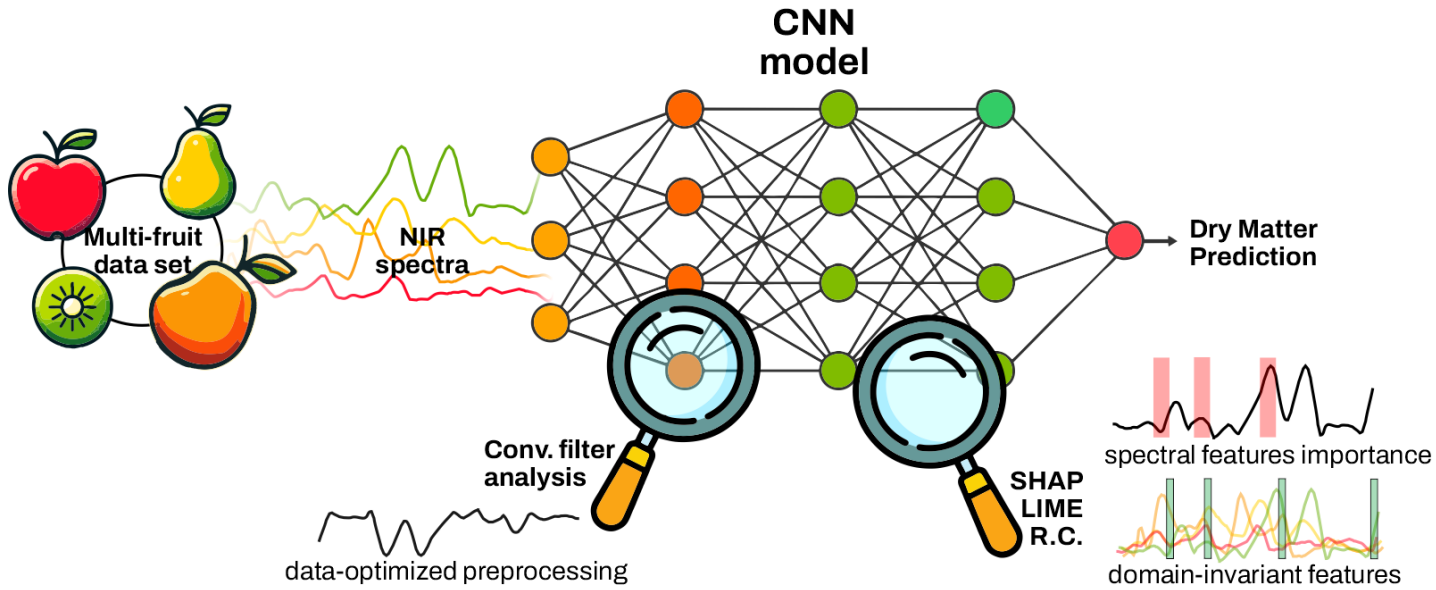
enbis European Network for Business and Industrial Statistics

ENBIS Spring Meeting 2025, Coimbra

Quality by Design (QbD) & Process Analytical Technology (PAT): Statistical, AI, and "Grey" Approaches



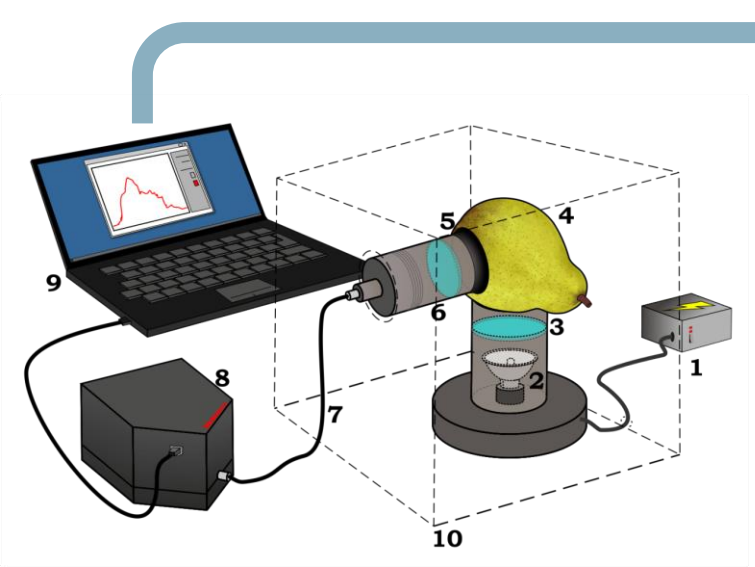
Exploring CNN architectures for NIR based chemometric tasks: the "Deep Tutti-Frutti" research case

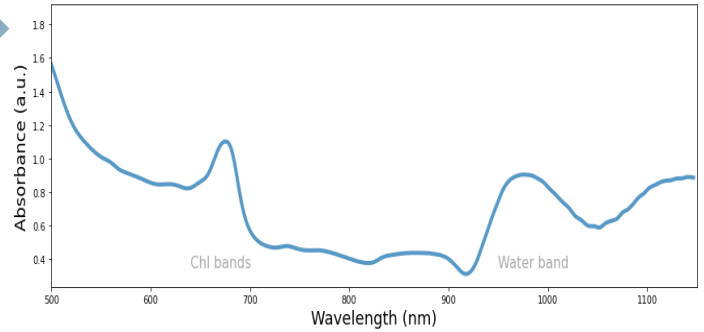
Author: Dário Passos

UAlg
UNIVERSIDADE DO ALGARVE
PHYSICS DEPARTMENT

Chemometric applications in the food industry (Fruit as an example)

NIR + Chemometrics → Non-destructive, cheap methodology and fast acquisition times are attractive points for food quality control





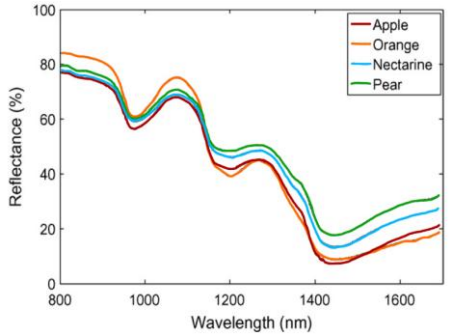
PREDICTED PROPERTIES:

- Dry matter content
- SSC (Soluble Solid Content)
- Firmness, Acidity, Protein, etc…
- Origin control

VARIABILITY SOURCES in DATA:

- Biological variability (chemistry)
- Harvest season (weather, fertilization, etc.)
- Measurement conditions (temperature, spectrometer, etc.)
- NIR (mostly overtones), low signal

Typical chemometric analysis pipeline



→ **Preprocessing** (SNV, MSC, 1d, 2d, smoothing, etc.)

