Explainable Artificial Intelligence "XAI"

FPM-ILIA

Sidi Ahmed Mahmoudi. sidi.mahmoudi@umons.ac.be

Context : AI & Deep Learning
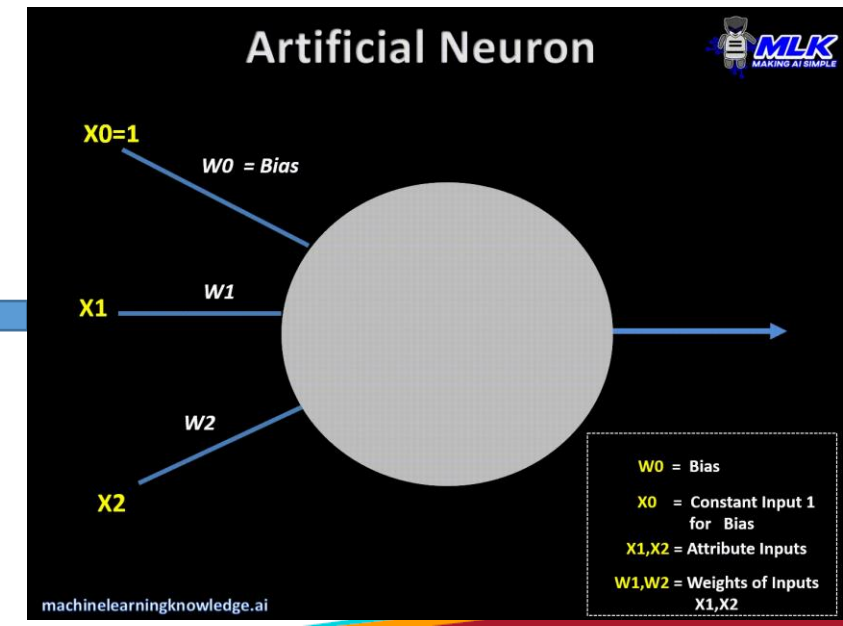