Tutorial 1 : Machine Learning with Orange

"Radiomics and AI in Molecular Imaging" - COMULIS Training school

Nicolas Captier & Louis Rebaud

LITO (U1288) Université Paris Saclay/Inserm/Institut Curie, Orsay, France

October 12, 2021

Medical University Vienna (MUW)

Institute For Applied Medical Physics (IFAMP)

COMULIS

COST

A few words about us

- Louis Rebaud, PhD student
- Université Paris-Saclay, Siemens Healthineers
- louis.rebaud@curie.fr

- Nicolas Captier, PhD student
- Université PSL, PRAIRIE
- nicolas.captier@curie.fr

"Whole-body biomarkers in Positron Emission Tomography (PET) imaging."

"Prediction of the response to immunotherapy in lung cancer with multi-modal data."

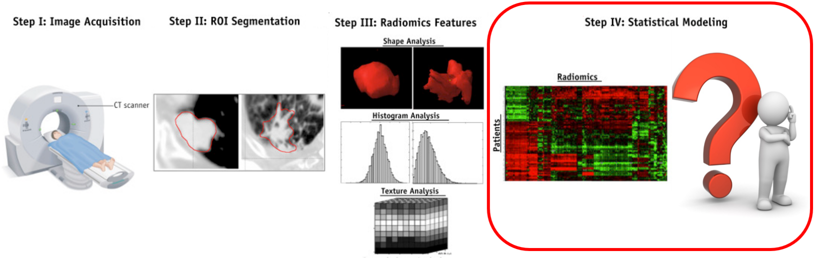
Subjects of interest

Medical imaging, radiomics, supervised learning, model interpretation and validation, encoding of medical data (linear methods, deep encoders...), inference with limited samples size...

Do not hesitate to contact us if you have any question !

Objectives of this tutorial

Once we have collected a sufficiently large radiomic data set, what can we do with it ?

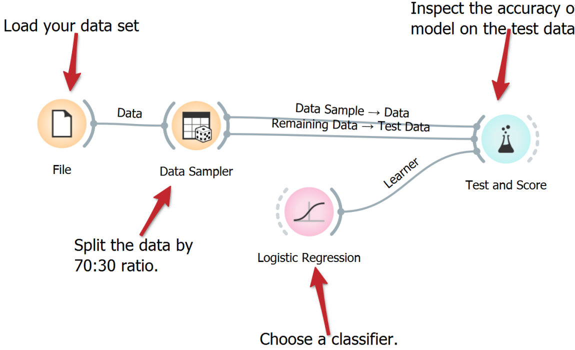

Step I: Image Acquisition
CT scanner
Step II: ROI Segmentation
Step III: Radiomics Features
Shape Analysis
Histogram Analysis
Texture Analysis
Step IV: Statistical Modeling
Radiomics
Patients

- Introduce you to a **free software** to perform data analysis with **no programming skills** required
- Give you **a bunch of good practices** when tackling a machine learning project.

This is just a starting point that will help you confidently exploring and analyzing your own data !

A few words about Orange

- Orange is an **open-source software** that allows you to build **complex and interactive** data analysis workflows.
- It provides the user **a wide range of computing tools** that can be manipulated and linked together **very intuitively**.

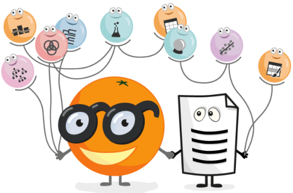


Load your data set

Inspect the accuracy of the model on the test data

Split the data by 70:30 ratio.

Choose a classifier.



We advise you to go exploring the vast and very instructive documentation on the Orange website.

Let's dive into our example !

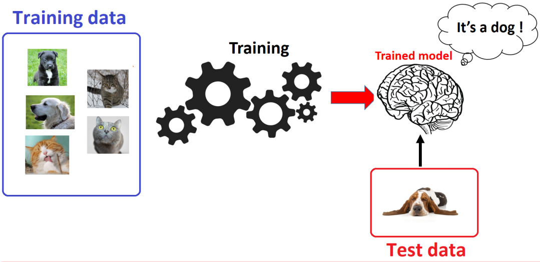
You first need to gather as much information as possible about your data set and clearly state the question to be addressed !

