TUM Data Innovation Lab

Munich Data Science Institute (MDSI)

Technical University of Munich

&

TUM Chair for Computer Aided Medical Procedures & Augmented Reality

&

Oxford Torr Vision Group

Final report of project:

Exploring the Potential of Causality and Deep Learning for Medical Imaging

Authors	Ronald Skorobogat, Daniel Bin Schmid, Shunyu Zhao, Leonard Waldmann
Mentors	Lennart Bastian (M.Sc.), Ashkan Khakzar (Dr.)
Project lead	Dr. Ricardo Acevedo Cabra (MDSI)
Supervisor	Prof. Dr. Massimo Fornasier (MDSI)

Feb 2024

Acknowledgements

We wish to express our appreciation to our supervisors Lennart Bastian and Ashkan Khakzar for their valuable feedback and guidance throughout this research project. Their expertise has been crucial in the development of this work. Additionally, we extend our thanks to Professor Uwe Siebert for the fruitful discussion that contributed significantly to the depth of our analysis. His insights have been highly beneficial in advancing our research.

Abstract

This project addresses the challenge of integrating causality with deep learning in healthcare. It focuses on two innovative approaches: the use of causal methods on top of zero-shot predictions from vision language models and the application of causally disentangled variational autoencoders for generating counterfactual data and for causally inspired downstream classification. To evaluate our approaches, we perform pathology classification and image generation on MIMIC-CXR as well as image generation on CelebA. Since the project is largely exploratory, it also provides insights into the challenges of merging deep learning with causal reasoning in medical applications. We provide theoretical discussions about the strengths and shortcomings of applications of causality, and highlight intersections where it can provide value.

Contents

1 Introduction

Developing algorithms for assisting medical experts has shown significant advances in recent years [16]. Deep learning algorithms have shown remarkable performance in healthcare [40], for example in enhancing x-ray analysis, which benefit from the surge in large-scale medical data collection efforts [8]. Despite its successes, the application of deep learning in this field faces significant hurdles related to the non-interpretability of its models, often termed as "black-box" models [50]. In particular, since medicine is a high-impact domain, algorithmic decision-making systems are highly scrutinized and have desirable properties like robustness, explainability and transparency, which pure deep learning models do not fulfill.

The causality framework as proposed by Judea Pearl [44] promises to resolve this lack of explainability and to improve robustness to distributional shifts [53, 52]. This is achieved by modelling causal relations instead of pure associations like conventional machine and deep learning algorithms do. Moreover, causality offers methods to accurately define and estimate the effect of treatments. Nevertheless, the application of causality in healthcare is not without its challenges, particularly in formalising and describing causal relationships of inherently complex mechanisms, which are naturally present in medical applications.

Combining the robustness of causality with the power of deep learning in the medical domain is a promising research direction. The goal of this report is to *explore* this triad of deep learning, causality, and medical applications. In particular, we seek to *identify* questions of medical utility and *solve* them such that both the benefits from causality and deep learning shine.

The rest of the report is structured as follows. In section 2.1, we motivate how to formally define causal effects and give an intuition of its benefit. In section 3, we briefly document the datasets we used for our experiments, including their causal interpretations. In section 4, we detail robustness experiments for using causality on top of vision-language feature extractors. In section 5, we dive into the experiments for generating counterfactual outcomes and enhancing predictions with causal variational models. Finally, we discuss the challenges we encountered and draw conclusions from our experiments in section 6 and section 7.

2 Potential of Causality

Classical statistics are built on properties of association like conditional probabilities or correlations. However, these concepts fail to answer or even formulate causal questions that are ubiquitous in daily life. The theory of causality as developed by Judea Pearl, [44] presents a framework that departs from the classical association-based statistics and provides a language to formally define and quantify causal relations between events. Since causal models rely on causal relations instead of spurious associative relations, they promise explainability and better generalisation to unseen data.

2.1 Association does not imply Causation

Let us consider a patient with a deadly disease where D denotes the severity of the disease, T the administration of a treatment and S the survival of the patient. We know that treatment T is only given to very sick patients, i.e. T depends on D, and treatment administration only slightly improves the chances of survival, i.e. T influences S. If we denote the variables as nodes and the influences between the variables as directed edges, we can depict the scenario by graph (a) in fig. 1. In this model, the administration of treatment is highly associated with the death of the patient, since the treatment is only given to patients that are likely to die and the treatment only has a slight chance to cure the patient. If we use any association-based method, e.g. correlation or machine learning methods, we observe that $P(S = 0|T = 1)$ is large and might draw the false conclusion that giving treatment *leads to* the patient's death. In this case, association and causation clearly are different concepts. In a medical context, relying on the false conclusion drawn from associative methods would have the detrimental consequence of deeming the treatment as counter-effective.

2.2 Formal Definition of Causation by Structural Causal Models

To formalize causal effects, let us have a look at a counterfactual definition of causality from philosophy.

Definition 1 (Philosophical Notion of Causation)**.** [22] X causes Y if and only if, without X, Y would not exist.

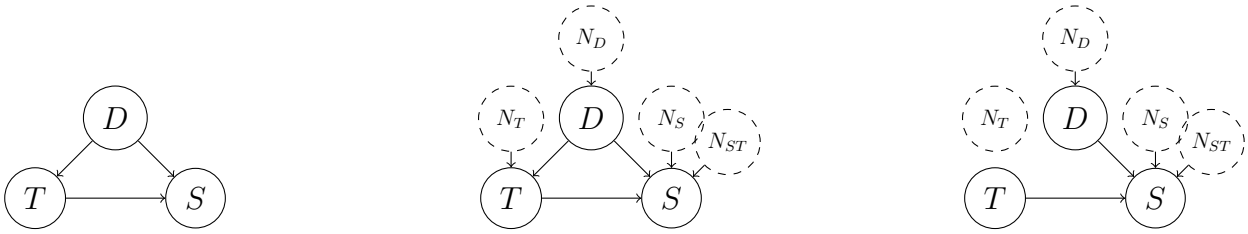
Extending this definition, X causes Y in 'environment' A, if in an identical 'environment' B with the only difference that we remove of 'vary' X, the observation Y changes. Let us define a convenient environment.

Definition 2 (SCM)**.** We model causal relationships with a *directed acyclic graph* (DAG) G, also called a *causal graph*, where nodes $X = \{X_1, ..., X_n\}, n \in \mathbb{N}$ denote events and directed edges denote cause-effect relations from parent to child. We quantify these cause-effect relations with deterministic functions $\mathcal{F} = \{f_1, ..., f_n\}$ and abstract away all randomness into *exogenous* random variables $N = \{N_1, ..., N_{n'}\}, n' \in \mathbb{N}$ with joint distribution p_N. Now,

$$X_j = f_j(\mathrm{pa}_j, N_{I_j}) \quad \text{where } I_j \subset \{1, ..., n'\} \tag{1}$$

and pa_j denotes the parents of X_j in G. Finally, a *structural causal model* (SCM) is the tuple $(X, N, p_N, \mathcal{F})$.

The joint distribution p over all $X \cup N$ is the *observational distribution* that the SCM *generates*. For simplicity, we usually assume that $I_j = \{j\}$, i.e. that we have observed all common causes of our variables, also known as *fully observed* SCMs. Given an SCM A, we now formalize the above notion of a causal effect of $X_i = T$ on $X_j = S$ by creating a new SCM B from A by removing all incoming edges into T and fixing T to a particular value t. The observational distribution of B is denoted by $p(S|\mathrm{do}(T = t))$ and called the *interventional distribution* (of A) when *intervening on* T. Using interventions, we can



(a) (b) (c)

Figure 1: Caption

now ask causal questions. Formally, we, e.g., define that there is a causal effect of T on S if T and S are dependent in the interventional SCM B. Let us formally model the scenario from section 2.1.

Example 1. *Let* $N_D \sim B(0.1), N_T \sim B(0.9), N_S \sim B(0.1), N_{ST} \sim B(0.2)$ *be independent Bernoulli variables and define*

$$D = N_D, \quad T = DN_T, \quad S = D(1-T)N_S + DTN_{ST} + (1-D). \tag{2}$$

The corresponding SCM is depicted by graph (b) in fig. 1. To determine the causal effect of T *on* S*, we intervene on* T *and obtain a new SCM as in graph (c) in fig. 1. We compute*

$$p(S = 1|do(T = 1)) = p(N_D = 0) + p(N_D = 1)p(N_S = 1) = 0.92 \tag{3}$$

To see that the interventional distribution is actually different from the associative conditional distribution, we compute

$$p(S = 1|T = 1) = \frac{p(S = 1 \cap T = 1)}{p(T = 1)} = \frac{0.1 \cdot 0.9 \cdot 0.2}{0.1 \cdot 0.9} = 0.20 \neq p(S = 0|do(T = 1)) \tag{4}$$

This is a precise mathematical formulation of the statement that association does not equal causation. Equation (3) *is so large since it also incorporates the non-sick patients, which is very likely. However, eq.* (4) *is so small since if we observe administration of treatment, we automatically know that the patient was sick so their survival chances are very low.*

Moreover, it turns out that we can obtain a simple graphical criterion: if there are only directed paths from T to S, association equals causation [44]. However, any undirected path between T and S *may* introduce *spurious correlations* that we need to account for with causal theory, e.g. the undirected path $T \leftarrow D \rightarrow S$. We note that there are other ways to define causal connections, e.g. the potential outcome framework or Granger causality, c.f. [17]. Apart from providing a language to ask and quantify causal questions, causality promises explainable models since we clearly map out our variables and define functional relations between them [53]. Moreover, by defining these independent causal mechanisms, causal methods are expected to be robust to distribution shifts, since the causal mechanisms do not change in a different domain [52].

