Multimodal Explainability via Latent Shift applied to COVID-19 stratification

Valerio Guarrasi[a], Lorenzo Tronchin[a], Domenico Albano[b,c], Eliodoro Faiella[d], Deborah Fazzini[e], Domiziana Santucci[a,d], Paolo Soda[a,f,*]

[a]*Unit of Computer Systems and Bioinformatics, Department of Engineering, University Campus Bio-Medico of Rome, Rome, Italy*
[b]*IRCCS Istituto Ortopedico Galeazzi, Milan, Italy*
[c]*Department of Biomedical, Surgical and Dental Sciences, Università degli Studi di Milano, Milan, Italy*
[d]*Department of Radiology, Sant'Anna Hospital, San Fermo della Battaglia, Como, Italy*
[e]*Department of Diagnostic Imaging and Stereotactic Radiosurgey, Centro Diagnostico Italiano S.p.A., Milan, Italy*
[f]*Department of Diagnostics and Intervention, Radiation Physics, Biomedical Engineering, Umeå University, Umeå, Sweden*

Abstract

We are witnessing a widespread adoption of artificial intelligence in healthcare. However, most of the advancements in deep learning in this area consider only unimodal data, neglecting other modalities. Their multimodal interpretation necessary for supporting diagnosis, prognosis and treatment decisions. In this work we present a deep architecture, which jointly learns modality reconstructions and sample classifications using tabular and imaging data. The explanation of the decision taken is computed by applying a latent shift that, simulates a counterfactual prediction revealing the features of each modality that contribute the most to the decision and a quantitative score indicating the modality importance. We validate our approach in the context of COVID-19 pandemic using the AIforCOVID dataset, which contains multimodal data for the early identification of patients at risk of severe outcome. The results show that the proposed method provides meaningful explanations without degrading the classification performance.

Keywords: XAI, Multimodal deep learning, Joint fusion, Classification, COVID-19

*Corresponding author
Email address: paolo.soda@umu.se (Paolo Soda)

1. Introduction

In the last decade, the practice of modern medicine has started to heavily rely on the utilization of data coming from multiple sources [1]. At the same time, artificial intelligence (AI) has achieved state-of-the-art results in various domains [2], healthcare included. Nevertheless, most of the deep neural networks applied to medical tasks consider only one data modality, neglecting information available in other sources. However, analyzing medical findings is multimodal by its very nature: a characteristic that, in turn, asks for developing AI approaches able to process data of different modalities [3, 4]. This has fostered the rise of multimodal deep learning (MDL) [4], which aims to develop learning models able to process and link information gathered from different modalities. MDL is a topic where researchers have investigated several methods to learn together multimodal information via early, late, and joint fusion [4], as it will be presented in section 2.

With the goal of reaching high performance, many complex AI models developed so far have a black-box nature [5], neglecting trust and transparency [6], two features that are of particular importance in biomedicine [7]. Indeed, a lack in explainability limits the application of AI models into the clinical practice. To overcome this limitation, in the last years large research efforts have been directed towards explainable AI (XAI), which aims to explain how a black-box model produces its outcomes.

In healthcare, the need of multimodal models and of explainability make multimodal explanations vital to develop robust and trustworthy AI models. This happens because multimodal models extract more comprehensive information than the unimodal models, so that their explanations could offer more insights into the available medical data. The multimodal setting of XAI explores the complementary and explanatory strengths of the different modalities, with the goal of obtaining better explanations that localize the relevant features and modalities [8]. Despite this relevance, to the best of our knowledge, the biomedical literature lacks of explainable deep multimodal models. For this reason we present a novel end-to-end multimodal architecture, with intrinsic explanations, that jointly learns modality reconstructions and multimodal classification using imaging and tabular modalities. With this architecture we extract

deep multimodal representations of the data; then, we apply a latent shift to simulate a counterfactual prediction, thus obtaining an intrinsic explanation that reveals how and why the model arrived at a particular decision.

We test our approach in the context of the COVID-19 pandemic using the AIforCOVID public dataset [9] for three reasons. First, the development of AI-based tools supporting COVID-19 prognosis exploiting multimodal data is still an open research issue addressed by few work in the literature [9, 10, 11, 12, 13, 14, 15], which will be discussed in section 2. Second, while there is a lack of multimodal XAI (MXAI) approaches in general, no one has proposed multimodal explanations to gain trust and transparency in COVID-19 prognosis. Third, the AIforCOVID dataset is the largest publicly available repository containing chest X-ray (CXR) images and clinical data collected at the time of hospitalization, further to clinical outcomes stratifying patients into those with and without risk of a severe disease progression [16]. It is worth noting that, in general, images and clinical data are two important sources of information in medicine. Indeed, the former allows radiologists to focus on visual evidence for both diagnostic and prognostic purposes, whereas the latter which are usually stores as tabular data, offer to clinicians a concise and multi-dimensional assessment of patients' health status. Hence, having both the modalities should be important to test MDL and MXAI approaches that would support the medical decision process.

The main contributions of our work are:

- The development of an intrinsic explainable architecture specifically designed for multimodal classification.
- The introduction of a joint learning approach that enables to simultaneously train both the data reconstruction and the classification tasks using tabular and imaging data.
- The proposal of a novel latent space counterfactual method that allows for explainability in both multimodal and unimodal contexts. It reveals the modalities and features that contribute the most to the decision-making process.
- The effective application of the proposed approach in the context of the COVID-

19 pandemic to early identify patients at risk of severe outcomes, using the publicly available AIforCOVID dataset.

- The validation of the proposed method, strengthened by a reader study with four radiologists, showing that it provides meaningful explanations without sacrificing the classification performance.

The rest of the manuscript is organized as follows. Section 2 introduces the state-of-the art of both MDL and MXAI. Then, section 3 presents our novel architecture and training procedure, and it explains the MXAI method extracting multimodal explanations. In section 4 we describe the dataset used to validate the methods, the pre-processing phase on the data, the implementation setup and the validation strategy adopted. Section 5 presents and discusses the obtained results, whilst section 6 provides concluding remarks.

2. Background

In this section we first present the state-of-the-art of MDL in the context of COVID-19 prognosis prediction by using images and tabular clinical data, whereas the interested readers can deepen MDL in healthcare in recent surveys [1]. Indeed the literature is quite large and its review is out of the scope of this work. Second, given the lack of MXAI approaches in COVID-19 prognosis [17], we will summarize MXAI research in healthcare, whilst the readers can refer to [18] for XAI applications on unimodal data in COVID-19 imaging. We will conclude this section by summarizing the motivations of this work.

2.1. MDL

There is a general consensus that medical images complemented by clinical data can help physicians, and radiologist in particular, better understanding the patient's state, thus advancing to a more informative decision making process [19]. In this respect, research in the field of MDL has been increasing [4] since multimodal data give the opportunity to train models that can learn the complex dynamics behind a disease.

The level at which the fusion of input modalities occurs in the network is usually distinguished into early, intermediate or late fusion [4]. Early fusion combines raw data or extracted features by the different modalities which are fed to a simple learner, whilst late fusion combines at the decision level the outputs of networks independently trained on each modality. Intermediate fusion learns a joint representation of different modalities at a shared representation layer, by propagating the loss back to the feature extractor network in an end-to-end manner.

The application of AI in COVID-19 using medical imaging and clinical data has mainly focused on discriminating patients suffering from COVID-19 pneumonia from those which are healthy or affected by different types of pneumonia [20]. Nevertheless, only four work investigated patients' stratification into mild and severe outcomes using such multimodal data [9, 13, 14, 15], which can be further divided in three using computed tomography (CT) scans [13, 14, 15] and one using CXR images [9].

In [13] the authors proposed a deep network using CT scans and 53 clinical features to detect the potential malignant progression of mild patients. Via a multi-layer perceptron (MLP) processing the clinical features only, their method gets an embedding that is then concatenated in an early fusion approach with the flattened CT scan. This new vector then fed a Long short-term memory (LSTM) network followed by a fully-connected (FC) network, which produces the output. On a private cohort of 199 patients, they obtained an accuracy of 79.20%. In [14] the authors used 130 clinical features and CT scans to discriminate between negative, mild and severe COVID-19 cases via an early fusion of a VGG-16 and a 7-layer FC network. They used a private dataset containing 1521 patients, achieving an accuracy equal to 81.10%. Fang et al applied joint fusion combining CT scans and 61 clinical features, which fed a deep network to predict COVID-19 malignant progression [15]. The approach extracts the abstract representations of CT images using a 3D ResNet, and of the clinical data via a FC network. These two embeddings are concatenated and given to an LSTM followed by another FC network. The whole architecture is training in an end-to-end modality. On a private dataset containing 1040 patients they achieved an accuracy equal to 87.70%.