→ **Model**
PLS*
SVM
PCR
Other ML
CNN

→ **Predicted Quantity** (regression or classification)

Literature supported claims of CNN advantages in NIR Chemometrics

CNN tend to outperform many traditional (linear) methods (PLS, SVM, MLR, etc.) especially in the presence of complex spectra (e.g. biological samples, **food products**, etc.).

The 4 reviews below present several tens of examples…

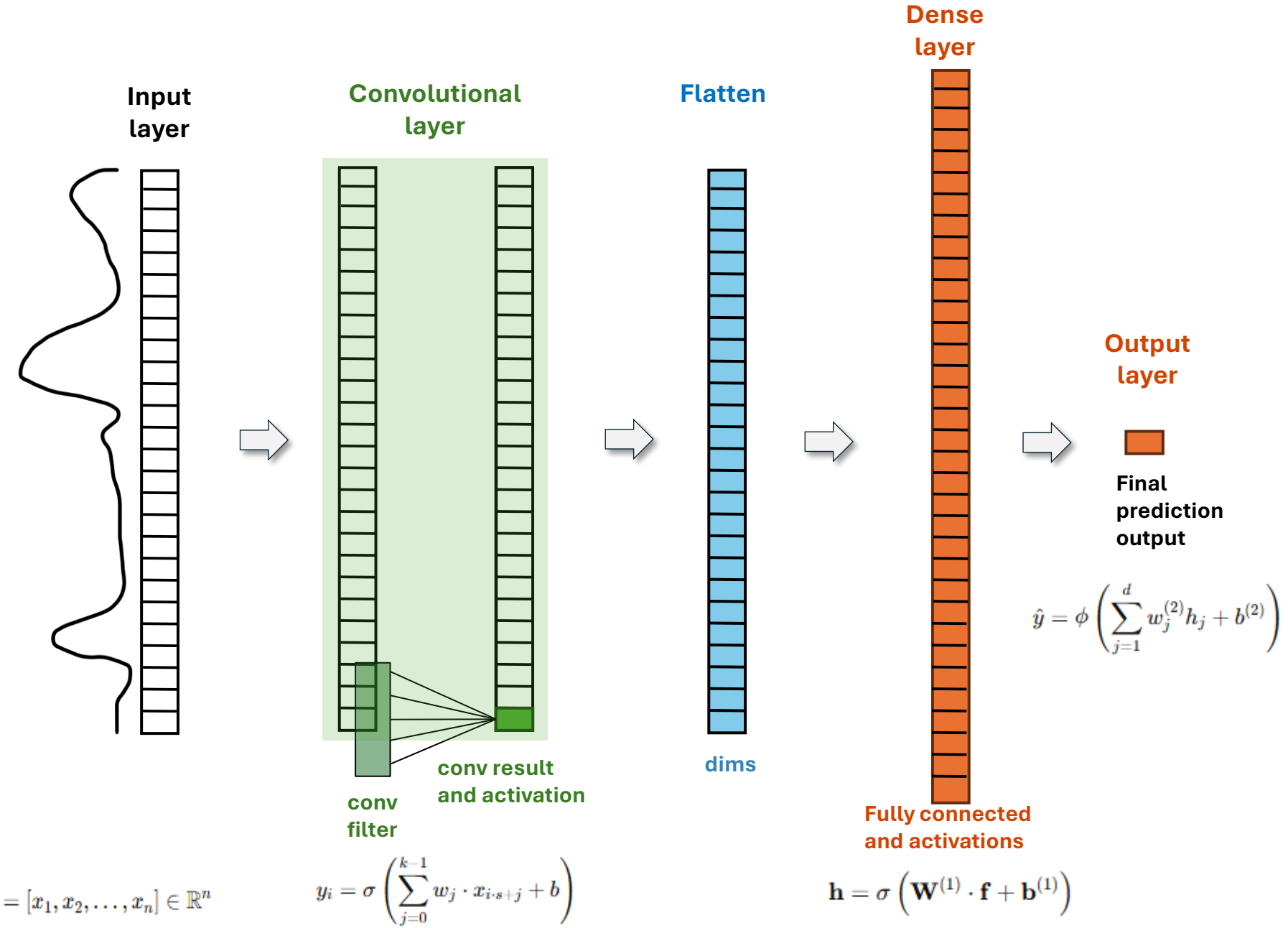
REF "X. Zhang et al 2021, Food and agro-product quality evaluation based on spectroscopy and deep learning: A review, Trends in Food Science & Technology Vol. 112, 431-441, https://doi.org/10.1016/j.tifs.2021.04.008"

REF "P. Mishra et al 2022,Deep learning for near-infrared spectral data modelling: Hypes and benefits, Trends in Analytical Chemistry, Vol 157, 116804, https://doi.org/10.1016/j.trac.2022.116804"

REF "J. Walsh et al 2023, Review: The evolution of chemometrics coupled with near infrared spectroscopy for fruit quality evaluation…, JNIR 2023, Vol. 31(3) 109–125, https://doi.org/10.1177/09670335231173140 "

REF "X. Zhang et al 2024, Advanced chemometrics toward robust spectral analysis for fruit quality evaluation, Trends in Food Science & Technology Vol. 150, 104612 https://doi.org/10.1016/j.tifs.2024.104612 "

Convolutional Neural Networks in 1D



The task: Dry Matter (DM) content in fruit

- DM is closely related to the quantity of sugar and starch content and therefore an overall indicator of fruit's quality
- Vis-NIR spectrometry and Chemometrics is nowadays widely used to assess DM in several fruits

  Walsh, K. *et al* 2020, PBT 168
- The most promising models for this task are based on convolutional neural networks (CNNs) with one to three conv. layers.

  Walsh, J. *et al* 2023, JNIRS Vol.3
- Current CNN downsides are related with model complexity and lack of a standard architecture for the task (results are usually data set dependent)

  Mishra, P. *et al* 2022, TrAC 157

  Luo, N. *et al* 2024, JFCA 128

The question

Can we develop a CNN architecture for the task of DM prediction that is data set / fruit independent?

