# Making AI trustworthy
## in multimodal and healthcare scenarios

**AI Responsabile e Affidabile**

**Rosa Sicilia**
Unit of Computer Systems and Bioinformatics, Department of Engineering,
University Campus Bio-Medico of Rome, Italy

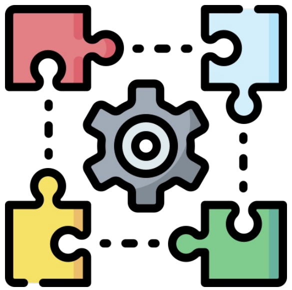
# Our directions

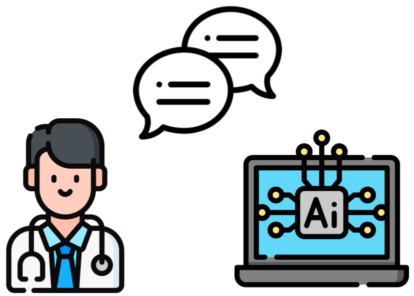
### Translating XAI to Multivariate Time Series

Boosted attention on TS classification models together with the need to explain them

### Multimodal XAI

Possibility to explore more complex deep architectures, combining unimodal networks, with an exacerbation of  the problem of understanding

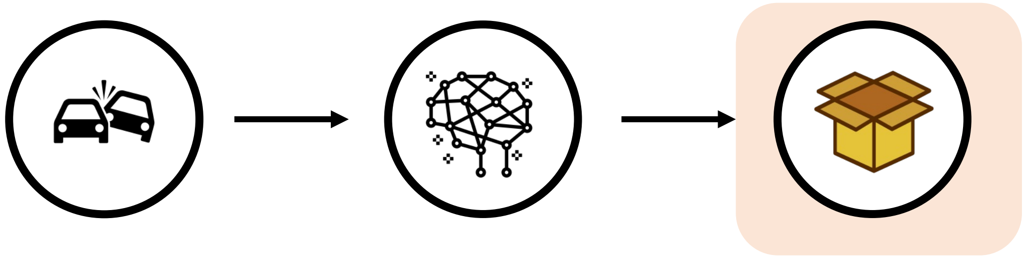





### Towards eXplainable Medical Concepts

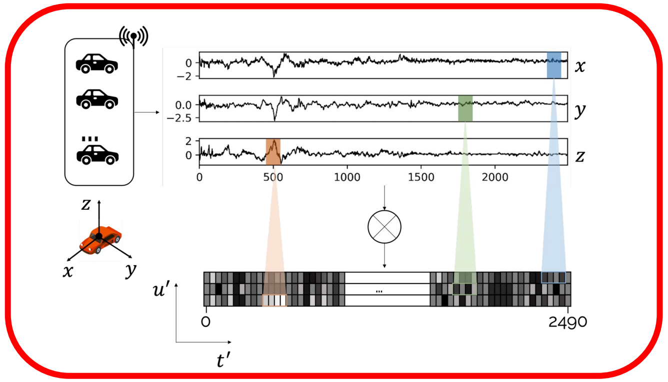
In the medical field identifying anatomical structures or tissue features that can be defined as relevant on an abstract scale is much more challenging and these elements may not be unambiguously defined

# Translating XAI to Multivariate Time Series



Explaining a **real-world multimodal task** of anomaly detection on telematics data from vehicles' black-box, where the available **modalities** are **acceleration** MTS and **velocity** UTS