- **Mono-center** cohort of **378 patients** diagnosed with **primary lung cancer**.
- **FDG PET/CT images** acquired 60 ± 5 min after injection (Siemens Biograph 6 LSO or General Electric Discovery690).
- Each lung lesion was semi-automatically defined on PET images with a **threshold of** 40% **SUVmax**.
- **43 PET textural features** computed with the LIFEx software.

Can we predict the histological subtype among primary lung cancers (adenocarcinomas vs squamous cell carcinomas + others) from PET images ?

[0] The data set comes from "Ability of FDG PET and CT radiomics features to differentiate between primary and metastatic lung lesions" - Kirienko et al. 2018

A critical step : saving aside test data before the analysis



We want our trained model to perform well on new unseen data !

How to evaluate the ability of our final model to generalize well when we have a single data set as input ?

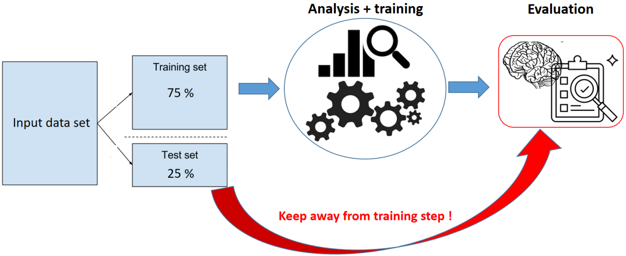
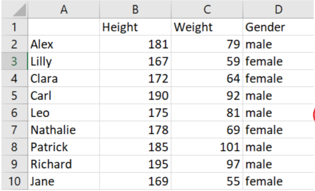


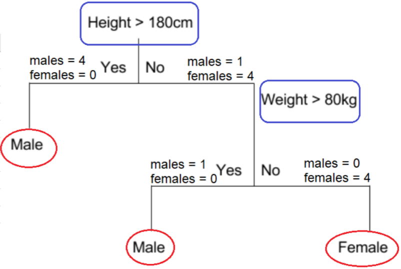
Figure 1: Train-test split pipeline (75% – 25%)

A first model : Decision Tree classifier - I

Do not use a function without having at least a rough idea of how it works !



	A	B	C	D
1		Height	Weight	Gender
2	Alex	181	79	male
3	Lilly	167	59	female
4	Clara	172	64	female
5	Carl	190	92	male
6	Leo	175	81	male
7	Nathalie	178	69	female
8	Patrick	185	101	male
9	Richard	195	97	male
10	Jane	169	55	female



- **Iteratively** selects the combination of feature and threshold (ex : Height, 180cm) that **best splits** the data into two distinct classes.
- The algorithm stops when all the leaves contain only one class or when a terminal condition is reached (ex : maximum depth, minimum number of samples for a node...).

A first model : Decision Tree classifier - II

List the advantages and the drawbacks of the method to see if it suits your data and identify what you should pay attention to.

Some advantages :

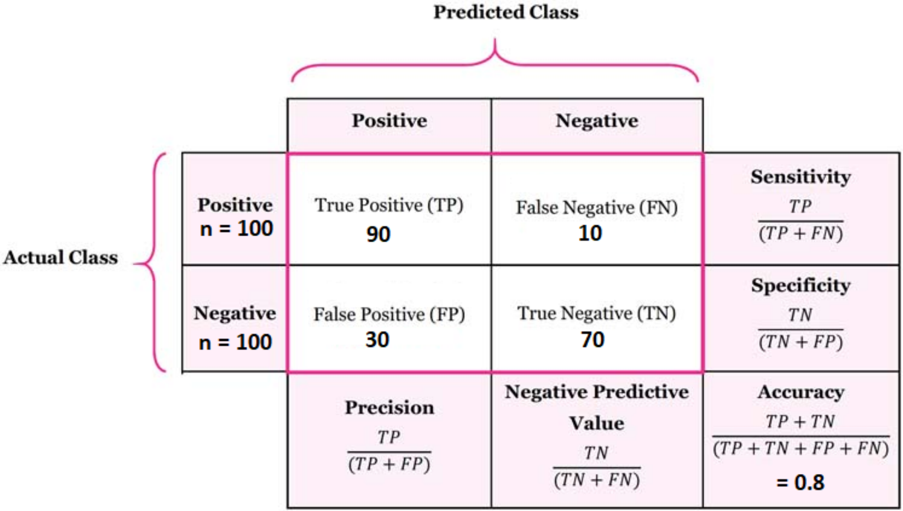
- ✓ **Simple to understand and to interpret**. Trees can be **visualised**.
- ✓ Requires **little data preparation** (ex : no normalization needed).