Hypothesis

"If we train a CNN model on a broad data set of different fruit, the intrinsic variability of the data set will eventually lead to a DL model that can generalize better. This means that we should be able to create a global CNN for the task to DM prediction. "

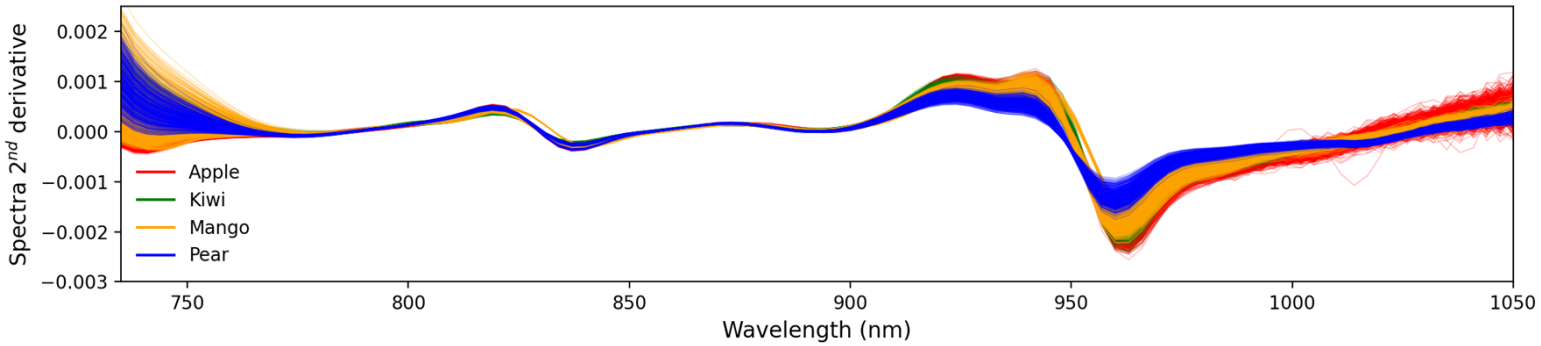
Study case: Deep tutti-frutti I and II

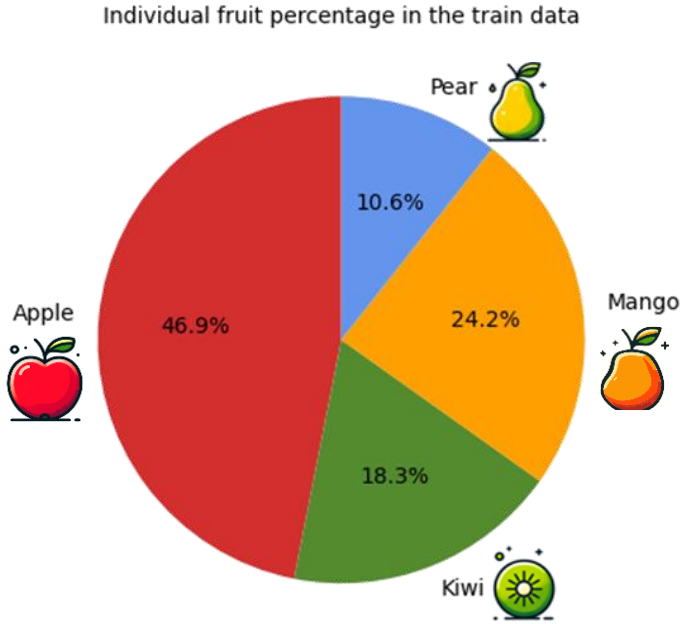
Dry matter prediction on a multi-fruit data set using CNNs

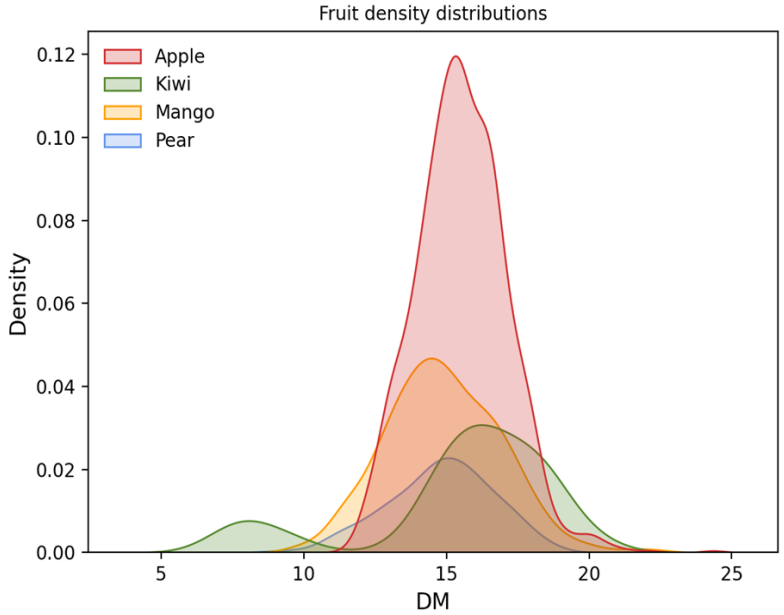
Passos, D., Mishra, P. **2023**, Chemo. Int. Lab. Syst. 243
D. Passos **2025**, Spec. Acta A: Mol. Biomol. Spec 337, 5, 126068

Multifruit dataset

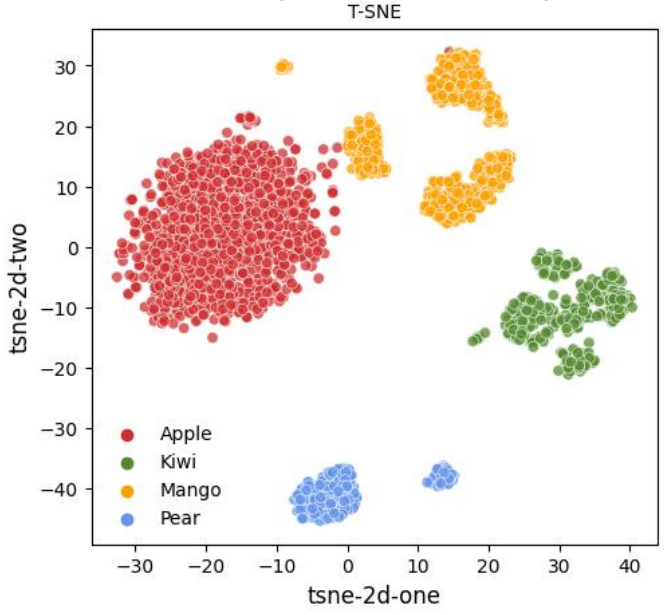
- 2nd derivative Vis-NIR spectra from: **Apple**, **Kiwi**, **Mango**, and **Pear**
  Range: 735 nm to 1050 nm from different Felix F750 spectrometers
  n = **2997**
- Dry matter content