## Materials and Methods

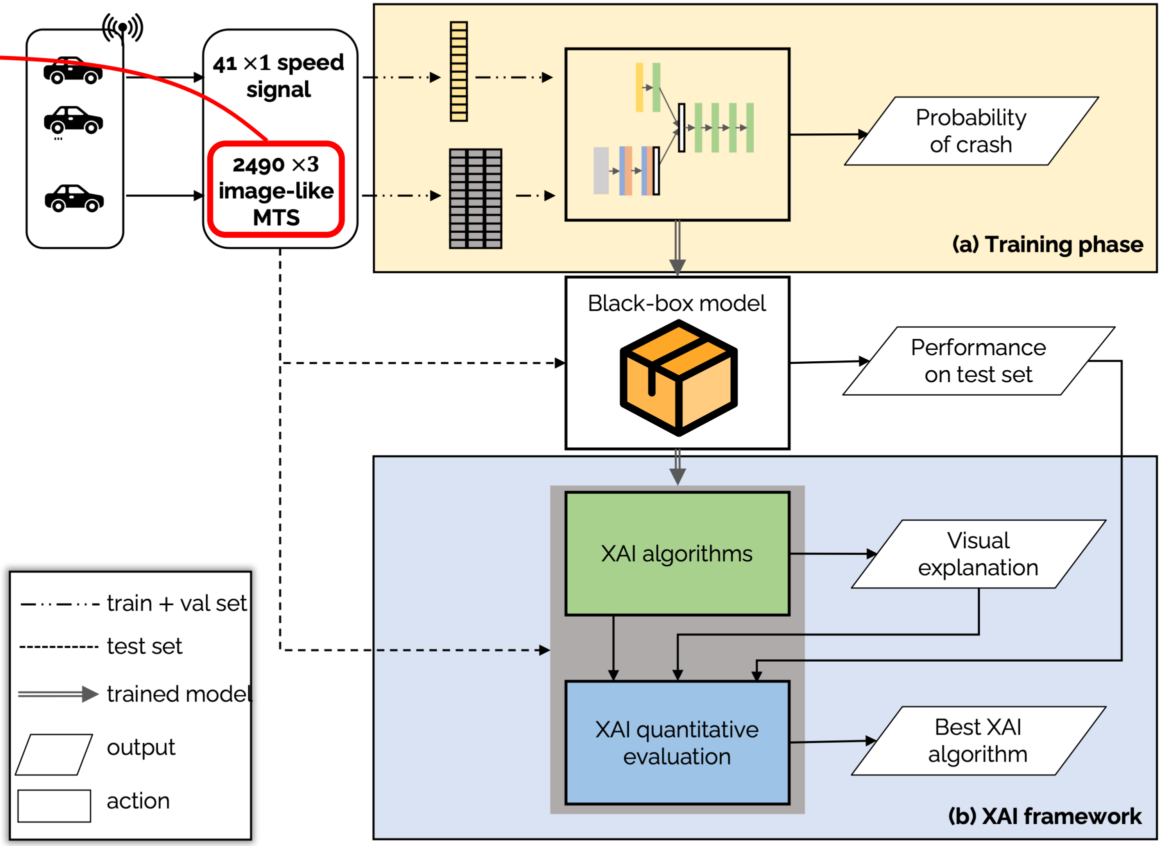



**81173** samples

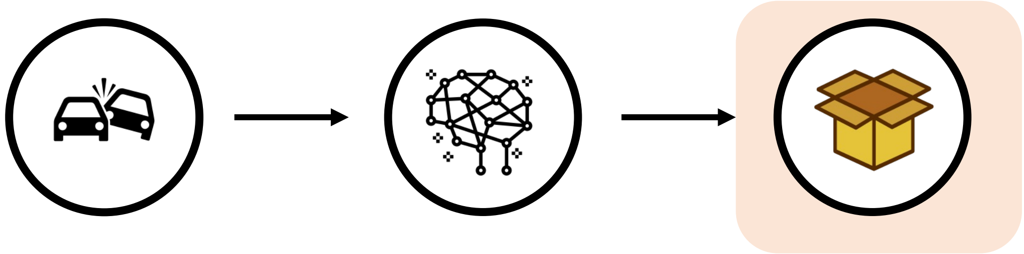
LIME

Grad-CAM

Integrated Gradients

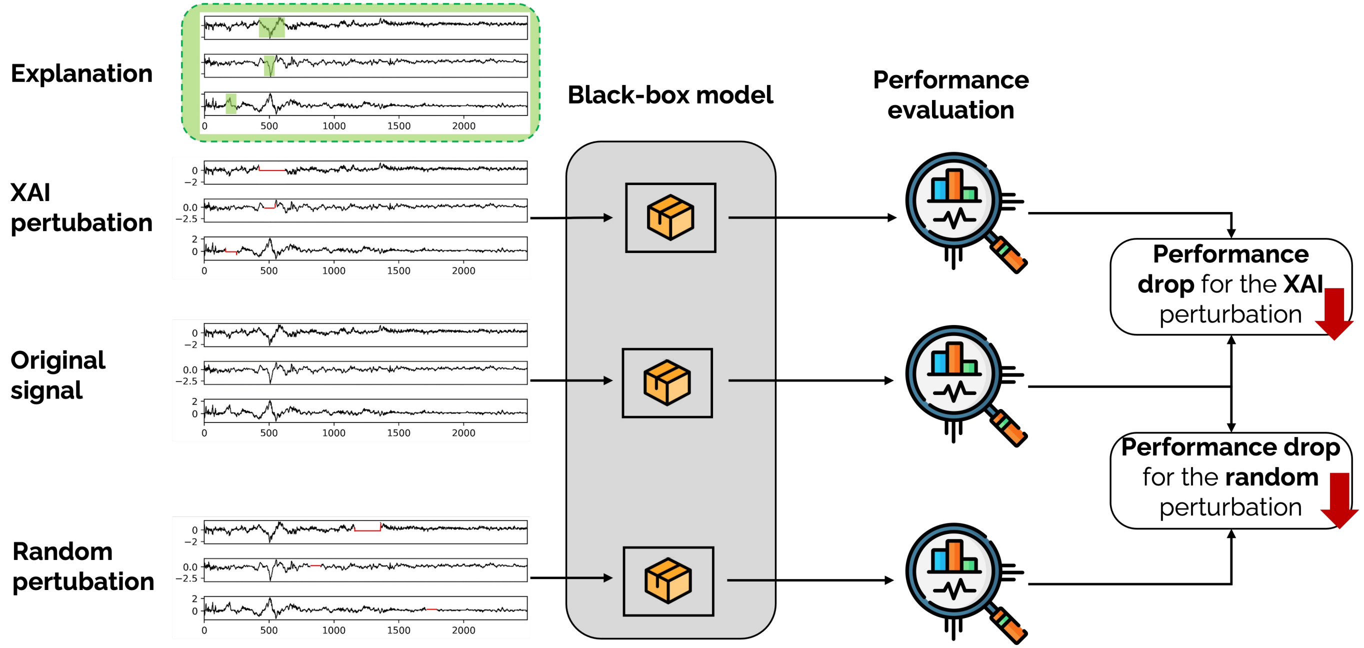


# Translating XAI to Multivariate Time Series



Explaining a **real-world multimodal task** of anomaly detection on telematics data from vehicles' black-box, where the available **modalities** are **acceleration** MTS and **velocity** UTS

