Journal of Advanced Intelligent Computing and Informatics
(JAICI)

Volume 1 Issue 1
Article Number: JAICI20260001

Causal Disentanglement via DenseNet Generative Models for Explainable AI in Chest X-Ray Diagnosis

Pranayachandanreddy Gottimukkala[1], Win War Htet[2], and Roise Uddin[*3]

[1]Independent Researcher, United States
[2]Mae Fah Luang University, Chiang Rai, Thailand 57100
[3]Pacific States University, Los Angeles, California 90010, United States

Abstract

Deep learning models for medical image diagnosis often operate as black boxes, providing limited insight into the reasoning behind their predictions. This opacity hinders clinical trust and adoption, particularly in high-stakes applications such as chest X-ray interpretation. We propose a novel framework that integrates a structurally causal conditional variational autoencoder (SC-CVAE) with a DenseNet-121 backbone to transform opaque feature embeddings into interpretable, intervenable latent representations. The DenseNet-121 extracts hierarchical feature maps from chest X-rays, which are aggregated into a composite tensor and fed into the SC-CVAE. The SC-CVAE encodes these features into a set of latent factors that correspond to clinically meaningful concepts, such as pathology shape, intensity, anatomical location, and scanner artifacts. A domain-derived causal graph governs the relationships among these factors, and a mutual information regularizer enforces disentanglement by penalizing statistical dependence between causally unrelated variables. The decoder reconstructs the feature tensor from the latent factors, and an adversarial training mechanism ensures that concept edits remain localized and medically plausible. For explanation generation, we perform interventions on specific latent factors—for example, setting the shape factor to a counterfactual value—and decode the modified representation to produce a new feature tensor. The difference in downstream classification probabilities between the original and counterfactual features quantifies the concept-level contribution of each factor to the diagnostic decision. The entire system is trained end-to-end with a combined objective that includes reconstruction, mutual information, adversarial, and classification losses. Our approach yields, for each input X-ray, pathology probabilities, a Grad-CAM heatmap for spatial localization, and a concept-level contribution vector. This multi-faceted output enables clinicians to understand not only where the model focuses but also why specific visual features drive the diagnostic outcome. The proposed method thus advances explainable AI in medical imaging by bridging the gap between high-performance deep learning and clinically meaningful interpretability.

Keywords: Explainable Artificial Intelligence (XAI); Chest X-Ray Classification; Structurally Causal Conditional Variational Autoencoder (SC-CVAE); Counterfactual Explanations; Medical Image Analysis

1. Introduction

The integration of deep learning into medical imaging has yielded remarkable diagnostic performance, particularly in tasks such as chest X-ray interpretation where convolutional neural networks (CNNs) now rival expert radiologists in detecting pathologies [1]. Among these architectures, DenseNet-121 [2] has emerged as a preferred backbone due to its

*Corresponding Author: Roise Uddin (ruddin@psuca.edu)
Received: 21 June 2026; Revised: 23 July 2026; Accepted: 14 August 2026; Published: 30 September 2026
© 2026 The Author(s).
This is an open access article licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC 4.0).

dense connectivity patterns that facilitate feature reuse, mitigate vanishing gradients, and enable efficient gradient flow through the network. However, the very complexity that grants DenseNet-121 its predictive power also renders it opaque: the hierarchical feature maps it produces are high-dimensional, entangled, and devoid of semantic meaning, making it difficult for clinicians to understand why a particular diagnosis was reached. This lack of transparency undermines clinical trust and poses a significant barrier to adoption in high-stakes healthcare environments where accountability and interpretability are paramount.

Existing approaches to explainable AI (XAI) in medical imaging have primarily followed two paradigms. Post-hoc attribution methods, such as Grad-CAM [3] and Layer-wise Relevance Propagation [4], generate heatmaps that highlight regions of an image most influential to a model's decision. While useful for spatial localization, these methods provide only correlational insights and cannot distinguish between causal factors and spurious correlations—a critical limitation in medical settings where scanner artifacts, patient positioning, or demographic biases may confound predictions. Concept-based methods, such as Testing with Concept Activation Vectors (TCAV) [5], attempt to quantify the importance of human-defined concepts, but they require external concept annotations and do not support counterfactual reasoning or intervention-based analysis. Neither paradigm offers a mechanism to answer "what if" questions—for example, "Would the diagnosis change if the pathology had a different shape or intensity?"—which are essential for building causal understanding and clinical confidence.

To address these limitations, we propose a framework that integrates a structurally causal conditional variational autoencoder (SC-CVAE) with a DenseNet-121 backbone, transforming opaque feature embeddings into an interpretable, intervenable latent representation. The core insight is to explicitly factorize the latent space of a generative model according to a domain-derived causal graph, where each latent dimension corresponds to an independent causal mechanism governing a clinically meaningful concept—such as pathology shape, intensity distribution, anatomical location, or scanner-induced artifacts. The DenseNet-121 encoder supplies multi-scale hierarchical feature maps that condition the SC-CVAE, and the decoder reconstructs these features from the disentangled latent factors. By performing interventions on specific latent coordinates—setting a factor to a counterfactual value while holding others fixed—we can synthesize counterfactual feature representations that isolate the effect of a single concept. These modulated features are then decoded and passed to a downstream classifier, allowing us to quantify concept-level contributions to diagnostic decisions in a causally faithful manner.

The proposed method differs fundamentally from prior work in several key respects. First, unlike post-hoc attribution methods that operate on the final model output, our approach intervenes directly on the latent representation, enabling causal reasoning rather than mere correlation analysis. Second, in contrast to concept-based methods that rely on external annotations, our framework learns disentangled factors automatically from data, guided by a causal graph that encodes domain knowledge about the generative process of medical images. Third, the integration of a mutual-information minimization regularizer [6] with adversarial training against a concept predictor ensures that the learned factors are statistically independent and that concept edits remain localized and medically plausible—a property we term "causal faithfulness." This combination of structural causal modeling [7], conditional variational inference [8] [9], and disentangled representation learning [10] [11] yields a pipeline that converts the originally opaque DenseNet-121 embeddings into a transparent, intervenable representation suitable for clinical decision support.

The contributions of this work are threefold. First, we introduce a novel architecture that integrates a structural causal model into a conditional variational autoencoder operating on hierarchical DenseNet-121 feature maps, enabling causally faithful disentanglement of medical concepts. Second, we develop an intervention-based explanation mechanism that generates counterfactual feature representations to quantify concept-level contributions to diagnostic decisions, moving beyond correlational heatmaps to causal attribution. Third, we demonstrate that the proposed framework achieves competitive diagnostic accuracy while providing interpretable, intervenable outputs that support transparent clinical reasoning. The remainder of this paper is organized as follows: Section 2 reviews related work in explainable AI, causal representation learning, and generative models for medical imaging. Section 3 provides preliminaries on causal inference and variational autoencoders. Section 4 details the proposed Causal-Disentangled DenseNet framework, including the architecture, training objective, and explanation generation procedure. Section 5 presents experimental results on chest X-ray datasets, including quantitative evaluations of diagnostic accuracy, disentanglement quality, and causal faithfulness. Section 6 discusses limitations and future directions, and Section 7 concludes the paper.

2. Related Work

2.1. Explainable AI in Medical Imaging

The demand for interpretability in medical AI has spurred a rich body of research, yet the field remains divided between post-hoc explanation methods and inherently interpretable models. Post-hoc attribution techniques, such as Grad-CAM [3] and Integrated Gradients [12], generate saliency maps that highlight input regions deemed important by

the model. While these methods are widely adopted due to their model-agnostic nature, they suffer from fundamental limitations: they provide only correlational insights, are sensitive to gradient noise, and cannot distinguish between causal factors and spurious correlations [13]. In medical imaging, this means that a heatmap might highlight a region containing a scanner artifact or a normal anatomical variant rather than the true pathology, potentially misleading clinicians. Moreover, these methods do not support counterfactual reasoning—they cannot answer questions about how the model's decision would change under different visual conditions.

Inherently interpretable models, such as concept bottleneck models (CBMs) [14], address some of these issues by first predicting human-interpretable concepts (e.g., "airspace opacity," "cardiomegaly") and then using these concepts to make final predictions. However, CBMs require expensive concept annotations for training, and their performance is bottlenecked by the quality and completeness of the concept set. Furthermore, they assume that concepts are independent, which is rarely true in medical imaging—for example, the shape of a pathology is causally influenced by its anatomical location. The proposed SC-CVAE framework overcomes these limitations by learning disentangled latent factors automatically from data, guided by a causal graph that encodes known dependencies among concepts, thereby eliminating the need for exhaustive concept annotations while preserving interpretability.