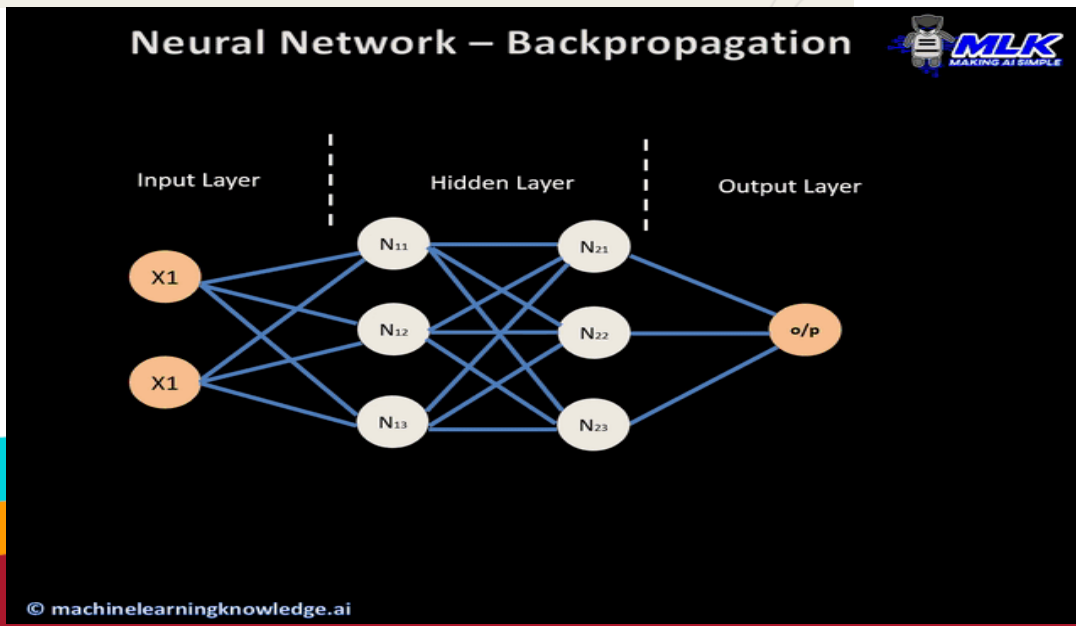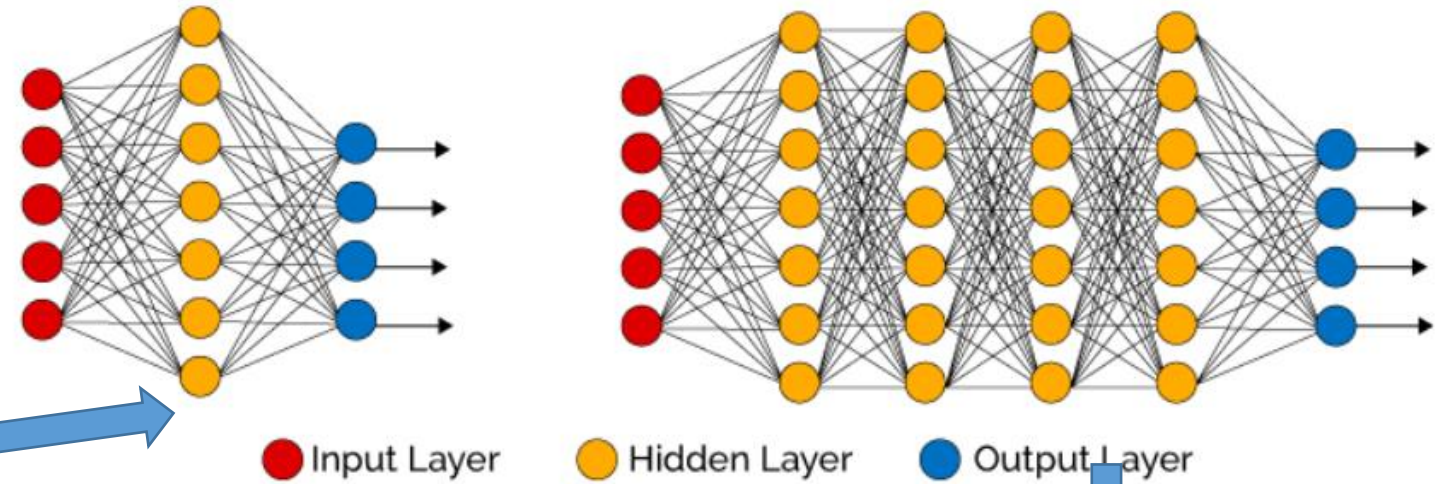
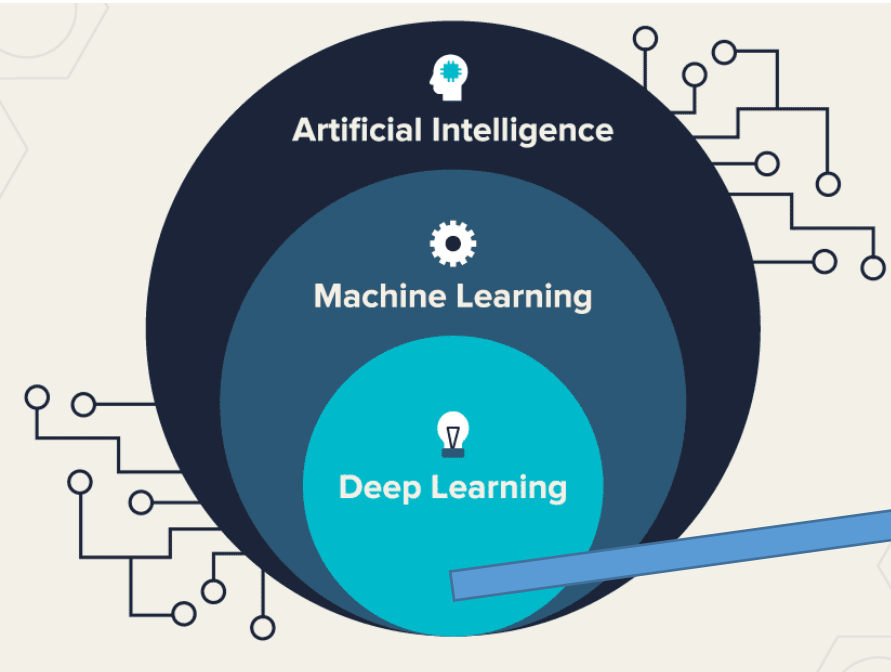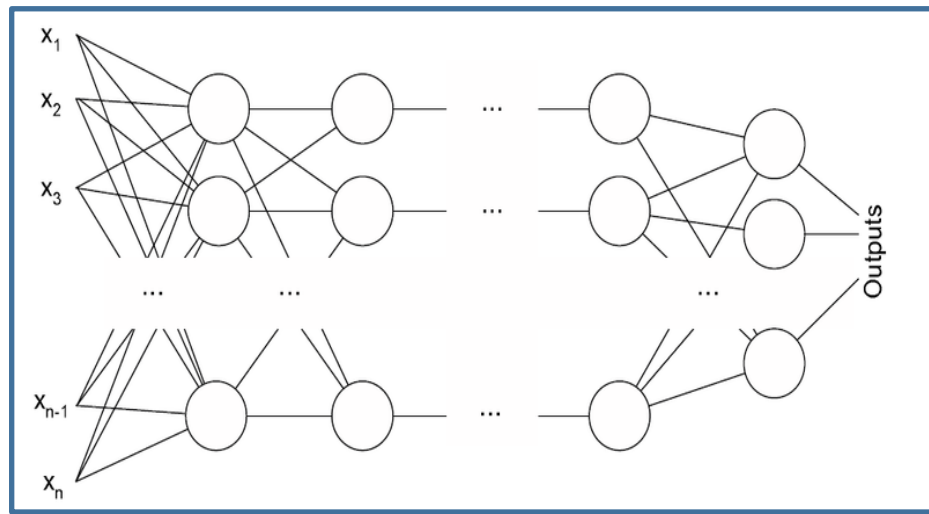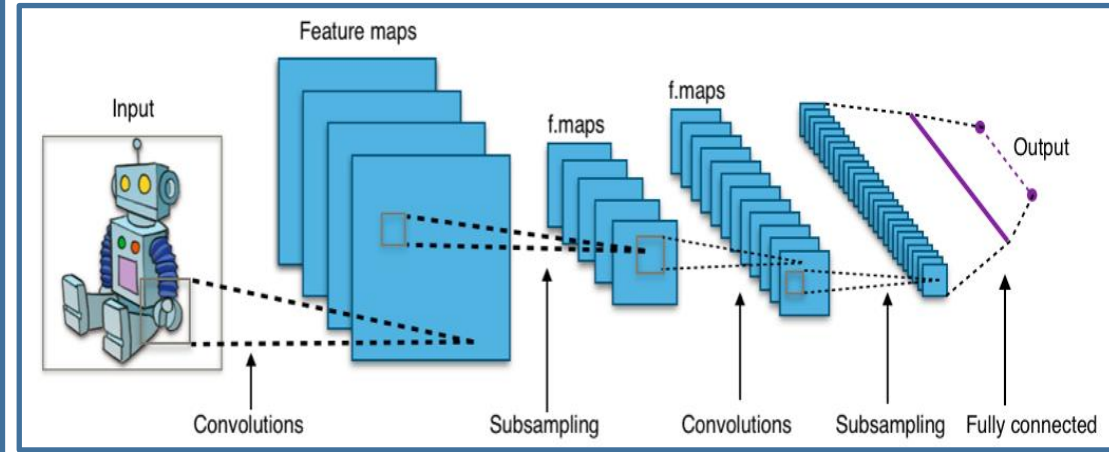