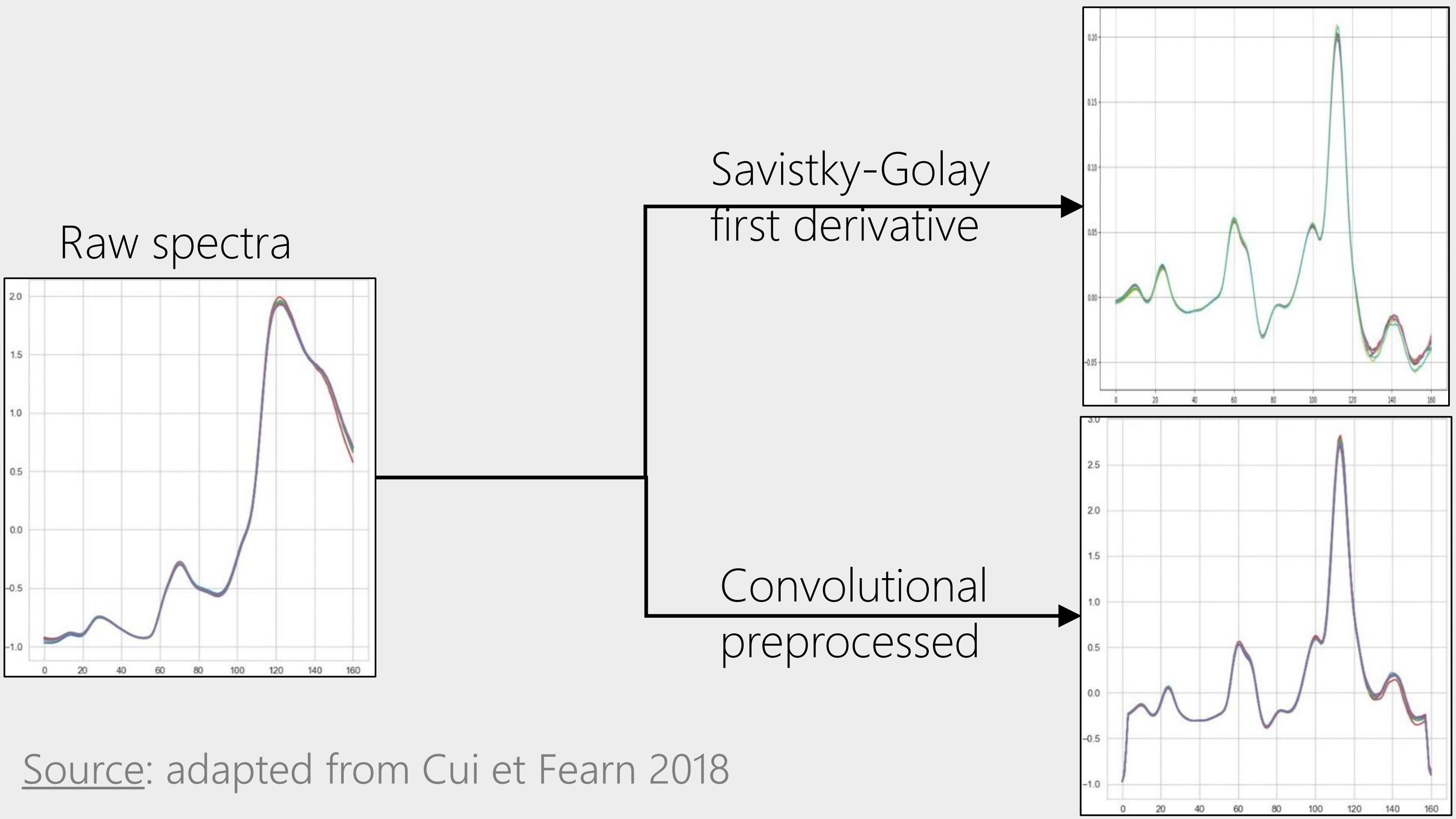
- PLS: Partial Least Squares
  - "Linear"
  - Loss of information
    - Reduction of dimensions (loss of part of the information: 1050 variables -> 2-48 components)
    - Applies only 1 pretreatment combination (optimal but incomplete) => loss of noise but also loss of information
  - Sensitive to outliers and especially spectral outliers
    - Spectral => arbitrary suppression of spectra based on distances (Euclidean, Mahalanobis…) unrelated to the information carried
    - Risk of loss of information