Some drawbacks :

- × Decision-tree can create over-complex trees that **do not generalise the data well** (especially when there are many features). This is called **overfitting**.
- × Decision trees can be **unstable** because small variations in the data might result in a completely different tree being generated.

The confusion matrix (simple performance metrics)

Choose your performance metric wisely! This choice should depend on your data and on your problematic.



		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive** **n = 100**	True Positive (TP) **90**	False Negative (FN) **10**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative **n = 100**	False Positive (FP) **30**	True Negative (TN) **70**	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$ **= 0.8**

Tutorial 2 : Machine Learning with Orange

"Radiomics and AI in Molecular Imaging" - COMULIS Training school

Nicolas Captier & Louis Rebaud

LITO (U1288) Université Paris Saclay/Inserm/Institut Curie, Orsay, France

October 13, 2021

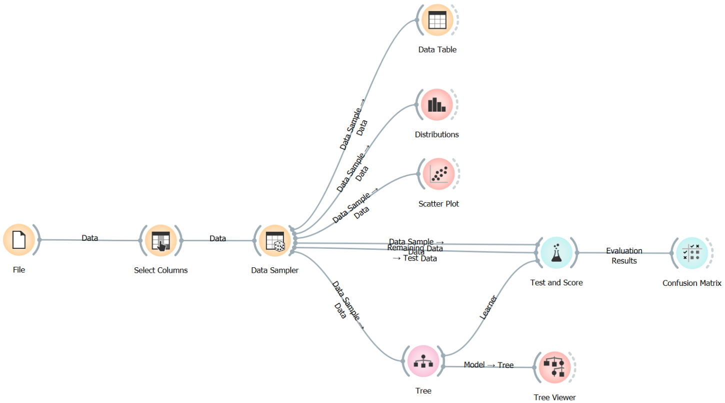
Medical University Vienna (MUW)

Institute For Applied Medical Physics (IFAMP)

COMULIS

COST

What we have done so far



- ✓ Thoroughly explored our data and clearly identified the problem we want to tackle
- ✓ Selected a suitable model and trained it with a training set resulting from a train-test split.
- ✓ Quantified the generalization ability of our trained model on an independent test set (remaining part of the train-test split).

There is still plenty of room for improvement !

Reminder: Our main question

Can we predict the histological subtype among primary lung cancers (adenocarcinomas vs squamous cell carcinomas + others) from PET images ?

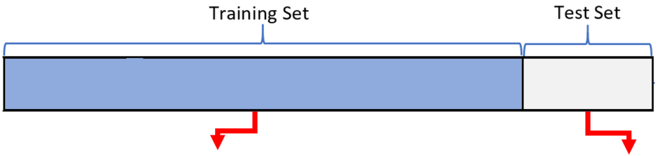
- Is train-test split the best method to confidently assess the generalization performance of our selected model and answer the question ?
- So far, we have only tested the decision tree classifier. There exist plenty more algorithms that are suited to our learning task and are likely to perform better. We should try them to answer the question more exhaustively !

To improve your pipeline and make it more relevant you need to clearly understand the tools and methods you are manipulating !

Plan

1. Cross-validation : efficient scheme for small data sets
2. Explore the ML models jungle : new opportunities and challenges
3. A step further : select a final model, assess its performance and interpret it !

The limitations of a single train-test split



Small training set may not be representative of all the possible cases (bad generalization ability of the fitted model !)