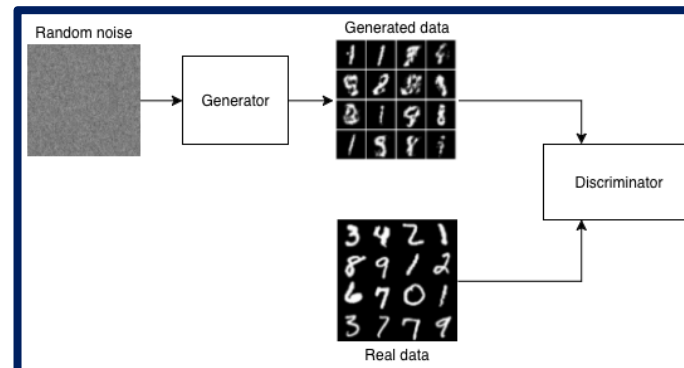# Context : Types of Deep Neural Networks
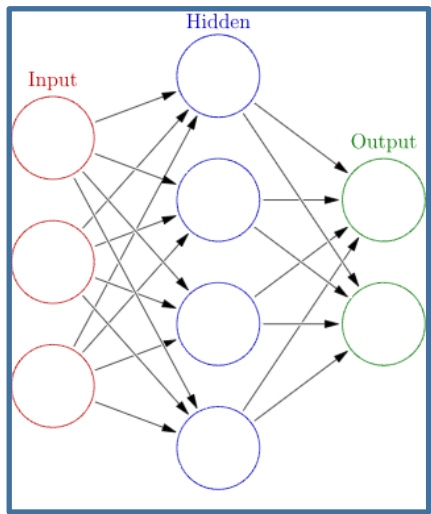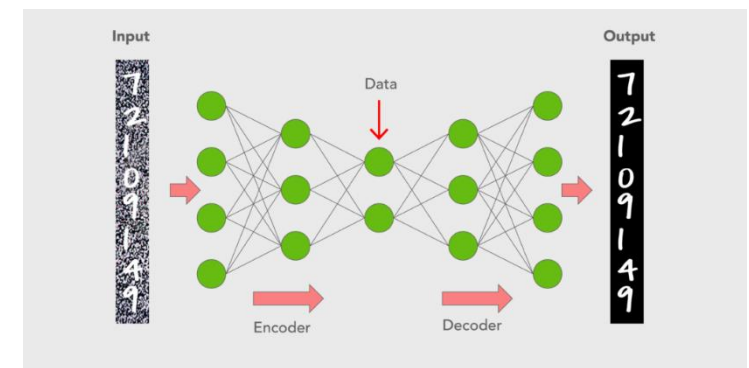