It is worth noting that CXR helps indicating abnormal formations of a large va-

riety of chest diseases by using a very small amount of radiation, whilst CT delivers a much higher detail level of the lungs' structures. Furthermore, X-ray equipment is much smaller, less complex, and with lower costs than CT scans; it also prevents other three limitations of CT imaging, i.e., the lack of available machines' slots, the difficulty of moving bedridden patients, and the long sanitation times. For these reasons, several authors indicated that CXR imaging fit well with the needs of COVID-19 pandemic [21]. Let us now focus on the only work that uses CXR images and clinical data for the COVID-19 prognosis [9]. There, the authors presented three different multimodal AI approaches, offering also baseline performance on the AIforCOVID dataset, which we will present in section 4 together with the best results attained. The first method, named as handcrafted approach, computed first-order and texture features from the images, which are then stacked in an early fusion fashion with the clinical features, then feeding different learners among which the Support Vector Machine resulted to be the best. The second approach, referred to as hybrid, combines automatic features extracted by a convolutional neural network with the clinical ones; then it runs a feature selection stage whose output is given to the learner. The hybrid approach achieved the best results using the GoogleNet and the Support Vector Machine classifier. The third approach, named as end-to-end, performs intermediate fusion of the two modalities by defining a multi-input network concatenating hidden vectors of the two modalities. This architecture contains three main branches: two process independently CXR scans and clinical features to get a small number of relevant and abstract features, while the third one concatenates such embeddings that is given to a FC network, which outputs the prognosis.

2.2. MXAI

High performing deep models are often black-boxes, which hide their decision-making process, making it hard to understand why a certain result is obtained. This has boosted the growth of XAI, and many unimodal methods have been proposed to extract explanations on how the model has interpreted the data [6] with applications to different fields. In particular, many authors agree that explanations are strongly recommended in medical applications [22], because this would help mapping explainability

with causability that, in turns, would allow practitioners to understand why a model came up with a result.

The only available review on MXAI [23] surveys its applications in computer vision and natural language processing, showing that MXAI lacks in the medical field. This is confirmed by the position paper [24], which states that in radiology there is a lack of integrative methods that, combining imaging and tabular data, provide explanations on the decisions taken. This confirms the need of multimodal explanations to capture the complexity of all the factors underlying a disease. Indeed, for a medical task to have a comprehensive global view of the data and of the system, an ideal MXAI method should be able to identify the importance of each modality and the importance of each unimodal feature.

The MXAI review [23], even if it does not focus on the medical field, is also interesting because it groups XAI algorithms adopting three different criteria. First, it focuses on the stage at which the XAI can be applied, identifying pre-modeling, during modelling and post-hoc modeling explanations [23]. As their names explain, the pre-modeling methods' explainability is included before the model development, during modelling include the models which are usually explainable by design and employ intrinsic methods, and post-hoc modelling is applied after the model is developed by extracting explanations via perturbations or backpropagation methods [25]. Second, with reference to the scope of the explanation, XAI models can be either local or global [23], depending if the explanation regards a single instance or the model as a whole. The third criterion deal with the dependency of the XAI algorithm, so that it distinguishes model specific and model agnostic explanations [23]. While the interested readers can deepen [23] to have more details, on the basis of this survey we observe that there is a lack in multimodal intrinsic explainability, i.e., methods able to return multimodal local explanations.

As mentioned in the forewords of this section, the analysis of the literature that uses multimodal data for COVID-19 prognosis shows that, to the best of our knowledge, none has investigated MXAI in this field yet.

2.3. Motivations

Multimodal settings have improved the predictive power of models in many applications thanks to the interaction of different modalities, via a richer representation with task-relevant features [8]. Nevertheless, this availability of information from different modality makes explainability a key necessity to reduce the opacity of the multimodal deep architectures [7]. This has recently fostered the raise of MXAI, which has mainly focused on computer vision and natural language processing. Indeed, the literature on XAI in medical applications has concentrated more on unimodal attribution methods, struggling in having explanations of neural networks working on multiple data sources. Therefore, developing multimodal methods for explainability is an urgent and open issue, also because the development of multimodal deep architectures in different healthcare applications asks for novel approaches to open such black boxes. In turn, this can help physicians, patients and regulators to trust the decisions taken. Among the several fields where MDL and MXAI can be applied, we test our methodology to the early identification of COVID-19 patients at risk of severe outcome using imaging and tabular data, because the survey of the literature presented hereinbefore shows that few work has addressed this challenge, despite the disruptive impact of this disease worldwide.

3. Methods

In this section we present a novel architecture that exploits joint learning, for which we design an intrinsic counterfactual MXAI approach to extract explanations of a classification task. In general, counterfactual explanations refer to a type of explanation that aims to comprehend the causes of an observed outcome by exploring alternative scenarios, which helps in gaining a deeper understanding of the causal relationships that led to the observed outcome [5]. Such multimodal explanations will permit users to understand not only the importance of each modality for each classification, but also the features which contributed the most to the decision for every single modality.

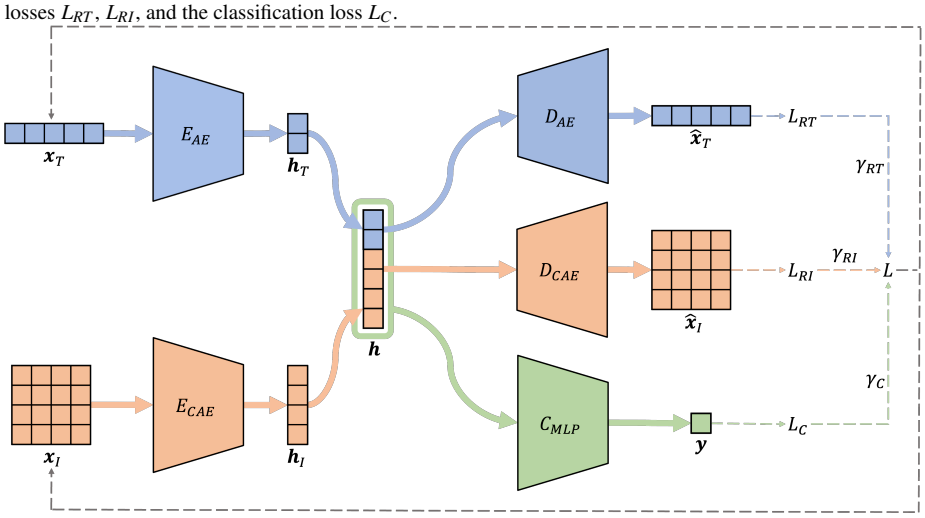
We first present the architecture of the multimodal model; second we focus on the training approach and, third, we detail the intrinsic MXAI method.

3.1. Notation

The notation used henceforth makes us of the following symbols:

- T and I are the tabular and the imaging modalities, respectively;
- $\boldsymbol{x}_T$ and $\boldsymbol{x}_I$ are the inputs for the tabular and imaging modality, respectively;
- AE and CAE are the autoencoder and convolution autoencoder which receive as input $\boldsymbol{x}_T$ and $\boldsymbol{x}_I$, respectively. Both are composed of an encoder E_{AE}, E_{CAE} and a decoder D_{AE}, D_{CAE}, respectively;
- $\boldsymbol{h}_T \in \mathbb{R}^n$ and $\boldsymbol{h}_I \in \mathbb{R}^m$ are the latent vectors of the AE and the CAE, respectively. Their concatenation produces the multimodal embedding $\boldsymbol{h} \in \mathbb{R}^{n+m}$;
- $\hat{\boldsymbol{x}}_T$ and $\hat{\boldsymbol{x}}_I$ are the outputs produced by the AE and the CAE, respectively, representing the reconstruction of the inputs $\boldsymbol{x}_T$ and $\boldsymbol{x}_I$;
- C_{MLP} is the multi-layer perceptron receiving the vector $\boldsymbol{h}$ to perform the classification;
- $\boldsymbol{y} \in \mathbb{R}^c$ is the output vector of C_{MLP}, which expresses the predicted posterior probability, with c being equal to the number of classes;
- L_T, L_I, L_C are the loss functions of the AE, CAE and C_{MLP}, respectively, whose linear combination results in L, weighted by the corresponding scalar parameters $\gamma_T \in \mathbb{R}$, $\gamma_I \in \mathbb{R}$, $\gamma_C \in \mathbb{R}$;
- $\boldsymbol{h}_T^\lambda \in \mathbb{R}^n$, $\boldsymbol{h}_I^\lambda \in \mathbb{R}^m$ and $\boldsymbol{h}^\lambda \in \mathbb{R}^{n+m}$ are the modified vector embeddings of $\boldsymbol{h}_T$, $\boldsymbol{h}_I$ and $\boldsymbol{h}$, respectively, regulated by scalar parameter $\lambda \in \mathbb{R}$;
- $\hat{\boldsymbol{x}}_T^\lambda$, $\hat{\boldsymbol{x}}_I^\lambda$ and $\boldsymbol{y}^\lambda \in \mathbb{R}^c$ are the outputs produced by D_{AE}, D_{CAE} and C_{MLP}, respectively, when the input is $\boldsymbol{h}_T^\lambda$, $\boldsymbol{h}_I^\lambda$ and $\boldsymbol{h}^\lambda$, respectively;
- $\Delta_T \in \mathbb{R}$ and $\Delta_I \in \mathbb{R}$ express the resulting modality importance comparing $\boldsymbol{h}_T$ with $\boldsymbol{h}_T^\lambda$ and $\boldsymbol{h}_I$ with $\boldsymbol{h}_I^\lambda$, respectively;
- $\hat{\boldsymbol{\Delta}}_T \in \mathbb{R}^n$, $\hat{\boldsymbol{\Delta}}_I \in \mathbb{R}^m$ express the resulting unimodal feature importance comparing $\hat{\boldsymbol{x}}_T$ with $\hat{\boldsymbol{x}}_T^\lambda$ and $\hat{\boldsymbol{x}}_I$ with $\hat{\boldsymbol{x}}_I^\lambda$, respectively.