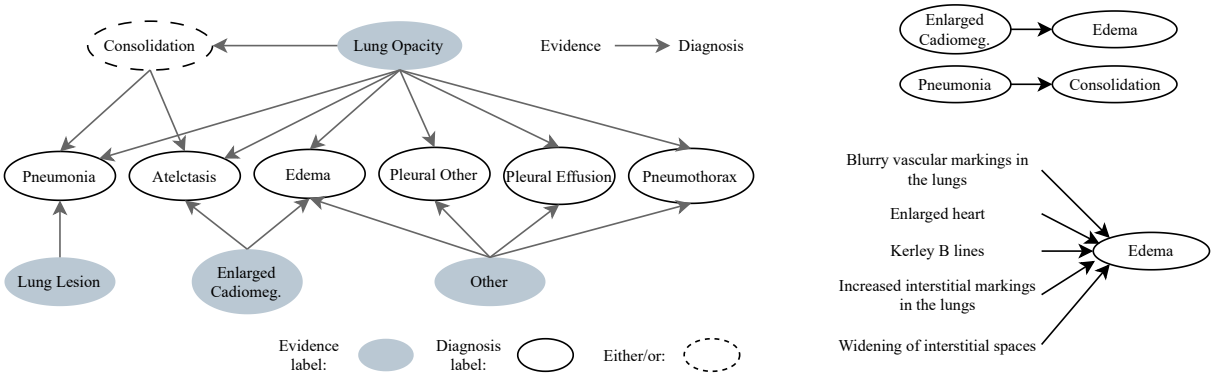


Figure 2: Radiology association modeling. On the left, the evidence graph proposed in [27]. On the top right, known causal relationships between pathologies. On the bottom right, an evidence graph for edema from [46].

3 Datasets for Causal Exploration

This section presents image datasets and how we impose causal interpretation of those images. For a theoretical discussion see section 6. We focus on the chest x-ray MIMIC-CXR-JPG dataset, but since the algorithms for disentanglement are not validated on radiology datasets, we additionally use CelebA and CausalCircuit for experimentation.

3.1 MIMIC-CXR-JPG

The MIMIC-CXR-JPG [25] dataset is a processed version of the original MIMIC-CXR [26] dataset and contains $377,110$ JPG Images of frontal and lateral chest x-ray images. Accompanying the images, the MIMIC-CXR dataset contains free-text reports from professional radiologists that contain a descriptive 'findings' and a commentary 'impressions' section. Resembling the computer science literature for chest x-rays, we are interested in 14 pathologies that professional radiologists usually identify in these x-ray images. As *weakly supervised* ground-truth labels for these pathologies we use the well-known CheXpert [24] label extractor that can deduce those 14 labels from the written reports with high AUC of > 0.9 across all pathologies except for atelectasis (0.85). Hereby, a pathology can be present, absent, or not mentioned. For 11 of these pathologies, [27] defined an evidence graph indicating associations if two pathologies appear jointly in a sentence in the report, which can be seen on the left of fig. 2. We note that these dependencies should not be interpreted as causal relationships, but as evidence relationships that influence the diagnosis of a trained radiologist [27]. Two actual causal relationships are shown in the top right of fig. 2: enlarged cardiomegaly can lead to peripheral edema due to increased pressure in the hepatic and systemic veins [1], and pneumonia can result in lung consolidation by filling the alveolar spaces with infectious exudate [15]. More on this theoretical discussion in section 6. The 3 pathologies not included in fig. 2 are 'no finding', 'support devices', and 'cardiomegaly'.

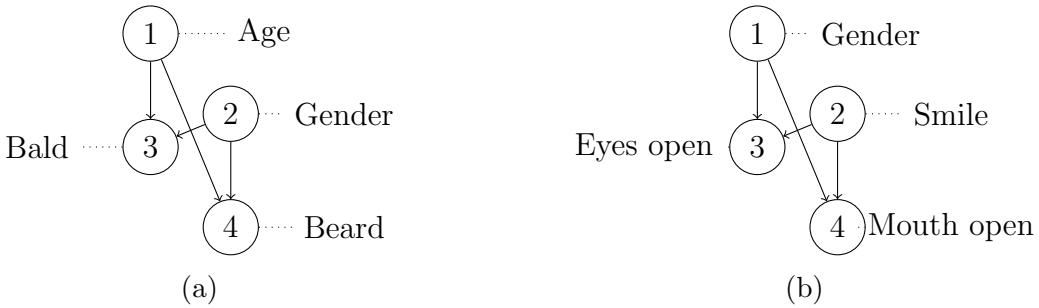


Figure 3: Ground-truth causal graphs on CelebA [36, 64].

3.2 CelebA

CelebA [36] is a dataset containing 200 thousand celebrity portraits, each with 40 labels, describing facial features or pose of the depicted celebrity. While there can not necessarily a causal graph between all labels drawn, previous work in causal disentanglement [64, 33] take subsets of the labels for which a ground truth causal relationship can be assumed. These ground truth causal graphs are shown in fig. 3. It should be noted that while the edges are given in this interpretation, the ground-truth structural equations are not. Therefore, our experiments include learning the structural equations.

3.3 Causal Circuit

CausalCircuit [7] is a synthetic dataset modeling the setup of turning on and off different light sources using light switches generated using MuJoCO [60], an open-source physics engine. The scene contains a robot arm that is placed on one of three light switches, one switch for blue, red, and green light, respectively. The dataset comes with a ground-truth data generating process, i.e. ground-truth structural equations. Note that we use a simplified version of the dataset adopted by [33]. In the simplified version, the data-generating process is given by

$$z_A \sim Uniform(0,1), \quad l_B \sim 1\{z_A < \frac{1}{3}\}, \quad l_G \sim 1\{\frac{1}{3} \leq z_A \leq \frac{2}{3}\} \tag{5}$$

where z_A is the arm position, and l_G and l_B are binary variables indicating whether the green or blue light is turned on, respectively. First, the arm position is sampled uniformly. Secondly, depending on the arm position (i.e. whether it is placed on the button for blue or for green light), either the blue or green light is turned on. Unlike in the original CausalCicruit dataset [7], the red light l_R is always turned on in the simplified version. The set of guiding variables now is defined as $\{z_A, l_R, l_G, l_B\}$.

4 Causality on Vision Language Model Zero-Shots

To leverage the full potential of causality, we first need to define the nodes of our graph. Because large graphs are neither scalable nor interpretable, we usually want to have a few semantically reasonable events that are humanly understandable. Since we have complex

multi-modal data, the main idea of this section is to use deep vision language encoders to extract those variables. Then, we can apply any method from causality theory. Here, we focus on classification and causal discovery. Essentially, this section leverages the idea of [46], frames it as a causal model, and explores the potential of this extension.

4.1 Background: Zero-shots with Deep Vision Language Models

The interest in vision language models has recently surged [68, Fig. 1]. A VLM is trained on a dataset of paired vision and text samples, i.e. $D = \{x_n^v, x_n^t\}_{n=1}^N$. We want to learn a vision encoder f^v and a text encoder f^t that map to the same-dimensional latent space of unit-norm vectors. The objective is to push the representations $z_n^v = f^v(x_n^v), z_n^t = f^t(x_n^t)$ of paired samples x_n^v, x_n^t close and of unpaired samples far away. Formally, let $B \subset \{1, ..., N\}$ and we define the *symmetric contrastive loss* [6]

$$\mathcal{L} := -\frac{1}{|B|}\left(\sum_{i\in B}\frac{\exp(z_i^t \cdot z_i^v/\tau)}{\sum_{j\in B}\exp(z_j^t \cdot z_i^v/\tau)} + \sum_{i\in B}\frac{\exp(z_i^t \cdot z_i^v/\tau)}{\sum_{j\in B}\exp(z_i^t \cdot z_j^v/\tau)}\right), \tag{6}$$

where $\tau > 0$ is a temperature parameter and $\cdot$ denotes the standard scalar product. Now, we can perform simple zero-shot classification. Given an image x^v and a class of interest, e.g. 'pneumonia', we design a prompt x^t that queries the existence of the class, e.g. 'There is pneumonia', and interpret the cosine similarity between their embeddings z^v, z^t, i.e. $z^v \cdot z^t \in [0, 1]$ as probability of presence of the class. To capture both presence and absence of a class, it is considered beneficial to use a positive and negative prompt, e.g. 'There is no penumonia', with corresponding text embeddings z_+^t, z_-^t [46] and compute the probability of presence as

$$\frac{\exp(z^v \cdot z_+^t)}{\exp(z^v \cdot z_+^t) + \exp(z^v \cdot z_-^t)}. \tag{7}$$

The advantage of these models is that they can be trained on large unlabelled datasets like captioned images from the internet without manual labelling. Although there usually is less data available in the medical imaging domain, there still are well-performing open-source VLMs for chest x-rays like BioVil [6]. More advanced models offer conversational interaction with a user based on chest a xray image [56, 45]. However, all of the above models learn associations. Modification of the attention layer to incorporate causal aspects has been discussed [67] as well as improving image captioning by causal inference [66]. What we want to investigate in this section is if we can add causality *on top of pre-trained* associative vision language models.

4.2 Experiments

We now apply zero-shot prompting to classify the 11 of the 14 *pathologies*, c.f. section 3, that appear in the dependency graph in fig. 2 on the left in a causal way. We denote these pathologies with P and use 60000 random samples from the mimic-cxr dataset, where the weakly supervised labels of P extracted with the Chexpert labeller serve as ground truth. For each pathology $p \in P$, we consider additional *observations* O_p that radiologists use to determine the presence of p as in [46]. Moreover, we consider *direct* zero-shot prediction

of $p \in P$ denoted by $d_p \in D$. We use eq. (7) for computing the node values where we prompt 'There is/are (no) p' for the direct prediction of $p \in P$ and 'There is/are (no) o indicating p' for $o \in O_p$ following [46]. We note that defining causality on static images can be difficult, since causal connections always require time passing between cause and effect [44]. Throughout this section we interpret the causal edges not as causal effects of the actual observations, but as evidence relationship that influence the diagnose of a trained radiologist, c.f. [27]. This theoretical and philosophical issue is further discussed in section 6.