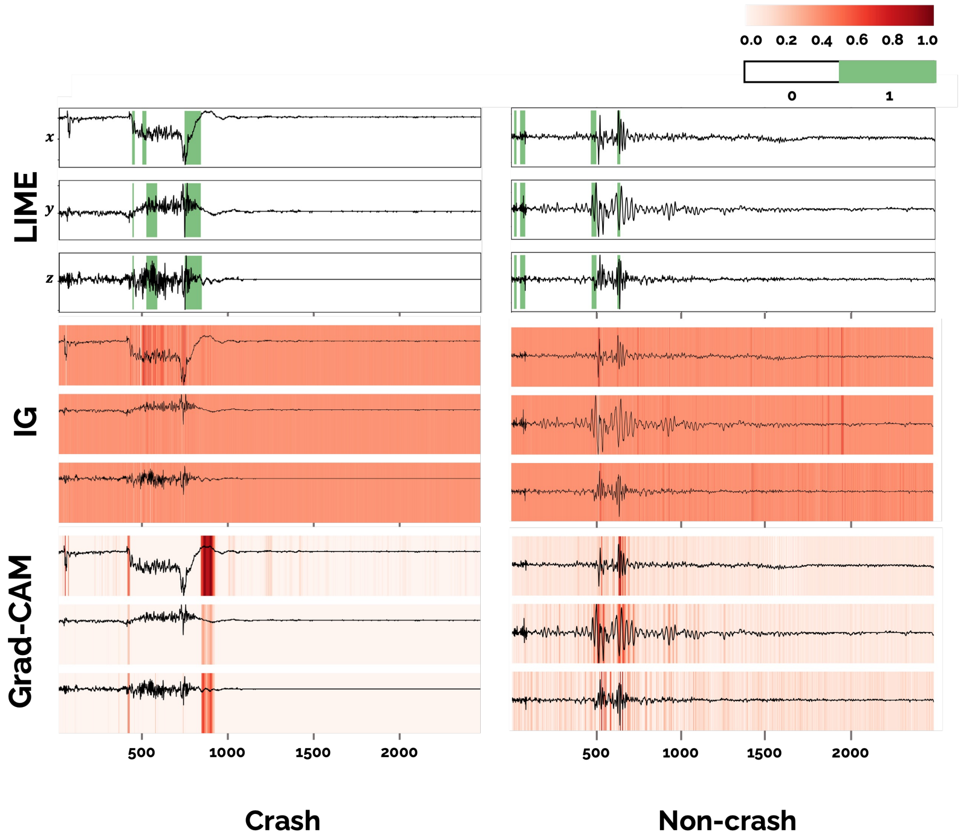
## Evaluation



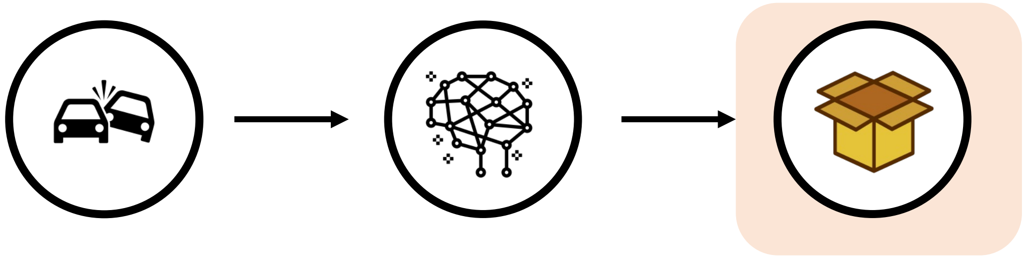

Drop XAI > Drop random

Drop XAI < Drop random

| Method | Drop LIME | Drop IG | Drop Grad-CAM | Drop Random |
|---|---|---|---|---|
| Zero | 54,3% | 58,2% | 20,1% | 0,7% |
| Swap | 13,8% | 15,2% | 6,5% | 14,3% |
| Mean | 10,0% | 5,5% | 2,7% | 3,6% |

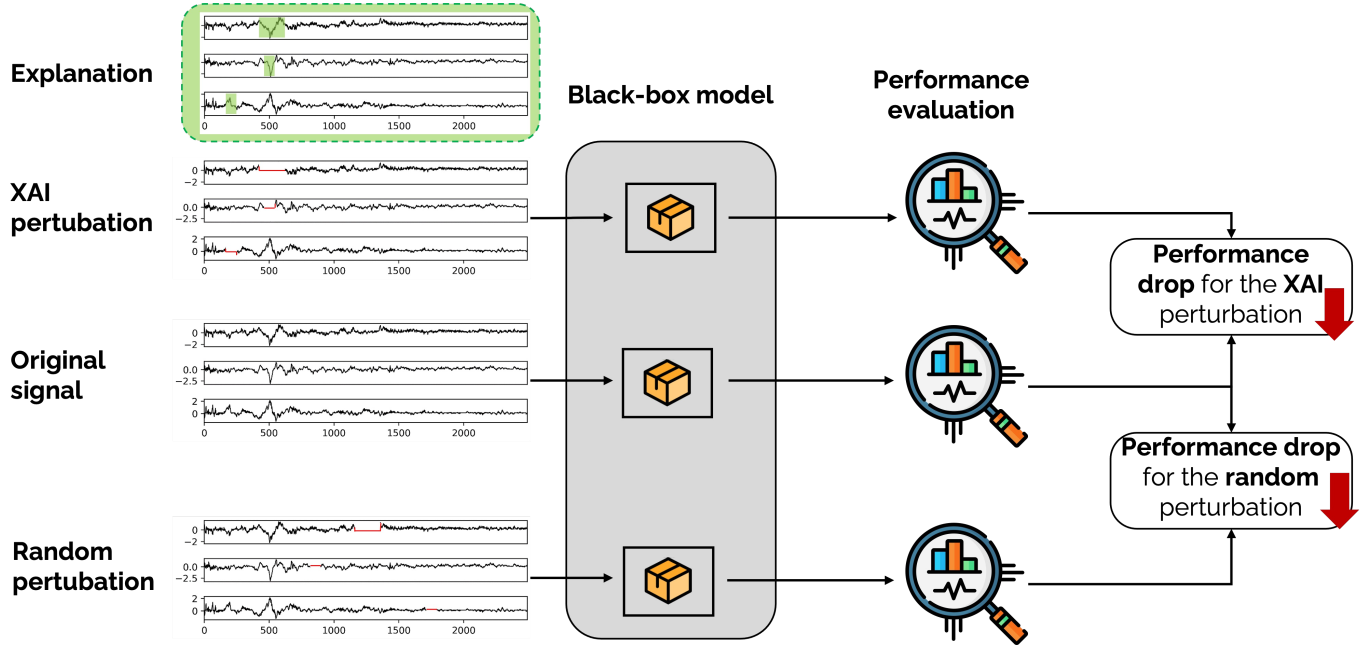


# Translating XAI to Multivariate Time Series



Explaining a **real-world multimodal task** of anomaly detection on telematics data from vehicles' black-box, where the available **modalities** are **acceleration** MTS and **velocity** UTS

## Evaluation




Drop XAI > Drop random

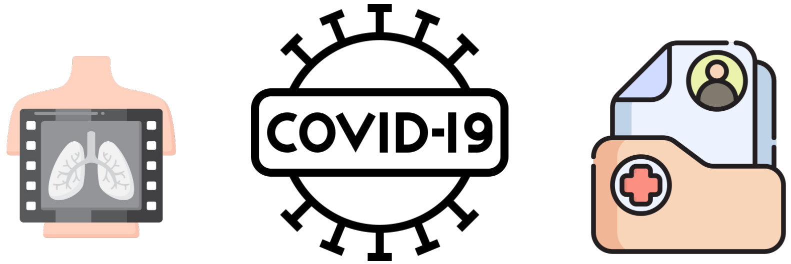
Drop XAI < Drop random

| Method | Drop LIME | Drop IG | Drop Grad-CAM | Drop Random |
|---|---|---|---|---|
| **Zero** | 54,3% | 58,2% | 20,1% | 0,7% |
| **Swap** | 13,8% | 15,2% | 6,5% | 14,3% |
| **Mean** | 10,0% | 5,5% | 2,7% | 3,6% |