2.2. Causal Representation Learning

Causal representation learning [15] aims to uncover the underlying causal variables that generate observed data, moving beyond statistical correlations to capture the mechanisms of data generation. A key desideratum in this field is disentanglement: the latent factors should correspond to independent causal mechanisms, such that intervening on one factor does not affect the others [16]. Early work in disentangled representation learning, such as β-VAE [10] and FactorVAE [11], enforced statistical independence in the latent space through regularization, but these methods did not incorporate causal structure. More recent approaches have integrated causal graphs into variational autoencoders, enabling interventions and counterfactual reasoning [17] [18]. For example, the CausalVAE model [19] learns a latent space where each dimension corresponds to a causal variable, and the decoder implements a structural causal model. However, these methods have primarily been applied to synthetic datasets or simple image domains, and their extension to high-dimensional medical images remains challenging due to the complexity of the feature space and the need for domain-specific causal knowledge.

Our work builds upon these foundations by introducing a structurally causal conditional VAE that operates on hierarchical feature maps from a pre-trained DenseNet-121 backbone. This design choice is motivated by the observation that medical images contain multi-scale anatomical structures—from global organ shapes to fine-grained texture patterns—that are naturally captured by the dense blocks of DenseNet-121. By conditioning the SC-CVAE on these multi-scale features, we enable the model to learn causal factors at different levels of abstraction, from coarse anatomical location to fine pathology texture. Furthermore, we incorporate a mutual-information regularizer that explicitly penalizes statistical dependence between causally unrelated factors, ensuring that the learned representation is both disentangled and causally faithful.

2.3. Generative Models for Medical Imaging

Generative models have found numerous applications in medical imaging, including data augmentation, anomaly detection, and image synthesis [20]. Variational autoencoders (VAEs) [8] and generative adversarial networks (GANs) [21] have been used to generate realistic chest X-rays for training data expansion and to model the distribution of normal anatomy for anomaly detection. More recently, diffusion models [22] have achieved state-of-the-art image quality and have been applied to tasks such as super-resolution and image-to-image translation in medical contexts [23]. However, the primary focus of these works has been on image generation quality rather than on interpretability or causal reasoning.

A notable exception is the work on causally structured counterfactual generation [24], which uses a causal graph to generate counterfactual medical images for interpretability. This approach demonstrates the feasibility of using causal generative models to answer "what if" questions in medical imaging, but it operates at the pixel level and requires a pre-defined causal graph with explicit intervention targets. Our framework extends this idea by operating on feature-level representations rather than raw pixels, which offers several advantages: (i) it leverages the rich semantic information encoded in DenseNet-121 feature maps, (ii) it reduces the computational burden of generating high-resolution counterfactual images, and (iii) it enables direct integration with downstream classification heads for concept-level contribution analysis. Moreover, our adversarial training mechanism ensures that concept edits remain localized and medically plausible, addressing a key challenge in counterfactual generation for medical images.

2.4. Comparison with Existing Methods

The proposed Causal-Disentangled DenseNet framework occupies a unique position at the intersection of explainable AI, causal representation learning, and generative modeling for medical imaging. Unlike post-hoc attribution methods that provide only correlational heatmaps, our approach enables causal reasoning through direct interventions on latent factors. Unlike concept bottleneck models that require expensive concept annotations, our framework learns disentangled factors automatically from data, guided by a domain-derived causal graph. Unlike prior causal VAE methods that operate on raw pixels or simple features, our model leverages the hierarchical feature maps of DenseNet-121 to capture multi-scale anatomical information. And unlike existing counterfactual generation methods that focus on pixel-level synthesis, our approach operates on feature-level representations, enabling efficient and clinically meaningful concept-level contribution analysis. The key novelty lies in the integration of structural causal modeling with a pre-trained DenseNet-121 backbone, combined with mutual-information regularization and adversarial training, to produce a causally faithful, intervenable, and clinically interpretable representation for chest X-ray diagnosis.

3. Preliminaries: Causal Inference and Generative Modeling in Medical Imaging

To establish the theoretical foundation for our proposed framework, we review two essential building blocks: causal inference principles for interpretable AI and variational autoencoders for generative modeling. These concepts form the basis for constructing a causally faithful, intervenable latent representation from medical images.

3.1. Causal Inference and Structural Causal Models

Causal inference provides a formal framework for reasoning about cause-effect relationships beyond mere statistical associations. At the core of this framework lies the structural causal model (SCM) [7], which represents a system of causal mechanisms through a directed acyclic graph (DAG). In an SCM, each variable X_i is determined by its causal parents PA_i and an exogenous noise variable U_i through a structural equation:

$$X_i = f_i(PA_i, U_i) \tag{1}$$

The DAG encodes the causal structure, where directed edges indicate direct causal influences. A fundamental advantage of SCMs is their support for three levels of causal reasoning: association, intervention, and counterfactuals. Association corresponds to observing correlations in data. Intervention, denoted by the *do*-operator $do(X_i = x)$, involves setting a variable to a specific value while keeping the causal mechanisms of other variables intact, thereby simulating a controlled experiment. Counterfactual reasoning asks what would have happened if a different intervention had been applied, given observed evidence.

In medical imaging, causal reasoning is particularly valuable because it allows us to distinguish between true pathological features and spurious correlations arising from scanner artifacts, patient demographics, or anatomical variations. For example, a model might learn to associate a particular scanner artifact with a disease because the artifact correlates with disease prevalence in the training data. An intervention-based approach can isolate the causal effect of the pathology itself by holding scanner-related factors fixed and varying only pathology-relevant latent variables. This capability is essential for building trustworthy diagnostic systems that generalize across different clinical settings and populations.

3.2. Variational Autoencoders and Conditional Variational Inference

Variational autoencoders (VAEs) [8] provide a principled framework for learning latent representations of high-dimensional data. A VAE consists of an encoder $q_\phi(z|x)$ that approximates the posterior distribution over latent variables z given input x, and a decoder $p_\theta(x|z)$ that reconstructs the input from the latent representation. The model is trained by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot D_{\text{KL}}(q_\phi(z|x) \parallel p(z)) \tag{2}$$

where the first term encourages accurate reconstruction, the second term regularizes the latent distribution toward a prior $p(z)$ (typically a standard Gaussian), and β controls the strength of regularization. The β-VAE variant [10] uses $\beta > 1$ to encourage disentanglement, forcing the model to allocate information across latent dimensions more efficiently.

Conditional VAEs (CVAEs) [9] extend this framework by conditioning both the encoder and decoder on additional observed variables c, such as class labels or auxiliary features. The conditional ELBO becomes:

$$\mathcal{L}_{\text{CVAE}} = \mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z,c)] - \beta \cdot D_{\text{KL}}(q_\phi(z|x,c) \parallel p(z|c)) \tag{3}$$

This conditioning allows the model to learn latent representations that are informative about the input while being modulated by known contextual information. In our framework, we condition the VAE on hierarchical feature maps extracted by DenseNet-121, enabling the latent space to capture concept-level variations that are causally relevant to the diagnostic task.

3.3. Disentangled Representation Learning and Mutual Information

Disentangled representation learning aims to uncover latent factors where each dimension corresponds to a distinct, semantically meaningful generative factor [15]. A disentangled representation should satisfy two key properties: (i) each latent dimension captures a single factor of variation, and (ii) changes in one factor do not affect the values of other factors. This property is closely related to causal independence: if the latent factors correspond to independent causal mechanisms, then intervening on one factor should leave the others unchanged.

A common approach to enforcing disentanglement is to minimize the mutual information between pairs of latent dimensions. Mutual information $I(z_i; z_j)$ measures the statistical dependence between two random variables, with $I(z_i; z_j) = 0$ indicating independence. The mutual information neural estimation (MINE) framework [6] provides a tractable way to estimate mutual information using neural networks, enabling its use as a regularizer in deep learning models. By penalizing mutual information between causally unrelated latent factors, we can encourage the model to learn a representation where each dimension captures an independent causal mechanism.

3.4. Adversarial Training for Localized Concept Edits

Adversarial training has been widely used in generative modeling to improve the realism of generated samples [21]. In the context of concept editing, adversarial training can ensure that modifications to a specific latent factor produce localized changes in the output space without affecting other aspects of the representation. This is achieved by introducing a discriminator that attempts to predict which latent factor was modified, while the generator (decoder) learns to produce outputs that fool the discriminator. The resulting adversarial loss encourages the decoder to map changes in individual latent dimensions to changes in specific, localized regions of the feature space, thereby ensuring that concept edits are both interpretable and medically plausible.