Figure 1: Schematic view of the multimodal deep architecture: for each instance, the input modalities $\boldsymbol{x}_T$ and $\boldsymbol{x}_I$ feed into their corresponding encoders E_{AE} and E_{CAE}, obtaining the unimodal embeddings $\boldsymbol{h}_T$ and $\boldsymbol{h}_I$, respectively. These embeddings are then concatenated into the multimodal embedding $\boldsymbol{h}$, which subsequently feeds into the decoders D_{AE}, D_{CAE}, and the classifier C_{MLP}. The resulting outputs are the reconstructions $\hat{\boldsymbol{x}}_T$, $\hat{\boldsymbol{x}}_I$, and classification $\boldsymbol{y}$, respectively. The model is trained by simultaneously minimizing the reconstruction losses L_{RT}, L_{RI}, and the classification loss L_C.



3.2. Architecture

Here we present the structure of the designed classification model that works with T and I. The proposed multimodal architecture consists of three blocks: an autoencoder (AE), a convolutional autoencoder (CAE), and a multi-layer perceptron classifier (C_{MLP}). As shown in Figure 1, the network has two inputs and three outputs. The tabular modality $\boldsymbol{x}_T$ feeds the AE, whereas the imaging modality $\boldsymbol{x}_I$ is given to the CAE. Both are composed of an encoder and a decoder, returning the reconstruction of the respective modality, $\hat{\boldsymbol{x}}_T$ and $\hat{\boldsymbol{x}}_I$. By concatenating the two embeddings $\boldsymbol{h}_T$ and $\boldsymbol{h}_I$ we get $\boldsymbol{h}$, which is given to the C_{MLP} classifier that returns the classification vector $\boldsymbol{y}$. The entire architecture is trained in an end-to-end manner, via a linear combination of three loss functions, two for reconstruction (L_T and L_I) and one for the classification (L_C).

This overview reveals that our framework jointly learns deep representations of imaging and tabular data to perform a classification task. Indeed, it learns a feature space with local modality structure able to be used for reconstruction, and it manipulates the combined feature space by incorporating a classification oriented loss.

Autoencoders. An AE and a CAE are artificial neural networks which learn an approximation to the identity function, with the goal of minimizing the distance between the outputs $\hat{\boldsymbol{x}}_T$, $\hat{\boldsymbol{x}}_I$ and the inputs $\boldsymbol{x}_T$, $\boldsymbol{x}_I$, respectively. The encoders E_{AE} and E_{CAE} compress the corresponding inputs $\boldsymbol{x}_T$, $\boldsymbol{x}_I$ to a latent space representation $\boldsymbol{h}_T$ and $\boldsymbol{h}_I$, using fully connected layers in the AE and convolutional layers in the CAE, respectively. The decoders D_{AE} and D_{CAE} use the bottleneck latent space representation $\boldsymbol{h}_T$ and $\boldsymbol{h}_I$ to reconstruct the inputs $\boldsymbol{x}_T$, $\boldsymbol{x}_I$ in $\hat{\boldsymbol{x}}_T$, $\hat{\boldsymbol{x}}_I$, respectively. Therefore:

$$\boldsymbol{h}_T = E_{AE}(\boldsymbol{x}_T) \tag{1}$$

$$\boldsymbol{h}_I = E_{CAE}(\boldsymbol{x}_I) \tag{2}$$

$$\hat{\boldsymbol{x}}_T = D_{AE}(\boldsymbol{h}_T) \tag{3}$$

$$\hat{\boldsymbol{x}}_I = D_{CAE}(\boldsymbol{h}_I) \tag{4}$$

When training the AE and the CAE we aim to minimize the distance between its inputs and outputs over all samples, using as reconstruction loss functions L_T and L_I for the tabular and imaging modalities, respectively. We constrain the dimension of latent spaces $\boldsymbol{h}_T$ and $\boldsymbol{h}_I$ to be lower than input data $\boldsymbol{x}_T$ and $\boldsymbol{x}_I$, respectively, forcing both the AE and the CAE to capture the most salient features of the data. This is a well-known approach to avoid identity mapping [26].

Classifier. The two embeddings $\boldsymbol{h}_T$ and $\boldsymbol{h}_I$ are concatenated in $\boldsymbol{h}$ and used as input to the fully connected C_{MLP}, which performs the classification task returning $\boldsymbol{y}$, So that:

$$\boldsymbol{y} = C_{MLP}(\boldsymbol{h}) \tag{5}$$

The goal of this block is to minimize the classification error with a classification loss L_C. Note that the final layer of the C_{MLP} uses the Softmax activation function, such that

$$\sum_{i=1}^{c} \boldsymbol{y}_i = 1 \tag{6}$$

This implies that $\boldsymbol{y}$ can be considered as an estimate of the posterior probability.

End-to-end training. In this way, the network's training can be back-propagated in an end-to-end manner, via a linear combination of the three loss functions:

$$L = \gamma_T L_T + \gamma_I L_I + \gamma_C L_C \tag{7}$$

where γ_T, γ_I, and γ_C are parameters, which regulate the importance of each loss function.

This approach has the beneficial effect of being able to learn embedded features in an end-to-end way, which are jointly used to perform data reconstruction and classification, minimizing the reconstruction loss of AE and the CAE and the classification loss of the C_{MLP}. Our key idea is that the co-learning of the AE, the CAE and the C_{MLP} is beneficial to learn features from the tabular and imaging modality to obtain a classification and a good reconstruction useful for the explainability, presented in section 3.4.

3.3. Three-stage training

Given the complex structure of the architecture proposed, we train the network with a three-stage procedure, which adapts the γ_T, γ_I, γ_C parameters in way to concentrate the training on different parts of the network. The three stages are:

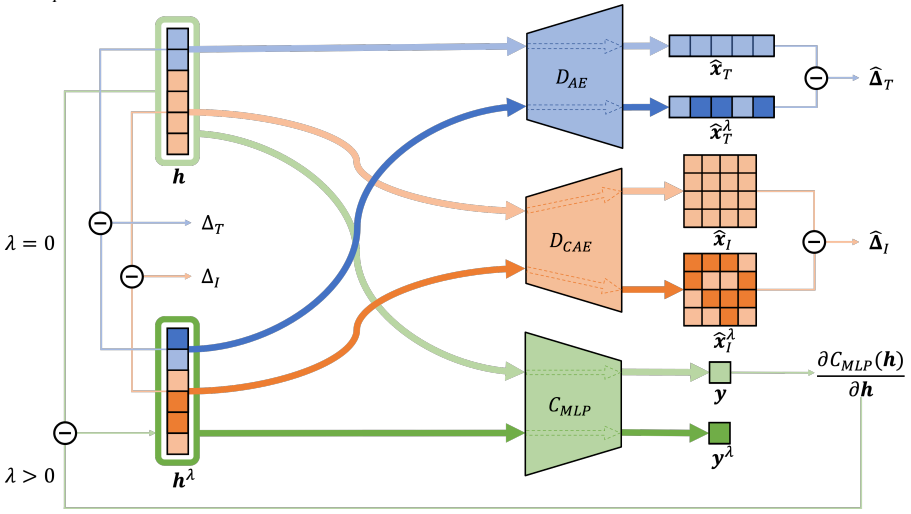
1. Setting $\gamma_T = 1$, $\gamma_I = 0$ and $\gamma_C = 0$ to train only the weights of the AE;
2. Setting $\gamma_T = 0$, $\gamma_I = 1$ and $\gamma_C = 0$ to train only the weights of the CAE;
3. Setting $\gamma_T = 1$, $\gamma_I = 1$ and $\gamma_C = 1$ to train all the weights of network.

The main idea is to help the training of the C_{MLP} classifier, giving initialization weights that constrict an optimal modality embedding for reconstruction, expressing a good summary of the data. Notice also that, given the architecture of the network, it is irrelevant if we invert stage 1 and 2 since the AE and CAE have no weights in common. In step 3 we decided to set all the parameters to 1 so that all the tasks would have equal weight.

3.4. MXAI

We use a gradient update, also referred to as latent shift, that can transform the latent representation of the inputs to exaggerate or curtail the features used for predic-

Figure 2: Schematic view of the MXAI framework: once the model is trained, each instance's multimodal embedding $\boldsymbol{h}$ feeds into the decoders D_{AE}, D_{CAE}, and the classifier C_{MLP}, according to colors that specify that portion of h is given to each network. The decoders and the classifier provide the original reconstructions $\hat{\boldsymbol{x}}_T$, $\hat{\boldsymbol{x}}_I$, and classification $\boldsymbol{y}$. Via the latent-shift method we obtain a $\lambda > 0$, which gives us a flip in the classification $\boldsymbol{y}^\lambda$ by feeding the shifted multimodal embedding $\boldsymbol{h}^\lambda$ to C_{MLP}. By feeding this new embedding to D_{AE} and D_{CAE}, we obtain new reconstructions $\hat{\boldsymbol{x}}_T^\lambda$ and $\hat{\boldsymbol{x}}_I^\lambda$. By comparing $\boldsymbol{h}$ with $\boldsymbol{h}^\lambda$, $\hat{\boldsymbol{x}}_T$ with $\hat{\boldsymbol{x}}_T^\lambda$, and $\hat{\boldsymbol{x}}_I$ with $\hat{\boldsymbol{x}}_I^\lambda$, we obtain the corresponding multimodal and unimodal explanations, respectively.



tion. Via the latent shift explanations we obtain both modality importance and feature importance for each prediction.