t-distributed **S**tochastic **N**eighbour **E**mbedding

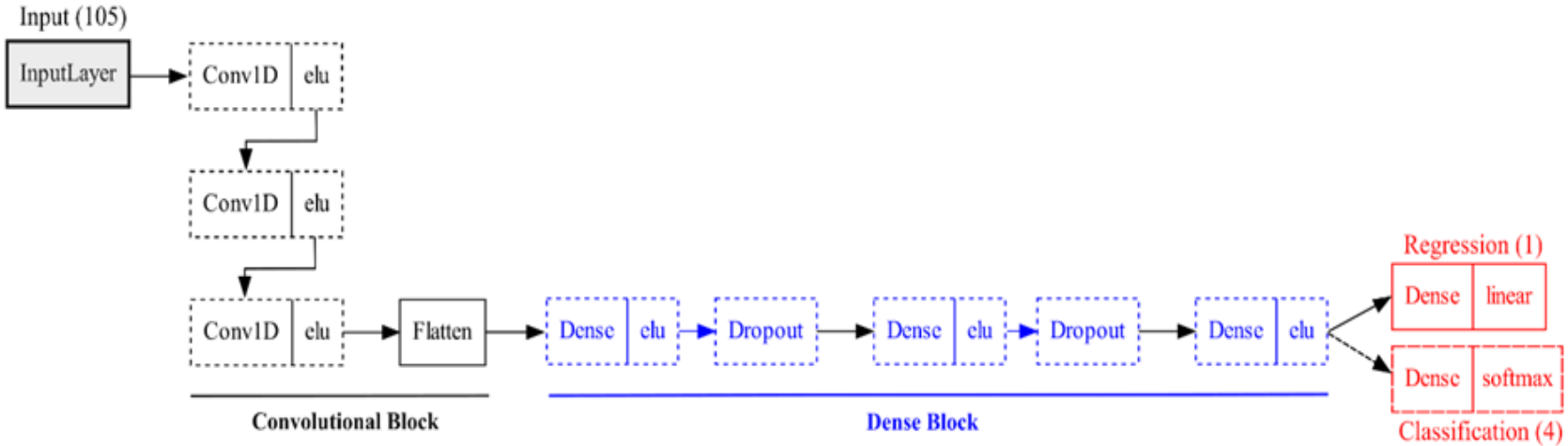


CNN optimization for 2 architectures: Regression (DM), Regression (DM) + Classification (Fruit)

Develop a hyperparameter optimization (HPO) and neural architecture search (NAS) pipeline that allows to find promising CNN architectures using:

Zela, A. *et al* 2018, ArXiv abs/1807.06906

- ❖ Randomization of initial conditions of training (different random weights initialization)
- ❖ Different training strategies (cross-fruit validation and 5-fold cross-validation)
- ❖ Several initial base architectures were explored allowing for a more guided HP optimization
- ❖ Additional validation on external data



Fixed hyperparameters:

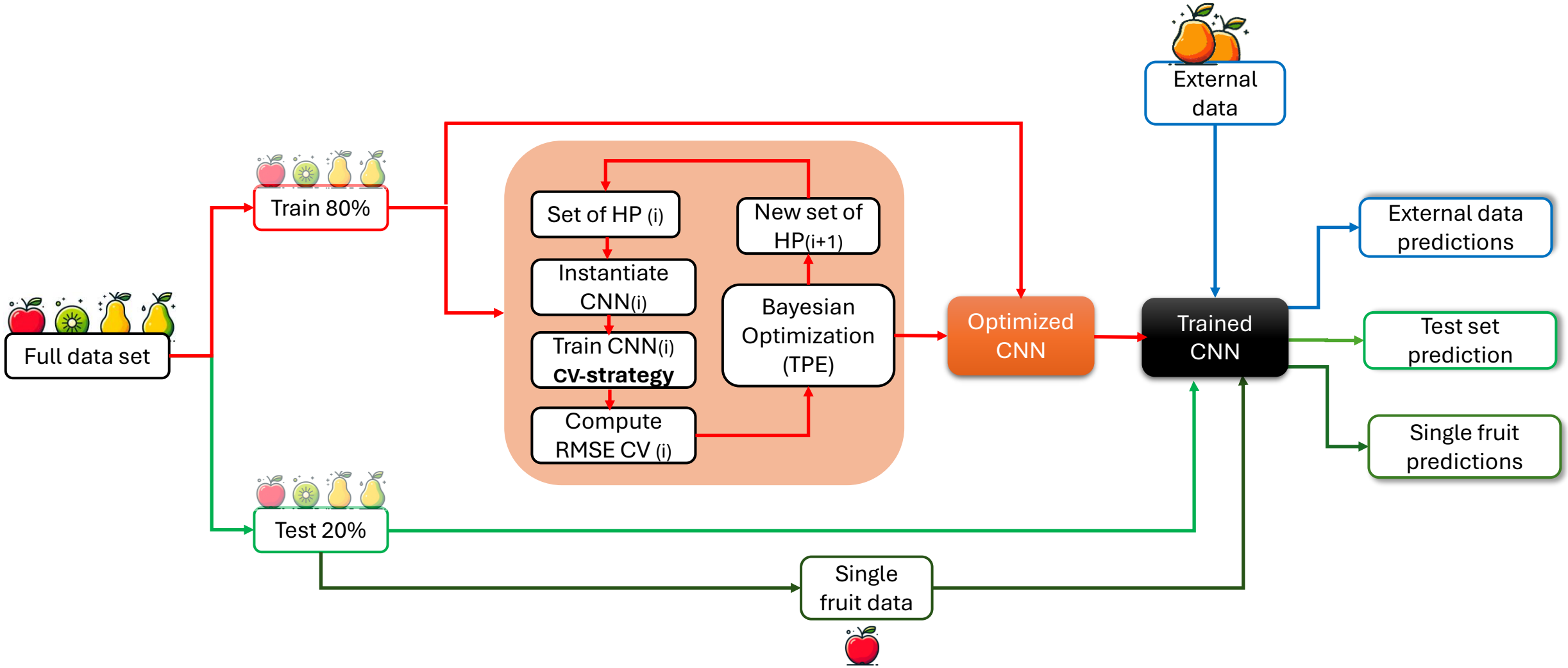
- Activation functions (ELU)
- Conv. filter stride = 1
- Padding = 'valid'
- Weight initialization = 'he_normal'

Optimizable hyperparameters:

- Number of Conv. layers (1-3)
- Number of conv. filters
- Width of conv. filters
- Number of Dense layers (1-3)
- Number of units in dense layers
- Use dropout regularization
- Dropout rate
- Strength of L2 regularization
- Mini-batch size

Models implemented in python using tensorflow.keras and optimized using the Optuna library