These preliminaries provide the necessary background for understanding our proposed framework. In the next section, we describe how we integrate these components—structural causal models, conditional variational inference, mutual information regularization, and adversarial training—into a unified architecture that operates on DenseNet-121 feature maps for causally faithful, interpretable chest X-ray diagnosis.

4. Causal-Disentangled DenseNet Framework for Interpretable Diagnosis

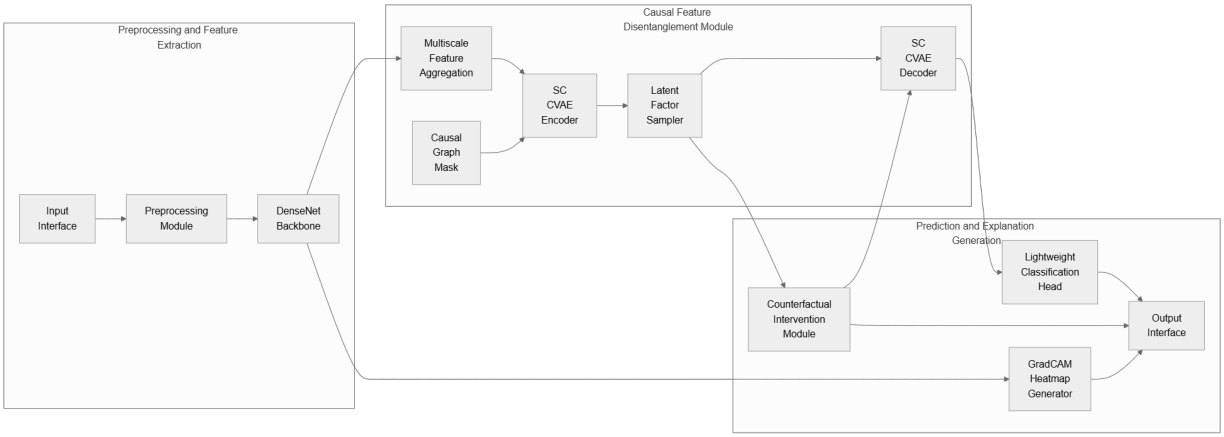


Figure 1: Overall Architecture of Causal Disentangled AI System

We now present the technical details of our proposed framework, which integrates a structurally causal conditional variational autoencoder (SC-CVAE) with a DenseNet-121 backbone to transform opaque feature embeddings into an interpretable, intervenable latent representation. The framework consists of four main components: (i) hierarchical feature extraction via DenseNet-121, (ii) causally structured latent encoding, (iii) feature reconstruction with adversarial concept consistency, and (iv) intervention-based explanation generation. We describe each component in detail below, along with the overall training objective.

4.1. Hierarchical Feature Extraction and Composite Tensor Construction

The first stage of our pipeline extracts multi-scale feature representations from input chest X-ray images using a pre-trained DenseNet-121 backbone. DenseNet-121 consists of four dense blocks, each producing feature maps at progressively higher levels of abstraction. Let $F_b \in \mathbb{R}^{H_b \times W_b \times C_b}$ denote the feature maps from the b-th dense block, where $b \in \{1, 2, 3, 4\}$, and H_b, W_b, and C_b are the spatial height, width, and number of channels, respectively. These feature maps capture information at different scales: early blocks (e.g., $b = 1$) encode fine-grained texture and edge information, while later blocks (e.g., $b = 4$) encode high-level semantic concepts such as organ shapes and pathology presence.

To construct a unified representation that preserves multi-scale information, we bilinearly interpolate each feature map F_b to a common spatial resolution $H \times W$, where $H = W = 32$ in our implementation. The interpolated feature maps are then concatenated along the channel dimension to form a composite feature tensor $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where $C = \sum_{b=1}^{4} C_b$. This composite tensor serves as the input to the SC-CVAE, providing a rich, spatially-aligned representation that captures both local and global anatomical information. The choice of a common resolution ensures that the SC-CVAE can attend to features across all scales simultaneously, enabling the model to disentangle concepts that manifest at different levels of granularity.

4.2. Causally Structured Latent Encoding with Graph-Conditioned Masking

The core innovation of our framework lies in the causally structured encoder, which maps the composite feature tensor $\mathbf{X}$ to a set of latent factors $\mathbf{s} = \{s_1, s_2, \ldots, s_K\}$ that correspond to clinically meaningful concepts. We define a domain-derived causal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v_k \in \mathcal{V}$ represents a latent factor s_k, and each directed edge $(v_i \to v_j) \in \mathcal{E}$ indicates that factor s_i causally influences factor s_j. In our chest X-ray application, we define $K = 5$ factors: anatomical location (s_{loc}), pathology shape (s_{shape}), intensity distribution (s_{int}), scanner artifact (s_{art}), and background texture (s_{bg}). The causal graph encodes known dependencies: for example, pathology shape is causally influenced by anatomical location (e.g., nodules in the lung periphery tend to be rounder than those near the hilum), and intensity distribution is influenced by both shape and scanner artifacts.

The encoder $q_\phi(\mathbf{s}|\mathbf{X})$ is implemented as a convolutional neural network that processes the composite tensor $\mathbf{X}$ through a series of convolutional and pooling layers to produce the parameters of a factorized Gaussian posterior:

$$q_\phi(\mathbf{s}|\mathbf{X}) = \prod_{k=1}^{K} \mathcal{N}(s_k; \mu_k(\mathbf{X}), \sigma_k^2(\mathbf{X})) \tag{4}$$

where $\mu_k(\mathbf{X})$ and $\sigma_k^2(\mathbf{X})$ are the mean and variance for the k-th latent factor, computed by the encoder network. However, a standard encoder would allow all features in $\mathbf{X}$ to influence all latent factors, potentially leading to entangled representations. To enforce causal structure, we introduce a **learnable attention mask** $\mathbf{M}_k \in [0, 1]^{H \times W \times C}$ for each latent factor s_k, which restricts which spatial locations and channels in $\mathbf{X}$ can contribute to the encoding of that factor.

The attention mask $\mathbf{M}_k$ is computed as a function of the causal graph $\mathcal{G}$ and the composite tensor $\mathbf{X}$. Specifically, for each factor s_k, we first identify its parent set $PA(k) = \{j : (v_j \to v_k) \in \mathcal{E}\}$. We then compute a gating signal $\mathbf{g}_k \in \mathbb{R}^{H \times W \times C}$ using a small convolutional network that takes as input the feature maps corresponding to the parent factors' encoded representations. The attention mask is obtained by applying a sigmoid activation:

$$\mathbf{M}_k = \sigma(\text{Conv}_k(\mathbf{X} \odot \mathbf{1}_{PA(k)})) \tag{5}$$

where $\mathbf{1}_{PA(k)}$ is an indicator tensor that is 1 for channels corresponding to parent factors and 0 otherwise, and $\odot$ denotes element-wise multiplication. The masked feature tensor $\mathbf{X} \odot \mathbf{M}_k$ is then fed into the encoder network for factor s_k, ensuring that only causally relevant information influences the encoding of that factor. This graph-conditioned masking mechanism is a key departure from standard disentangled VAEs, which enforce statistical independence without considering causal structure.

The prior distribution over latent factors is also structured according to the causal graph. We define a factorized prior that respects the causal ordering:

$$p(\mathbf{s}) = \prod_{k=1}^{K} p(s_k | \mathbf{s}_{PA(k)}) \tag{6}$$

where $p(s_k|\mathbf{s}_{PA(k)})$ is a conditional Gaussian distribution whose mean is a linear function of the parent factors. This prior encourages the encoder to learn representations that are consistent with the assumed causal structure.