Latent shift. The only requisite to apply latent shift to a network is of having all the network components, which receive the latent vector $\boldsymbol{h}$, to be differentiable. With the *AE*, a *CAE* and a C_{MLP} we satisfy this requisite and, in addition, they are simple to implement and train. Once these components are trained, we extract the explanation as shown in Figure 2. Multimodal input instances $\boldsymbol{x}_T$ and $\boldsymbol{x}_I$ are encoded producing the multimodal latent representations $\boldsymbol{h}_T$ and $\boldsymbol{h}_I$, which are combined into $\boldsymbol{h}$, as already described. Perturbations of this latent embedding are computed via

$$\boldsymbol{h}^\lambda = \boldsymbol{h} - \lambda \frac{\partial C_{MLP}(\boldsymbol{h})}{\partial \boldsymbol{h}} \tag{8}$$

where $\lambda \in \mathbb{R}$ is a parameter establishing how much the original embedding is modified. With $\lambda > 0$, we expect that $C_{MLP}(\boldsymbol{h}^{\lambda})$ would provide a prediction $\boldsymbol{y}^{\lambda}$ so that

$$\max(\boldsymbol{y}) \geq \boldsymbol{y}^{\lambda}(\arg\max(\boldsymbol{y})) \tag{9}$$

This implies that, as λ increases, we expect a flip of the predicted class. In other words, guided by the direction of variation of the output in the latent space determined by the gradient of network output, we are interested in determining the λ value for which a classification label flip occurs. With too small values of λ, for the smoothness principle, the difference between the original modality input and the reconstruction will not be large enough to change the prediction of the model. On the contrary, too large values of λ would distort the reconstruction so much that it will not be useful for explainability. Thus, to find the value of λ where the class flip occurs, we use an iterative search that, starting from $\lambda = 0$, and using a fixed step heuristically set to 10, increases λ until $\boldsymbol{y}^{\lambda} \neq \boldsymbol{y}$.

We produce λ-shifted counterfactual multimodal reconstructions $\boldsymbol{x}_T^{\lambda}$, $\boldsymbol{x}_I^{\lambda}$ and output probabilities $\boldsymbol{y}^{\lambda}$ defined as:

$$\hat{\boldsymbol{x}}_T^{\lambda} = D_{AE}(\boldsymbol{h}_T^{\lambda}) \tag{10}$$

$$\hat{\boldsymbol{x}}_I^{\lambda} = D_{CAE}(\boldsymbol{h}_I^{\lambda}) \tag{11}$$

$$\boldsymbol{y}^{\lambda} = C_{MLP}(\boldsymbol{h}^{\lambda}) \tag{12}$$

where $\boldsymbol{h}_T^{\lambda}$ and $\boldsymbol{h}_I^{\lambda}$ are given by:

$$\boldsymbol{h}_T^{\lambda} = \boldsymbol{h}_T - \lambda \frac{\partial C_{MLP}(\boldsymbol{h})}{\partial \boldsymbol{h}_T} \tag{13}$$

$$\boldsymbol{h}_I^{\lambda} = \boldsymbol{h}_I - \lambda \frac{\partial C_{MLP}(\boldsymbol{h})}{\partial \boldsymbol{h}_I} \tag{14}$$

so that $\boldsymbol{h}^{\lambda}$ is the concatenation of $\boldsymbol{h}_T^{\lambda}$ and $\boldsymbol{h}_I^{\lambda}$.

It is worth noting that finding an informative latent space relies on the quality of the AE and CAE. This justifies even more the use of the three-stage training, facilitating the training of the AE and CAE.

Modality importance. Since the multimodal embedding $\boldsymbol{h}$ is composed by the concatenation of the unimodal embeddings $\boldsymbol{h}_T$ and $\boldsymbol{h}_I$, we know to which modality each element of $\boldsymbol{h}$ is associated to. Once calculated $\boldsymbol{h}_T^\lambda$ and $\boldsymbol{h}_I^\lambda$, we compute the modality normalized absolute differences, to understand how much each element has been shifted:

$$\Delta_T = \frac{\|\boldsymbol{h}_T - \boldsymbol{h}_T^\lambda\|_1}{n} \tag{15}$$

$$\Delta_I = \frac{\|\boldsymbol{h}_I - \boldsymbol{h}_I^\lambda\|_1}{m} \tag{16}$$

where n and m denote the number of elements in each vector, and $\|.\|_1$ denotes the l_1-*norm*. Hence, Δ_T and Δ_I express the importance of each modality: the more a modality embedding has changed, the more important it is for the classification of a given sample.

Feature importance. Similar to the modality importance, we now focus on an approach to reveal which features per modality are more important for the classification of a certain instance. Using the shifted reconstructions $\hat{\boldsymbol{x}}_T^\lambda$ and $\hat{\boldsymbol{x}}_I^\lambda$, we compute the absolute differences with the original reconstructions $\hat{\boldsymbol{x}}_T$ and $\hat{\boldsymbol{x}}_I$

$$\hat{\boldsymbol{\Delta}}_T = |\hat{\boldsymbol{x}}_T - \hat{\boldsymbol{x}}_T^\lambda| \tag{17}$$

$$\hat{\boldsymbol{\Delta}}_I = |\hat{\boldsymbol{x}}_I - \hat{\boldsymbol{x}}_I^\lambda| \tag{18}$$

Note that $\hat{\boldsymbol{\Delta}}_T$ and $\hat{\boldsymbol{\Delta}}_I$ make us understand for each feature how much it has changed for the classification shift. The more a feature changes, the more important it is for the classification. This works for both modalities, resulting $\hat{\boldsymbol{\Delta}_T}$ to be an importance vector for the tabular modality and $\hat{\boldsymbol{\Delta}}_I$ to be an importance matrix for the imaging modality.

Putting our method in the taxonomy. Following the taxonomy introduced in section 2 and originally presented in [23], our proposal is an hybrid between during and post-hoc modelling as it exploits a specific network architecture using both perturbations and backpropagation methods to extract the explanations. In particular our method uses counterfactual explanations, specifying the minimal desired changes required to flip

the decision, mapping the class-specific and discriminative features of each modality. In addition, our MXAI method is local, since we are interested in explaining how the model functions at instance level. Finally, our method is model-specific since its architecture is constructed in a way to output the explanations.

4. Experimental configuration

In this section we introduce the used dataset and how the two modalities are pre-processed to train the network. Then, we deepen the validation phase of the proposed MXAI method where we conducted a reader study with four COVID-19 expert radiologists.

4.1. Dataset

For the last two years the world has been struck by the COVID-19 pandemic causing millions of cases and deaths. During this period, many researchers practitioners and companies have developed novel AI methods and tools to combat the rising of the pandemic by deepening the virus's understanding. Many studies have focused their attention on unimodal data using CXR, CT or clinical examinations to replace or to supplement the reverse transcriptase-polymerase chain reaction tests. But given the multimodal nature of medicine, both imaging data and clinical information can help radiologists and practitioners on determining the source of symptoms, stratifying the disease severity, and establishing the best treatment plan for the patient's specific needs.

We use the AIforCOVID imaging archive [9] because it is the only publicly available multimodal dataset on COVID-19 stratification, as shown in the survey [16]. The archive includes clinical data (tabular modality) and CXR scans (imaging modality) of 820 patients recorded from six different Italian hospitals. In particular, there are 120, 104, 31, 139, 101, and 325 patients per hospital. The interested readers can refer to [9, 27, 28, 29, 30] for further details.

The patients' data were collected at the time of hospitalization if the TR-PCR test resulted positive to the SARS-CoV-2 infection. All the patients were assigned to the mild or severe class, on the basis of the clinical outcome. The mild group includes

384 patients who were either sent back to domiciliary isolation or hospitalized without any ventilatory support, whereas the severe group is composed of 436 patients who required non-invasive ventilation support, intensive care unit admission, or those who died. Furthermore, any AI model trained on the AIforCOVID dataset is exposed to a diverse range of patient populations since it incorporates data from multiple centers, which should help ensure that the model is more generalizable and applicable to a wider extent.

4.2. Pre-processing

We applied the same pre-processing procedure and validation approach presented in [9] to avoid any performance bias, which are now briefly summarized for the sake of presentation.

Tabular data. We use the 34 clinical descriptors indicated in [9] which are not direct indicators of the prognosis. Missing data were imputed using the mean and the mode for continuous and categorical variables, respectively. A min-max scaler was applied along the variables to have the features all in the same range $[0, 1]$.