4.2.1 Robust Classification

Model We assume that the observed pathologies are generated by some unknown fully-observed SCM $(P \cup O, N, p_N, \mathcal{F})$ where $O := \cup_{p \in P} O_p$. Let $p \in P$ a pathology we want to predict. Since $p = f_p(\text{pa}_p, N_p)$, we need to know f_p, pa_p and N_p. We additionally make the simplifying assumption that $f_p(\text{pa}_p, N_p) = \sigma(f'_p(\text{pa}_p) + N_p)$ where σ is the sigmoid function and $N_p \sim N(0, 1)$. Given f'_p and pa_p, we clearly have the maximum likelihood estimator $\hat{p} = \sigma(f'_p(\text{pa}_p))$. To define the parents of p, i.e., which variables are causes of p, we use domain knowledge to define different sets of edges:

$$\mathcal{P} := \{\text{all edges from the dependency graph fig. 2}\} \tag{8}$$

$$\mathcal{O} := \{o \to p | \forall p \in P \ \forall o \in O_p\} \tag{9}$$

$$\mathcal{D} := \{d_p \to p | \forall p \in P\} \tag{10}$$

Sets $\mathcal{P}$ and $\mathcal{O}$ are verified by professional radiologists [27, 46]. Any node $d_p \in D$ indicate the confidence of BioVil that p is present, thus, d_p influences our believe that p is present. Using these edges, we can define different causal graphs for our SCM. To define the structural equations, we estimate them by supervised learning where we predict p using its parents pa_p with any model. Overall, we need to train n models, where $n \leq |P|$ is the number of nodes in the graph that have any parents. Since for any $i \to j \in \mathcal{P} \cup \mathcal{O} \cup \mathcal{D}$, we have $j \in P$, we always use the ground-truth pathology values as target for training our models. If we train a model m that has other pathologies P' as parents, we first train the models for those pathologies and use their predictions for training m instead of using the ground truth values to let m adjust to the previous model. If $p' \in P'$ has no parents, e.g. when only using $\mathcal{P}$, we set p' to the zero-shot prediction d_p. We note that causality in our setting essentially boils down to feature selection.

Technical details. We use 48000 random samples as training dataset and 12000 random samples as test dataset. Since the positive observations of the pathologies are very infrequent, require to have ≥ 2 positive observations of each pathology in both train and test dataset. For training any supervised model, we use the AutoML library PyCaret[1] that automatically fits several supervised learning algorithms and returns the one that performs best on a hold-out validation set. We use the Area Under the Receiver Operating Characteristic Curve (AUC) as evaluation metric since it reflects imbalanced classes and is also used in the literature for pathology classification. To combat the imbalanced data, we apply the pre-implemented Synthetic Minority Oversampling Technique (SMOTE) [10] from the PyCaret library.

[1] https://github.com/pycaret/pycaret

	Zero-Shot		Causation			Association		
Pathology	$\mathcal{D}$	Xplainer	$\mathcal{O}$	$\mathcal{OP}$	$\mathcal{OPD}$	$\mathcal{O}$	$\mathcal{OP}$	$\mathcal{OPD}$
En. Cardio.	0.643	0.644	0.699	0.676	0.669	0.693	**0.709**	**0.709**
Cardio.	0.728	0.708	0.740	0.740	0.757	**0.793**	0.786	0.791
Lung Opacity	0.707	0.681	0.753	0.740	0.732	0.757	0.757	**0.758**
Lung Lesion	0.652	0.612	0.724	0.744	0.759	0.758	0.745	**0.785**
Edema	0.767	0.786	0.863	0.877	0.868	**0.880**	0.875	0.865
Consolid.	0.775	0.760	**0.817**	0.803	0.805	0.812	0.816	**0.817**
Pneumonia	0.679	0.680	0.722	0.704	**0.745**	0.726	0.719	0.737
Atelectasis	0.664	0.646	0.701	0.764	0.774	0.793	**0.794**	0.786
Pneumoth.	0.725	0.693	0.815	0.796	0.853	0.852	0.849	**0.875**
Pleural Eff.	0.839	0.795	0.847	0.850	0.872	0.891	**0.905**	0.902
Pleural Oth.	0.510	0.614	0.797	**0.889**	0.818	0.831	0.802	0.803
Fracture	0.496	0.515	0.649	0.625	0.633	0.613	0.663	**0.695**
mean	0.682	0.678	0.761	0.767	0.774	0.783	0.785	**0.794**

Table 1: AUC values for predicting pathologies in rows with models in columns, where bold number denotes maximum in each row. Target values were binarized. The $\cup$ symbol was dropped for simplification in the header.

Baselines. We compare our models to the zero-shot Xplainer [2] that predicts $p \in P$ by averaging the presence probabilities for all $o \in O_p$. Additionally, we train a standard association based model for each $p \in P$ with PyCaret, where we drop all nodes that have no edges and predict p using all remaining nodes except p. Essentially, this corresponds to ignoring the causal feature selection and train on all available data. We note, that when only using the edges in $\mathcal{D}$, each pathology p depends on exactly d_p so we just set the presence probability of the direct prediction as zero-shot prediction of the pathology.

Classification results. The results can be seen in table 1. As expected, $\mathcal{O}$ performs better than Xplainer since $\mathcal{O}$ learns a model on the observations instead of just averaging them. Adding $\mathcal{P}$ or $\mathcal{D}$ does not change the results by much. Overall, purely association-based approaches perform slightly better than the zero-shot and causal approaches investigated, but the results are not clearly conclusive.

Distribution shift results. That association performs better in prediction on in-distribution data with lots of training data than causation is not surprising since the causality essentially acts as feature selection here and more features are always better. One of the main arguments of causality is its potential for generalisation to out of distribution data. To investigate this, we introduce a distribution shift by splitting the dataset by gender, age, and ethnicity. We perform the same training procedure as before where the training and test set size varies based on the split criterion. We only evaluate $\mathcal{OPD}$ since it performed best in the classification before. The results are in table 2. Again, the association approaches performs better, however, the causal approach performs slightly better in the age split. A potential explanation is that the introduced distribution shift was not big enough for the causal method to show its potential. This is supported by the

	Causation ($\mathcal{POD}$)			Association ($\mathcal{POD}$)		
split	in	out	diff	in	out	diff
gender	0.764	0.746	-2.35%	0.776	0.762	-1.80%
age	0.797	0.702	-11.91%	0.857	0.747	-12.83%
ethnicity	0.774	0.742	-4.13%	0.836	0.813	-2.75%

Table 2: Mean AUC score of predicting pathologies where 'in' is a hold-out in-distribution, 'test' is a hold-out out-of-distribution, and 'diff' is out/in $-$ 1. The split criterion varies for each row.

small differences in AUC score between in- and out-of-distribution test set. Another potential reason is that our underlying causal graphs are incorrect or do not capture enough information to satisfy the fully observed SCM assumption. We will investigate this in the next section.

4.2.2 Causal Discovery

Instead of setting edges with expert knowledge, we can also use algorithms to learn the causal structure from data, which is known as *causal discovery*. Similarly to above, we are particularly interested in the causal parents of the pathologies. Unfortunately, the causal graph cannot be uniquely identified from the data generating distribution in general, i.e. there exist different SCMs that generate the same observational distribution. Thus, causal discovery algorithms usually return a set of DAGs, also known as *equivalence class* E, that could all have generated the observed distribution. In this case, we define the *potential parents* of node i as $\text{ppa}_i^E := \{j | \exists G \in E : j \to i \in G\}$. For an overview of causal discovery algorithms, see here [62]. Since we only have few nodes, make no further assumptions on the SCM and do not exclude the possibility of hidden confounding, we decided to use the standard fast causal inference (FCI) algorithm [59]. For any $\phi \subset P$ let E_ϕ denote the equivalence class found by the FCI algorithm over the nodes $\phi \cup \bigcup_{p\in\phi} O_p$ with $\alpha = 0.01$. To measure the agreement of the equivalence classes with the edges $\mathcal{P} \cup \mathcal{O}$ from the previous section, we define the metric

$$m_n = \frac{\sum_{\phi\subset P,|\phi|=n} \sum_{p\in\phi} |\{o \in O_p | o \in \text{ppa}_p^{E_\phi}\}|}{\sum_{\phi\subset P,|\phi|=n} \sum_{p\in\phi} |O_p|} \quad \in [0,1] \tag{11}$$

for $n \in \{1, ..., 12\}$. Unfortunately, we observe $m_1 = 0.194, m_2 = 0.074$, and $m_3 = 0.019$ indicating that our graph from the previous section is a poor fit for the data. In fig. 4 on the left, we see a good scenario for $n = 1$ where 4 out of the 7 observations could potentially be parents of p. In fig. 4 in the middle, we see a bad scenario were only 1 edge is potentially correct and the others are not. Finally, we use FCI with $\alpha = 0.01$ on P. The result can be seen in fig. 4 on the right. We observe that essentially all arrows are bi-directional, which corresponds to the presence of a hidden confounder between the variables in the FCI algorithm. Hence, the pathology graph does not represent the observed data well. This might explain the bad AUC score of the causal P model in table 1.

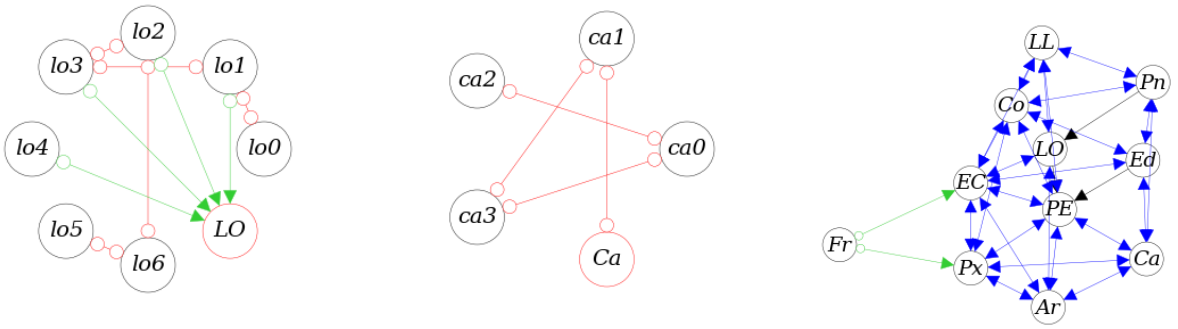


Figure 4: Causal discovery with FCI and $\alpha = 0.01$. The left figure shows discovery of Lung Opacity and its observations. The middle figure shows discovery of Cardiomegaly and its observations. The right figure shows causal discovery on the whole pathology graph.