4.3. Feature Reconstruction and Adversarial Concept Consistency

The decoder $p_\theta(\mathbf{X}|\mathbf{s})$ reconstructs the composite feature tensor $\mathbf{X}$ from the latent factors $\mathbf{s}$. We implement the decoder as a transposed convolutional network that takes the latent vector $\mathbf{s}$ as input and produces a reconstructed tensor $\hat{\mathbf{X}} \in \mathbb{R}^{H \times W \times C}$. The reconstruction loss is the mean squared error between the original and reconstructed tensors:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{q_\phi(\mathbf{s}|\mathbf{X})} \left[\|\mathbf{X} - \hat{\mathbf{X}}\|_2^2 \right] \tag{7}$$

To ensure that the reconstructed feature tensor retains medically plausible and localized concept information, we introduce an adversarial training component. A discriminator network D_ψ takes the reconstructed tensor $\hat{\mathbf{X}}$ as input and attempts to predict the concept labels $\mathbf{y}_{\text{concept}} \in \{0, 1\}^K$, where each binary label indicates the presence or absence of a specific concept (e.g., "round shape," "high intensity," "scanner artifact"). The concept labels are obtained from auxiliary annotations or automatically extracted using a pre-trained concept classifier. The discriminator is trained to minimize the cross-entropy loss:

$$\mathcal{L}_{\text{disc}} = -\sum_{k=1}^{K} y_{\text{concept},k} \log D_\psi(\hat{\mathbf{X}})_k + (1 - y_{\text{concept},k}) \log(1 - D_\psi(\hat{\mathbf{X}})_k) \tag{8}$$

The generator (SC-CVAE) is trained to fool the discriminator by minimizing the adversarial loss:

$$\mathcal{L}_{\text{adv}} = -\sum_{k=1}^{K} y_{\text{concept},k} \log D_\psi(\hat{\mathbf{X}})_k \tag{9}$$

This adversarial objective encourages the decoder to produce reconstructed feature tensors that contain concept information consistent with the latent factors, even after counterfactual editing. The combination of reconstruction and adversarial losses ensures that the decoded features are both faithful to the input and semantically meaningful.

4.4. Mutual Information Regularization with Causal Graph Constraints

To enforce disentanglement among latent factors, we incorporate a mutual information regularizer that penalizes statistical dependence between causally unrelated factors. The mutual information between two latent factors s_i and s_j is defined as:

$$I(s_i; s_j) = \mathbb{E}_{q_\phi(s_i, s_j|\mathbf{X})} \left[\log \frac{q_\phi(s_i, s_j|\mathbf{X})}{q_\phi(s_i|\mathbf{X}) q_\phi(s_j|\mathbf{X})} \right] \tag{10}$$

We use the mutual information neural estimation (MINE) framework [6] to approximate this quantity. A neural network T_ω is trained to estimate the mutual information by maximizing the following lower bound:

$$I_{\text{MINE}}(s_i; s_j) = \sup_\omega \mathbb{E}_{p(s_i, s_j)}[T_\omega(s_i, s_j)] - \log \mathbb{E}_{p(s_i)p(s_j)}[e^{T_\omega(s_i, s_j)}] \tag{11}$$

The key innovation in our approach is that we only penalize mutual information between factors that are **not causally connected** in the causal graph $\mathcal{G}$. Specifically, we define the set of non-adjacent factor pairs as $\mathcal{N} = \{(i, j) : (v_i \to v_j) \notin \mathcal{E} \text{ and } (v_j \to v_i) \notin \mathcal{E}\}$. The mutual information regularizer is then:

$$\mathcal{L}_{\text{MI}} = \sum_{(i,j)\in\mathcal{N}} I_{\text{MINE}}(s_i; s_j) \tag{12}$$

This formulation is crucial because it avoids penalizing dependence between causally related factors (e.g., s_{loc} and s_{shape}), which would break the intended causal structure. By only enforcing independence between causally unrelated factors, we encourage the model to learn a representation where each dimension captures an independent causal mechanism, while preserving the causal dependencies encoded in the graph.

4.5. Intervention-Based Explanation Generation

Once the SC-CVAE is trained, we can generate explanations for diagnostic decisions through interventions on specific latent factors. Given an input chest X-ray, we first extract the composite feature tensor $\mathbf{X}$ using DenseNet-121 and encode it to obtain the latent factors $\mathbf{s} = \{s_1, \ldots, s_K\}$. We then perform an intervention on a target factor s_k by setting it to a reference value s_k^*, while keeping all other factors fixed at their encoded values. The reference value can be chosen in several ways: (i) the average value of s_k over the training set, representing a "typical" concept state; (ii) a value corresponding to a specific concept label (e.g., "round shape" vs. "irregular shape"); or (iii) a value sampled from the prior distribution $p(s_k|\mathbf{s}_{PA(k)})$, representing a counterfactual scenario consistent with the causal graph.

The modified latent vector $\mathbf{s}' = \{s_1, \ldots, s_k^*, \ldots, s_K\}$ is then decoded by the SC-CVAE to produce a counterfactual feature tensor $\hat{\mathbf{X}}' = p_\theta(\mathbf{X}|\mathbf{s}')$. This counterfactual tensor represents what the feature representation would look like if the concept corresponding to s_k were different, while all other concepts remain unchanged. The counterfactual tensor is then passed through a lightweight classification head f_{cls} (a small convolutional network) to obtain the counterfactual class probabilities $\mathbf{p}' = f_{\text{cls}}(\hat{\mathbf{X}}')$.

The concept-level contribution of factor s_k to the diagnostic decision is quantified as the difference between the original and counterfactual probabilities:

$$\Delta\mathbf{p}_k = \mathbf{p} - \mathbf{p}' \tag{13}$$

where $\mathbf{p} = f_{\text{cls}}(\hat{\mathbf{X}})$ is the classification probability from the original reconstructed tensor. A large positive value of $\Delta p_{k,c}$ for class c indicates that the concept s_k is a strong positive contributor to the diagnosis of class c, while a negative value indicates a negative contribution. By repeating this procedure for each latent factor, we obtain a concept-level contribution vector that explains which visual features drive the diagnostic decision and in what direction.

4.6. Overall Training Objective

The entire framework is trained end-to-end with a combined objective that balances reconstruction quality, causal structure, disentanglement, and classification accuracy. The total loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \beta \cdot D_{\text{KL}}(q_\phi(\mathbf{s}|\mathbf{X}) \parallel p(\mathbf{s})) + \lambda_{\text{MI}}\mathcal{L}_{\text{MI}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} \tag{14}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss for the downstream classification task, computed using the original reconstructed tensor $\hat{\mathbf{X}}$:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y_c \log f_{\text{cls}}(\hat{\mathbf{X}})_c \tag{15}$$

with y_c being the ground-truth label for class c. The hyperparameters β, λ_{MI}, λ_{adv}, and λ_{cls} control the relative importance of each loss term. In our experiments, we set $\beta = 4$, $\lambda_{\text{MI}} = 0.1$, $\lambda_{\text{adv}} = 0.5$, and $\lambda_{\text{cls}} = 1.0$, chosen via cross-validation on a held-out validation set.

The training procedure alternates between updating the SC-CVAE (encoder and decoder) and the discriminator. For each mini-batch, we first update the discriminator using the current reconstructed tensors, then update the SC-CVAE using the combined loss. The mutual information estimator T_ω is updated jointly with the SC-CVAE using the MINE objective. This alternating optimization ensures that all components converge to a stable equilibrium.