CNN NAS+Hyperparameter Optimization Loop



Training strategies for HPO

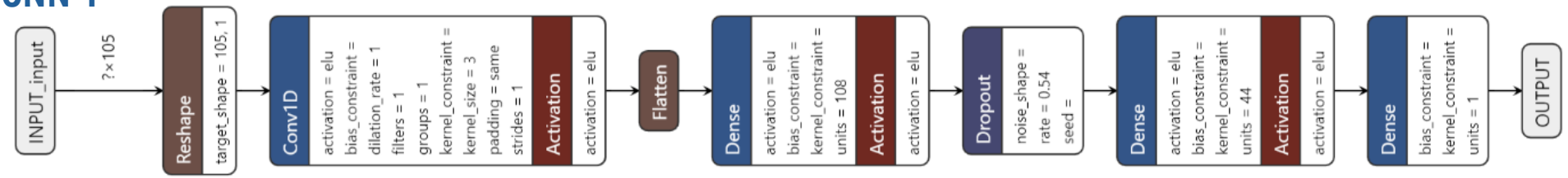
Strategy 1 – Cross-fruit validation, fixed seed in weights initialization
Strategy 2 – 5-fold cross validation, fixed seed in weights initialization
Strategy 3 – 5-fold cross validation, random seed inweights initialization

9 different CNN architectures generated

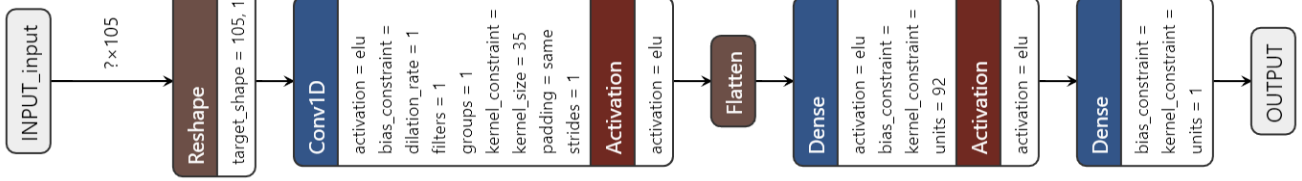
CNN NAS and HPO done for different objective generated several (9) architectures

Different number of layers, different number filters, etc…

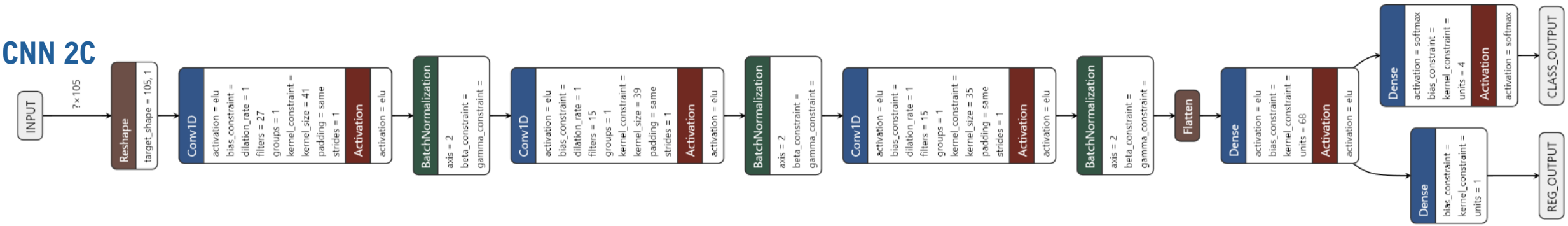
CNN 1



CNN 1B



CNN 2C



Models performance similar with little variance between results but enough to rank them (best RMSE = 0,61 ; PLS RMSE = 0,89)

Code available at: https://github.com/dario-passos/DeepLearning_for_VIS-NIR_Spectra

Additional chemometric models for comparison

- PLS (LV = 7), optimized using 5-fold cross validation
- LW-PLS (LV = 13 and λ = 0.5), optimized using 5-fold cross validation

Results (multifruit test set)

Model	HPO objective function	Single model on Train Set		Single model on Test Set		Average of 10 models on Train Set		Average of 10 models on Test Set		Ensemble of 10 models on Test Set	
	RMSECV	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE ± std**	**R^2**	**RMSE ± std**	**R^2**	**RMSE**	**R^2**
CNN-R_v1	1.109	0.98	0.794	0.99	0.785	0.818 ± 0.003	0.856	0.791 ± 0.004	0.863	0.791	0.863
CNN-R_v1B	0.665	0.613	0.919	0.619	0.916	0.629 ± 0.009	0.915	0.631 ± 0.013	0.913	0.602	0.921
CNN-R_v1C	0.658	0.607	0.921	0.625	0.914	0.635 ± 0.023	0.913	0.643 ± 0.023	0.909	0.604	0.92
CNN-R_v1D	0.649	0.612	0.919	0.611	0.918	0.624 ± 0.035	0.916	0.622 ± 0.027	0.915	0.596	0.922
CNN-R_v1E	**0.643**	**0.565**	**0.931**	**0.605**	**0.920**	**0.549 ± 0.009**	**0.935**	**0.614 ± 0.011**	**0.917**	**0.595**	**0.922**
CNN-R_v1F	0.985	0.843	0.847	0.861	0.837	0.774 ± 0.053	0.871	0.771 ± 0.052	0.869	0.699	0.893
CNN-RC_v2	0.664	0.614	0.919	0.625	0.914	0.632 ± 0.008	0.914	0.634 ± 0.008	0.912	0.599	0.921
CNN-RC_v2B	0.659	0.611	0.920	0.618	0.916	0.620 ± 0.012	0.917	0.635 ± 0.015	0.912	0.599	0.921
CNN-RC_v2C	0.645	0.589	0.925	0.612	0.918	0.565 ± 0.030	0.931	0.695 ± 0.027	0.894	0.612	0.918
PLS (LV = 7)	0.914	0.904	0.824	0.892	0.826						
LW-PLS (LV = 13 and $\lambda = 0.5$)	0.699	0.515	0.943	0.687	0.897						

Accuracy for class of fruit ~ 99% for all classification models

Results (single fruit test set)

Global CNN (trained on multifruit data) vs **Global PLS** (trained on multifruit data) vs **Individual PLS** (trained on individual fruit data)

Model	Apple		Kiwi		Mango		Pear	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Model CNN-R_v1E	0.580	0.857	**0.736**	**0.951**	**0.557**	**0.917**	**0.565**	**0.887**
Global PLS (7)	0.745	0.821	1.328	0.878	0.694	0.872	0.944	0.693
Individual PLS	**0.574**	**0.861**	0.908	0.927	0.678	0.879	0.616	0.872
#LV = (5, 8, 7, 7)								