ANN

MLP (Multilayer Perceptron)

CNN
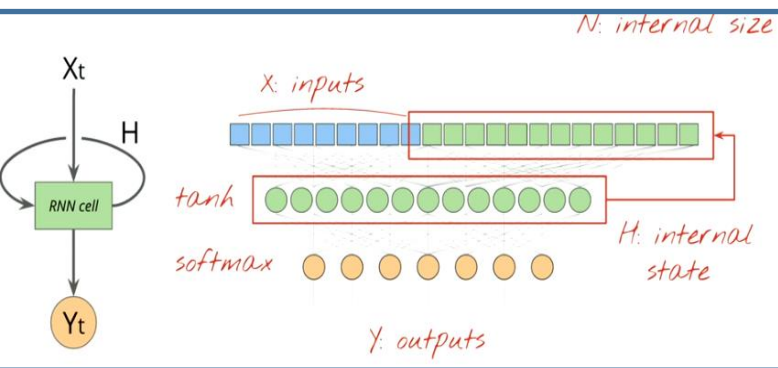
**RNN**

GAN

Auto encoders

. . .

Problem

High success of Deep Learning due to :

- **High availability of data :** Big Data & Cloud Computing

- **Computation Power:** HPC, GPU, TPU, etc.

- **Transfer Learning :** between models' weights

- **Regularization technics :** solve overfitting problems

- **Frameworks:** Tensorflow, Pytorch, Paddle, MxNet, etc.

- **Several architectures :MLP, CNN, RNN, LSTM, GAN, etc.**

What about explainability & interpretability ?