Imaging data. This modality consists of CXR scans, which were processed by extracting the segmentation mask of lungs, using a U-Net trained on two non-COVID-19 lung datasets [31, 32]. The mask was used to extrapolate the minimum squared bounding box containing both lungs. The extracted box was then resized to 224×224 matrix, and normalized with a min-max scaler bringing the pixel values in the range $[0, 1]$.

4.3. Implementation setup

Here we describe the architectures of the three blocks of the model, as well as the parameters and settings used during training.

The AE's input and output layers consist of 34 (one for each feature) and 2 (one for each class) neurons, respectively. Its encoder and decoder are composed of fully connected hidden layers activated by ReLU functions. We opted to use such architectures since these feed-forward networks are able to learn a low-dimensional representation

before being fused with the other modality [33]. Both E_{AE} and D_{AE} have 6 hidden layers and $n = 8$. The loss function L_T is the mean squared error (MSE).

The *CAE* is a 2D ResNet101 [34], which we selected because of the skip connections that mitigate the problem of the vanishing gradient, ensuring high fidelity image reconstruction. Both the input and the output of the network have a 224×224 dimension, so that the dimension of the embedding is $m = 4608$. The dimensions the embeddings of both the *AE* and the *CAE* were chosen small enough to prevent the curse of dimensionality. To facilitate the reconstruction training, this model was pretrained trained on 4 different CXR datasets [35], that in total account for a total of 674525 scans. For consistency, the corresponding loss function L_I is the MSE.

Let us now focus on the C_{MLP} classifier: its input and output layers consist of 4616 and 2 neurons (one for each class), respectively. It is composed of 7 fully connected hidden layers (with 512, 256, 128, 64, 32, 16, 8 neurons, respectively) activated by ReLU functions, with a Softmax activation in the output layer. We design this structure to gradually learn the classification from the multimodal embedding. The loss function L_C is the cross-entropy.

For all the three stages of the training, introduced in section 3.3, we adopt the same training procedure of [9], now summarized. To prevent overfitting of the *CAE*, we applied the following image random transformations: horizontal or vertical shift ($-20 \leq$ pixels ≤ 20), random zoom ($0.9 \leq$ factor ≤ 1.1), vertical flip, random rotation ($-15° \leq$ angle $\leq 15°$), and elastic transform ($20 \leq \alpha \leq 40$, $\sigma = 7$). No augmentation was applied on the tabular data. The loss functions L_T, L_I, L_C and L are regulated by an Adam optimizer with an initial learning rate of 0.001, which is scheduled to reduce by an order of magnitude every time the minimum validation loss does not change for 10 consecutive epochs. To prevent overtraining and overfitting we fixed the number of maximum epochs to 300, with an early stopping of 25 epochs on the validation loss.

4.4. Validation approach

To understand the robustness of the model we trained the network in 10-fold stratified cross-validation (CV), and leave-one-center-out CV (LOCO), following the same experimental procedure described in [9], thus ensuring a fair competition between the

approaches. In CV, the fold distribution of the training, validation and testing sets is 70%-20%-10%, respectively. In LOCO validation we study how the models generalize to different data sources, since in each fold the test set contains all the samples belonging to only one of the six hospitals that, of course, are not in the training and validation sets.

All the experiments were performed using a batch size of 16 on a NVIDIA TESLA A100 GPU with 32 GB of memory, using PyTorch as the deep learning library.

4.5. Sanity check

To study the validity of the proposed MXAI method, we conducted a reader study with four radiologists assessing the prognosis of 96 patients randomly extracted. Each radiologist has more than 10 years of experience. The radiologists R_1, R_2, R_3, R_4 were presented with a survey that has two aims. The first is to compare our method's classification performance with the one of human experts. The second is to understand if the importance metrics Δ_T, Δ_I, $\hat{\mathbf{\Delta}}_T$ and $\hat{\mathbf{\Delta}}_I$, returned by our method, are coherent with the ones selected by the radiologists, which we denote as $\Delta_T^{R_i}$, $\Delta_I^{R_i}$, $\hat{\mathbf{\Delta}}_T^{R_i}$ and $\hat{\mathbf{\Delta}}_I^{R_i}$. In particular, $\Delta_T^{R_i}$, $\Delta_I^{R_i}$ are the modality importance for the i^{th} radiologist, and $\hat{\mathbf{\Delta}}_T^{R_i}$, $\hat{\mathbf{\Delta}}_I^{R_i}$ are the unimodal feature importance vector and matrix for the i^{th} radiologist. The survey was executed in double-blind, where no interaction between the radiologists was permitted.

In the survey, each radiologist observed both data modalities at the same time for each patient and performs the prognosis task. Afterwards, the radiologists have to attribute an importance score, on a scale from 1 to 5, indicating how much significant each modality was for the prognosis task. The grading of the scores are: 1 insignificant, 2 a bit significant, 3 neutral, 4 significant, 5 important. A Softmax activation is applied constraining such values on the range $[0, 1]$, where 0 means that the considered modality has no importance and 1 attributes the maximum importance. As mentioned before, these modality importance are denoted as $\Delta_T^{R_i}$, $\Delta_I^{R_i}$. The radiologist has the possibility to attribute the same importance to each modality if he/she believes that, for that patient, both modalities had the same impact in the decision.

Then, to understand the most important features for each modality, we asked the

radiologist to select the clinical variables and to segment the areas of interest in the X-ray image most useful to stratify the patient, collecting $\hat{\mathbf{\Delta}}_T^{R_i}$ and $\hat{\mathbf{\Delta}}_I^{R_i}$, which are boolean, where elements equal to 1 correspond to important features. On both $\hat{\mathbf{\Delta}}_T^{R_i}$ and $\hat{\mathbf{\Delta}}_I^{R_i}$ a min-max normalization is applied on each instance, putting the elements on the range $[0, 1]$, where 0 means that the considered feature has no importance and 1 attributes the max importance.

In the case of modality importance, we would expect a high intersection between the information reported by the radiologists $\Delta_T^{R_i}$ and $\Delta_I^{R_i}$ with the output of our method Δ_T and Δ_I, respectively; the same holds in the case of unimodal feature importance, when comparing $\hat{\mathbf{\Delta}}_T^{R_i}$ and $\hat{\mathbf{\Delta}}_I^{R_i}$ with $\hat{\mathbf{\Delta}}_T$ and $\hat{\mathbf{\Delta}}_I$, respectively.

The surveys were executed on Google Forms, and the tool utilized to show and to segment the CXR scans was ITK-SNAP [36].

4.6. Statistical analysis

The accuracy, the sensitivity and the specificity are the evaluation metrics used to assess the classification performance, as in [9]. To asses if there exists a difference between the performance of our model and the baseline model we apply the one-way ANOVA and, to interpret the statistical significance, we used the pairwise Tukey test with a Bonferroni p-value correction at $\alpha = 0.05$.

As described at the end of the previous section, we validate the MXAI modality importance Δ_T, Δ_I and the feature importance $\hat{\mathbf{\Delta}}_T$, $\hat{\mathbf{\Delta}}_I$, comparing them with the importance proposed by the radiologists $\Delta_T^{R_i}$, $\Delta_I^{R_i}$ and $\hat{\mathbf{\Delta}}_T^{R_i}$, $\hat{\mathbf{\Delta}}_I^{R_i}$, respectively.

For the modality importance, we calculate the Pearson correlation ρ, and the paired sample t-test between the vector of importance modality (Δ_T and Δ_I) over the instances given by our method, with the corresponding importance ($\Delta_T^{R_i}$ and $\Delta_I^{R_i}$) vector reported by the radiologists. With the resulting statistics we can comprehend the measure of dependency between our method and the radiologists. The higher ρ, the more our explanations are coherent with the importance scores reported by the radiologists, if the resulting t-test is not statistically significant (p-value > 0.05).

Turning our attention to the feature importance, we compute the intersection over union (IoU) between the feature importance proposed by the radiologists ($\hat{\mathbf{\Delta}}_T^{R_i}$ and $\hat{\mathbf{\Delta}}_I^{R_i}$)

and the importance resulting from the latent shift ($\hat{\mathbf{\Delta}}_T$ and $\hat{\mathbf{\Delta}}_I$), respectively. For the tabular modality IoU_T, we take the important features presented by the radiologists $\hat{\mathbf{\Delta}}_T^{R_i}$ and the binarized feature vector $\hat{\mathbf{\Delta}}_T^b$ (such that the values of $\hat{\mathbf{\Delta}}_T$ which are < 0.5 are set to 0 and the ones ≥ 0.5 are set to 1), and compute

$$IoU_T = \frac{\hat{\mathbf{\Delta}}_T^{R_i} \cap \hat{\mathbf{\Delta}}_T^b}{\hat{\mathbf{\Delta}}_T^{R_i} \cup \hat{\mathbf{\Delta}}_T^b} \tag{19}$$

The higher the metric, the more concurrences there are between our method and the human annotation. Similarly, when analyzing the imaging modality IoU_I we take the segmented mask returned by the radiologists $\hat{\mathbf{\Delta}}_I^{R_i}$ and the binarized attribution map $\hat{\mathbf{\Delta}}_I^b$ (such that the values of $\hat{\mathbf{\Delta}}_I$ which are < 0.5 are set to 0 and the ones ≥ 0.5 are set to 1), and compute

$$IoU_I = \frac{\hat{\mathbf{\Delta}}_I^{R_i} \cap \hat{\mathbf{\Delta}}_I^b}{\hat{\mathbf{\Delta}}_I^{R_i} \cup \hat{\mathbf{\Delta}}_I^b} \tag{20}$$

As before, the higher the metric, the more concurrences there are between our method and the human annotations.