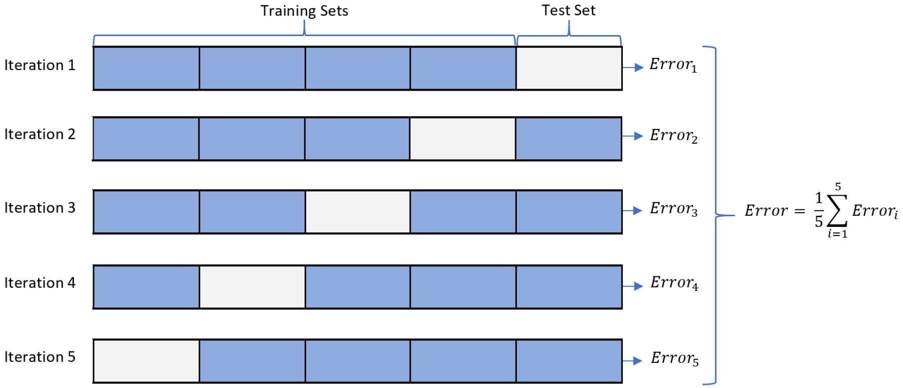
Small test set may not be representative of all the possible cases (optimistic or pessimistic assessments !)

For a small data set size and a complex question, the assessment of the performance may be very variable with respect to the train-test split !

Here we estimate the generalization ability of a model trained with **this particular training subset** (i.e 75% of the whole data) !

Can we find a way to use more efficiently **all the available data** and produce a **reliable** assessment of the performance of the model on unseen data (for a **model trained on all the available data**) ?

Cross-validation : use all the data through multiple splits



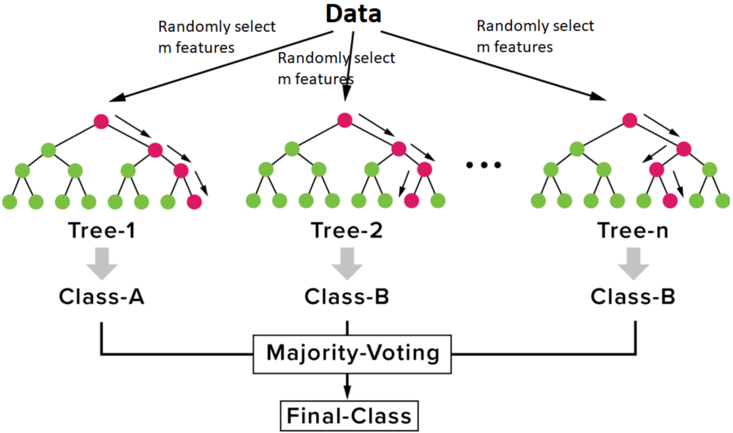
There is no clear answer for the choice of the number of folds. Usually 5 or 10 folds offer a good compromise for computational burden, variance of the estimate and bias.

Plan

1. Cross-validation : efficient scheme for small data sets
2. Explore the ML models jungle : new opportunities and challenges
3. A step further : select a final model, assess its performance and interpret it !

Random forest classifier

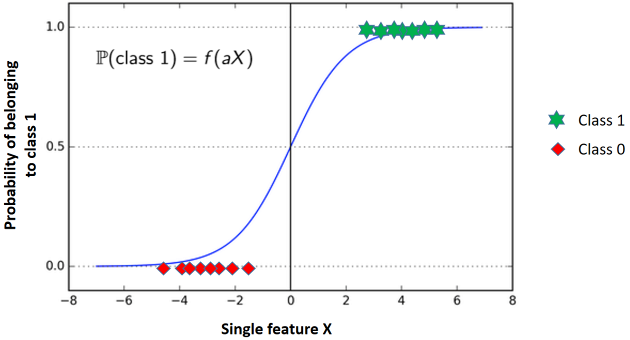
Ensembling strategy : a large number of relatively uncorrelated weak learners (trees) operating as a committee will outperform any of the individual constituent models.


Data
Randomly select m features
Randomly select m features
Randomly select m features
...
Tree-1
Tree-2
Tree-n
Class-A
Class-B
Class-B
Majority-Voting
Final-Class

Logistic regression

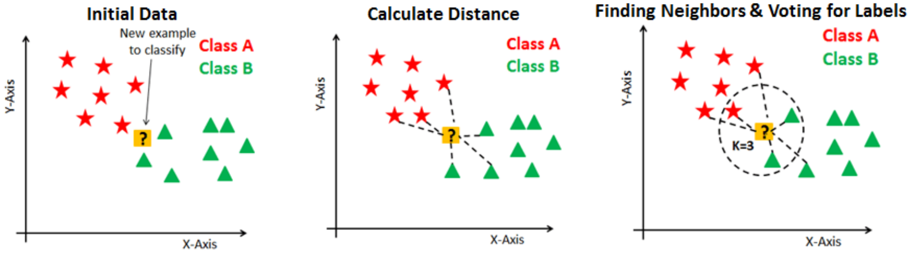
The logistic regression assumes a relationship between the probability of belonging to one class and a linear combination of the features.