- **Can we trust Deep Learning Models ?**

- **Whey ? When ?**

- **How can we select a model over another ?**

Input → Black Box → Output

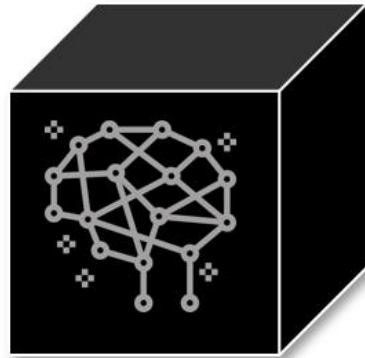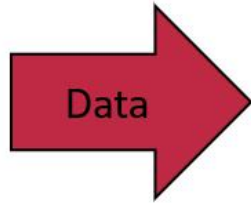*Internal behavior of the code is unknown*
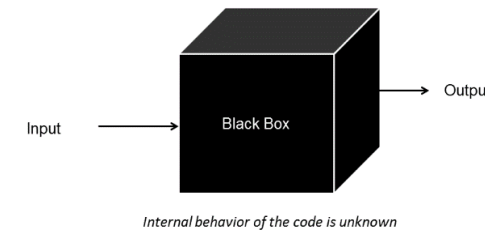
Explainabilty : statement or account that makes something clear; a **reason or justification** given for **an action or belief**

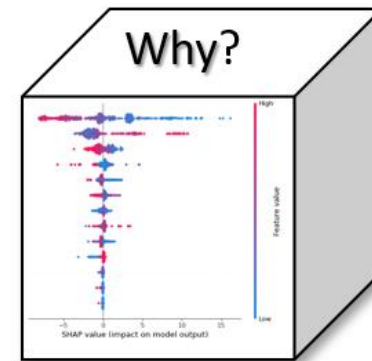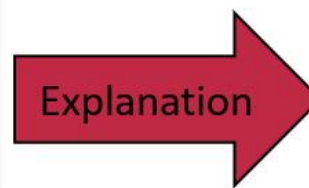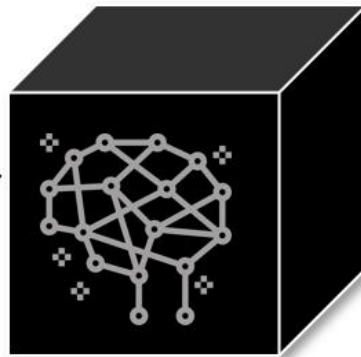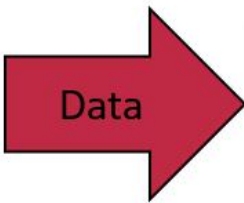**Interpretability :** the ability to explain or to present in **understandable terms to a human**.

Solution : Explainable Deep Learning "XAI"