5. Results and discussion

In this section we present the results obtained, dividing the discussion into five subsections that, in order, deal with the classification and reconstruction performance, the three-stage training assessment, MXAI performance, an ablation study and forthcoming clinical applications.

5.1. Classification and reconstruction performance

Table 1 shows the classification performance and its columns specify: the learning model (human radiologists included), the validation approach (CV, LOCO or the reduced set of images used for the survey), the evaluation metrics employed, i.e., accuracy, sensitivity and specificity, for which we show the mean and the standard deviation of the metric in CV and in LOCO. The table is organized into four horizontal sections: the first, the second and the fourth report the performance attained by learning models, whereas the third shows the performance of the four radiologists. In particular, the first

Table 1: Classification performance.

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
Our proposal (one-stage training)	CV	70.38±1.78	72.57±1.72	68.62±1.12
	LOCO	68.35±1.17	70.92±1.08	62.16±1.89
	Survey	70.48	72.51	68.67

section shows results of our proposal and the second includes the best baseline model presented by [9], which were attained by the hybrid approach.

Given that our method aims to increase the explainability of the model and not necessarily increase the performance, we first verify that with the co-learning we do not have a drop in performance with respect to the baseline [9]. The results show that our model, even if it co-learns two tasks at once, obtains a performance similar to [9]. In fact, the differences between our model and the baseline on all metrics, in both CV and LOCO, are not statistically significant (p-value > 0.05). This suggests that our method is resilient to the notion that a decrease in performance is required to obtain better explanations.

We now compare our results with those of the radiologists: our proposal provides larger accuracy and sensitivity, while the specificity is lower. This happens because predicting the prognosis of a patient affected by COVID-19 is a difficult task, giving a hint that AI could aid practitioners in the decision-making process.

Let us recall that our method is jointly trained to classify and reconstruct the inputs via the autoencoders: for this reason in Table 2 we show the MSE of AE and CAE, i.e., the two autoencoders working with the T and I modalities, respectively. As in Table 1, here we have a similar row-column organization. Turning our attention to the first section of this table, it is worth noting that the small values of the MSE confirm the high quality of the reconstruction for both modalities. This ensures that our MXAI method

Table 2: Reconstruction performance.

Model	**Validation**	**Modality**	**MSE**
Our proposal (three-stage training)	CV	*T*	0.04±0.01
	LOCO	*T*	0.05±0.02
	Survey	*T*	0.04
	CV	*I*	0.03±0.01
	LOCO	*I*	0.04±0.02
	Survey	*I*	0.03
Our proposal (one-stage training)	CV	*T*	0.09±0.04
	LOCO	*T*	0.11±0.05
	Survey	*T*	0.09
	CV	*I*	0.07±0.02
	LOCO	*I*	0.09±0.03
	Survey	*I*	0.07

Table 3: ρ (on the lower triangular) and the corresponding t-test p-value (upper triangular) of the modality importance, computed for each pair between our model and the radiologists.

	Our proposal	R_1	R_2	R_3	R_4
Our proposal	-	0.26	0.50	0.37	0.42
R_1	0.78	-	0.28	0.32	0.45
R_2	0.84	0.90	-	0.35	0.50
R_3	0.77	0.82	0.78	-	0.41
R_4	0.79	0.77	0.83	0.85	-

can provide good interpretability since it relies on the quality of such reconstructions.

5.2. Three-stage training assessment

To validate the three-stage training introduced in section 3.3, we compare the classification and reconstruction performance with those attained adopting a one-stage training, which consists of skipping phases 1 and 2 of our method and directly train, in an end-to-end manner, the entire network with the combined loss L, without any pre-training. The corresponding results are shown in the last section of Tables 1 and 2. In the case of classification performance (Table 1), we observe that the one-stage training provides lower performance than our three-stage proposal, whatever the performance metric and whatever the validation approach. Furthermore, such performance differences are all statistically significant (p-value < 0.01). Similar considerations hold in the case of reconstruction performance (Table 2). These results confirm the usefulness of the three-stage training procedure, which aids the multimodal joint model in converging to a better solution.

Table 4: IoU_T (lower triangular) and the IoU_I (upper triangular) of the feature importance, computed for each pair between our model and the radiologists.

	Our proposal	R_1	R_2	R_3	R_4
Our proposal	-	59.62±3.13	62.97±2.61	63.77±2.25	64.63±2.76
R_1	52.37±3.12	-	60.56±2.78	63.42±3.39	61.81±3.28
R_2	54.72±2.86	52.37±2.78	-	60.98±3.76	62.44±2.99
R_3	53.52±3.21	54.69±3.42	51.23±2.98	-	63.36±3.64
R_4	51.31±2.69	52.73±3.31	54.66±2.79	55.43±3.05	-

5.3. MXAI performance

We now focus on validating the explanations provided by our proposal. Specifically, using the patients included in the survey, we compare the modality and feature explanations of our model to the importance reported by the radiologists.

Before going deep with the results, let us recall that in section 3.4, we formally put in relationship the counterfactual explanations with the data (equations 15, 16, 17, 18). Indeed, in the case of modality importance (Δ_T and Δ_I), a counterfactual explanation highlights how large is the perturbation of the abstract representation of the clinical features or of the images caught by the latent space (equations 15 and 16). In the case of the importance of each feature ($\hat{\mathbf{\Delta}}_T$ and $\hat{\mathbf{\Delta}}_I$), a counterfactual explanation works at level of each descriptor: for clinical data it represents how large is the variation between the original and the reconstructed clinical information (equation 17), whereas for imaging data it measures pixels variations (equation 18). The quantities defined in such four equations are then considered in the sanity check (section 4.5), which we introduced to validate the MXAI method.

According to section 4.6, where we explain how we quantitatively compare the explanations provided by the model and those provided by the four radiologists, Table 3 shows the Pearson correlation ρ and the corresponding t-test p-values computed between the importance vector of a modality reported by our method and the importance vector reported by each radiologist. These results reveal a high measure of dependency between our method and the radiologists and among the radiologists, suggesting that our model gives reasonable modality importance while producing the prognosis.

Let us now consider the unimodal feature importance: in this case Table 4 shows the IoU_T and IoU_I for each possible pair between our model and the radiologists. As

Table 5: Comparison of IoU_T and IoU_I of the feature importance, computed between XAI methods and the radiologists. The values in the table are the mean and standard deviation of IoU across all radiologists.

XAI Method	IoU_T	IoU_I
Our Proposal	52.98±2.97	62.75±2.69
Integrated Gradients	53.10±3.05	63.20±2.83
LIME	52.70±3.22	62.90±2.98
SHAP	53.05±3.12	63.10±2.91

mentioned in section 4.6, these metrics permit us understand the coherence between the returned feature importance. These scores not only show that the radiologists have a fairly high degree of intersection of important features among each other, but also that the degree of intersection is of the same magnitude even with our proposal. This implies that our model concentrates on the relevant features of each modality when making the decision on the prognosis.

As stated in section 2, there is currently a lack of multimodal XAI methods in the literature. Therefore, to further validate the performance of our explanations, we compared the unimodal explanations generated by our method with other well-established XAI methods, namely Integrated Gradients, LIME, and SHAP [6]. We selected these methods because they can be applied to both tabular and imaging modalities, they are model-agnostic, i.e., they can be used with any model irrespective of its underlying architecture, and they all offer local explanations. Specifically, we extract the explanations by utilizing the C_{MLP}, E_{AE}, and E_{CAE} modules for each modality, respectively. In Table 5, we present the average IoU_T and IoU_I across the feature importance scores from all the radiologists for both our proposal and the competing methods. The results demonstrate that our unimodal explanations are not statistically different from these XAI methods (p-value > 0.05), indicating that our approach is coherent with state-of-the-art methods from a unimodal perspective. Furthermore, it is worth noting that our proposed method not only introduces unimodal explanations but also incorporates multimodal explanations, a novel feature that is not available in existing XAI techniques.

5.4. Ablation study

We now discuss what happens when only one modality is available. To this end, we ran two experiments: in the first we removed the AE and we trained again the

Table 6: Classification performance in the ablation study

Model	**Validation**	**Modality**	**Accuracy** (%)	**Sensitivity** (%)	**Specificity** (%)
Our Proposal	CV	*T*	75.78±0.75	76.63±0.71	74.74±1.02
	LOCO	*T*	73.48±3.20	69.87±3.11	79.56±8.60
	Survey	*T*	75.69	76.65	74.71
	CV	*I*	74.14±1.03	74.57±1.78	73.96±1.21
	LOCO	*I*	70.46±1.06	72.03±1.02	69.55±1.61
	Survey	*I*	74.32	74.42	73.88
AIforCOVID [9]	CV	*T*	75.70±0.80	76.00±0.70	75.40±1.10
	LOCO	*T*	73.40±4.40	69.90±15.80	79.50±13.60
	CV	*I*	74.20±1.00	74.80±1.90	73.80±1.30
	LOCO	*I*	70.50±1.00	72.00±1.10	69.60±1.50

Table 7: Reconstruction performance of the ablation study.