Context

- AI: Artificial Intelligence / DL: Deep learning
  - Management of overfitting designed in advance as inherent to DL methods
  - Noise is information: all features and spectral outliers are useful
  - Data augmentation is more efficient as the introduction of noisy spectra does not "harm" the performance of the algorithm
  - No need to choose between combinations of pretreatments (APA)

Context

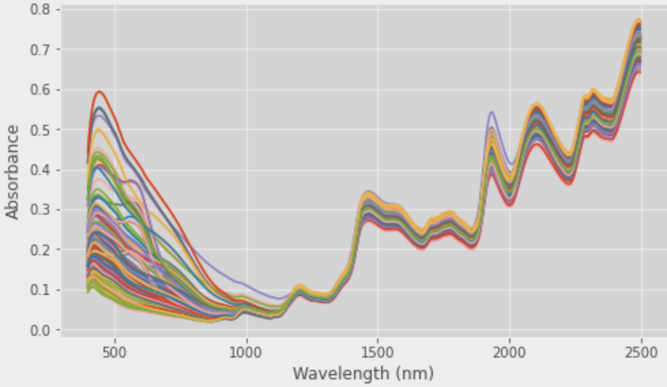
- CNN: Convolutional Neural Network
  - Reduce noise due to measurement conditions of spectra (convolutional layer acts like super-pretreatment)