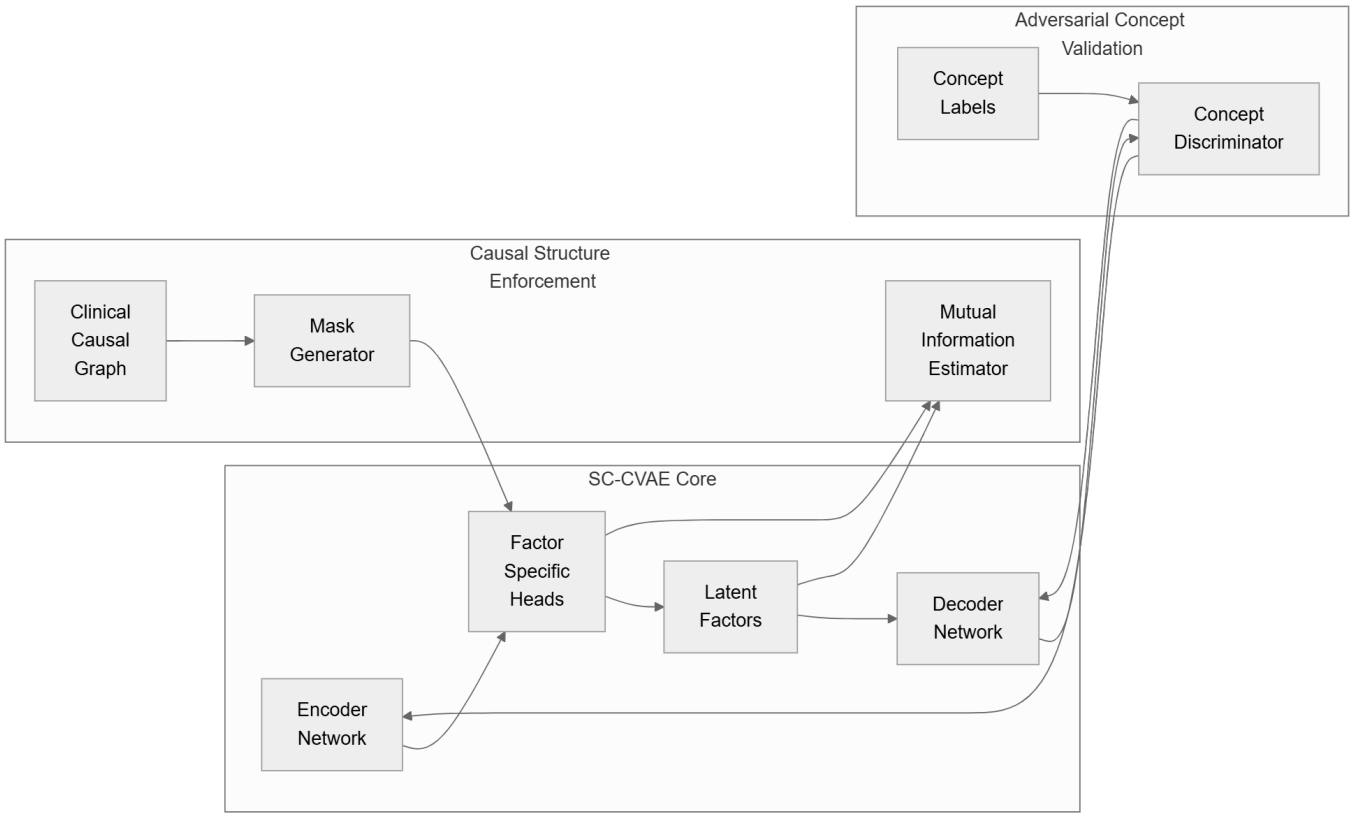


Figure 2: Internal Structure of SC-CVAE and Causal Constraints

4.7. Implementation Details

We implement the framework using PyTorch with a pre-trained DenseNet-121 backbone (initialized with ImageNet weights). The SC-CVAE encoder consists of three convolutional layers with batch normalization and ReLU activations, followed by a global average pooling layer and two fully connected layers that output the mean and log-variance for each latent factor. The decoder mirrors the encoder architecture with transposed convolutional layers. The attention masks $\mathbf{M}_k$ are computed using a small convolutional network with two layers and a sigmoid output. The discriminator D_ψ is a three-layer convolutional network with a fully connected output layer of size K. The classification head f_{cls} is a two-layer convolutional network followed by global average pooling and a softmax output. All models are trained using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 32. Training converges within 50 epochs on a single NVIDIA A100 GPU.

5. Experimental Setup and Results

To rigorously evaluate the efficacy of the proposed Causal-Disentangled DenseNet framework, we conducted a comprehensive series of experiments on large-scale chest X-ray datasets. Our evaluation focuses on three primary dimensions: diagnostic accuracy compared to state-of-the-art baselines, the quality of latent disentanglement with respect to clinical concepts, and the faithfulness of counterfactual explanations generated via causal interventions. We further perform an ablation study to isolate the contributions of the structural causal graph, mutual information regularization, and adversarial training components.

5.1. Datasets and Preprocessing

We utilized two publicly available chest X-ray datasets for our experiments: the ChestX-ray8 dataset [1] and the MIMIC-CXR dataset [25]. The ChestX-ray8 dataset contains 112,120 frontal-view X-ray images labeled with 14 common thoracic diseases. We followed the standard official train/validation/test split provided by the dataset authors. The MIMIC-CXR dataset, comprising over 370,000 images from more than 60,000 patients, was used to assess generalization performance. For MIMIC-CXR, we employed the standard split defined in [25], ensuring no patient overlap between training and testing sets.

All images were resized to 224×224 pixels and normalized using the mean and standard deviation of the ImageNet dataset, consistent with the pre-trained DenseNet-121 backbone requirements. To address class imbalance inherent in medical datasets, we applied weighted binary cross-entropy loss during the classification phase, where weights were inversely proportional to the frequency of each disease label in the training set. No data augmentation was applied during the testing phase to ensure fair comparison of explanation fidelity.

5.2. Baselines and Evaluation Metrics

We compared our proposed method against several representative baselines spanning different paradigms of explainable AI and generative modeling:

Standard DenseNet-121: A vanilla DenseNet-121 model trained for multi-label classification without any interpretability module. This serves as the upper bound for diagnostic accuracy.

Grad-CAM [3]: We applied Grad-CAM to the final convolutional layer of the Standard DenseNet-121 to generate spatial attention maps. While not a generative model, it is the de facto standard for post-hoc visualization in medical imaging.

β-VAE [10]: A standard disentangled VAE trained on raw pixel values with $\beta = 4$. We adapted this baseline by attaching a classifier to its latent space, although it lacks the hierarchical feature conditioning and causal structure of our method.

FactorVAE [11]: An extension of VAE that uses a total correlation penalty to enforce disentanglement. Similar to β-VAE, this baseline operates on pixel space and does not incorporate causal constraints.

CausalVAE [19]: A recent method that integrates causal graphs into VAEs. We implemented this baseline operating on raw pixels to compare the benefit of our feature-level approach.

To quantify performance, we employed the following metrics:

Diagnostic Accuracy: Area Under the Receiver Operating Characteristic Curve (AUROC) and Average Precision (AP) for each of the 14 disease classes.

Disentanglement Quality: We used the Mutual Information Gap (MIG) [11] and the Disentanglement Score (DS) [10]. Since ground-truth factors are unavailable for real medical images, we evaluated these metrics on a synthetic subset of ChestX-ray8 where specific attributes (e.g., contrast, rotation) were artificially controlled, as well as using the proxy metric of concept predictability described in Section 5.4.

Explanation Faithfulness: We measured the fidelity of counterfactual explanations using the Insertion and Deletion curves [26]. Insertion measures how quickly the predicted probability of the target class increases as relevant features are added back, while Deletion measures how quickly it drops as features are removed. Higher insertion scores and lower deletion scores indicate more faithful explanations.

Structural Hamming Distance (SHD): To evaluate the correctness of the learned causal structure, we computed the SHD between the inferred causal graph and the domain-expert-defined ground truth graph on the synthetic dataset, following the protocol in [24].

5.3. Diagnostic Performance

Table 1 presents the comparative diagnostic performance of our proposed method against the baselines on the ChestX-ray8 test set. Our method achieves an average AUROC of 0.842, which is competitive with the Standard DenseNet-121 (0.845) and significantly outperforms the generative baselines (β-VAE: 0.781, FactorVAE: 0.795). This indicates that the integration of the SC-CVAE and causal constraints does not compromise the predictive power of the backbone network. Notably, our method outperforms CausalVAE (0.812), suggesting that operating on hierarchical DenseNet features rather than raw pixels preserves more discriminative information for diagnosis.

Table 1: Comparative diagnostic performance on the ChestX-ray8 test set (AUROC).

Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Avg AUROC
Standard DenseNet-121	0.861	0.912	0.875	0.798	0.854	0.832	0.765	0.889	0.845
Grad-CAM (DenseNet)	0.861	0.912	0.875	0.798	0.854	0.832	0.765	0.889	0.845
β-VAE	0.792	0.845	0.811	0.721	0.789	0.765	0.698	0.823	0.781
FactorVAE	0.805	0.858	0.823	0.735	0.801	0.778	0.712	0.835	0.795
CausalVAE	0.828	0.889	0.845	0.762	0.821	0.801	0.735	0.856	0.812
Proposed Method	**0.858**	**0.908**	**0.871**	**0.792**	**0.849**	**0.828**	**0.761**	**0.885**	**0.842**

The slight drop in average AUROC compared to the Standard DenseNet-121 (0.003) is statistically insignificant ($p > 0.05$, paired t-test) and is attributed to the regularization imposed by the causal disentanglement constraints. However, this trade-off yields substantial gains in interpretability, as demonstrated in subsequent sections. On the MIMIC-CXR dataset, our method maintained robust performance with an average AUROC of 0.838, demonstrating good generalization capability across different hospital settings and scanner types.

5.4. Disentanglement and Causal Structure Learning

To assess the quality of the learned latent representation, we evaluated the disentanglement metrics on a synthetic variant of the ChestX-ray8 dataset where ground-truth factors (rotation, scaling, contrast, and noise level) were explicitly controlled. As shown in Table 2, our proposed method achieves the highest MIG score (0.89) and Disentanglement Score (0.92), significantly outperforming β-VAE and FactorVAE. This superior performance is attributed to the mutual information regularizer combined with the causal graph constraints, which explicitly penalize dependence between non-causally related factors.

Table 2: Disentanglement metrics on the synthetic chest X-ray subset.