XAI : perturbation and relevance-based approaches

Occlusion Visualization
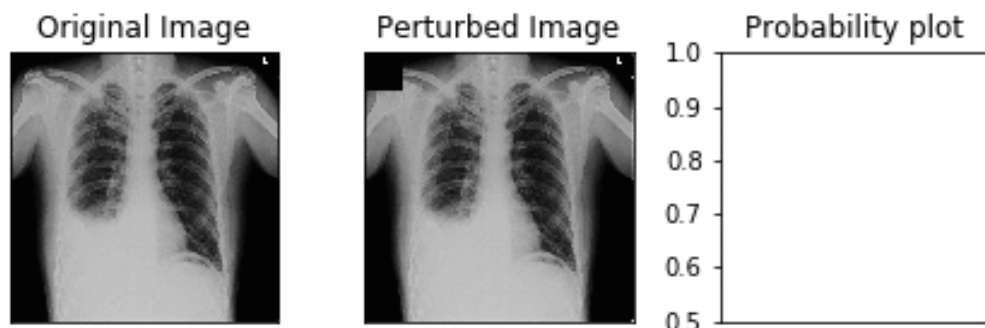
Relavance-based

Grad-Cam

Class Activation Mapping

Source : https://arxiv.org/abs/1512.04150
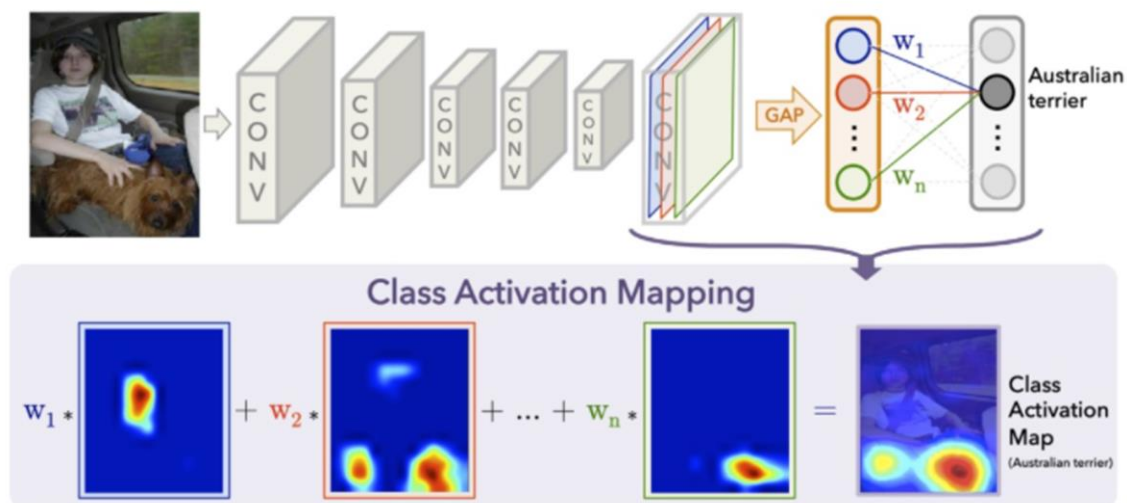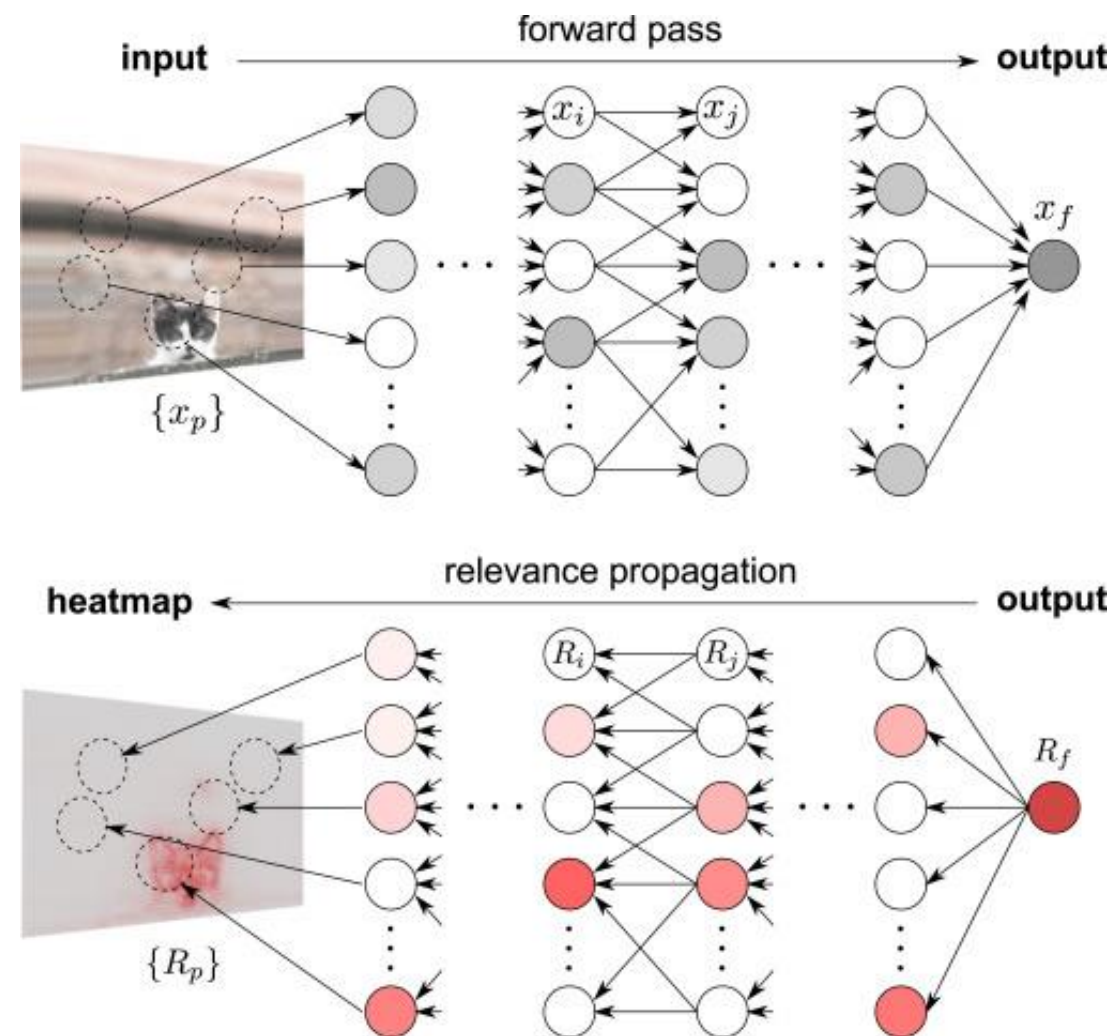
Example : XAI for medical images classification "Covid-19"