5 Causal Variational Models

In this section, we investigate the application of integrating causal aspects into the latent space of variational models, i.e. VAEs and a relation to diffusion models. Our methods are grounded in advancements in causal disentanglement which are mainly centered around VAEs, see section 5.1. Exploiting the full potential of VAEs, we dive into two core applications, its generative and discriminative utility. Specifically, we explore how causality can assist in counterfactual image generation utilizing causal mechanisms [32] in section 5.2 (i.e. exploring the generative aspect), and the benefits of causal disentanglement for downstream classification [54, 4] in section 5.3 (i.e. the discriminative aspect).

5.1 Background

Causal representation learning has been gaining traction in the deep learning community [53]. Pioneering work leverages VAEs to semantically disentangle the latent space into causally related factors [54, 17, 64]. Hereby, it proves to be useful that VAEs enforce a distribution with the kl-divergence term. It allows for generating images by sampling from the enforced distribution in the latent space, or, e.g., for an explicit formulation to compute total correlation in the latent space which guides the feature extracting process to extract independent and disentangled factors [11, 65]. The latent space distribution and reconstruction is enforced by maximising the evidence lower bound (ELBO)

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z)\right] - \mathrm{KL}\left[q_\phi(z|x) \| p(z)\right] \tag{12}$$

where the full generative model $p_\theta(x, z) = p_\theta(x|z)p(z)$ consists of an inference model $q_\phi(z|x)$ that approximates its true posterior $p_\theta(z|x)$, and a generative model $p_\theta(x|z)$ which reconstructs the input space. $x \in \mathbb{R}^d$ is an observed input vector, $z \in \mathbb{R}^n$ is referred to as a latent vector and $\theta \in \Theta$ is the parameter vector that characterises the neural network.

Disentanglement. *Disentanglement* in machine learning is the process of decomposing the latent structures of data into distinct and interpretable factors of variation [5].

This representation allows for individual dimensions in the latent space to correspond to single generative factors, where alterations in one dimension are invariant to changes in others [29]. On the other hand, *causal disentanglement* takes this concept a step further by aligning the disentangled factors with the causal mechanisms that generate the data. Unlike traditional disentanglement, which may simply separate factors based on statistical independence or lack of correlation, causal disentanglement seeks to understand and represent the generative process of the data in terms of causal relationships. This means that the latent factors must not necessarily be independent, and also can reflect the causal structure of the domain, i.e. how changes in one factor can lead to changes in others [64].

Identifiability of VAEs. By experimental evidence and a theoretical proof by drawing connections to nonlinear independent component analysis (ICA), it has been shown that disentanglement is intractable without the usage of guiding variables [28, 37], which induced a shift from unsupervised disentanglement to semi-supervised methods. In particular, it has been shown that a VAE trained in an unsupervised manner is not *identifiable*, i.e.

$$\forall(\theta, \theta^*) : p_\theta(x) = p_{\theta^*}(x) \implies \theta = \theta^* \tag{13}$$

does not hold. Hence, multiple distinct latent representations can generate the same data distribution.

Causal disentanglement in identifiable VAE. The relationship between causal disentanglement and identifiability (as defined in eq. (13)) is foundational: identifiability is a prerequisite for causal disentanglement; without the former, it would be challenging to claim that the latent variables represent distinct causal factors. A model that ensures identifiability guides the way for causal disentanglement by establishing a one-to-one correspondence between latent variables and causal factors. This correspondence allows for meaningful interventions and manipulations within the latent space, facilitating causal inference and reasoning. Formally speaking, by introducing an additionally observed variable $u \in \mathbb{R}^n$ to make the priors conditional, then, identifiability as in eq. (13) can be shown [23], which is reflected in a modified ELBO:

$$\mathbb{E}_{p_D}\left[\log p_\theta(\mathbf{x}|\mathbf{u})\right] \geq \mathcal{L}(\theta, \phi) := \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{u})\right] - \mathrm{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u}) \| p_\theta(\mathbf{z}|\mathbf{u})\right). \tag{14}$$

Related work in causal disentanglement. Recently, there has been a significant growth in works in causal disentanglement [64, 33, 21, 34] and this paragraph gives a brief overview. CausalVAE [64] assumes that the SCM in the latent space is a linear additive noise model (ANM), and learns the causal graph as an adjacency matrix through the continuous optimization loss objective from NOTEARS [69], partially as a pre-training step on top of the guiding variables used for disentangling the latent space. A similar concept can be seen in [34] where the graph encoding the relationship between variables is assumed to be sparse, or in GraphVAE [21] where the SCM is imposed on the chain-structured latent space of LadderVAE [58]. ICM-VAE assumes a given causal graph as prior [33] and models the structural equations of the SCM with flow-based diffeomorphic functions. CMCL [39] distinguishes between a set of variables u and z in the latent space,

where there can only exist causal relations $u \to z$ but not vice versa. Using a generative adversarial network (GAN) for the generation has been explored by CausalGAN [31] or DEAR [54]. Under the setting of no and only partial supervision for the guiding variables, existing work include SCADI [42] or [7]. Another line of work studies the related direction of "object-centric learning", where factors in the latent space are matched to visual objects in the image [38, 63]. Note that the disentangled factors in the latent space are often modelled as scalar variables instead of vectors [39, 64]. While it is argued that existing approaches can be generalised to the vector-valued case [54, 33], a thorough experimental evaluation is missing. Extending the scalar-valued case to vector-valued disentangled representation might pose non-trivial difficulties [65]. Only assuming a partial given causal graph is addressed by GOGGLE [35] for the application of generating tabular data.

5.2 Causal DiffuseVAE

As outlined before, we investigate the generative utility of VAEs in this section. We introduce a novel method *Causal DiffuseVAE*, that combines supervised VAEs and a causal layer similar to CausalVAE with the noise-reducing DiffuseVAE [43]. We first introduce the architecture in detail and then investigate results.

5.2.1 Motivation

Modern generative models such as diffusion models or VAEs are prone to producing erroneous outputs since they have no underlying understanding of the data. By integrating causality, we try to mitigate this issue, while also enabling precise interventions on selected causal factors. This approach enhances the accuracy of generated outputs, highlighting the primary contributions of our method. By learning a distribution for each latent variable, VAE models enable generative sampling from the latent space, thus producing new outputs from the same data distribution. However, many challenges arise from the sampling process. Conventional VAEs have a highly entangled latent space, as mentioned in section 5.1, thus, manipulating one single latent factor affects multiple generative factors. For example, for the CelebA dataset, conventional VAEs cannot disentangle the generative factor for eyes and smile. This is a core limitation, which we address with our supervised VAE model by using supervision signals in the form of labels.

5.2.2 Model Architecture.

Model Overview Given images $\mathcal{D} = \{x_j\}_{j=1}^{M}$ and $N \in \mathbb{N}$ semantic concepts within these images, we want to generate new images similar to $\mathcal{D}$ under manipulation of the semantic concepts. For each image x_j, we assume a corresponding binary label vector $y^j \in \{0,1\}^N$ of ground truth concepts. In the case of CelebA, we have, e.g., $N = 40$ concepts like gender or smile. We want to represent each concept i by a $d \in \mathbb{N}$ dimensional vector z_i in the latent space of our VAE resulting in the $(N \times d)$-dimensional latent space $Z = \{z_1, ..., z_N\}$. Finally, our model consists of the encoder $E : \mathcal{X} \to \mathbb{R}^{N \times d}$, an auxiliary classifier $f_{cls} : \mathbb{R}^{N \times d} \to \{0,1\}^N$, our causal layer, decoder $D : \mathbb{R}^{N \times d} \to \mathcal{X}$, and the diffusion model. The complete image of our architecture can be found in fig. 5.

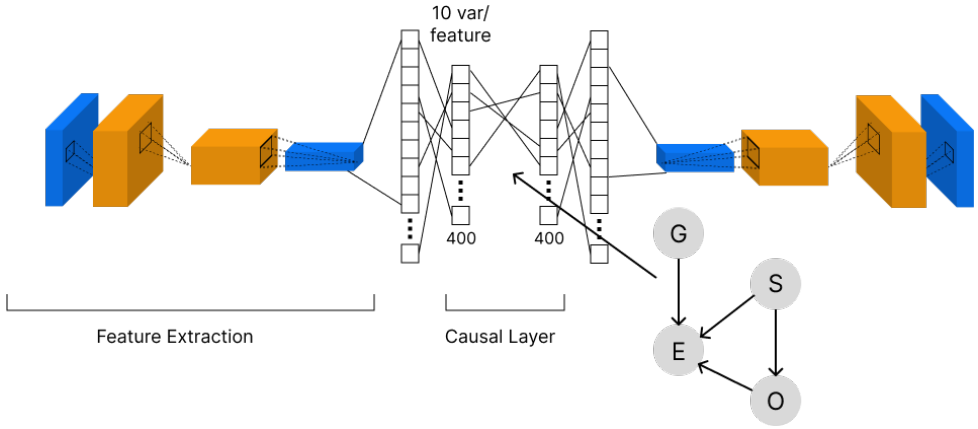
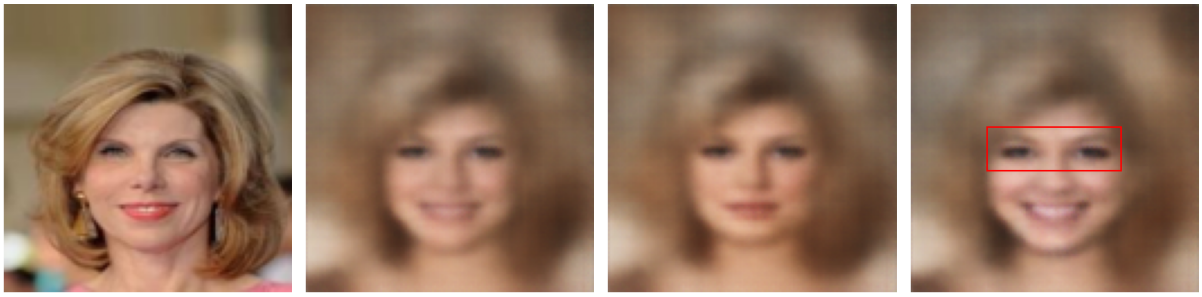


Figure 5: Causal supervised VAE architecture. The architecture consists of convolutional layers for feature extraction. Afterwards, the latent space is modelled in a manner equivalent to conventional VAES. The refined, but flexible causal graph is utilized in the causal layer. Subsequently, the latent feature vectors are causally transformed to model the aforementioned SCM.