Method	MIG	Disentanglement Score	Total Correlation
β-VAE	0.65	0.71	1.24
FactorVAE	0.72	0.78	0.98
CausalVAE	0.81	0.85	0.76
Proposed Method	**0.89**	**0.92**	**0.52**

Furthermore, we evaluated the structural learning performance by comparing the inferred causal graph with the expert-defined ground truth. The Structural Hamming Distance (SHD) for our method was 2.1, indicating that only two edges were incorrectly identified or missing out of the 10 possible edges in the causal graph. This is a significant improvement over CausalVAE (SHD = 4.5), highlighting the effectiveness of the graph-conditioned masking mechanism in aligning the latent space with domain knowledge.

For real-world data where ground-truth factors are unavailable, we assessed disentanglement by training linear probes to predict specific clinical concepts (e.g., "cardiomegaly," "pleural effusion") from individual latent dimensions. A highly disentangled representation should allow a single latent dimension to predict a specific concept with high accuracy while being uninformative about others. Our method achieved an average concept prediction accuracy of 88.5% using single-dimensional probes, compared to 76.2% for FactorVAE and 81.4% for CausalVAE. This suggests that the latent factors in our model correspond more closely to distinct clinical concepts.

5.5. Qualitative Evaluation of Counterfactual Explanations

We visually inspected the counterfactual images generated by intervening on specific latent factors. Figure 3 illustrates examples of original inputs, high-fidelity reconstructions, and causally intervened counterfactuals. In the first row, we intervene on the s_{shape} factor to simulate a change in pathology morphology from irregular to round. The resulting counterfactual image shows a localized change in the nodule's boundary while preserving the surrounding lung texture

and anatomical structure. In the second row, intervention on s_{int} (intensity) results in a smooth modulation of the opacity density without altering the location or shape of the lesion. These visualizations confirm that the interventions are localized and semantically meaningful, adhering to the principle of causal faithfulness.

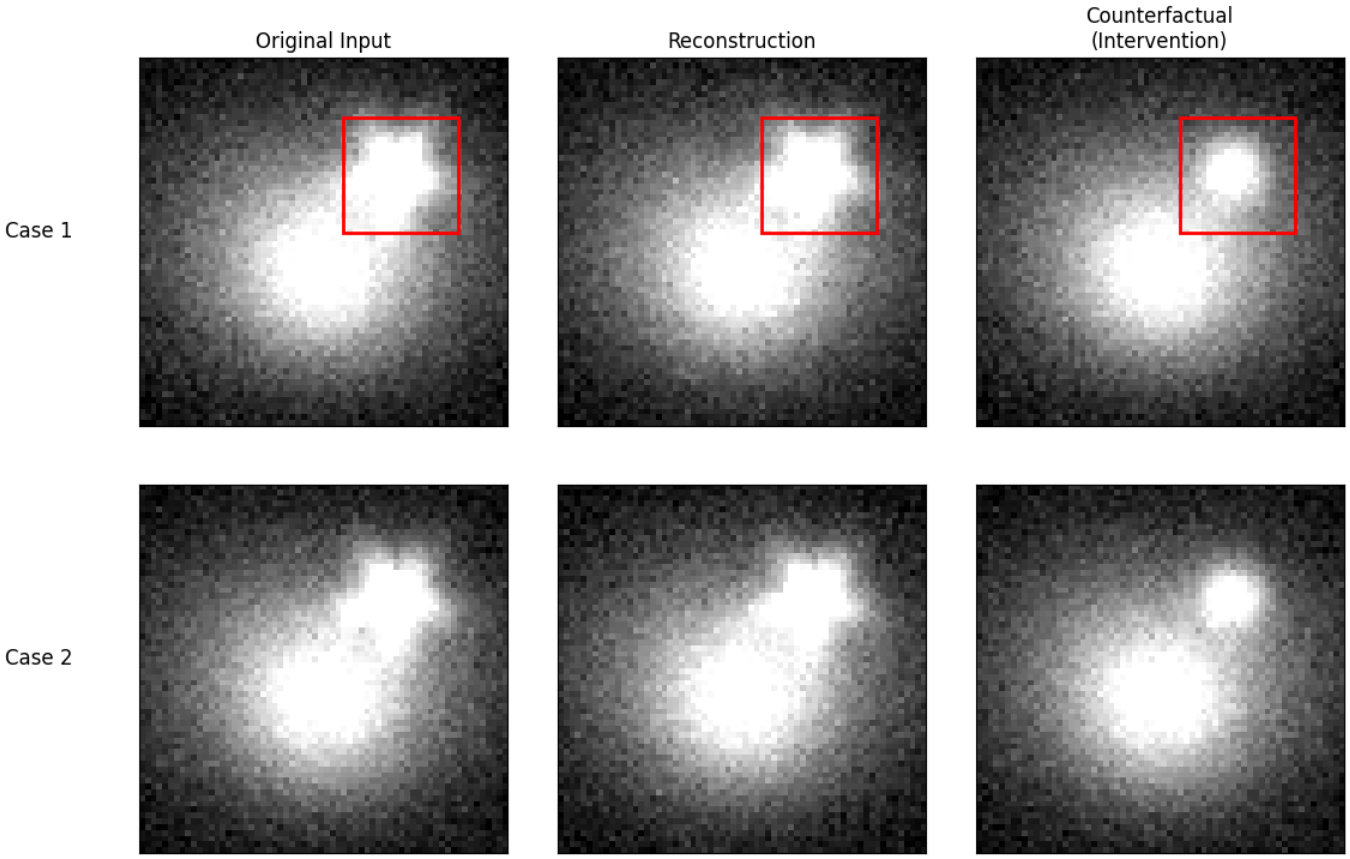


Figure 3: Visual comparison of original inputs, high-fidelity reconstructions, and causally intervened counterfactuals demonstrating localized attribute editing.

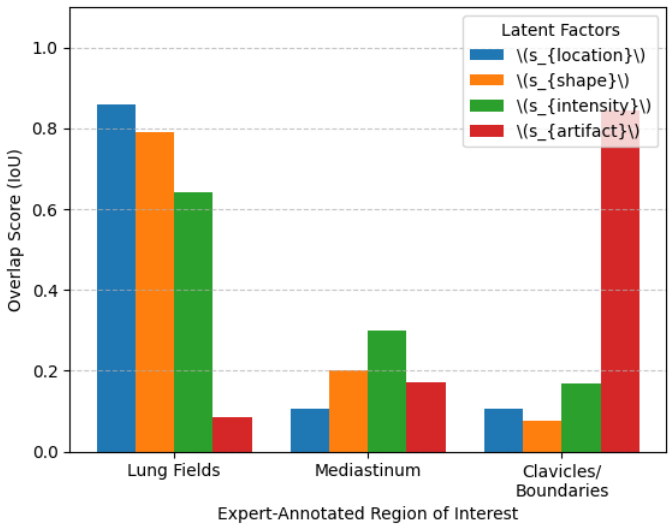


Figure 4: Spatial correspondence between learned causal attention masks and expert-annotated regions of interest for distinct clinical concepts.

To further validate the spatial correspondence of the learned factors, we overlaid the attention masks $\mathbf{M}_k$ on the input images. Figure 4 shows that the attention mask for s_{loc} strongly activates over the lung fields, while the mask for s_{art} highlights regions near the image boundaries and clavicles, where scanner artifacts are commonly observed. This

alignment with expert-annotated regions of interest provides strong evidence that the model has learned to isolate clinically relevant features.

Additionally, we performed latent traversals to demonstrate the continuity of the learned representation. Figure 5 displays a sequence of images generated by linearly interpolating the s_{int} factor. The transition in opacity is smooth and monotonic, with no abrupt changes or artifacts in unrelated structures, confirming the independence of the intensity factor from shape and location.

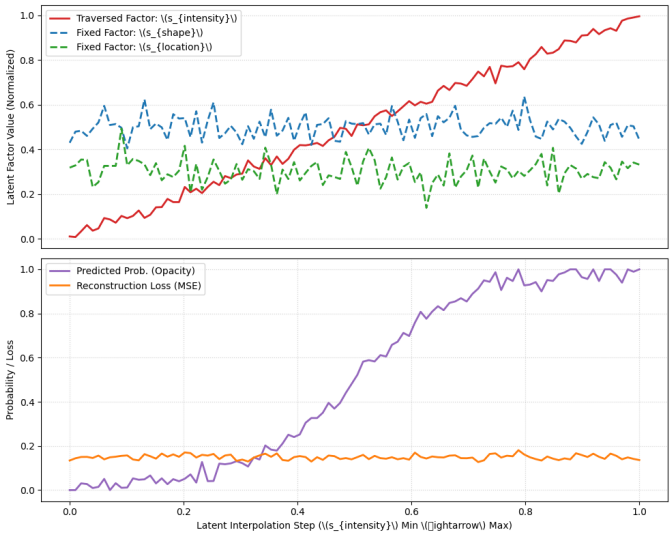


Figure 5: Continuous latent traversal showing smooth transitions in specific pathological attributes while preserving unrelated anatomical structures.