Source: adapted from Cui et Fearn 2018

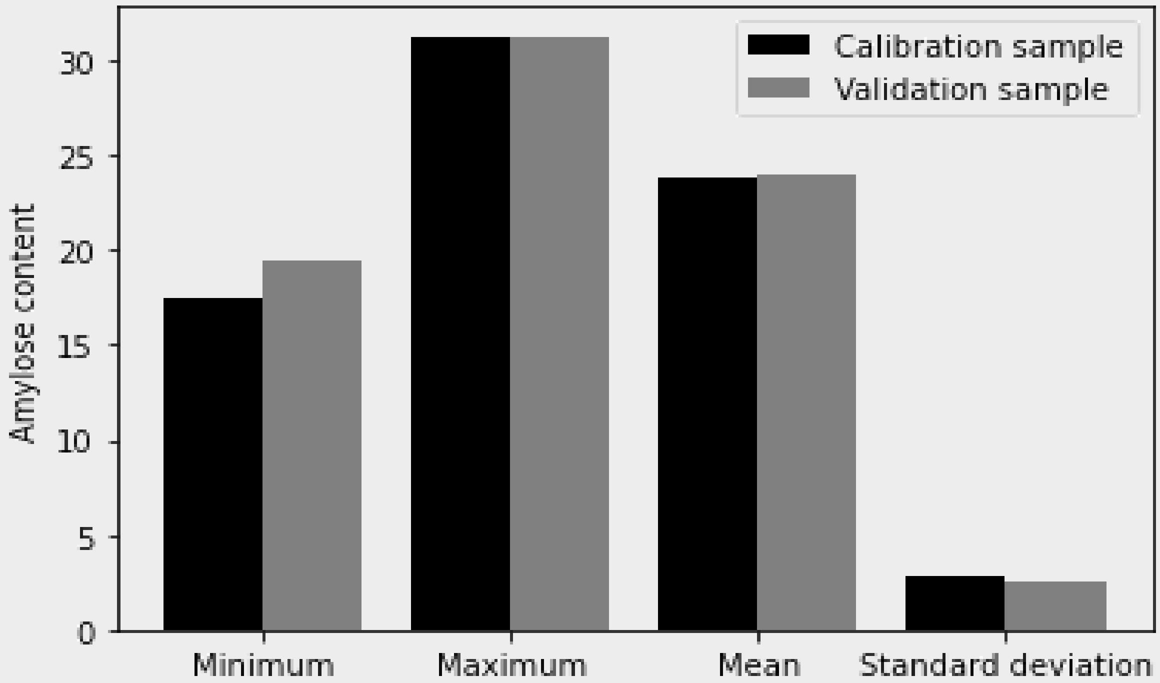
Methods

- Sample preparation
  - 21 genotypes (*D. alata*)
  - Peeled tubers, dried, ground, sieved
  - 93 samples
- Reference measurements of amylose (INRAE, UR Astro)
  - Colorimetry adapted from ISO-6647 and calibrated with DSC measurements
- NIRS measurement
  - FOSS NIRsystems 6500 (INRAE, URZ)
  - 1050 absorbance values from 400 to 2498 nm
  - 2 repetitions per sample (186 spectra)



Methods

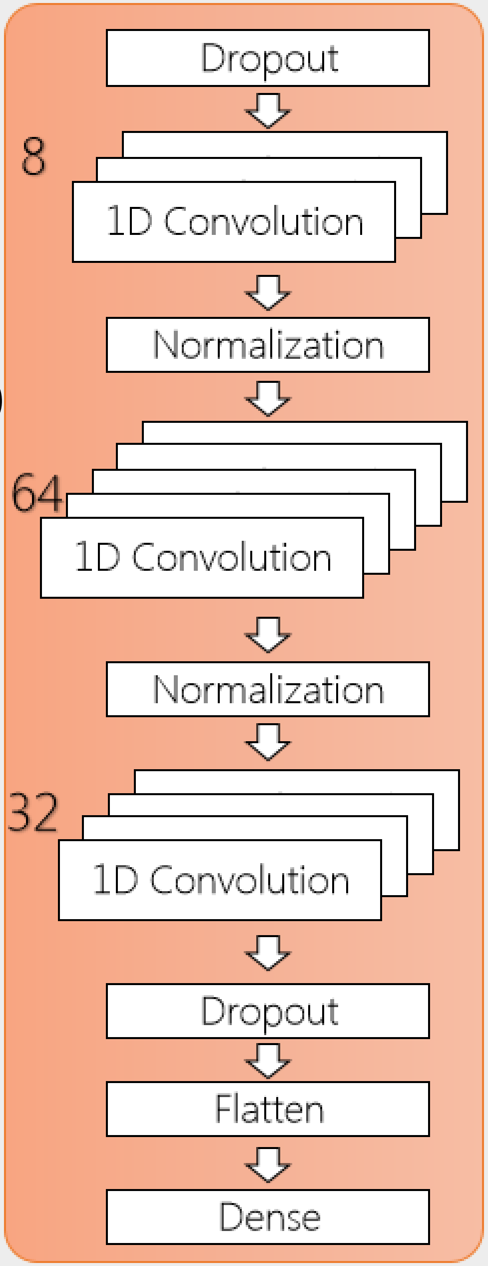
- Raw spectra +12 pretreatments used (gaussian, SavGol, MSC, SNV, Haar...) and their combination two by two => 157 possible datasets
- Separation into calibration (3/4) and validation (1/4) sets with Kennard-Stone