Results (2 external mango data sets)

1448 samples (Australian mango data set, season 2018)

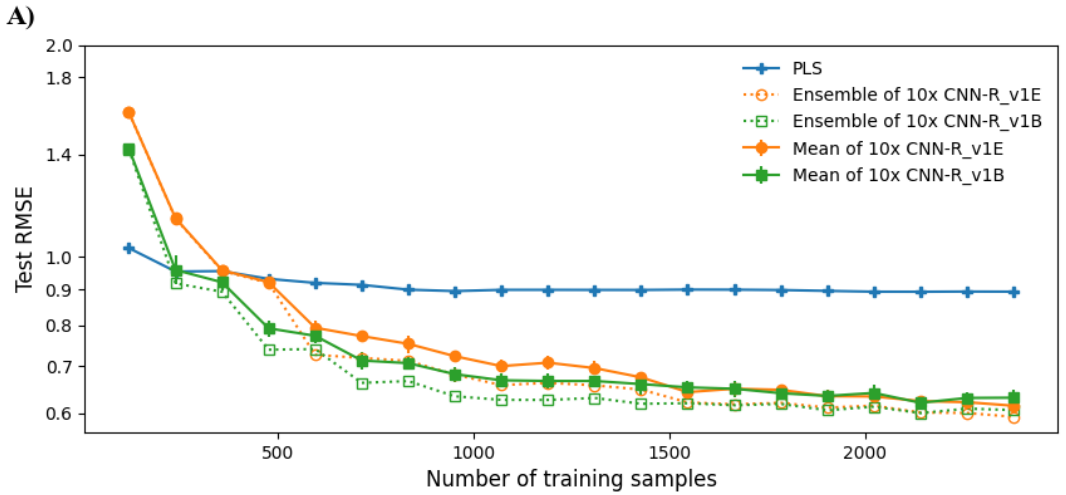
Anderson, N., *et al* 2020, PBT 168

965 samples (mango from Brazil, season 2023)

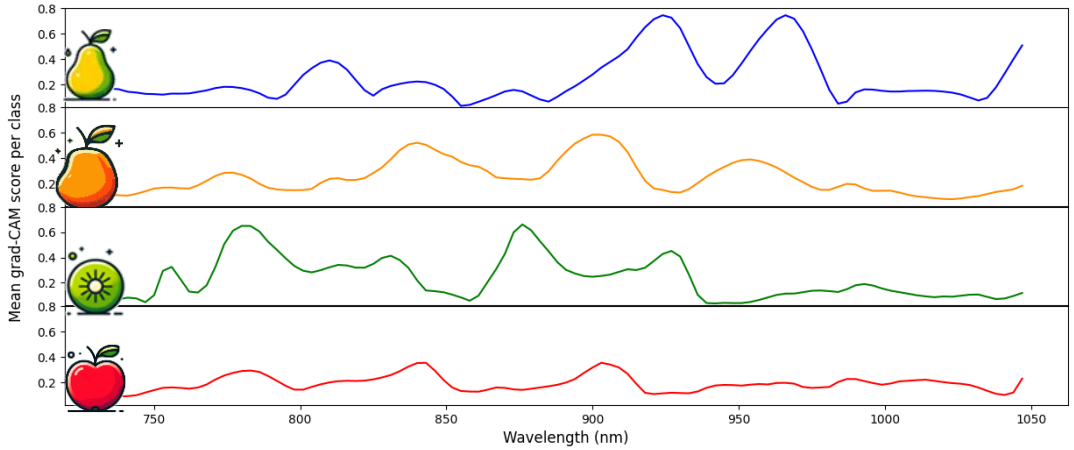
Puneet Mishra, Wageningen University and Research

Model	Single model on Anderson's mango test set		Average of 10 models on Anderson's mango test set		Ensemble of 10 models on Anderson's mango test set		Single model on new Puneet mango data set		Average of 10 models on new Puneet mango set		Ensemble of 10 models on new Puneet mango set	
	RMSE	**R2**	**RMSE ± STD**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE ± STD**	**R2**	**RMSE**	**R2**
CNN-R_v1B	1.172	0.807	1.214 ± 0.097	0.792	1.173	0.807	1.309	0.616	1.311 ± 0.095	0.613	1.286	0.631
CNN-R_v1C	1.186	0.803	1.419 ± 0.271	0.707	1.349	0.744	1.416	0.551	1.280 ± 0.105	0.630	1.242	0.654
CNN-R_v1D	1.215	0.793	1.395 ± 0.056	0.726	1.358	0.741	1.276	0.635	1.319 ± 0.066	0.609	1.286	0.630
CNN-R_v1E	1.358	0.741	1.315 ± 0.110	0.756	1.284	0.769	1.201	0.677	1.247 ± 0.036	0.651	1.234	0.659
PLS (LV=7)	1.331	0.803					1.535	0.512				