## Challenges and perspectives

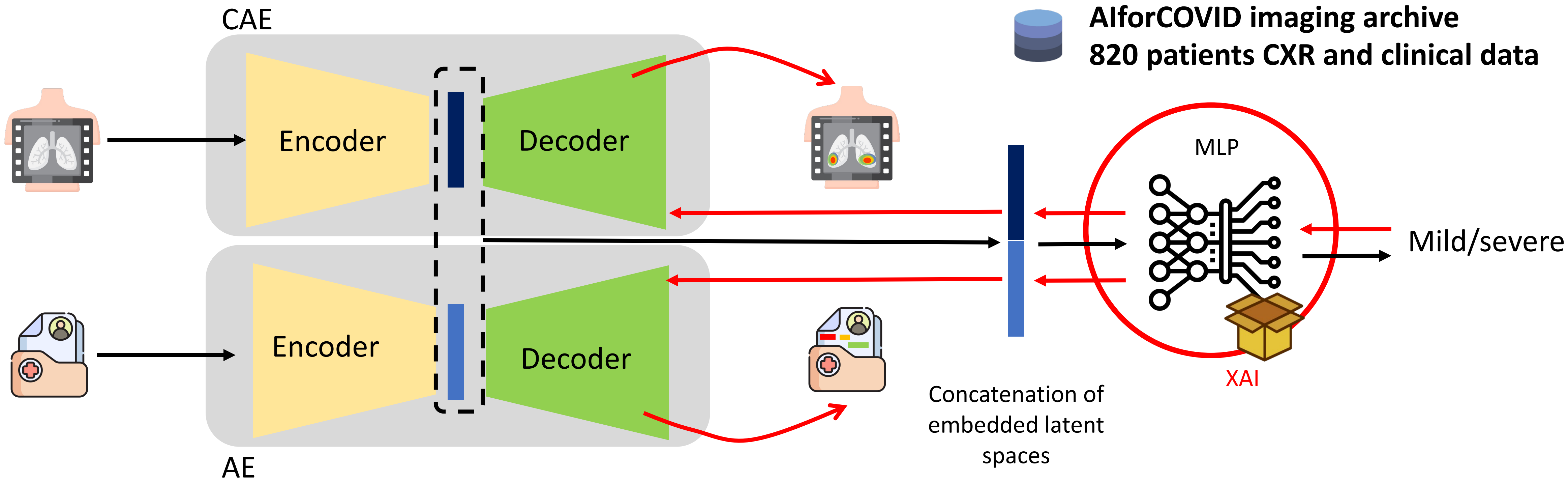
- More **human-interpretable** representations
- Developing a **multimodal XAI** method able to explain both signals available

# Multimodal XAI



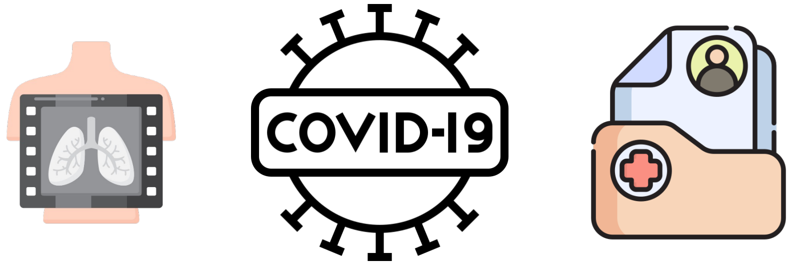

Supervised **multimodal fusion** applied to early identify **patients at risk of the severe outcome**, like intensive care or death, among those affected by **SARS-CoV-2**, and using chest X-ray (CXR) scans and clinical data.

## Materials and Methods



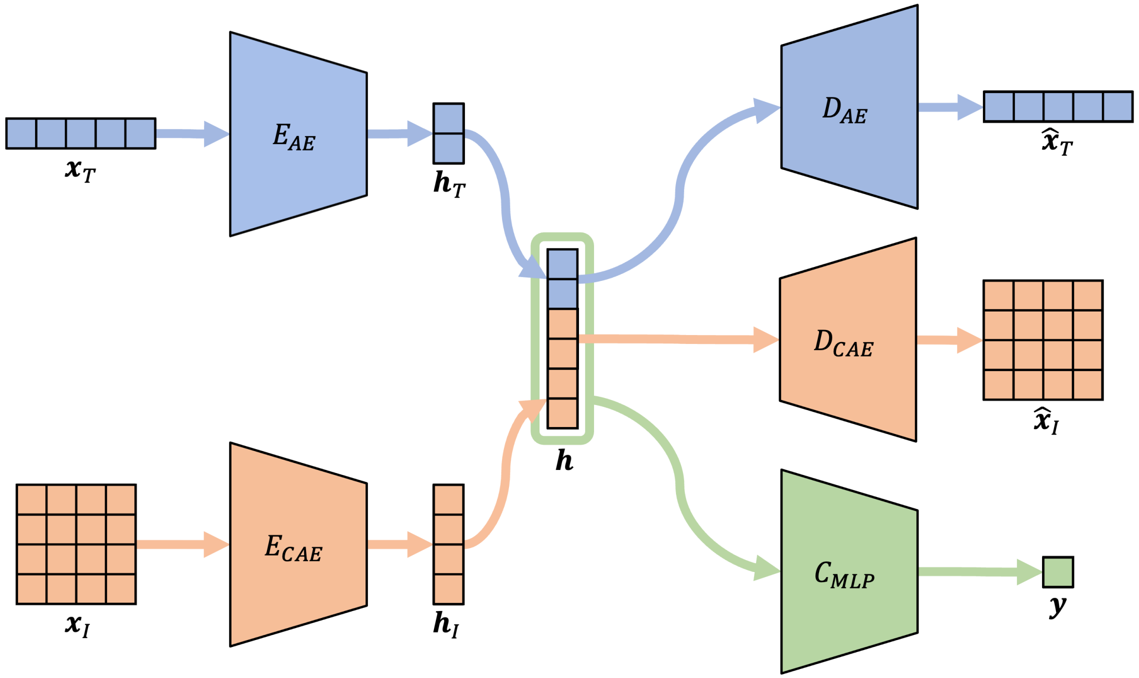

*End-to-end training*

# Multimodal XAI




Supervised **multimodal fusion** applied to early identify **patients at risk of the severe outcome**, like intensive care or death, among those affected by **SARS-CoV-2**, and using chest X-ray (CXR) scans and clinical data.