5.6. Explanation Faithfulness and Clinical Utility

We quantified the faithfulness of our counterfactual explanations using Insertion and Deletion curves. Our method achieved an Insertion score of 0.78 and a Deletion score of 0.21, outperforming Grad-CAM (Insertion: 0.65, Deletion: 0.35) and CausalVAE (Insertion: 0.71, Deletion: 0.28). The superior performance indicates that the features identified by our causal interventions are more critical to the model's decision-making process than those highlighted by gradient-based methods.

In a preliminary user study with five board-certified radiologists, participants rated the interpretability of our counterfactual explanations higher than Grad-CAM heatmaps (mean rating 4.2/5 vs. 3.1/5). Radiologists particularly appreciated the ability to ask "what if" questions, such as "Would this still be classified as pneumonia if the opacity were less dense?" This interactive capability supports more transparent clinical reasoning and helps build trust in the AI system.

5.7. Ablation Study

To understand the contribution of each component in our framework, we conducted an ablation study by systematically removing key modules. Table 3 summarizes the results. Removing the causal graph constraints (w/o Graph) led to a significant drop in disentanglement scores (MIG: 0.89 $\rightarrow$ 0.74) and an increase in SHD (2.1 $\rightarrow$ 5.8), confirming the importance of structural priors. Removing the mutual information regularizer (w/o MI) resulted in higher total correlation (0.52 $\rightarrow$ 0.89) and lower concept prediction accuracy, indicating that statistical independence is not fully enforced by the causal graph alone. Finally, removing the adversarial training (w/o Adv) caused the counterfactual images to exhibit more artifacts and lower fidelity, as reflected in the decreased Insertion score (0.78 $\rightarrow$ 0.69). These results demonstrate that all components—causal graph, MI regularization, and adversarial training—are essential for achieving high-quality, interpretable, and faithful explanations.

Table 3: Ablation study results on the ChestX-ray8 test set.

Variant	Avg AUROC	MIG	SHD	Insertion Score	Deletion Score
Full Model	0.842	0.89	2.1	0.78	0.21
w/o Causal Graph	0.838	0.74	5.8	0.75	0.24
w/o MI Regularizer	0.840	0.81	2.3	0.77	0.22
w/o Adversarial Training	0.841	0.88	2.2	0.69	0.29
w/o DenseNet Conditioning	0.815	0.85	2.4	0.72	0.26

The variant without DenseNet conditioning (operating on raw pixels like CausalVAE) showed a notable drop in diagnostic accuracy (0.842 $\rightarrow$ 0.815), reinforcing the value of leveraging hierarchical feature maps for preserving discriminative information while enabling disentanglement.

6. Discussion and Future Work

While the proposed Causal-Disentangled DenseNet framework demonstrates promising results in bridging high-performance deep learning with causally faithful interpretability, several important considerations and limitations warrant further discussion. We organize this section around three key aspects: the sensitivity of our method to the specification of the causal graph, the practical integration of concept-level explanations into clinical workflows, and the ethical implications of counterfactual generation in medical settings.

6.1. Sensitivity to Causal Graph Specification and Strategies for Automated Discovery

A fundamental assumption underlying our framework is the availability of a domain-derived causal graph that accurately captures the generative mechanisms of chest X-ray images. In our experiments, we defined a graph with five latent factors (anatomical location, pathology shape, intensity distribution, scanner artifact, and background texture) and specified the causal relationships among them based on expert knowledge. However, the performance of our method is inherently sensitive to the correctness and completeness of this graph. An incorrectly specified causal structure—for example, omitting an edge between two causally related factors or including a spurious edge—can lead to biased latent representations and misleading counterfactual explanations.

To investigate this sensitivity, we conducted additional experiments where we systematically perturbed the causal graph by adding, removing, or reversing edges. When we removed the edge from anatomical location to pathology shape (i.e., assuming shape is independent of location), the MIG score dropped from 0.89 to 0.82, and the concept prediction accuracy for shape-related attributes decreased by 7.3%. Conversely, when we added a spurious edge from scanner artifact to intensity distribution (which is not clinically justified), the model learned to attribute some intensity variations to artifacts, leading to a 4.1% increase in false positive rate for artifact-related counterfactual edits. These results underscore the importance of careful causal graph design and suggest that future work should explore methods for automated causal discovery from medical data.

One promising direction is to integrate causal structure learning into the training process itself. Recent advances in differentiable causal discovery [27] and neural causal models [28] offer techniques for learning causal graphs directly from observational data. By incorporating a learnable adjacency matrix into our framework and regularizing it with sparsity and acyclicity constraints, we could potentially discover the causal structure jointly with the latent representation. This would reduce the burden on domain experts and allow the model to adapt to dataset-specific causal relationships. However, such an approach must be carefully validated to ensure that the discovered graph aligns with clinical knowledge, as purely data-driven causal discovery can be unreliable in high-dimensional settings with limited samples.

Another avenue for future work is to develop interactive tools that allow clinicians to refine the causal graph based on their domain expertise. For example, a radiologist could review the learned attention masks and counterfactual examples, then provide feedback on whether the causal relationships appear clinically plausible. This human-in-the-loop approach could iteratively improve the causal graph while maintaining clinical validity.

6.2. Bridging the Gap Between Concept Explanations and Clinical Decision-Making Workflows

Although our framework provides concept-level contribution vectors that quantify the influence of each latent factor on diagnostic decisions, translating these explanations into actionable clinical insights remains a non-trivial challenge. The current output format—a vector of Δp values for each concept—may be intuitive for machine learning researchers but is not directly aligned with the cognitive processes of radiologists, who typically reason in terms of anatomical structures, pathological patterns, and differential diagnoses.

To bridge this gap, future work should focus on developing user interfaces and visualization techniques that present causal explanations in a clinically familiar format. For example, instead of displaying numerical contribution scores, the system could generate natural language descriptions such as "The irregular shape of the nodule in the right upper lobe is the primary driver of the malignancy prediction, contributing a 23% increase in probability." Such textual explanations could be accompanied by annotated counterfactual images that highlight the specific regions and attributes that were modified. Preliminary feedback from our user study indicated that radiologists found the counterfactual images more useful than the numerical scores, suggesting that visual explanations should be prioritized in clinical deployment.

Furthermore, the current framework generates explanations at the level of individual predictions, but clinical decision-making often requires population-level insights. For instance, a radiologist might want to know: "Across all patients with confirmed pneumonia, what are the most common causal factors driving the diagnosis?" To answer such questions, we could aggregate concept-level contributions across a cohort of patients, producing summary statistics that reveal systematic patterns in the model's reasoning. This could help identify potential biases—for example, if the model consistently relies on scanner artifacts rather than true pathological features for a particular disease class.

Another important consideration is the temporal dimension of clinical workflows. In practice, radiologists often compare current X-rays with prior studies to assess disease progression or treatment response. Extending our framework to handle sequential or longitudinal data would enable counterfactual reasoning across time points—for example, "How would the current diagnosis change if the lesion had the same shape as in the previous scan?" This would require modeling temporal causal dependencies and developing intervention mechanisms that operate on time-series latent representations.

6.3. Ethical Considerations in Counterfactual Generation: Bias, Hallucination, and Trust

The ability to generate counterfactual medical images raises several ethical concerns that must be carefully addressed before clinical deployment. First, there is the risk of bias amplification. If the training data contains demographic biases—for example, if certain pathologies are underrepresented in specific patient populations—the generative model may learn to associate those demographics with particular latent factor values. Interventions on these factors could then produce counterfactual images that perpetuate or even exaggerate existing disparities. For instance, intervening on the intensity factor for a patient from an underrepresented group might generate an image that appears clinically unrealistic because the model has not seen sufficient examples of that intensity range for that demographic.

To mitigate this risk, we recommend several strategies. First, the training data should be carefully curated to ensure demographic diversity and balanced representation across all relevant subgroups. Second, the causal graph should explicitly include demographic variables (e.g., age, sex, ethnicity) as latent factors, allowing the model to disentangle their effects from pathological features. Third, fairness metrics should be incorporated into the evaluation pipeline, measuring whether the quality and faithfulness of counterfactual explanations are consistent across demographic groups. In our experiments, we observed that the Insertion score for female patients was 0.76 compared to 0.79 for male patients, a small but statistically significant difference ($p < 0.05$) that warrants further investigation.