(a) (b) (c) (d)

Figure 6: This figure illustrates interventions performed on the latent space. The first image is the original ground truth. Image b. is the VAE output. In figure c. we intervene on the variable for smiling by setting it to zero. The final picture shows the effect of the intervention on smiling set to a higher value.

Causal Layer. To integrate causality into our VAE, we define a causal layer similar to CausalVAE. In particular, we define a causal graph G over the N semantic concepts. On CelebA, for example, we utilize the ground truth causal graphs depicted in fig. 3 with e.g. the causal relationship that 'smile' $\rightarrow$ 'mouth open'. Using the causal graph, we generate a new causal representation $z^c \in \mathbb{R}^{N \times d}$ in the latent space by re-computing each z_i according to the data-generating process of the causal graph. To obtain disentanglement, we use the supervision signal m and mask out all non-present concepts. The supervision signal is either the ground truth label $m^j = y_j$ or the output of the auxiliary classifier $m = f_{cls}(z)$. Formally, for an image x_j we compute

$$z_i^c = f((m \cdot Z)_{\text{an}_i^G}) + z_i \tag{15}$$

where $Z = E(x_j)$, $(m \cdot Z)_{\text{an}_i^G} := \{m_i z_i | X_i \in \text{an}_i^G\}$, and f is a 1-layer multi-layer perceptron. Adding z_i is crucial, since it carries essential information about its variable i, including

(a) (b) (c) (d)

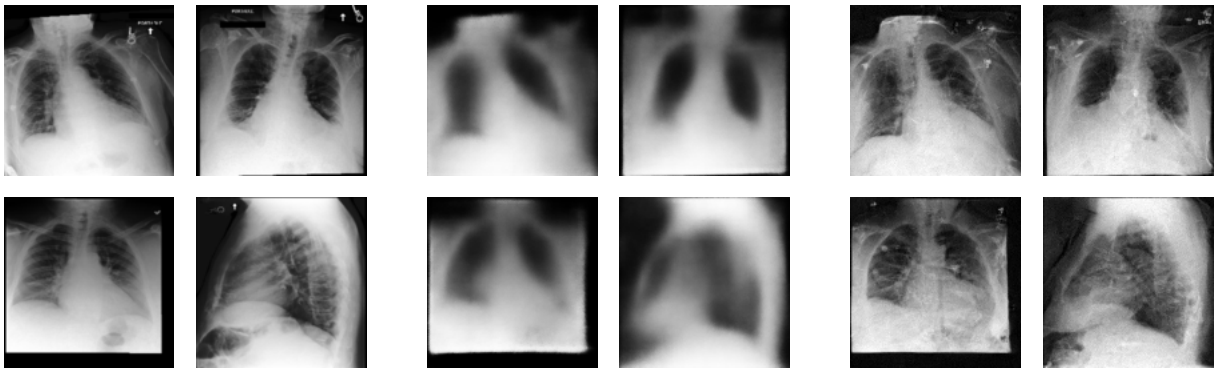
Figure 7: This figure illustrates interventions performed on the latent space. The results are generated utilizing $\gamma = 0.01$. The first image is the original input. Image b. is the VAE output. Figure c. show an intervention where we set the gender to female.. The final picture shows the effect of intervening on smiling.

aspects that are not directly modelled by the data-generating process outlined by the SCM and ensuring that no information is lost. Compared to CausalVAE, we use the ancestral set an_i^G instead of only the parents pa_i^G which allows us to incorporate longer-range ancestry. Additionally, we fix the causal graph from prior knowledge , whereas CausalVAE learns the weights of the VAE jointly with the causal graph. We can interpret z_i as exogenous variables, as they originate from the encoding process of the image. Here, we interpret the image as the main exogenous factor, which is then disentangled by the VAE encoder, resulting in the set of variables $\{z_1, \ldots, z_N\}$ where each z_i corresponds to the exogenous noise of endogenous variable X_i. The z_i^c can be interpreted as the transformed latent variables after the causal mechanism. While this formulation does not closely resemble the definition of a SCM, it integrates a novel causal mechanism into the latent space, resulting in the ability to generate causally plausible counterfactual images.

Training Objective. We train the VAE and the diffusion model in two separate steps. Given $x_j \in \mathcal{D}$ with the ground truth labels y^j, we first encode x_j to $z = E(x_j) \in \mathbb{R}^{N \times d}$ in the forward pass. From z, we classify $\hat{y}^j := f_{cls}(z)$ where f_{cls} is an auxiliary classifier. For masking, we use either $m := y^j$ or $m := \hat{y}^j$ and apply the causal layer on z yielding $z^c \in \mathbb{R}^{N \times d}$. Finally, we feed z^c to the decoder and obtain the reconstruction $\hat{x_j}$. We train the weights of the encoder, decoder, and the classifier (if $m = \hat{y}^j$) using the objective of the semi-supervised VAE from [30], i.e.

$$\mathcal{L}(x_j, y^j) = \alpha \cdot \text{MSE}(x_j, \hat{x}_j) + \beta \cdot D_{\text{KL}} + \gamma \cdot \text{BCE}(y^j, m). \tag{16}$$

The purpose of the last component, the binary cross-entropy loss, is to train f_{cls} to perform supervised multi-label classification corresponding to the labelled features. The values of $\alpha, \beta, \gamma > 0$ are hyper-parameters that correspond to the reconstruction loss, the KL-divergence and the classification loss, respectively. Note that changing one hyper-parameter affects the quality of the other aspects. For example increasing γ will make the network focus more on accurate classification, rather than accurate reconstruction. As the second training step, we freeze the pre-trained weights of the VAE and train a diffusion model on top, similarly to DiffuseVAE [43]. The outputs of the decoder are used



(a) (b) (c)

Figure 8: Example outputs of DiffuseVAE [43] on MIMIC-CXR. a) Original images; b) supervised variational autoencoder (SVAE) output; c) Diffusion model reconstructions.

as input the diffusion model. The diffusion model's objective is to yield more accurate reconstructions.

5.2.3 Experiments

CelebA. We train our model on CelebA using the $N = 40$ concepts with $d = 10$ latent dimensions per concept. As causal graph, we set only the edges of fig. 3. The reason is that with simpler graphs, it is easier to verify the learned causal mechanism. In fig. 6 and fig. 7, two interventions are performed to generate counterfactual outcomes. Note that for training the model that produces the outcomes in fig. 6, we use the ground truth labels for masking, whereas in fig. 7, we use the predictions of the auxiliary classifier with $\gamma = 0.001$ for masking. In the first figure, by intervening on the latent space vector corresponding to 'smile', the woman exhibits different levels of smiling. Firstly, this illustrates our model's disentanglement capability. Secondly, by observing the other images, the relationships defined by the causal graph for CelebA, mentioned in section 3, can also be observed. For example, by intervening on the 'smile' variable, the eyes are causally affected. We can see the effect in fig. 6. Another example of intervening on the gender can be seen in fig. 7. The 'gender' causally affects the 'eyes open'. In addition, by intervening on the 'smile' parent, both parents causally affect the child variable 'eyes open' in the last output. This corresponds to the causal mechanism learned by the model.

MIMIC-CXR. We train our model on MIMIC-CXR without utilizing a causal graph as input, hence setting $\gamma = 0$ and skipping masking and the causal layer. The results utilizing the latter dataset can be seen in fig. 8. The figure illustrates the VAE reconstruction as well as the final diffusion model output. We can observe that the diffusion model successfully reconstructs the sharpness of the blurry reconstruction of the VAE, demonstrating the capability of the two-stage DiffuseVAE pipeline to accurately reconstruct images for radiology. Note that the diffusion model naturally introduces artifacts, also known as hallucinations. Nevertheless, we argue that since the diffusion model does not deviate from the information present in the VAE reconstruction, these artifacts will

be reduced as long as the VAE generations entails the relevant information. The output of the VAE can be scaled up using hierarchical VAEs, e.g. NVAE [61].

5.3 Classification From Latent Space

Next to the generative aspect of causally disentangled VAEs, in this section, we explore the potential of causal disentanglement for the downstream pathology classification on MIMIC-CXR using ICM-VAE. Interestingly, the discriminative and the generative aspect do not need to be separate components, but can work in interplay with each other. Particularly, the generative aspect can be used to explain downstream predictions using latent traversals [18, 2, 4]. For instance, Atad et al. deploy StyleGAN to explain predictions with counterfactual generations [2]. We do not go into detail in how the generative aspect of VAEs can be leveraged to explain its predictions, but we highlight its benefit by referring to the existing work.
Moreover, seeing that disentanglement improves robustness of predictions, for example, with ICAM [4] or DirVAE [18], integrating causal aspects of disentanglement can bring further improvements, conjecturing that disentanglement of causal factors may be more data efficient and potentially invariant to distribution shifts [54]. The conducted experiment attempts to delve into the advantage of the causal aspect, when applied to MIMIC-CXR.