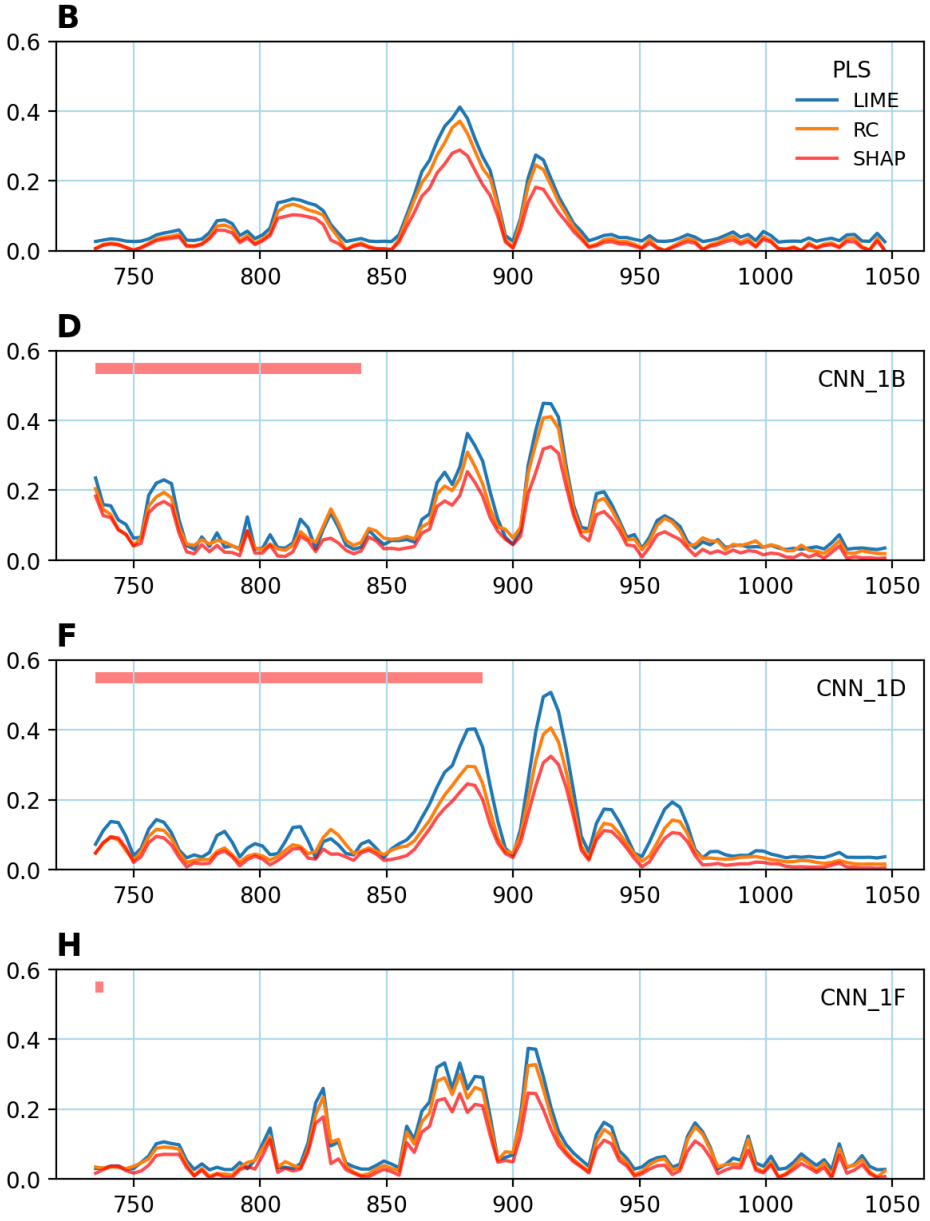
Model performance vs training set size



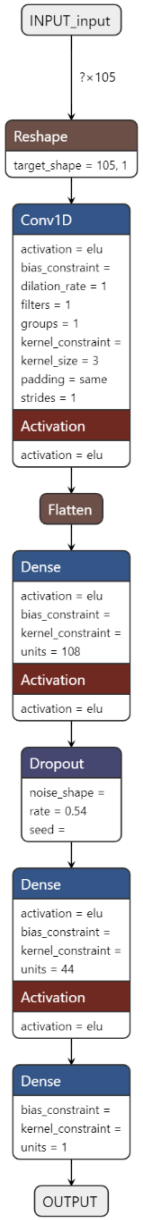
Interpretability: GradCAM scores per fruit class (CNN1B)



Interpretability (Shapley values, LIME, RC)

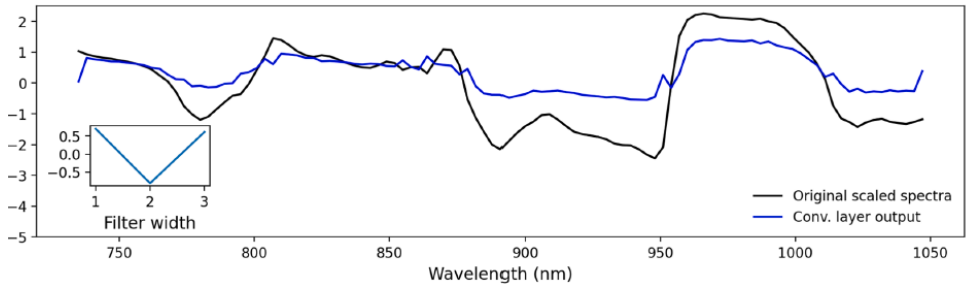


CNN filter analysis: same architecture (CNN1/B) optimized under CV strategies



Cross fruit validation

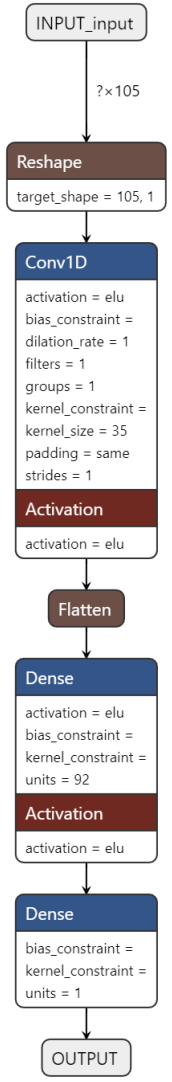
Calibration **Validation**



Filter width = 3, the CNN relied in narrow bands (less relevant band overlap between Cal and Val samples)

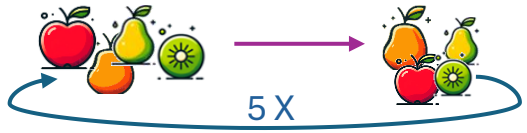
Use of **Dropout layer (0,54)** and **L2 regularization = 0,0735**

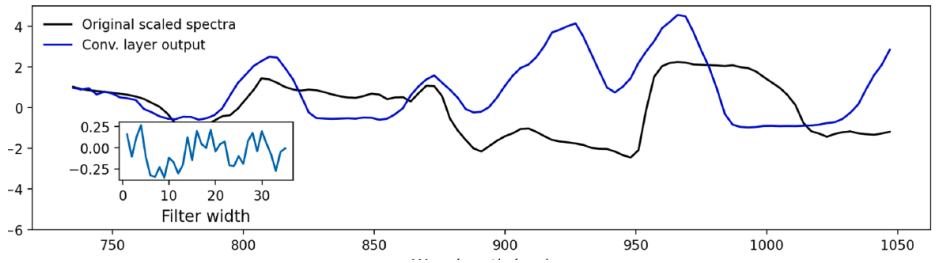
Need to use **two Dense layers**, indicates more difficulty to combine features from Cal samples to predict Val samples



5-fold cross validation

Calibration **Validation**





Filter width = 35, the CNN relied in wider spectral regions and resembles a high-pass/edge detection filter that tends to emphasize transitions or amplitude changes

No Dropout layer and lower **L2 regularization = 0,0295**

Only one Dense layer used

CNN explainable “advantages” in NIR (supported by the literature)

i) Convolutional layers perform a “sort of” data-driven optimized preprocessing of the input data and sometimes even mimic classical preprocessings (derivatives, smoothing, etc.)

ii) In heterogeneous datasets (e.g. multiple batches, samples types, etc.) CNNs tend to rely on robust features across sample groups (informative variable selection across different sample domains)

iii) Certain types of peak shifts (e.g. related to temperature fluctuations) are easier to account for due to filter operations or the use of pooling layers.

iv) In deeper models, CNNs learn hierarchical abstractions by nonlinearly combining low-level features. With piecewise-linear activations (e.g. ReLU), certain architectures effectively implement a piecewise-linear fit in the transformed feature space.

v) Multitask predictions and easier fusion of data from different modalities (e.g. spectra + image).