$$\mathbb{P}\left(\text{the sample belongs to class 1}\right) = f\left(a_0 + a_1 \text{feature}_1 + \ldots + a_p \text{feature}_p\right)$$



* It is often recommended to normalize data as a pre-processing step. It will help the interpretation of the fitted model.

k-nearest neighbors classifier



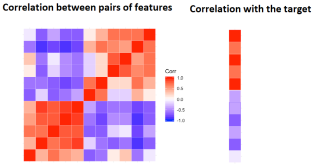
- A small k will fit the training data very well but is likely to overfit whereas a large k may perform poorly.
- The choice of the metric to compute distances is key !

⚠ With many features euclidean distance is no more informative, (k-nearest neighbors are not particularly closer than any other points). **We need to first reduce the number of features !**

A small glimpse into the world of dimensionality reduction

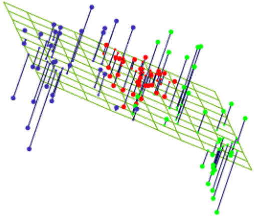
Select or **create** a **reduced number of features** that best retain some meaningful information from the original data (ex : the predictive information)

Feature selection with filter methods



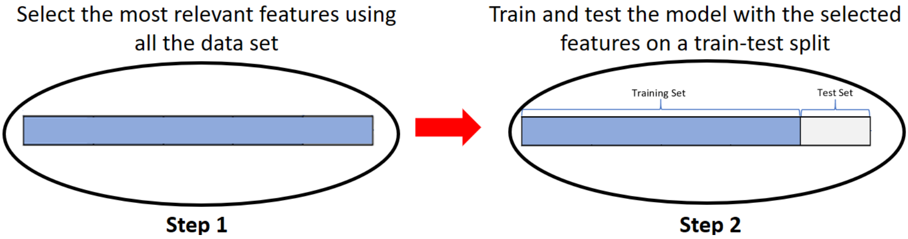
Fast correlation based filter : Weights each feature depending on its correlation with the target. Then, it only selects a subset in order to avoid any redundancy between

Principal Component Analysis



The goal of PCA is to create a reduced number of new features (**linear combination** of original features) so that the data is **as little distorted as possible**.

Something you should pay attention to : data leakage



Select the most relevant features using all the data set

Train and test the model with the selected features on a train-test split

Step 1

Step 2

Figure 1: A simple example of data leakage

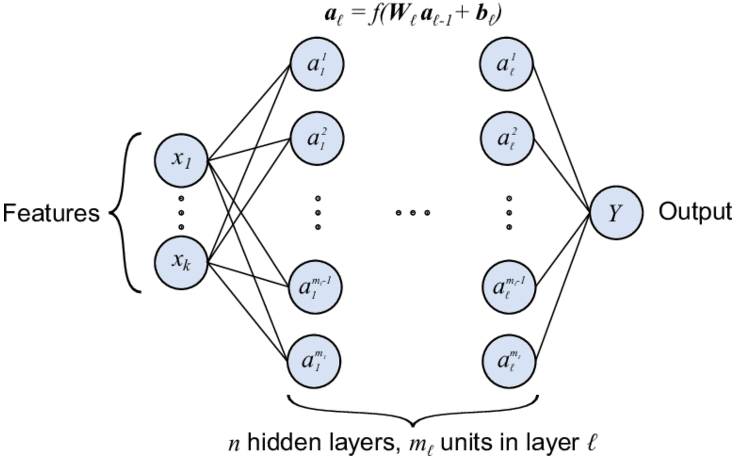
The tested model has an **unfair advantage** : it has already seen the test data during the feature selection step and selected the features on the basis of all the samples. It may lead to an **optimistic** assessment of its performance on completely unseen data !

Whatever the operation you include in your predictive pipeline, training and test data should never be mixed with each other !

Within a cross-validation scheme pre-processing should be rerun for each fold.