Model	**Data**	**Modality**	**MSE**
Our Proposal	CV	*T*	0.03±0.01
	LOCO	*T*	0.04±0.02
	Survey	*T*	0.03
	CV	*I*	0.02±0.01
	LOCO	*I*	0.03±0.02
	Survey	*I*	0.02

other part of the model, i.e., we worked only with the imaging modality disregarding the tabular clinical data. In the second we flipped the ablation, removing the *CAE* and, thus, we did not consider the images. Table 6 shows the results we achieved in the case of experiments ran in CV, LOCO and using the survey images. Furthermore, the third columns specifies the modality used. As before, we also show the hybrid baseline model presented in [9], which is trained only on one modality. The results of our proposal shows that using only the tabular clinical data or only the imaging data provides similar results, which do not statistically differ from each other, whatever the performance score considered (p-value > 0.05). Furthermore, in comparison with the performance of the full multimodal approach (first section of Table 1), we notice that both unimodal models report a statistically significant drop in performance (p-value < 0.01), whatever the validation approach or the performance score. As before, we notice that our model, even if it co-learns two tasks at once, obtains similar performance to the hybrid model, i.e., the best baseline model of [9]. In fact, the difference between our model and the baseline model on all metrics, in both CV and LOCO, is not statistically significant (p-value > 0.05).

Table 8: IoU_T (lower triangular) and the IoU_I (upper triangular) of the feature importance for models trained in ablation, computed for each pair between our model and the radiologists.

	Our proposal	R_1	R_2	R_3	R_4
Our proposal	-	60.31±2.98	62.45±2.31	63.57±2.49	64.22±3.58
R_1	53.63±2.45	-	60.56±2.20	62.10±3.10	61.90±2.98
R_2	54.65±2.95	52.45±3.21	-	61.07±3.30	62.90±3.20
R_3	54.49±3.03	53.99±3.32	50.89±3.00	-	62.88±3.50
R_4	50.40±3.01	52.41±3.40	52.54±3.11	55.30±2.87	-

For completeness, Table 7 shows the reconstruction results in terms of MSE for each modality. In comparison with Table 2, as expected, we notice that the reconstruction error has significantly decreased (p-value > 0.05) becaause the model can focus on one modality at a time, making it easier to learn an efficient embedding mapping.

As before, we also investigate the results in terms of explainability. Straightforwardly, in this case the modality importance does not make sense, so that Table 8 reports the IoU_T and the IoU_I of new unimodal feature importance. These results show that, even if we have a drop in classification and reconstruction performance, the explanations are consistent between the radiologists and between our method and the radiologists, suggesting that the MXAI method is robust to a missing modality [37, 38].

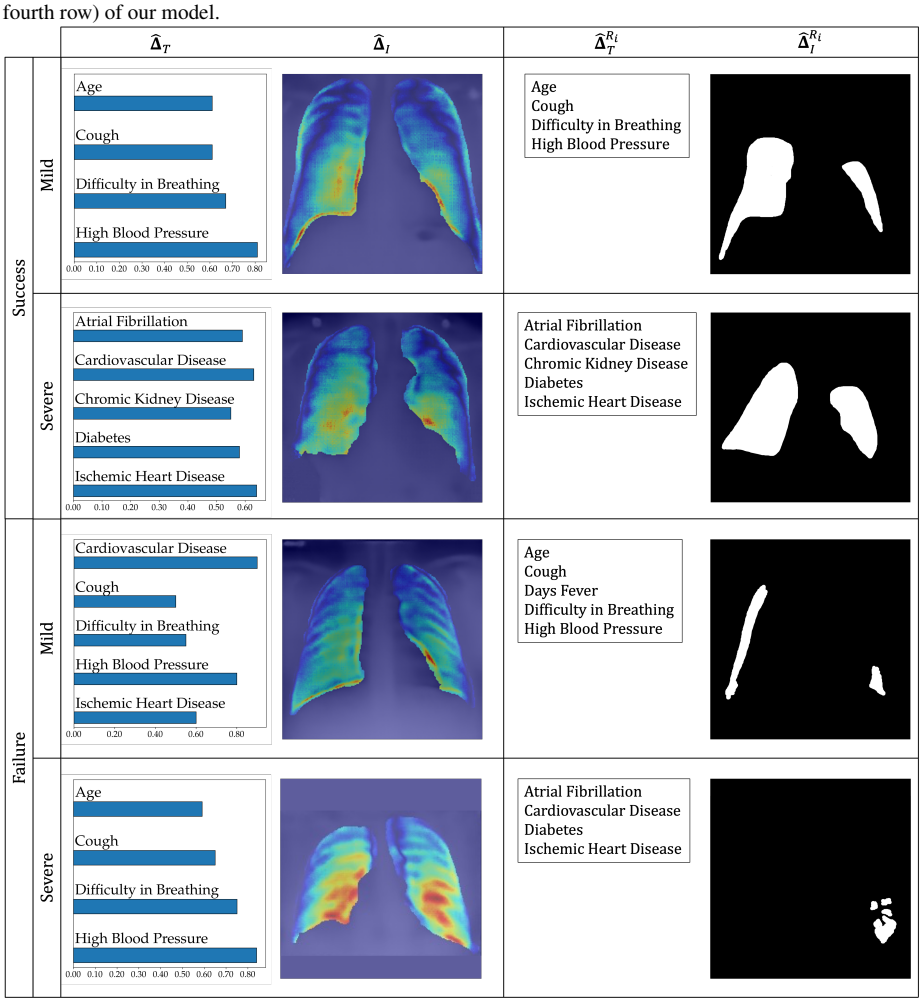
5.5. Clinical perspective and case studies

In a clinical practice scenario, we believe that our AI system can serve as a precursor to subsequent multimodal research for predicting the evolution of COVID-19. Its classifications and explanations can assist radiologists in performing prognosis tasks.

Indeed, on the one side, in [39] the authors showed that the rise of X-ray severity over the course of COVID-19 infection increases the sensitivity of COVID-19 detection using CXR over time (55% at $\leq$ 2 days to 79% at $>$ 11 days), whilst it decreases the specificity (83% at $\leq$ 2 days to 70% at $>$ 11 days). On the other side, as Table 1 shows, our proposal provides a larger sensitivity than the radiologists, suggesting that it can anticipate the evolution of positive COVID-19 cases, that is in an initial phase of the disease when the patient accesses the emergency department, our approach achieves a sensitivity (78.56%) equal to that which the X-ray alone shows after several days.

Furthermore, the proposed architecture has the beneficial feature to offer transpar-

Figure 3: Four case studies: for each we show the feature importance indicated by our proposal ($\hat{\mathbf{\Delta}}_T$ and $\hat{\mathbf{\Delta}}_I$) and the corresponding important features indicated by radiologists ($\hat{\mathbf{\Delta}}_T^{R_i}$ and $\hat{\mathbf{\Delta}}_I^{R_i}$) for the tabular and the imaging modalities, respectively. The rows show examples of patients with mild (first and third row) and severe (second and fourth row) outcomes, for both success (first and second row) and failure cases (third and fourth row) of our model.



ent decisions since, for each patient, the radiologists can observe at the same time the original data (both the clinical features and the X-ray image), the modality importance Δ_T and Δ_I, and the unimodal feature importance $\hat{\mathbf{\Delta}}_T$ and $\hat{\mathbf{\Delta}}_I$. With Δ_T and Δ_I, the radiologists would be guided to understand on which modality to concentrate more on. Instead, with $\hat{\mathbf{\Delta}}_T$ and $\hat{\mathbf{\Delta}}_I$ we guide the radiologist to concentrate on certain clinical characteristics and on specific areas of the X-ray scan.

Figure 3 presents four case studies. It is organized in four columns: the first two show the feature importance indicated by our proposal ($\hat{\mathbf{\Delta}}_T$ and $\hat{\mathbf{\Delta}}_I$), whereas the third and the fourth show the corresponding important features indicated by radiologists ($\hat{\mathbf{\Delta}}_T^{R_i}$ and $\hat{\mathbf{\Delta}}_I^{R_i}$), for the tabular and the imaging modalities, respectively. The tabular clinical data importance $\hat{\mathbf{\Delta}}_T$ is represented as a bar-plot on the $\hat{\mathbf{\Delta}}_T$ column, so that the longer the bar, the more important the clinical variable is. For the sake of visualization, we only show the features part of $\hat{\mathbf{\Delta}}_T^b$, keeping the magnitude of the importance computed according to equation 17. The X-ray image importance map $\hat{\mathbf{\Delta}}_I$ is shown as a heatmap on the original scan, which represents the relevance of each pixel in the image for the prognosis task on a color scale ranging from blue (low importance) to red (high importance) on the $\hat{\mathbf{\Delta}}_I$ column. Looking at the figure by row, the first group of rows shows two success cases, where our classifier correctly classifies the patient's outcome, whereas the second group of rows shows two failure cases, where our classifier incorrectly classifies the patient's outcome. In both cases of classification success or failure, we present an example from both the mild and severe classes. Note that all these cases have been correctly classified by the radiologist. In the success examples, we note a strong agreement between our proposal and the radiologist for both modalities. Specifically, in both mild and severe cases, the tabular features are the same, and the most important pixels for the model largely overlap with the area of interest segmented by the radiologist. In the failure cases, our model assigns importance to tabular features and to image regions that are different from those highlighted by the radiologist. We speculate that this may be the reason for the model incorrect predictions. In particular, in the mild case, only three out of five most important tabular features coincide with those suggested by the human expert, whilst for the severe case there is no overlap. Additionally, if we turn our attention to the images, both the mild and severe cases

exhibited a very low intersection between the most important regions for the models and the manually segmented areas.