Experimental Setup. The encoder of ICM-VAE generates a latent representation $z \in \mathbb{R}^{l_z}$, where $l_z \in \mathbb{N}_1$ represents the dimensionality of the latent space. Consider $U := \{u_1, \dots, u_n\}$ as the set of semantically meaningful guiding variables, with $n \in \mathbb{N}_1$ (for instance, $n = 4$ for the CausalCircuit dataset as mentioned in section 3). Note that l_z must be divisible by n by definition. Each guiding variable u_i is associated with a distinct subspace $z_{sub}^{(u_i)} \in \mathbb{R}^{\frac{l}{n}}$ within the latent space, where $i \in [n]$, and $z_{sub}^{(u_i)}$ specifically disentangles the latent space for the variable u_i. Thus, we get

$$z = (z_{sub}^{(u_1)}, \dots, z_{sub}^{(u_n)})^T \in \mathbb{R}^{l_z} \tag{17}$$

Here, we leverage the pathology labels from CheXPert as guiding variables, as detailed in section 3.1. We adopt a 4-variable causal graph, as depicted in fig. 2, employing the

Mode	Latent subspace	s_{latent}	F1	AUC
Enlarged Cardiomegaly	$z_{sub}^{(E.C.)}$	32	0.676	0.652
Edema	$z_{sub}^{(E.)}$	32	0.729	0.695
Pneumonia	$z_{sub}^{(P.)}$	32	0.674	0.650
Consolidation	$z_{sub}^{(C.)}$	32	0.667	0.638
Causal	$z_{sub}^{(E.C.)}, z_{sub}^{(E.)}$	64	0.739	0.709
Baseline	z	128	0.740	0.711

Table 3: AUC and F1 scores for classifying edema from the latent space of ICM-VAE. s_latent denotes the size of the latent subspace used for the classification.

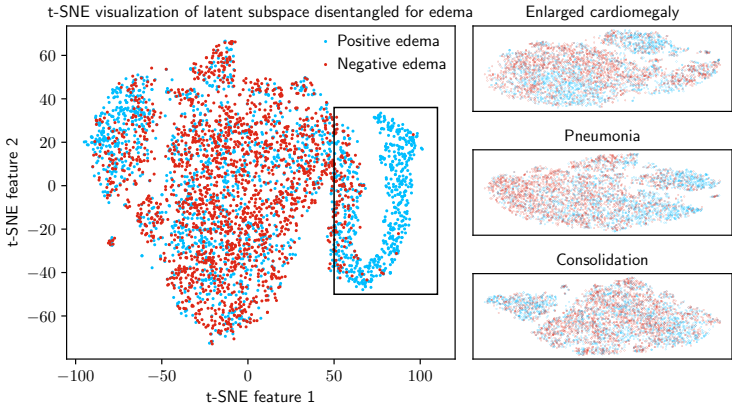


Figure 9: Latent space t-SNE visualization of ICM-VAE applied on MIMIC-CXR.

causal prior of two causal relationships, namely enlarged cardiomegaly (E.C.) → edema (E.) and pneumonia (P.) → consolidation (C.), see section 3. Consequently, we select

$$U_{CXR} := \{(E.C.), (E.), (P.), (C.)\} \tag{18}$$

as the set of guiding variables. Recall from eq. (17) that $z_{sub}^{t_{descr}}$ defines the latent subspace of guiding variable $t_{descr} \in U_{CXR}$. Our objective is to classify edema using the latent representation z, employing logistic regression for this task. The results are shown in table 3. As expected, utilizing the entire latent space for classification yields the highest AUC due to the maximal inclusion of input variables. The principle of disentanglement suggests that the subspace specific to edema, $z_{sub}^{(E.)}$, contains the most concentrated information for edema classification. This hypothesis is empirically validated by our results, which show the highest AUC when classifying from $z_{sub}^{(E.)}$, with a notable margin of 0.4 compared to other subspaces. Interestingly, incorporating the latent subspace associated with edema's causal parent, enlarged cardiomegaly, nearly matches the baseline performance achieved by the full latent space, with only a marginal difference of 0.002, despite utilizing only half of the latent dimensions. These findings underscore the enriched information content within $z_{sub}^{(E.)}$ for edema classification. Furthermore, the inclusion of causally related subspaces enhances performance while maintaining robustness and interpretability, thanks to the causal connections among guiding variables.

Latent Space Analysis. The classification outcomes underscore the disentanglement process's impact on latent space characteristics. This section delves into the latent space z structure to uncover further insights. Benchmarking and analyzing the disentanglement properties of latent spaces are challenging because of the complex characteristics of latent spaces. Dimension reduction techniques such as t-SNE may not always preserve the semantic integrity of guiding variables in their most significant components. Despite these concerns, our application of t-SNE revealed clustering tendencies within the latent subspaces, as illustrated in fig. 9. The two-dimensional t-SNE scatter plot illustrates the distribution of each latent subspace $z_{sub}^{(t_{descr})}$ for $t_{descr} \in U_{CXR}$. Labels are applied based on the positive and not positive edema (noting that 'not positive' is distinct from 'negative', as clarified in section 3.1). The t-SNE plot for $z_{sub}^{(E.)}$ uniquely exhibits a clear clustering of positive edema cases, unlike other subspaces, suggesting that edema-related

information predominantly influences the variation within $z_{sub}^{(E.)}$. This clustering effect can likely be attributed to the disentanglement process enhancing the specificity and density of information. However, the observed overlap within $z_{sub}^{(E.)}$'s points may be due to t-SNE's limitations for comprehensive latent space analysis and potential shortcomings in hyperparameter tuning and architectural decisions for ICM-VAE's application to the MIMIC-CXR dataset.

6 Discussion

The approaches previously described aim to advance approaches combining deep learning and causality in the medical domain. However, there are fundamental gaps that restrict their effectiveness.

General limitations of causality. Although the causality framework is promising in theory, there are many pitfalls in practice. Most importantly, the assumptions we make a very strong and usually hard to verify, e.g., the assumption of a fully observed SCM essentially postulates that we can 'map out the whole universe', i.e. we *know and observe all causes* of our variables of interest. Although we might test for existence of confounders, we have no reasonable possibility but domain knowledge to know the confounders. There exist many theoretical results on causal inference with hidden confounding but at the cost of further restrictions [48]. Especially fig. 4 hints that we have hidden confounding in our case. We have seen that causal discovery itself can in general only recover equivalence classes that are insufficient for many downstream tasks. To break up these symmetries, we often make difficult to verify assumptions like additive noise models. Additionally, causal discovery does not scale well because of the super-exponential growth of possible graphs [62] and a single wrongly oriented edge can completely change the causal effects. Moreover, we can never verify causal effects on real-world data since we can never observe two different outcomes, e.g., both what happens if we administer and not administer treatment to a patient. This is also called *fundamental problem of causal inference* in the potential outcome framework [13].

Causal Effects in Static Images. A cause must always have occurred before its effect in time. If static images capture processes, humans can identify the process and interpret causal meaning into those images. Take an image of a tree on a sunny day as example. It is clear, that the sun casts a shadow of the tree since it radiates light beams that the tree intersects. However, in static chest x-ray images, has an 'enlarged heart' caused Edema, has Edema caused an 'enlarged heart', or did they gradually co-develop? This question might not even be of interest if we only want to *predict* whether Edema is present. This is why we opted to interpret the labels as the 'observation of an enlarged heart' influences the 'diagnosis of Edema'. It is difficult to asses to which extent this framing deteriorates

Leveraging Additional Information. So far, we have considered all images to be independent. However, the MIMIC-CXR-JPG dataset actually contains multiple images from the same patients taken at different time points. Utilizing this information could further improve the methods, e.g. temporal BioVil [3]. We note that causal methods over

time come with additional challenges like time-dependent confounding, which cannot be handled by standard adjustment methods. Proposed methods are e.g. g-computations, for an overview see [41]. Additionally, we could leverage additional data about patients like the administration of drugs or similar from the HAIM datatset [57]. However, since there are lots of variables in HAIM, this would require additional medical domain knowledge.

Conflicting Goals. While causality requires to explicitly state and in the best case check assumptions, deep learning only contains implicit inductive biases and focuses on empirical performance metrics on a test hold-out set. This can create a conflict of interest, where a method that incorporates both causality and deep learning neither fulfills strict causal assumptions nor performs well on empirical metrics. This happened, e.g., in section 4, where we did not use a strict causal graph to generate substantially worse predictions than in the deep learning literature. Interestingly, causal disentanglement presents a field that emerged with the motivation of combining the benefits of causality and deep learning. However, this currently only seems to work in very simple settings, see next paragraph.

Causal Disentanglement For Classification. Utilizing disentangled VAEs for classification has been shown to improve robustness in downstream classification, compared to classification from the latent space of non-disentangled VAEs [4, 18]. In addition, it has been conjectured that classifying from *causally* disentangled factors further improves robustness of downstream classification [54]. However, due to the best of our knowledge, there currently exist no concrete theoretical nor empirical evidence to back this direction. In section 5.3 we have seen that the this causal prediction is comparable but still lacks behind just using the whole latent space. We think that the chosen setting of chest x-rays is too complicated and the chosen ground truth graph too simple for current state-of-the art methods to perform well. We also note that the datasets from the literature of causal representation learning usually have few very clear causal relations or even are synthetic datasets.

Causal Disentanglement for Generation. Counterfactual generation is a natural application for causally disentangled VAEs because VAEs naturally are generative models. Indeed, counterfactual image generation from causal VAEs is starting to gain traction [32]. However, for real-world applications, we still see a large gap between state-of-the-art generative artificial intelligence (AI) such as stable diffusion [49] and causal generators, especially in the generative output quality. Causal generators might pose the advantage of introducing the dimension of causal coherence to state-of-the-art generative models and might assist in reducing hallucinations. However, the overall success seems to be highly dependent on a good causal model.

7 Conclusion

We have motivated how incorporating causality into deep learning methods can yield robust results that are essential in the medical domain. We explored applying causal methods on top of zero-shot VLMs and presented a proof-of-concept for image generation

and prediction with causal variational models. However, both methods fell short compared to pure deep learning state-of-the-art literature without showing significant causal benefits. As a potential reason, we conjectured that not satisfying causal assumptions and limited knowledge or bad modelling of the data-generating process. Especially, the 14 pathologies do not seem to capture all necessary information. Creating a more sophisticated graph with more labels and correct underlying causal effects is possible but requires medical domain experts. Additionally, we observed that the MIMIC dataset is more complicated than the datasets typically used in the causal representation learning literature. While the time frame of the project did not suffice to fully polish our methods, the obtained results encourage further exploration and tuning.

References

[1] Hina Amin and Waqas J Siddiqui. "Cardiomegaly". In: *StatPearls [internet]*. StatPearls Publishing, 2021.

[2] Matan Atad et al. "Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan". In: *arXiv preprint arXiv:2207.07553* (2022).

[3] Shruthi Bannur et al. *Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing*. 2023. DOI: 10.48550/ARXIV.2301.04558.