## Materials and Methods




**AIforCOVID imaging archive**
**820 patients CXR and clinical data**

Modalities: Tabular ($T$) and Imaging ($I$)

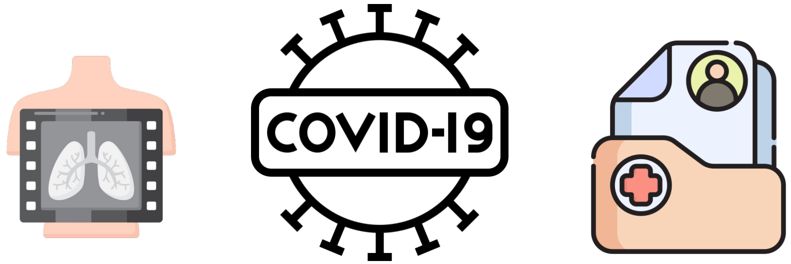
Inputs: $\boldsymbol{x}_T$ and $\boldsymbol{x}_I$

Embeddings: $\boldsymbol{h}_T$, $\boldsymbol{h}_I$ and $\boldsymbol{h}$ (concatenation)

Outputs:

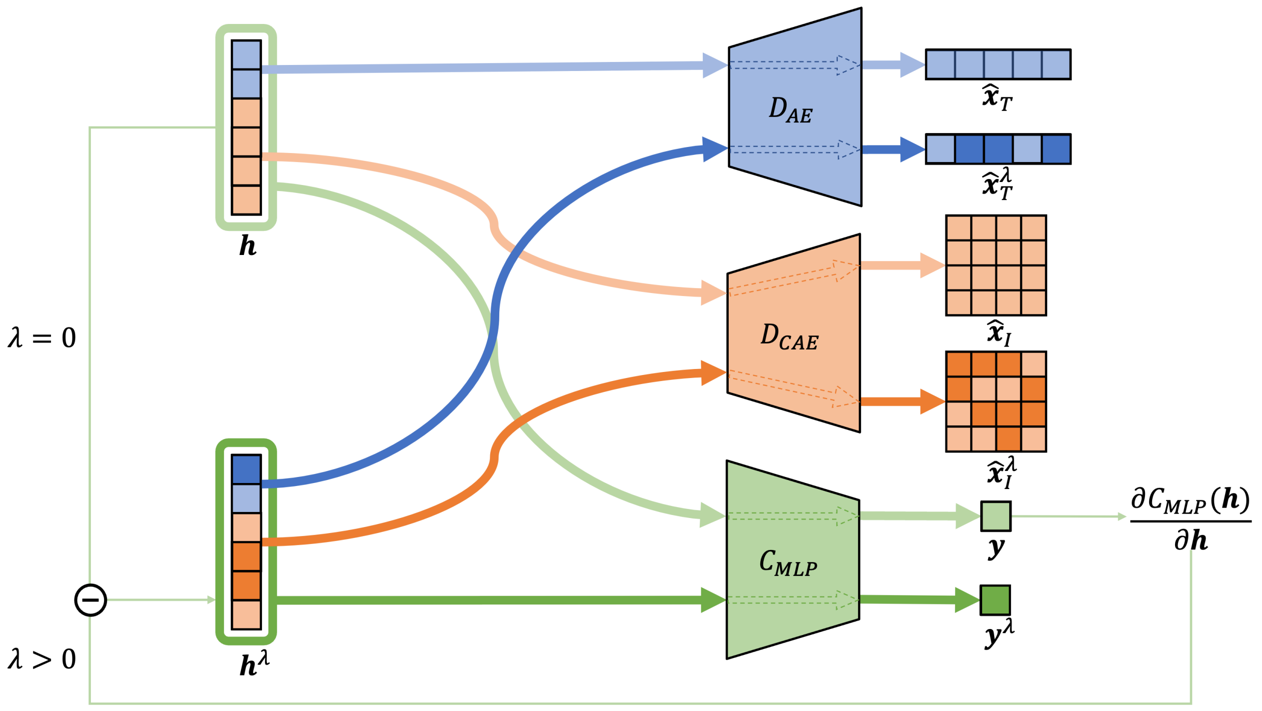
- Reconstruction: $\hat{\boldsymbol{x}}_T$, $\hat{\boldsymbol{x}}_I$
- Classification: $\boldsymbol{y}$

# Multimodal XAI




Supervised **multimodal fusion** applied to early identify **patients at risk of the severe outcome**, like intensive care or death, among those affected by **SARS-CoV-2**, and using chest X-ray (CXR) scans and clinical data.

## Materials and Methods




**AIforCOVID imaging archive**
**820 patients CXR and clinical data**

**λ-shifted counterfactual** multimodal reconstructions and output:

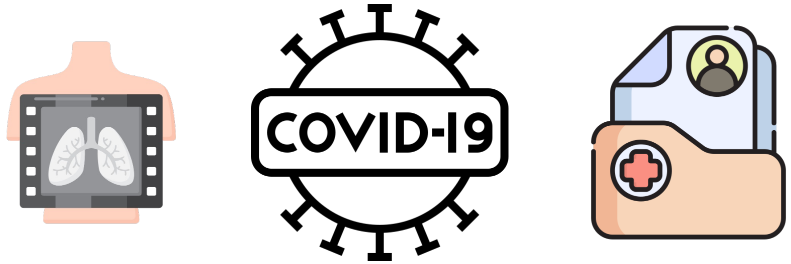
$$\hat{x}_T^\lambda = D_{AE}(h_T^\lambda)$$

$$\hat{x}_I^\lambda = D_{CAE}(h_I^\lambda)$$

$$y^\lambda = C_{MLP}(h^\lambda)$$

as λ increases, we expect a **flip** of the predicted class.