6. Conclusion

In this work we presented an end-to-end multimodal architecture that jointly learns modality reconstructions and multimodal classification using tabular clinical and imaging data. With respect to the literature using such modalities for medical tasks, we deem that our method is the only one which offers intrinsic model-specific local multimodal explanations. In particular, multimodal explanations are computed by exploiting the latent space learnt by jointly training the end-to-end architecture and using a latent shift-based counterfactual method. We tested our approach in the context of the COVID-19 pandemic using the AIforCOVID public dataset, which includes both X-ray and clinical data. The extensive quantitative experimentation shows that the latent space retains features useful to succeed both in a reconstruction and classification task and, thus, resulting in an informative space for the latent-shift method. Moreover, the sanity check, although very time-consuming, was very useful since it showed a high intersection between the explanations provided by the method and those of the radiologists, both for the modality and the feature importance.

A reflection on this work highlights two main limitations. The first is that the reliability of the explanations the method produces is constrained by its reliance on the classification and reconstruction performance of the model. As all components of the method are data-driven and there is no dedicated module for explainability, the generalizability of the explanations could be limited by the quality of the data. In this respect, we plan to evaluate the effectiveness of our methodology on different datasets from different domains and with different data types. The second limitation stems from noticing that, although our proposal identifies the importance of each modality and the importance of each unimodal feature per sample, it does not find out high-level concepts, including those expert-based. To cope with this issue we deem that concept knowledge mining [40, 41] could be a viable solution that we plan to investigate in future work to enable human experts to better understand how the prediction is identified

by the model.

Acknowledgements

The computations of this work were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973. This work was partially supported by: i) "University-Industry Educational Centre in Advanced Biomedical 411 and Medical Informatics (CEBMI)" (Grant agreement no. 612462-EPP-1-2019-1-SK-EPPKA2-KA, 412 Educational, Audiovisual and Culture Executive Agency of the European Union); ii) PNRR MUR project PE0000013-FAIR; iii) FONDO PER LA CRESCITA SOSTENIBILE (F.C.S.) Bando Accordo Innovazione DM 24/5/2017 (Ministero delle Imprese e del Made in Italy), under the project entitled "Piattaforma per la Medicina di Precisione. Intelligenza Artificiale e Diagnostica Clinica Integrata" (CUP B89J23000580005).

Author Contributions

V.G.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. L.T.: Conceptualization, Methodology, Software, Validation, Formal analysis, Data Curation. D.A.: Validation. E.F.: Validation. D.F.: Validation. D.S.: Validation. P.S.: Conceptualization, Methodology, Validation, Formal analysis, Writing - Review & Editing, Supervision, Funding acquisition.

References

[1] S.-C. Huang, et al., Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, NPJ digital medicine 3 (1) (2020) 1–9.

[2] C. M. Caruso, et al., A multimodal ensemble driven by multiobjective optimisation to predict overall survival in non-small-cell lung cancer, Journal of Imaging 8 (11) (2022) 298.

[3] S. Cipollari, et al., Convolutional neural networks for automated classification of prostate multiparametric magnetic resonance imaging based on image quality, Journal of Magnetic Resonance Imaging 55 (2) (2022) 480–490.

[4] T. Baltrušaitis, et al., Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2) (2018) 423–443.

[5] E. Tjoa, et al., A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE Transactions on Neural Networks and Learning Systems 32 (11) (2020) 4793–4813.

[6] A. B. Arrieta, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115.

[7] D. Ramachandram, et al., Deep multimodal learning: A survey on recent advances and trends, IEEE Signal Processing Magazine 34 (6) (2017) 96–108.

[8] D. H. Park, et al., Multimodal explanations: Justifying decisions and pointing to the evidence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8779–8788.

[9] P. Soda, et al., AIforCOVID: predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays: an italian multicentre study, Medical Image Analysis 74 (2021) 102216.

[10] A. Signoroni, et al., BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset, Medical Image Analysis 71 (2021) 102046.

[11] J. S. Zhu, et al., Deep-learning artificial intelligence analysis of clinical variables predicts mortality in COVID-19 patients, Journal of the American College of Emergency Physicians Open 1 (6) (2020) 1364–1373.

[12] H. Al-Najjar, et al., A classifier prediction model to predict the status of Coronavirus COVID-19 patients in South Korea, European Review for Medical and Pharmacological Sciences (2020).

[13] X. Bai, et al., Predicting covid-19 malignant progression with ai techniques, MedRxiv (2020) 2020–03.

[14] W. Ning, et al., Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning, Nature Biomedical Engineering 4 (12) (2020) 1197–1207.

[15] C. Fang, et al., Deep learning for predicting COVID-19 malignant progression, Medical Image Analysis 72 (2021) 102096.

[16] B. G. Santa Cruz, et al., Public COVID-19 x-ray datasets and their impact on model bias–a systematic review of a significant problem, Medical Image Analysis 74 (2021) 102225.

[17] G. Fiscon, et al., Assessing the impact of data-driven limitations on tracing and forecasting the outbreak dynamics of COVID-19, Computers in Biology and Medicine 135 (2021) 104657.

[18] K. M. Abiodun, et al., Explainable AI for fighting COVID-19 pandemic: opportunities, challenges, and future prospects, Computational Intelligence for COVID-19 and Future Pandemics (2022) 315–332.

[19] W. W. Boonn, et al., Radiologist use of and perceived need for patient data access, Journal of Digital Imaging 22 (4) (2009) 357–362.

[20] L. Wynants, et al., Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal, BMJ 369 (2020).

[21] M. Roberts, et al., Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nature Machine Intelligence 3 (3) (2021) 199–217.

[22] R. Guidotti, et al., A survey of methods for explaining black box models, ACM Computing Surveys (CSUR) 51 (5) (2018) 1–42.

[23] G. Joshi, et al., A review on explainability in multimodal deep neural nets, IEEE Access 9 (2021) 59800–59821.

[24] A. Holzinger, Explainable AI and multi-modal causability in medicine, i-com 19 (3) (2020) 171–179.

[25] R. R. Selvaraju, et al., Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[26] N. Bahadur, et al., Dimension estimation using autoencoders and application, in: Deep Learning Applications, Volume 3, Springer, 2022, pp. 95–121.

[27] V. Guarrasi, et al., Pareto optimization of deep networks for COVID-19 diagnosis from chest X-rays, Pattern Recognition 121 (2022) 108242.

[28] V. Guarrasi, et al., Optimized fusion of CNNs to diagnose pulmonary diseases on chest X-rays, in: International Conference on Image Analysis and Processing, Springer, 2022, pp. 197–209.

[29] V. Guarrasi, et al., A multi-expert system to detect COVID-19 cases in X-ray images, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2021, pp. 395–400.

[30] V. Guarrasi, et al., Multi-objective optimization determines when, which and how to fuse deep networks: An application to predict covid-19 outcomes, Computers in Biology and Medicine 154 (2023) 106625.

[31] J. Shiraishi, et al., Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules, American Journal of Roentgenology 174 (1) (2000) 71–74.

[32] S. Jaeger, et al., Two public chest X-ray datasets for computer-aided screening of pulmonary diseases, Quantitative Imaging in Medicine and Surgery 4 (6) (2014) 475.

[33] X. Glorot, et al., Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial

Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[34] K. He, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[35] J. P. Cohen, et al., TorchXRayVision: A library of chest X-ray datasets and models, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2022, pp. 231–249.

[36] P. A. Yushkevich, et al., ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2016, pp. 3342–3345.

[37] A. Rofena, et al., A deep learning approach for virtual contrast enhancement in Contrast Enhanced Spectral Mammography, Computerized Medical Imaging and Graphics 116 (2024) 102398.

[38] C. M. Caruso, et al., A deep learning approach for overall survival prediction in lung cancer with missing values, Computer Methods and Programs in Biomedicine (2024) 108308.

[39] S. Stephanie, et al., Determinants of Chest X-ray sensitivity for COVID-19: a multi-institutional study in the United States, Radiology: Cardiothoracic Imaging 2 (5) (2020).

[40] Z. Qi, et al., Towards more explainability: concept knowledge mining network for event recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3857–3865.

[41] M. Graziani, et al., Concept attribution: Explaining CNN decisions to physicians, Computers in biology and medicine 123 (2020) 103865.