[4] Cher Bass et al. "ICAM: interpretable classification via disentangled representations and feature attribution mapping". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 7697–7709.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[6] Benedikt Boecking et al. "Making the Most of Text Semantics to Improve Biomedical Visionâ€"Language Processing". In: *Computer Vision â€" ECCV 2022*. Springer Nature Switzerland, 2022, 1â€"21. DOI: 10.1007/978-3-031-20059-5_1.

[7] Johann Brehmer et al. "Weakly supervised causal representation learning". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022. eprint: 2203.16437. URL: https://arxiv.org/abs/2203.16437.

[8] Erdi Çallı et al. "Deep learning for chest X-ray analysis: A survey". In: *Medical Image Analysis* 72 (2021), p. 102125.

[9] Pierre Chambon et al. "RoentGen: vision-language foundation model for chest x-ray generation". In: *arXiv preprint arXiv:2211.12737* (2022).

[10] N. V. Chawla et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (June 2002), 321â€"357. ISSN: 1076-9757. DOI: 10.1613/jair.953. URL: http://dx.doi.org/10.1613/jair.953.

[11] Ricky TQ Chen et al. "Isolating sources of disentanglement in variational autoencoders". In: *Advances in neural information processing systems* 31 (2018).

[12] Rewon Child. "Very deep vaes generalize autoregressive models and can outperform them on images". In: *arXiv preprint arXiv:2011.10650* (2020).

[13] Alicia Curth et al. "Really Doing Great at Estimating CATE? A Critical Look at ML Benchmarking Practices in Treatment Effect Estimation". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Vol. 1. Curran, 2021. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/2a79ea27c279e471f4d180b08d62b00a-Paper-round2.pdf.

[14] Bin Dai and David Wipf. "Diagnosing and enhancing VAE models". In: *arXiv preprint arXiv:1903.05789* (2019).

[15] EMCrit Project. *Thoracic Radiology - Consolidation*. https://emcrit.org/. Accessed: 2024-02-03. 2023.

[16] Andre Esteva et al. “A guide to deep learning in healthcare”. In: *Nature medicine* 25.1 (2019), pp. 24–29.

[17] Gaël Gendron, Michael Witbrock, and Gillian Dobbie. “A Survey of Methods, Challenges and Perspectives in Causality”. In: *arXiv preprint arXiv:2302.00293* (2023).

[18] Rachael Harkness et al. “Learning disentangled representations for explainable chest X-ray classification using Dirichlet VAEs”. In: *arXiv preprint arXiv:2302.02979* (2023).

[19] Mohammad Havaei et al. “Conditional generation of medical images via disentangled adversarial inference”. In: *Medical Image Analysis* 72 (2021), p. 102106.

[20] Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. “Efficientvdvae: Less is more”. In: *arXiv preprint arXiv:2203.13751* (2022).

[21] Jiawei He et al. “Variational autoencoders with jointly optimized latent dependency structure”. In: *International conference on learning representations*. 2018.

[22] David Hume. *An Enquiry concerning Human Understanding*. 1748.

[23] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. “Nonlinear ICA using auxiliary variables and generalized contrastive learning”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 859–868.

[24] Jeremy Irvin et al. *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*. 2019. arXiv: 1901.07031.

[25] A Johnson et al. *MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0)*. 2019. DOI: /10.13026/8360-t248.

[26] Alistair EW Johnson et al. “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports”. In: *Scientific data* 6.1 (2019), p. 317.

[27] Maxime Kayser et al. *Explaining Chest X-ray Pathologies in Natural Language*. 2022. arXiv: 2207.04343 [cs.CV].

[28] Ilyes Khemakhem et al. “Variational autoencoders and nonlinear ica: A unifying framework”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2207–2217.

[29] Hyunjik Kim and Andriy Mnih. “Disentangling by factorising”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2649–2658.

[30] Diederik P. Kingma et al. *Semi-Supervised Learning with Deep Generative Models*. 2014. arXiv: 1406.5298 [cs.LG].

[31] Murat Kocaoglu et al. “Causalgan: Learning causal implicit generative models with adversarial training”. In: *arXiv preprint arXiv:1709.02023* (2017).

[32] Aneesh Komanduri et al. “From Identifiable Causal Representations to Controllable Counterfactual Generation: A Survey on Causal Generative Modeling”. In: *arXiv preprint arXiv:2310.11011* (2023).

[33] Aneesh Komanduri et al. “Learning Causally Disentangled Representations via the Principle of Independent Causal Mechanisms”. In: *arXiv preprint arXiv:2306.01213* (2023).

[34] Sébastien Lachapelle et al. "Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA". In: *Conference on Causal Learning and Reasoning*. PMLR. 2022, pp. 428–484.

[35] Tennison Liu et al. "GOGGLE: Generative modelling for tabular data by learning relational structure". In: *The Eleventh International Conference on Learning Representations*. 2022.

[36] Ziwei Liu et al. "Large-scale celebfaces attributes (celeba) dataset". In: *Retrieved August* 15.2018 (2018), p. 11.

[37] Francesco Locatello et al. "Challenging common assumptions in the unsupervised learning of disentangled representations". In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.

[38] Amin Mansouri et al. "Object-centric architectures enable efficient causal representation learning". In: *arXiv preprint arXiv:2310.19054* (2023).

[39] Haiyi Mao et al. "Towards cross-modal causal structure and representation learning". In: *Machine Learning for Health*. PMLR. 2022, pp. 120–140.

[40] Riccardo Miotto et al. "Deep learning for healthcare: review, opportunities and challenges". In: *Briefings in bioinformatics* 19.6 (2018), pp. 1236–1246.

[41] Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. "An introduction to g methods". In: *International journal of epidemiology* 46.2 (2017), pp. 756–762.

[42] Heejeong Nam. "SCADI: Self-supervised Causal Disentanglement in Latent Variable Models". In: *arXiv preprint arXiv:2311.06567* (2023).

[43] Kushagra Pandey et al. "Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents". In: *arXiv preprint arXiv:2201.00308* (2022).

[44] Judea Pearl. *Causality*. Cambridge university press, 2009.

[45] Chantal Pellegrini et al. "RaDialog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance". In: *arXiv preprint arXiv:2311.18681* (2023).

[46] Chantal Pellegrini et al. "Xplainer: From X-Ray Observations to Explainable Zero-Shot Diagnosis". In: *arXiv preprint arXiv:2303.13391* (2023).

[47] Konpat Preechakul et al. "Diffusion autoencoders: Toward a meaningful and decodable representation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10619–10629.

[48] Thomas S. Richardson et al. "Nested Markov properties for acyclic directed mixed graphs". In: *The Annals of Statistics* 51.1 (2023). URL: http://dx.doi.org/10.1214/22-AOS2253.

[49] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. arXiv: 2112.10752 [cs.CV].

[50] Zohaib Salahuddin et al. "Transparency of deep neural networks for medical image analysis: A review of interpretability methods". In: *Computers in biology and medicine* 140 (2022), p. 105111.

[51] Pedro Sanchez and Sotirios A Tsaftaris. "Diffusion causal models for counterfactual estimation". In: *arXiv preprint arXiv:2202.10166* (2022).

[52] Bernhard Schölkopf. "Causality for machine learning". In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 2022, pp. 765–804.

[53] Bernhard Schölkopf et al. "Toward causal representation learning". In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634.

[54] Xinwei Shen et al. "Weakly supervised disentangled generative causal representation learning". In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 10994–11048.

[55] Yasin Shokrollahi et al. "A comprehensive review of generative AI in healthcare". In: *arXiv preprint arXiv:2310.00795* (2023).

[56] Karan Singhal et al. "Large language models encode clinical knowledge". In: *Nature* 620.7972 (Aug. 2023), pp. 172–180. DOI: 10.1038/s41586-023-06291-2. URL: https://doi.org/10.1038/s41586-023-06291-2.

[57] Luis R. Soenksen et al. "Integrated multimodal artificial intelligence framework for healthcare applications". In: *npj Digital Medicine* 5.1 (Sept. 2022), p. 149. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00689-4. URL: https://doi.org/10.1038/s41746-022-00689-4.

[58] Casper Kaae Sønderby et al. "Ladder variational autoencoders". In: *Advances in neural information processing systems* 29 (2016).

[59] Peter Spirtes. "An Anytime Algorithm for Causal Inference". In: *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*. Vol. R3. Proceedings of Machine Learning Research. PMLR, 2001, pp. 278–285. URL: https://proceedings.mlr.press/r3/spirtes01a.html.

[60] Emanuel Todorov, Tom Erez, and Yuval Tassa. "Mujoco: A physics engine for model-based control". In: *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE. 2012, pp. 5026–5033.

[61] Arash Vahdat and Jan Kautz. "NVAE: A deep hierarchical variational autoencoder". In: *Advances in neural information processing systems* 33 (2020), pp. 19667–19679.

[62] Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. *D'ya like DAGs? A Survey on Structure Learning and Causal Discovery*. 2021. arXiv: 2103.02582 [cs.LG].

[63] Yi-Fu Wu, Minseung Lee, and Sungjin Ahn. "Object-Centric Semantic Vector Quantization". In: *Causal Representation Learning Workshop at NeurIPS 2023*. 2023.

[64] Mengyue Yang et al. "Causalvae: Disentangled representation learning via neural structural causal models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 9593–9602.

[65] Tao Yang et al. "Vector-based Representation is the Key: A Study on Disentanglement and Compositional Generalization". In: *arXiv preprint arXiv:2305.18063* (2023).

[66] Xu Yang, Hanwang Zhang, and Jianfei Cai. "Deconfounded Image Captioning: A Causal Retrospect". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022). DOI: 10.1109/tpami.2021.3121705. URL: http://dx.doi.org/10.1109/TPAMI.2021.3121705.

[67] Xu Yang et al. *Causal Attention for Vision-Language Tasks*. 2021. arXiv: 2103.03493 [cs.CV].

[68] Jingyi Zhang et al. *Vision-Language Models for Vision Tasks: A Survey*. 2023. arXiv: 2304.00685 [cs.CV].

[69] Xun Zheng et al. "Dags with no tears: Continuous optimization for structure learning". In: *Advances in neural information processing systems* 31 (2018).