# Multimodal XAI



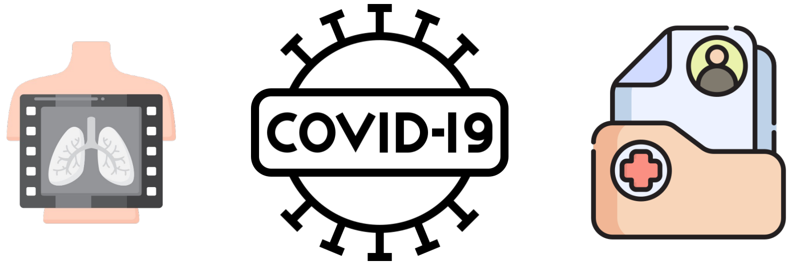

Supervised **multimodal fusion** applied to early identify **patients at risk of the severe outcome**, like intensive care or death, among those affected by **SARS-CoV-2**, and using chest X-ray (CXR) scans and clinical data.

## AI Evaluation

| Model | Validation | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| Our proposal (three-stage training) | CV | 76.75±5.32 | 78.58±6.48 | 74.55±5.86 |
| | LOCO | 74.21±6.08 | 76.73±18.88 | 68.40±15.46 |
| | Survey | 76.77 | 78.54 | 74.57 |
| AIforCOVID [9] | CV | 76.90±5.40 | 78.80±6.40 | 74.70±5.90 |
| | LOCO | 74.30±6.10 | 76.90±18.90 | 68.50±15.50 |
| $R_1$ | Survey | 68.75 | 43.75 | 93.75 |
| $R_2$ | Survey | 72.92 | 70.83 | 75.00 |
| $R_3$ | Survey | 76.04 | 70.83 | 81.25 |
| $R_4$ | Survey | 72.92 | 62.50 | 83.33 |

***No significant decrease with respect to literature***

# Multimodal XAI



Supervised **multimodal fusion** applied to early identify **patients at risk of the severe outcome**, like intensive care or death, among those affected by **SARS-CoV-2**, and using chest X-ray (CXR) scans and clinical data.

## XAI Evaluation

***High intersection between the multimodal explanation and the experts ground truth***

- The modality normalized absolute differences:

$$\Delta_T = \frac{||\boldsymbol{h}_T - \boldsymbol{h}_T^{\lambda}||_1}{n}$$

$$\Delta_I = \frac{||\boldsymbol{h}_I - \boldsymbol{h}_I^{\lambda}||_1}{m}$$

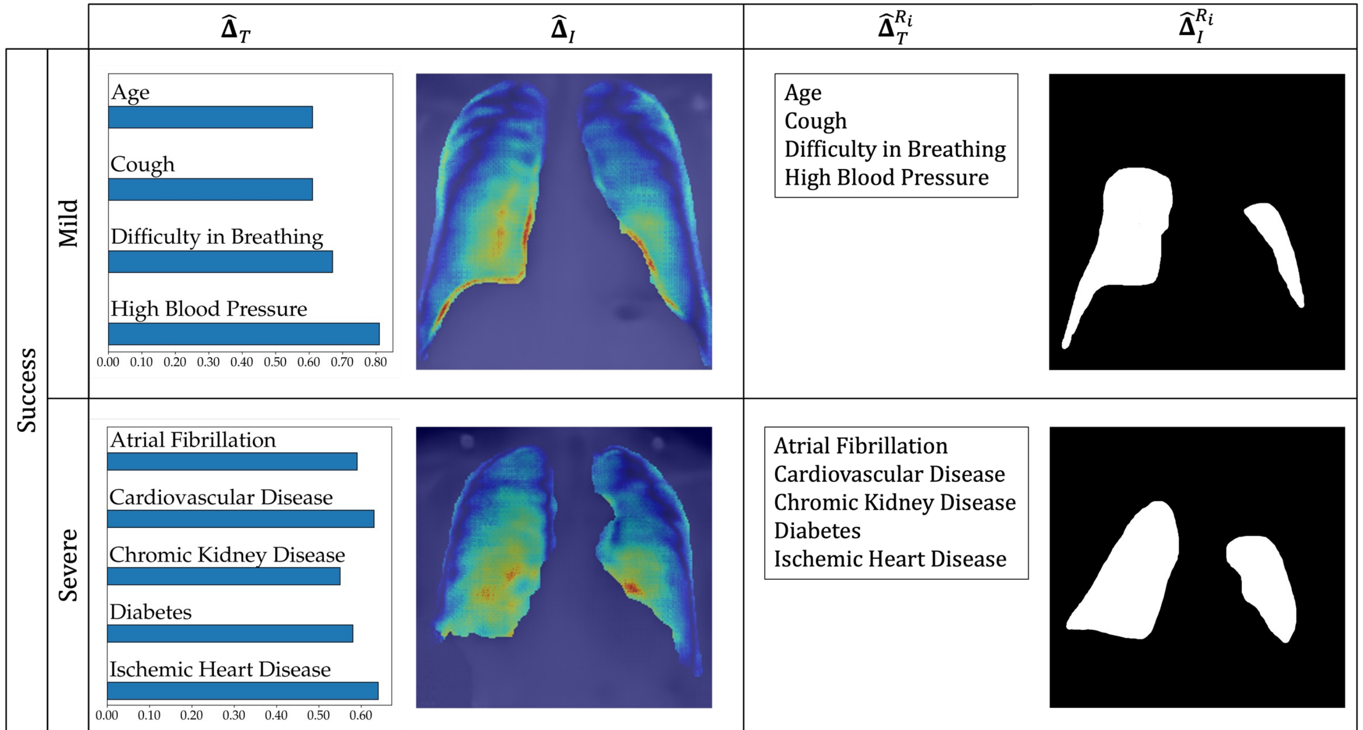
*The more a **modality embedding** has changed, the more important it is for the classification of a given sample.*

- Feature absolute distance, to understand how much each **feature** has shifted:

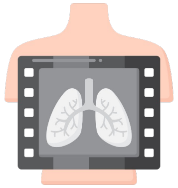
$$\hat{\boldsymbol{\Delta}}_T = |\hat{\boldsymbol{x}}_T - \hat{\boldsymbol{x}}_T^{\lambda}|$$

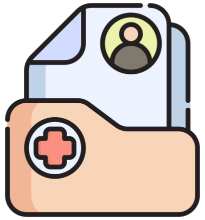
$$\hat{\boldsymbol{\Delta}}_I = |\hat{\boldsymbol{x}}_I - \hat{\boldsymbol{x}}_I^{\lambda}|$$