“Cui & Fearn 2018, Chem. Intel. Lab. Sys.182 (2018) 9–20, https://doi.org/10.1016/j.chemolab.2018.07.008 “

Filters works as 1st derivative preprocessing

“D. Passos 2025, Spec. Acta A: Mol. Biomol. Spec 337, 5, 126068, https://doi.org/10.1016/j.saa.2025.126068 “

Some filters work as symmetric 2nd derivative, others as data-specific hierarchical feature extractors…

“X. Zhang 2020, Analytica Chimica Acta 1119, 41e51, https://doi.org/10.1016/j.aca.2020.03.055 “

Some filters work as baseline removal…

“J. Martins et al 2023,Postharv Bio. Tech, 199, 112281, https://doi.org/10.1016/j.postharvbio.2023.112281 “

Predict temperature to enhance SSC predictions…

Conclusions

PROS

- Optimized global CNN models (different architectures) provide lower RMSECV in DM prediction than global PLS and global LW-PLS
- CNN model trained for regression and classification present 99% accuracy in fruit classification without notable decrease in performance on the regression task.
- Global CNN (trained on multifruit data) surpasses local PLS model in 3 out of 4 fruit sub-sets
- The optimized CNNs show good generalization when applied to truly external data sets (2 different mango spectral samples).
- Performance of CNNs (on regression) surpass PLS when the number of training samples increases above 400 ~ 500.
- Interpretability is possible using techniques such as Shapley values, Grad-CAM scores, etc.

CONS

- Optimization is computationally demanding
- Still no silver bullet, i.e., we still haven't found the perfect CNN architecture for this type of task (assuming it exists!)
- Shallow models, maybe not complex enough to capture all the variability in this type of multi-fruit data set?

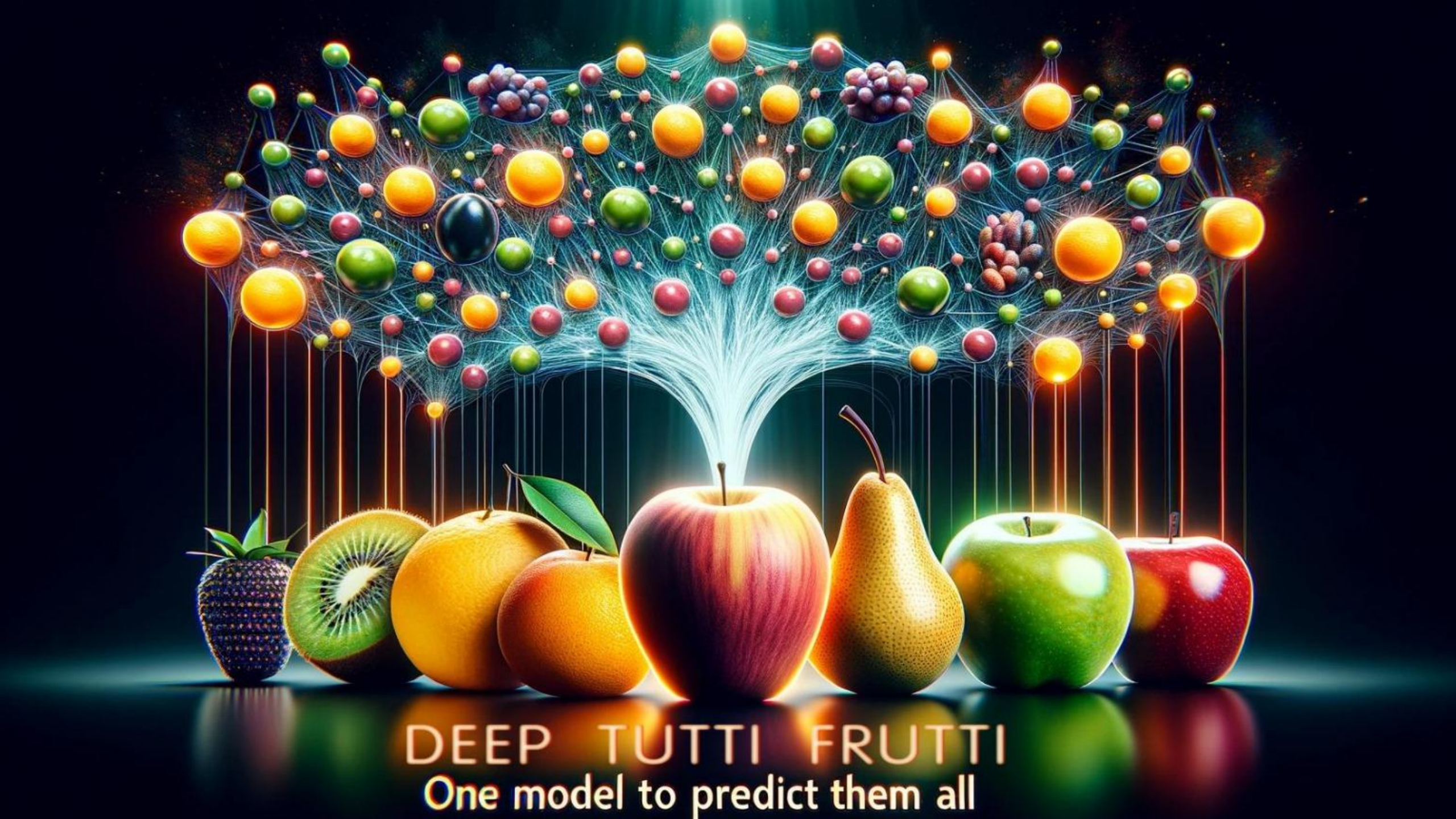
Code available at: https://github.com/dario-passos/DeepLearning_for_VIS-NIR_Spectra

sensAIfood project (IG 19145)

Artificial intelligence methods for spectral data processing to solve food fraud and authenticity issues

https://www.uco.es/investigacion/proyectos/sensaifood/

https://www.cost.eu/actions/IG19145/


DEEP TUTTI FRUTTI
One model to predict them all