A Related Work in Image Generation

Deep learning based image generation methods such as stable-diffusion [49] have shown tremendous success in diverse applications ranging from content creation to healthcare [55]. In the realm of generative AI, VAEs are known to produce sub-optimal blurry images due to their inherent limitations in capturing high-frequency details and the tendency of the variational inference process to average out variations in the data, leading to a loss of sharpness and detail in the generated images [14]. This limitation has been addressed by NVAE [61] and VDVAE [12, 20] by hierarchically scaling up the number of layers. While this increases the image generation quality, it comes with the cost of long and instable training [20]. Interestingly, while hierarchical VAEs hide the advantage of a small and controllable latent space as it grows by hierarchically added layer, it is shown that the first latent space still is the main bottleneck controlling the generative factors of the produced images [61].
Apart from VAE only approaches, existing work combines the benefits of a controllable latent space in VAEs and high generation quality of diffusion models. Highlighted works are DiffuseVAE [43] and DiffVAE [47], which form two-stage approaches of generating a base generation with a VAE, and enhancing the output with a second-stage diffusion model. We emphasise, that in their concepts, the VAE component is modular and can be replaced by a causally disentangled VAE.
In medical imaging, RoentGen [9] show-cases how specialised latent diffusion models can generate visually convincing x-ray images, and that using them for data augmentation provides valuable extra training data for increasing downstream classification performance. It is also argued that controllable generation in medical imaging presents a relevant challenge [19]. Related to a causally assisted diffusion model is Diff-SCM [51]. Diff-SCM is, however, restricted to bivariate causal graphs, which strongly limits their application potential. For a recent survey in causal generative models, see [32].

B Comments on Identifiability

$\sim$-Identifiability. Recall the definition of identifiability property:

$$\forall(\theta, \theta^*) : p_\theta(x) = p_{\theta^*}(x) \implies \theta = \theta^* \tag{19}$$

In the identifiable VAE, the crucial thing is to prove its architecture can satisfy this property and clarify what kinds of assumptions make it identifiable. In order to achieve this goal, [28] introduces the equivalence class concept and reasonably relax the definition of identifiability:

Definition 3. The generative model is $\sim$-identifiable if $p_\theta(\mathbf{x}) = p_{\hat{\theta}}(\mathbf{x}) \Rightarrow \theta \sim \hat{\theta}$.

The generative model of iVAE is defined by $\mathbf{x} \in \mathbb{R}^d$, and $\mathbf{u} \in \mathbb{R}^m$ be two observed random variables, and $\mathbf{z} \in \mathbb{R}^n$ (lower-dimensional, $n \leq d$) a latent variable. Let $\Theta = (f, T, \Lambda)$ be the parameters of the following conditional generative model: $p_\Theta(\mathbf{x}, \mathbf{z}|\mathbf{u}) = p_f(\mathbf{x}|\mathbf{z})p_{T,\Lambda}(\mathbf{z}|\mathbf{u})$ In such framework, we can define the equivalence relation in formal as well:

Definition 4. The equivalence relation $\sim$ on Θ defined as follows:

$$(f, T, \Lambda) \sim (\tilde{f}, \tilde{T}, \tilde{\Lambda}) \iff \exists \mathcal{A}, c \left[T(f^{-1}(x)) = \mathcal{A}T(\tilde{f}^{-1}(x)) + c, \forall x \in \mathcal{X} \right]$$

where $\mathcal{A}$ is an $n \times n$ matrix and c is a vector. If $\mathcal{A}$ is invertible, we denote this relation by $\sim_{\mathcal{A}}$. If $\mathcal{A}$ is a block permutation[3] matrix, we denote it by $\sim_P$.

iVAE model is $\sim_{\mathcal{A}}$-identifiable and it is by equating the joint distributions of the latent variables from two different models.

Variation in Causal VAE and ICM-VAE. CausalVAE and ICM-VAE are both advanced variations of iVAE framework, while both focused on integrating causal inference into Variational Autoencoders, differ in their approach and emphasis on the causal structure within the latent space. The architecture of a Causal VAE mirrors that of a standard VAE with an encoder, a decoder, and a latent space. However, the latent space in a Causal VAE is structured according to a hypothesized causal model by adding two layers causal layer and mask layer. Similar as iVAE, the CausalVAE framework[64]uses the same auxiliary variable $u \in \mathbb{R}^n$ i.e labels of concepts associated with the causal concepts as supervising signals. The generative model of CausalVAE is defined as $p_\Theta(\mathbf{x}, \mathbf{z}, \boldsymbol{\epsilon}|\mathbf{u}) = p_\Theta(\mathbf{x}|\mathbf{z}, \mathbf{u})p_\Theta(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u})$, where $\Theta = (f, h, C, T, \lambda)$. In order to get ELBO, CausalVAE defines the joint prior $p_\theta(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u})$ for latent variables $\mathbf{z}$ and $\boldsymbol{\epsilon}$ as $p_\theta(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u}) = p_\theta(\boldsymbol{\epsilon})p_\theta(\mathbf{z}|\mathbf{u})$, where $p_\theta(\boldsymbol{\epsilon}) = \mathcal{N}(0, I)$ and the prior of latent endogenous variables $p_\theta(\mathbf{z}|\mathbf{u})$ is a factorized Gaussian distribution conditioning on the additional observation $\mathbf{u}$, i.e. $p_\theta(\mathbf{z}|\mathbf{u}) = \prod_i p_\theta(z_i|\mathbf{u}_i)$; $\quad p_\theta(z_i|\mathbf{u}_i) = \mathcal{N}(\lambda_1(\mathbf{u}_i), \lambda_2^2(\mathbf{u}_i))$. The distribution has two sufficient statistics, the mean and variance of $\mathbf{z}$, which are denoted by sufficient statistics $T(\mathbf{z}) = (\mu(\mathbf{z}), \sigma^2(\mathbf{z})) = (T_{1,1}(z_1), \ldots, T_{n,2}(z_n))$. Then it is sufficient to show that the ELBO of this framework is

$$\mathbb{E}_{q_\mathbf{x}} [\log p_\theta(\mathbf{x}|\mathbf{u})] \geq \mathcal{L}(\theta, \phi) := \mathbb{E}_{q_\mathbf{x}} \left[\mathbb{E}_{\mathbf{e}, \mathbf{z} \sim q_\phi} [\log p_\theta(\mathbf{x}|\mathbf{z}, \boldsymbol{\epsilon}, \mathbf{u})] - KL [q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u}) \| p_\theta(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u})] \right]$$

Based on this formulation, CausalVAE can be identifiable:

Theorem B.1. *If all the observed data are from the generative models and following assumptions hold:*

1. *The set $\{\mathbf{x} \in \mathcal{X} | \phi_\epsilon(\mathbf{x}) = 0\}$ has measure zero, where ϕ_ϵ is the characteristic function of the density p_ϵ defined in the decoding and encoding processes $\mathbf{x} = f(\mathbf{z}) + \boldsymbol{\xi}, \quad \boldsymbol{\epsilon} = h(\mathbf{x}, \mathbf{u}) + \boldsymbol{\zeta}$, where $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are the vectors of independent noise with probability densities p_ξ and q_ξ.*

2. *The decoder function f is differentiable and the Jacobian matrix of f is of full rank.*

3. *The sufficient statistics $T_{i,s}(\mathbf{z}_i) \neq 0$ almost everywhere for all $1 \leq i \leq n$ and $1 \leq s \leq 2$, where $T_{i,s}(\mathbf{z}_i)$ is the sth statistic of variable $\mathbf{z}_i$.*

4. *The additional observations $\mathbf{u}_i \neq 0$.*

Then the parameters $\theta = (f, h, C, T, \lambda)$ are $\sim$-identifiable.

On the other hand, in models where causal factors are segregated, identical marginal distributions of the latent variables, $p_\theta(\mathbf{z})$ and $p_{\hat{\theta}}(\mathbf{z})$, iVAE does not guarantee the equivalence of the conditional distributions $p_\theta(\mathbf{z}_i|\mathbf{pa}_i)$ and $p_{\hat{\theta}}(\mathbf{z}_i|\mathbf{pa}_i)$. To solve this problem, the ICM-VAE (Independent Causal Mechanisms-VAE)[33] leverages flow-based diffeomorphic functions, which are smooth, invertible mappings ensuring that the structure of the latent space retains a meaningful and interpretable form. The concept of independent causal mechanisms refers to a set of underlying processes in a causal system where each mechanism that generates a particular variable is conditionally independent of the mechanisms that generate other variables, given its direct causes. This commit to a diffeomorphic structure helps in ensuring that each independent causal mechanism can be isolated and identified in a way that aligns with the principles of causal inference.

C Experimental details of section 4

For each model we learn, we consider the following models in their default PyCaret configuration: Logistic Regression, Light Gradient Boosting Machine, Random Forest Classifier, Extra Trees Classifier, Naive Bayes, Gradient Boosting Classifier, Linear Discriminant Analysis, Ada Boost Classifier, Decision Tree Classifier, Quadratic Discriminant Analysis, Dummy Classifier. However, we exclude support vector machines and k-nearest neighbors since they do not allow to compute prediction probabilities. Although using hand-crafted functions could further improve the performance, using the same training condition make our approaches comparable.

Pathology	Abbreviation in section 4.2.1	Abbreviation in section 4.2.2
Enlarged Cardiomediastinum	En. Cardio.	EC
Cardiomegaly	Cardio.	Ca
Lung Opacity	-	LO
Lung Lesion	-	LL
Edema	-	Ed
Consolidation	Consolid.	Co
Pneumonia	-	Pn
Atelectasis	-	Ar
Pneumothorax	Pneumoth.	Px
Pleural Effusion	Pleural Eff.	PE
Pleural Other	Pleural Oth.	PO
Fracture	-	Fr

Table 4: Abbreviations of pathologies used in section 4.

Split	train	test
gender	Male	Female
age	< 40	> 70
ethnicity	BLACK/AFRICAN AMERICAN, HISPANIC/LATINO	WHITE, ASIAN, UNKNOWN, UNABLE TO DETERMINE

Table 5: Splits in section 4.2.1.