*The more a **feature changes**, the more important it is for the classification.*



# Multimodal XAI



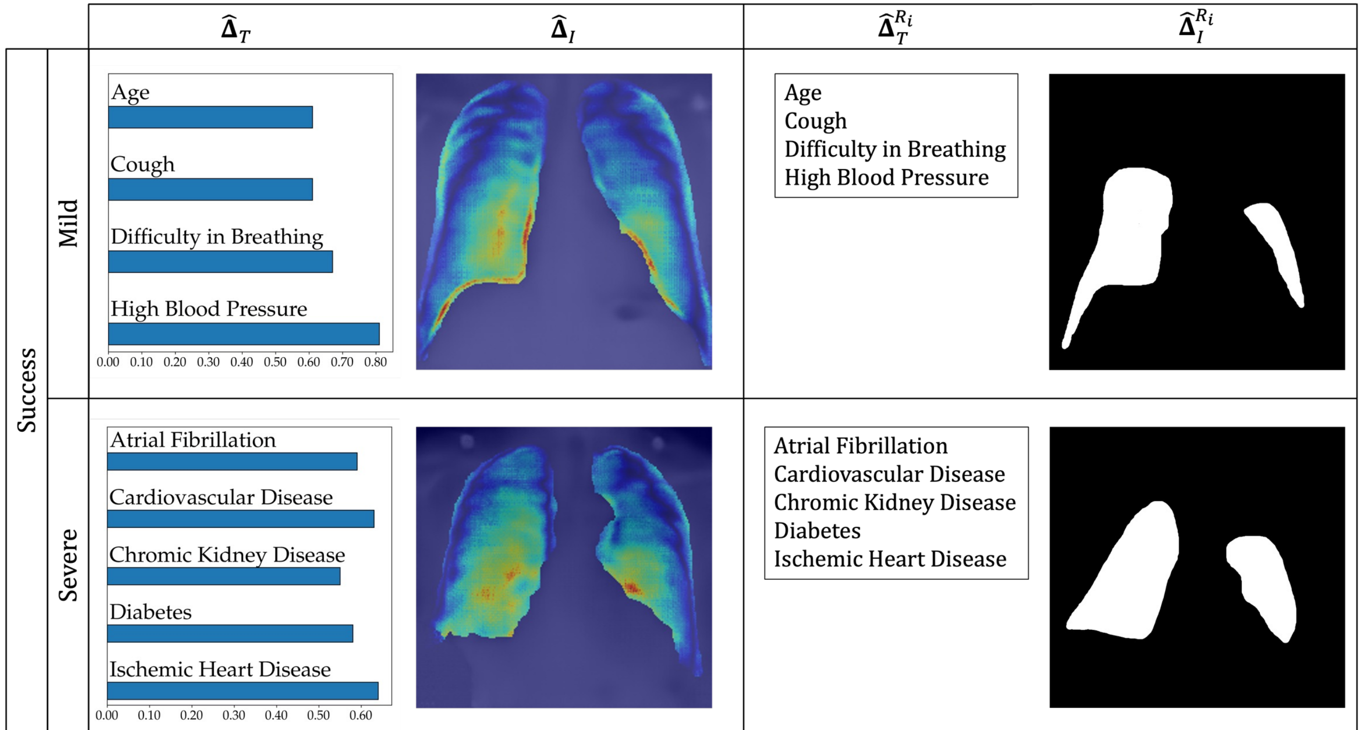
Supervised **multimodal fusion** applied to early identify **patients at risk of the severe outcome**, like intensive care or death, among those affected by **SARS-CoV-2**, and using chest X-ray (CXR) scans and clinical data.

## XAI Evaluation

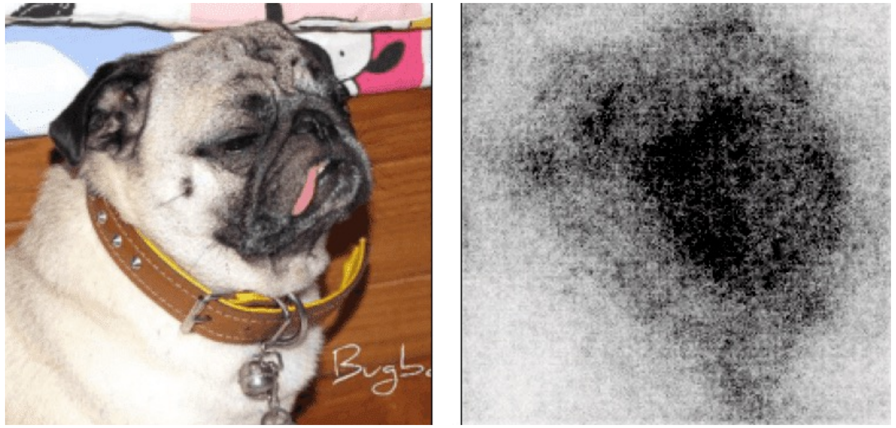
***High intersection between the multimodal explanation and the experts ground truth***

### Challenges and perspectives

- **More modalities** at play
- To tackle the problem of **missing modalities** especially from the explanation view point.