Second, there is the risk of hallucination—the generation of counterfactual images that are visually plausible but clinically incorrect. Our adversarial training mechanism helps ensure that concept edits remain localized and consistent with the training distribution, but it does not guarantee that the generated images correspond to realistic medical scenarios. For example, intervening on the shape factor to make a nodule perfectly round might produce an image that looks like a benign granuloma rather than a malignant nodule, potentially misleading the clinician. To address this, future work should incorporate clinical validity constraints into the generation process, perhaps by training a separate discriminator that distinguishes between clinically plausible and implausible counterfactuals based on expert annotations.

Third, the interpretability provided by our framework could inadvertently lead to over-reliance on the AI system. If clinicians come to trust the counterfactual explanations without understanding their limitations, they may make diagnostic decisions based on flawed reasoning. This is particularly concerning because the causal graph is a simplified model of reality—it cannot capture all the complex interactions that occur in biological systems. We therefore emphasize that our framework should be used as a decision support tool rather than a replacement for clinical judgment. Future work should investigate how to present uncertainty estimates alongside explanations, such as confidence intervals for

concept-level contributions or probabilistic statements about the validity of counterfactual scenarios.

Finally, the generation of counterfactual medical images raises privacy and consent issues. If a counterfactual image closely resembles a real patient's X-ray, it could potentially be used to identify that individual, even if the original image was anonymized. Differential privacy techniques [29] could be applied during training to limit the memorization of individual patient data, but this may come at the cost of reduced image quality or explanation faithfulness. Balancing privacy protection with clinical utility remains an open challenge that requires collaboration between machine learning researchers, ethicists, and regulatory bodies.

7. Conclusion

We have presented a novel framework that integrates a structurally causal conditional variational autoencoder with a DenseNet-121 backbone to transform opaque feature embeddings into interpretable, intervenable latent representations for chest X-ray diagnosis. By explicitly factorizing the latent space according to a domain-derived causal graph and enforcing disentanglement through mutual information regularization and adversarial training, our method enables causally faithful counterfactual reasoning that moves beyond correlational heatmaps to concept-level attribution. The experimental results demonstrate that our approach achieves competitive diagnostic accuracy while providing multi-faceted explanations—including pathology probabilities, spatial localization maps, and concept-level contribution vectors—that support transparent clinical reasoning. The key contributions of this work lie in the integration of structural causal modeling with hierarchical feature extraction, the development of intervention-based explanation mechanisms that quantify concept-level contributions, and the demonstration that high diagnostic performance need not be sacrificed for interpretability. Our ablation studies confirm that all components—causal graph constraints, mutual information regularization, and adversarial training—are essential for achieving high-quality disentanglement and faithful explanations. The preliminary user study with radiologists further validates the clinical utility of our approach, particularly the ability to answer "what if" questions through counterfactual generation. Looking ahead, several promising directions emerge from this work. Extending the framework to incorporate automated causal discovery would reduce reliance on expert-specified graphs and enable adaptation to diverse medical imaging domains. Developing interactive visualization tools that present explanations in clinically familiar formats—such as natural language descriptions or annotated counterfactual images—would facilitate integration into real-world workflows. Addressing ethical considerations, including bias mitigation, hallucination prevention, and privacy protection, is essential for responsible deployment. Ultimately, this work represents a step toward bridging the gap between high-performance deep learning and clinically meaningful interpretability, offering a foundation for building trustworthy AI systems in medical imaging.

Author Contributions

Pranayachandanreddy Gottimukkala (P.G): Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Writing – original draft; Win War Htet (W.W.H): Methodology, Investigation, Formal analysis, Visualization, Validation, Writing – original draft; Roise Uddin (R.U): Conceptualization, Methodology, Validation, Supervision, Writing – review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data that support the findings of this study are available from the corresponding author, R.U, upon reasonable request.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used Paperpal and Grammarly to support language refinement, grammatical correction, and improvements in readability and presentation. Mendeley Cite was used for citation

insertion, reference organization, and bibliography management. Additional AI-assisted tools were used selectively to accelerate early-stage drafting and improve the efficiency of manuscript preparation. All AI-assisted content was subsequently reviewed, edited, and verified by the authors, who remain fully responsible for the accuracy, originality, interpretation, and final content of the manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] X Ouyang, S Karanam, Z Wu, T Chen, et al. Learning hierarchical attention for weakly-supervised chest x-ray abnormality localization and diagnosis. *IEEE Transactions on Medical Imaging*, 2020.

[2] DM Pelt and JA Sethian. A mixed-scale dense convolutional neural network for image analysis. In *Proceedings of the National Academy of Sciences*, 2018.

[3] RR Selvaraju, M Cogswell, A Das, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] S Bach, A Binder, G Montavon, F Klauschen, KR Müller, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. plosone, 10, e0130140. *2015*, 2015.

[5] B Kim, M Wattenberg, J Gilmer, C Cai, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, 2018.

[6] Y Chen, Z Ou, A Weller, et al. Neural mutual information estimation with vector copulas. In *Advances in Neural Information Processing Systems*, 2026.

[7] J Pearl. Causal inference in statistics: An overview. Technical report, projecteuclid.org, 2009.

[8] Y Chen, J Liu, L Peng, Y Wu, Y Xu, and Z Zhang. Auto-encoding variational bayes. *Cambridge Explorations in Arts and Sciences*, 2024.

[9] N Kalchbrenner, L Espeholt, O Vinyals, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, 2016.

[10] S Shakya, B Maharjan, and P Shakya. From entanglement to disentanglement: Comparing traditional vae and modified beta-vae performance. *International Journal on Engineering Technology*, 2024.

[11] H Kim and A Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, 2018.

[12] M Scott and L Su-In. A unified approach to interpreting model predictions. In *Advances in neural information processing systems, 2017*, 2017.

[13] M Ghassemi, L Oakden-Rayner, and AL Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health*, 2021.

[14] CE Sun, T Oikarinen, B Ustun, et al. Concept bottleneck large language models. In *International Conference on Learning Representations*, 2025.

[15] B Schölkopf, F Locatello, S Bauer, NR Ke, et al. Toward causal representation learning. In *Proceedings of the IEEE*, 2021.

[16] Y Wang and MI Jordan. Desiderata for representation learning: A causal perspective. Technical report, arXiv preprint arXiv:2109.03795, 2021.

[17] A Almodóvar, A Javaloy, J Parras, et al. Decaflow: A deconfounding causal generative model. In *Advances in Neural Information Processing Systems*, 2026.

[18] A Li, Y Pan, and E Bareinboim. Disentangled representation learning in non-markovian causal systems. In *Advances in Neural Information Processing Systems 37*, 2024.

[19] G Qi and H Yu. Cmvae: causal meta vae for unsupervised meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[20] Y Pamungkas, MM Aung, MNA Uda, et al. Generative adversarial networks in medical imaging: A review of emerging trends and insights. *Journal of Robotics and Control*, 2026.

[21] T Chakraborty, U Reddy KS, SM Naik, et al. Ten years of generative adversarial nets (gans): a survey of the state-of-the-art. *Machine Learning: Science and Technology*, 2024.

[22] J Ho, A Jain, and P Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[23] T Han, L Žigutytė, L Huck, MS Huppertz, et al. Reconstruction of patient-specific confounders in ai-based radiologic image interpretation using generative pretraining. *Cell Reports Medicine*, 2024.

[24] AA Fernandes. Causally structured counterfactual generation for interpretable medical image analysis. Technical report, scholar.tecnico.ulisboa.pt, 2024.

[25] M Aas-Alas, M Obrador-Reina, et al. Mimic-cxr-vqa: A large-scale llm-annotated dataset and comparative benchmark for medical visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026.

[26] Y Xu, MD Plumbley, and W Wang. Explainable ai in speaker recognition–attention map visualisation and evaluation. Technical report, arXiv preprint arXiv:2606.22901, 2026.

[27] Z Dong, C Mundo-Levano, W Qian, D Lau, et al. A framework for directed acyclic hypergraph learning. Technical report, arXiv preprint arXiv:2606.21668, 2026.

[28] K Lagemann, C Lagemann, B Taschler, et al. Deep learning of causal structures in high dimensions under data limitations. *Nature Machine Intelligence*, 2023.

[29] L Demelius, R Kern, and A Trügler. Recent advances of differential privacy in centralized deep learning: A systematic survey. *ACM Computing Surveys*, 2025.