Methods

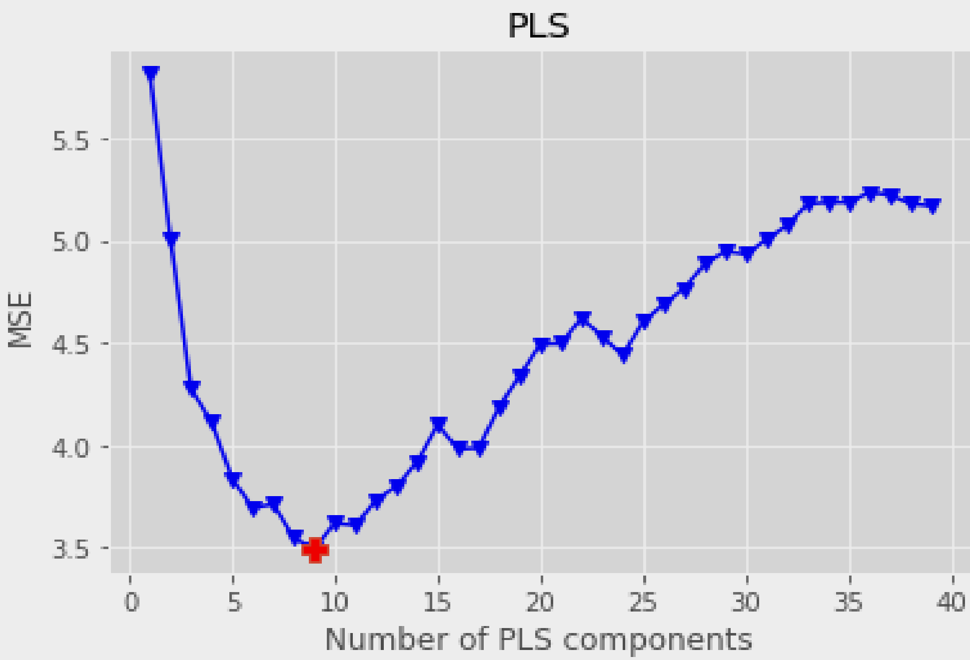
- PLS
  - Cross-validation for
    - the number of components to retain (up to 40)
    - the best combination of pretreatment (among the 157)
  - Python and *scikit-learn*
- CNN
  - Python, *keras*, *tensorflow*
  - Feature augmentation: 2nd order pretreatment combinations (157 data sets)

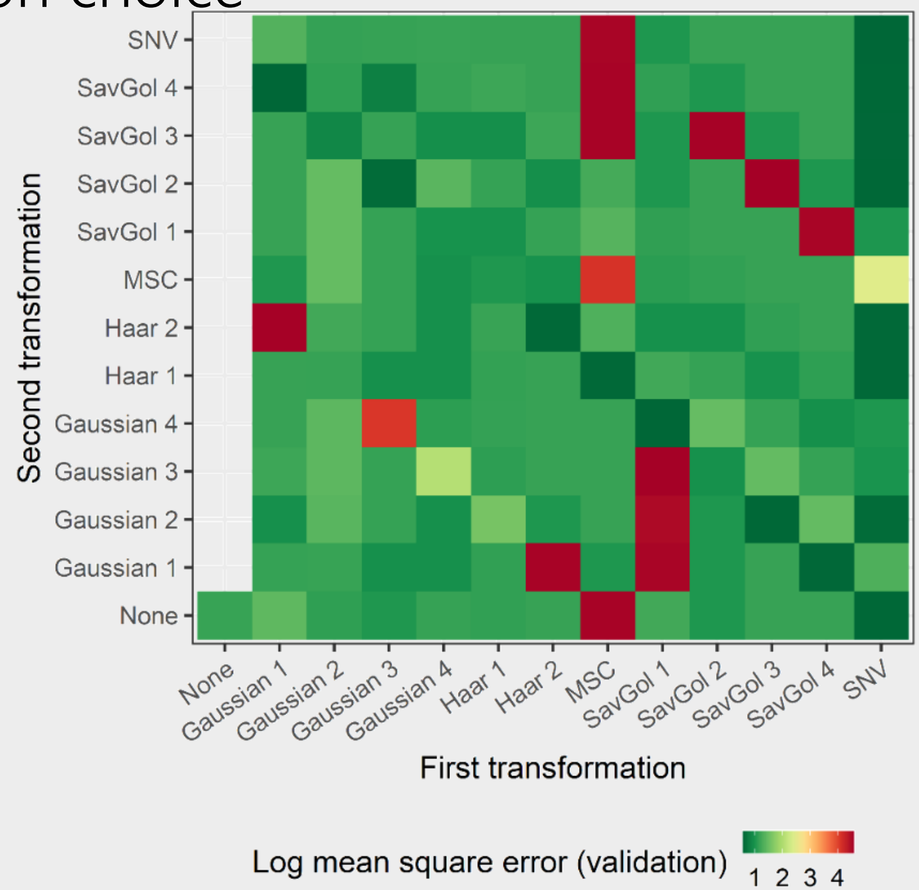
  => 157*1050=164850 features
  - Data augmentation and noise generation (140x5=700 synthetic spectra)