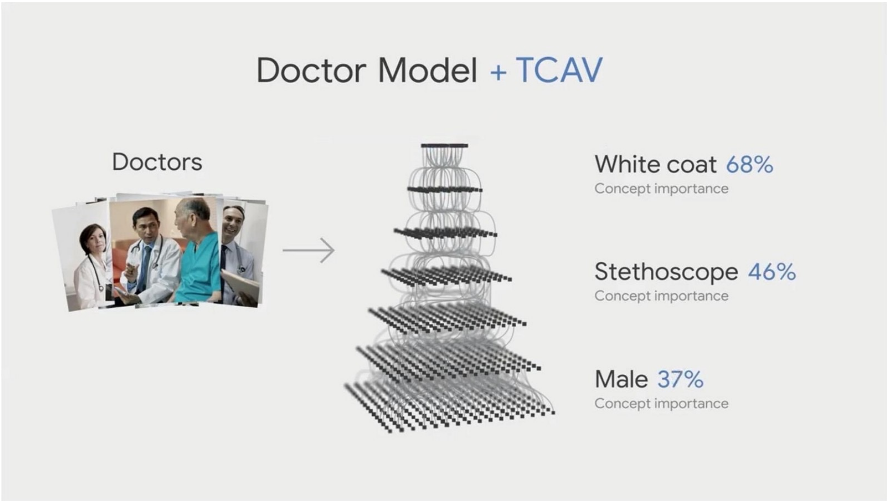


# Towards eXplainable Medical Concepts



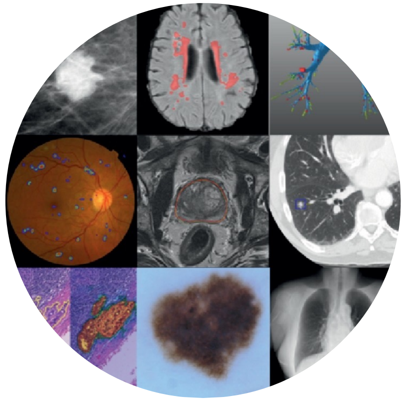
## Saliency Map

- ✓ Interpret pixel map of the decision
- ✗ Lack of texture-level explanation




## TCAV and the Concept-Based Interpretability

- ✓ Interpretation of Human-friendly Concepts defined by users
- ✗ No intuitive way to define medical concepts



# Towards eXplainable Medical Concepts

- **191** Patients
- **22384** CT slices
- **Retrospective** Clinical features

**Medical Concepts Extraction**
Automatic identification of common texture information related to the micro and macro structural properties of biomedical tissue.

**Challenges**
- **High** images **complexity**
- **Subjectiveness** in experts Interpretations

# Towards eXplainable Medical Concepts

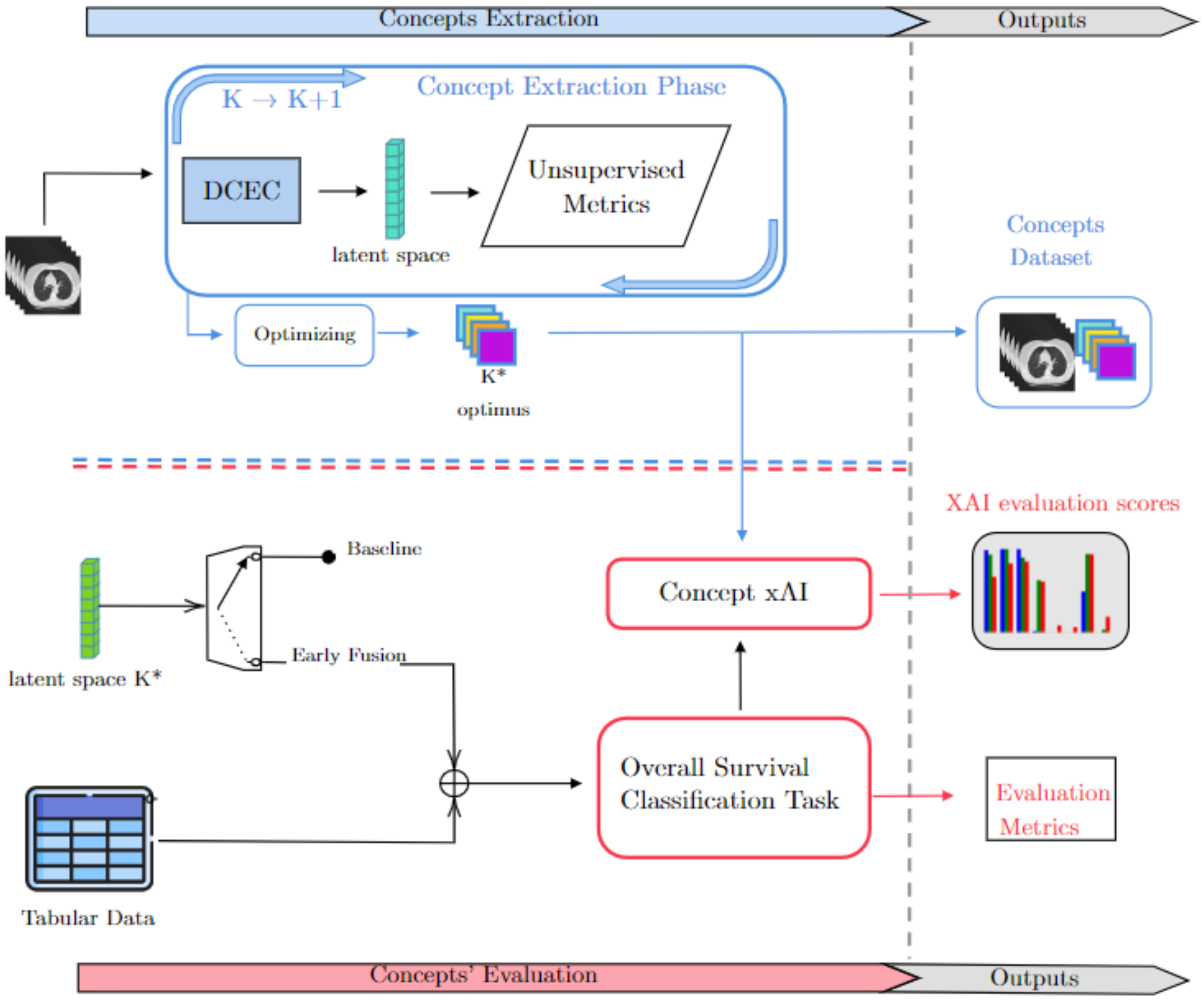
- **191** Patients
- **22384** CT slices
- **Retrospective** Clinical features

**Medical Concepts Extraction**
Automatic identification of common texture information related to the micro and macro structural properties of biomedical tissue.

**Challenges**
- **High** images **complexity**
- **Subjectiveness** in experts Interpretations

*Adoption of a framework based on* ***Deep Clustering*** *model and* ***Concepts Attribution XAI*** *methods in order to find the* ***best explainable groups of image in terms of meaningful semantics concepts.***

**SEMantic EXtraction Framework (SEM-EX)**



# Thanks for your time

For any doubt or suggestion

Rosa Sicilia, r.sicilia@unicampus.it
Assistant Professor (RTDA)
Computer Systems & Bioinformatics Laboratory
Department of Engineering, University Campus Bio-Medico of Rome