Data

- 03 classes: covid-19, normal, pneumonia

- Public dataset $\simeq$ 3000 images

- Local dataset $\simeq$ 1000 images

- Data augmentation $\simeq$ 4000 images

- Total size $\simeq$ **8000** images

Model development

- Images classification using DNN and CNN

- Transfer learning from pretrained models

- Optimization: regularization, dropout, etc.
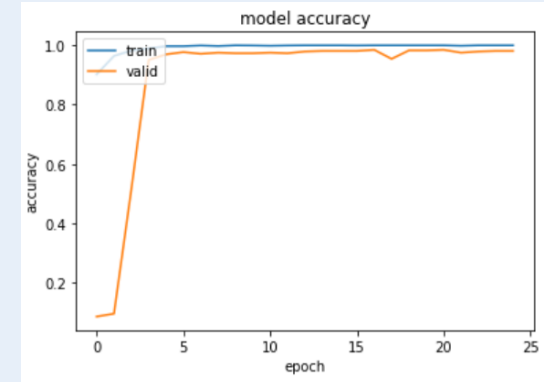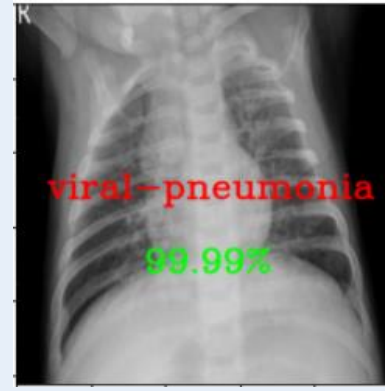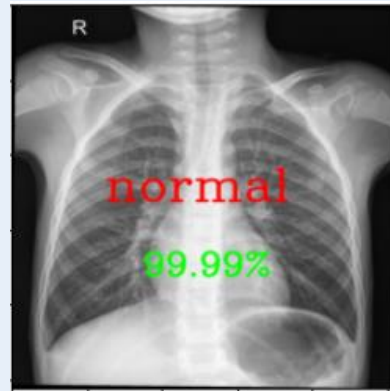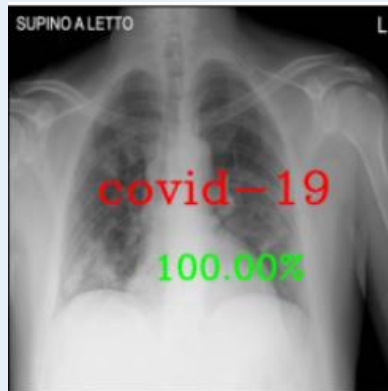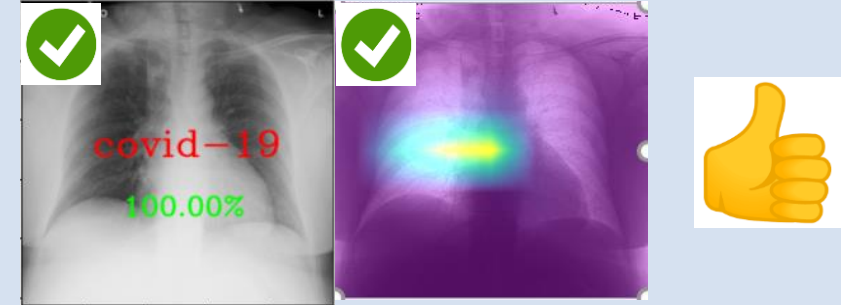
- Cross-validation, etc.
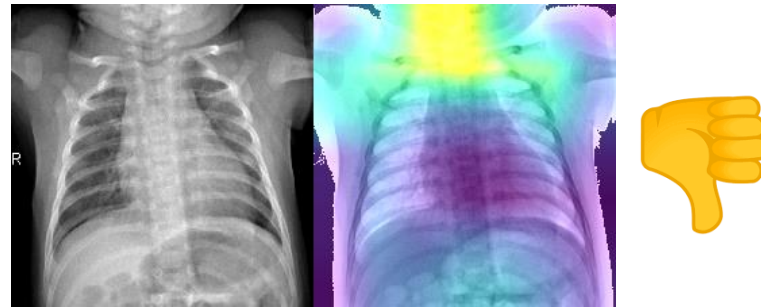
- Test_accuracy : **96,30%**