Results

- PLS optimization by cross-validation
  - Number of principal components
  - Pretreatment combination choice

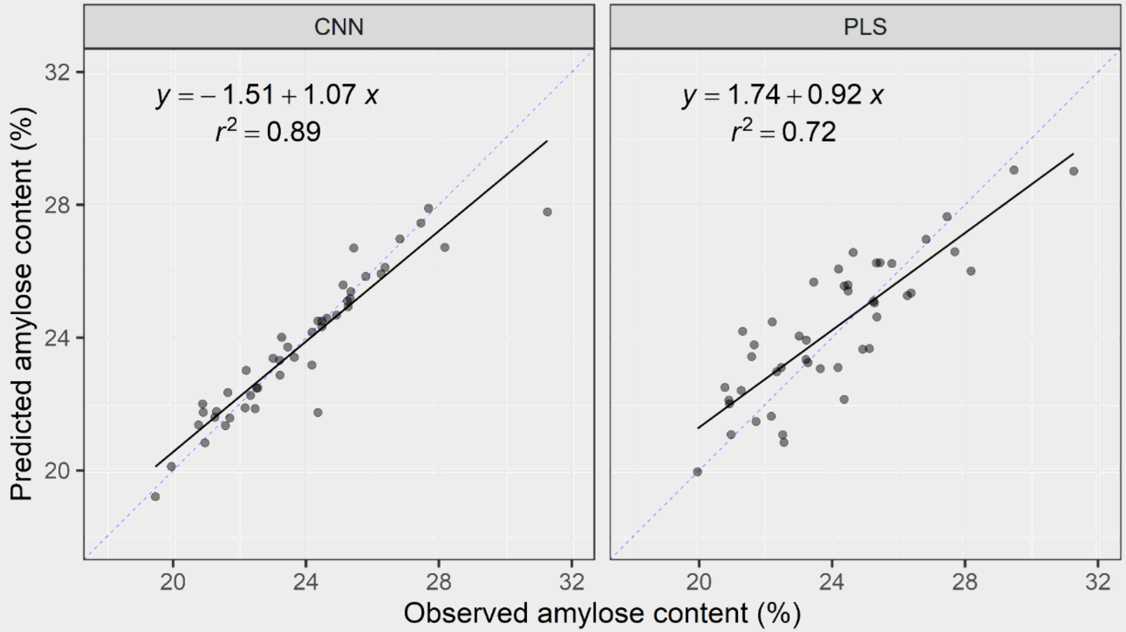




Results

- Comparaison PLS – CNN performance during validation step

Model	SEc	RMSEc	RMSEv	R^2v	RPD
PLS (Gaussian 1 + SavGol 4)	2.84	1.09	1.33	0.72	2.13
CNN	2.84	0.18	0.81	0.88	3.49



Perspectives

- External validation to test robustness (in progress)
- Tansfer learning
- Data augmentation using Variational AutoEncodeur (VAE) and Conditional Variational AutoEncodeur (CVAE)
- Model ensembling