Neural networks

Neural networks are not magical, they simply consists in a succession of linear operations followed by the application of non-linear functions.



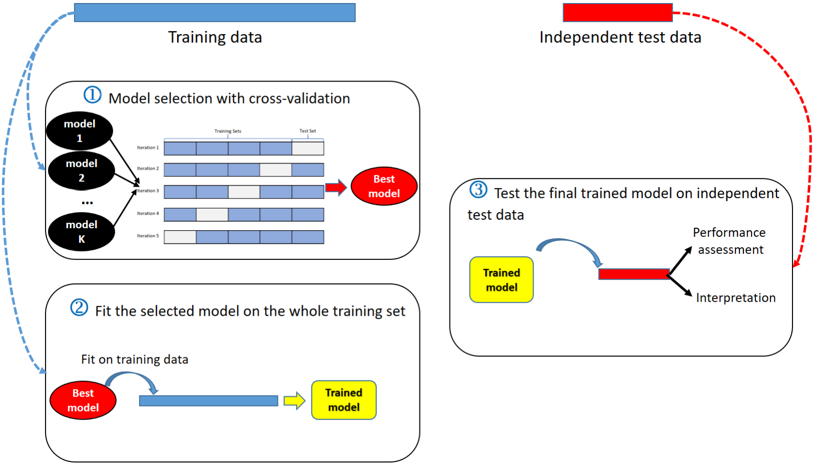
Training a neural network is not that straightforward (choice of hyperparameters, control of overfitting...) !

Plan

1. Cross-validation : efficient scheme for small data sets
2. Explore the ML models jungle : new opportunities and challenges
3. A step further : select a final model, assess its performance and interpret it !

Build a final predictive model : training, validation and test

Can we **build a final predictive model** for the histological subtype among primary lung cancers (adenocarcinomas vs squamous cell carcinomas + others) and **assess its generalization ability** ?



Training data

Independent test data

① Model selection with cross-validation

② Fit the selected model on the whole training set

Fit on training data

③ Test the final trained model on independent test data

A few words about our test data

You should gather as much information as possible on your test set to have a clear idea on the question you are trying to answer (ex : Does my model perform well on data from the same center ? from another center ? with different machines ? ...)

- 89 lung cancer patients extracted from TCIA[1] (62 adenocarcinomas and 27 squamous cell carcinomas + others).
- **FDG PET/CT images** acquired 70.4±24.9 min after injection (Biograph 64 mCT Siemens).
- Each lung lesion was semi-automatically defined on PET images with a **threshold of** 40% **SUVmax**.
- **43 PET textural features** computed with the LIFEx software (the same as our training data).

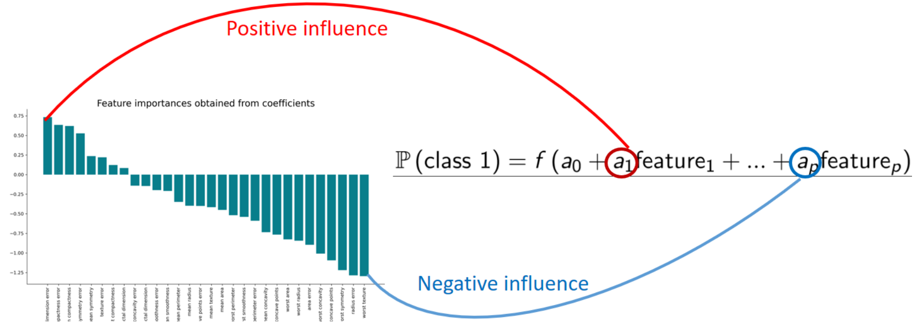
[1] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70224216

A glimpse into the world of model interpretation

Two main methods :

- **Model specific tools** - It requires a deep understanding of the model that you are manipulating
- **Model agnostic tools** - It requires a deep understanding of these tools (the main concepts behind them and the assumptions they make)

You should keep in mind that the interpretation that you obtain at the end is data and model dependent !



$$\mathbb{P}\left(\text{class } 1\right) = f\left(a_0 + a_1 \text{feature}_1 + \ldots + a_p \text{feature}_p\right)$$

Thank you for your attention !

LIFEx soft
Local Image Feature Extraction
A next generation of medical image features