Example: XAI for medical images classification "Covid-19"

- Data bias : 2 types of detected bias thanks to XAI

- classification based on transcribed letters on the X-ray radiographs

- The X-ray images that represent the normal class : children's radiographs
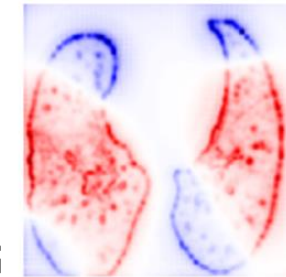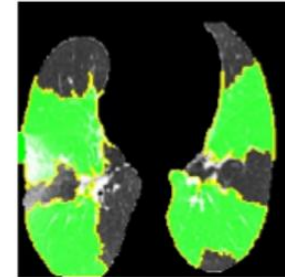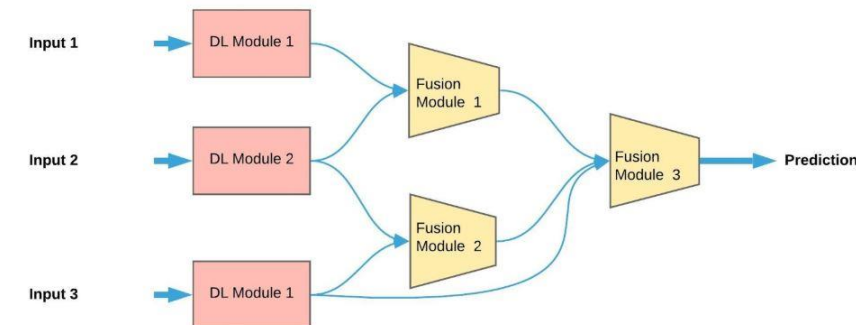

Biais N° 01 :

Biais N° 02 :

(a) Input   (j) GradCAM   (p) LIME

- XAI for AI **models validation and bias detection**

(c) LRP PresetAFlat (d) LIME Proxy Model

- XAI can be applied on **different types of Deep Neural Networks**

- XAI can applied to different applications : **images classification, object detection, text recognition, action recognition, etc.**

- **Challenge : evaluation of XAI methods** in order to define the appropriate approach

- **Challenge :** application of XAI on a **multimodal learning a**

Contact : Sidi Ahmed Mahmoudi
Sidi.mahmoudi@umons.ac.be