Learning disentangled representations for explainable chest X-ray classification using Dirichlet VAEs

Rachael Harkness[1,2], Alejandro F Frangi[1,2], Kieran Zucker[2,3], and Nishant Ravikumar[1,2]

[1]CISTIB Centre for Computational Imaging and Simulation Technologies in Biomedicine,
School of Computing
[2]University of Leeds, Leeds, United Kingdom
[3]LIMR Leeds Institute of Medical Research

1. ABSTRACT

This study explores the use of the Dirichlet Variational Autoencoder (DirVAE) for learning disentangled latent representations of chest X-ray (CXR) images. Our working hypothesis is that distributional sparsity, as facilitated by the Dirichlet prior, will encourage disentangled feature learning for the complex task of multi-label classification of CXR images. The DirVAE is trained using CXR images from the CheXpert database, and the predictive capacity of multi-modal latent representations learned by DirVAE models is investigated through implementation of an auxiliary multi-label classification task, with a view to enforce separation of latent factors according to class-specific features. The predictive performance and explainability of the latent space learned using the DirVAE were quantitatively and qualitatively assessed, respectively, and compared with a standard Gaussian prior-VAE (GVAE). We introduce a new approach for explainable multi-label classification in which we conduct gradient-guided latent traversals for each class of interest. Study findings indicate that the DirVAE is able to disentangle latent factors into class-specific visual features, a property not afforded by the GVAE, and achieve a marginal increase in predictive performance relative to GVAE. We generate visual examples to show that our explainability method, when applied to the trained DirVAE, is able to highlight regions in CXR images that are clinically relevant to the class(es) of interest and additionally, can identify cases where classification relies on spurious feature correlations.

Keywords: Multi-label classification, chest X-ray images, variational autoencoders, Dirichlet distribution, disentanglement, explainability

2. INTRODUCTION

Medical image interpretation is a complex and challenging task requiring in-depth understanding of anatomy and physiology, and years of education and experience. In particular, the interpretation of chest X-rays (CXRs), a 2D projection of 3D thoracic organs/structures, can be especially challenging. The resulting 'overlapping' tissue features makes identification of object boundaries challenging due to insufficient tissue contrast, impeding detection of abnormalities. Additionally, pre-existing conditions and comorbidities are common in patients that require acute chest X-ray exams, or are frequently invited to screenings (e.g. chronic obstructive pulmonary disorder and lung cancer screenings). Consequently, in practice multiple pathologies are often observed in a single exam. Co-occurring disease features can appear similarly and can often create diagnostic uncertainty for both expert radiologists and deep learning systems. The learning of explainable representations of clinically complex medical images, where anatomical and extrinsic image acquisition-related features are disentangled from relevant disease features, is a crucial step in overcoming these challenges.

In this study, we propose to use variational autoencoders (VAEs) to learn disentangled multi-modal latent representations of CXRs for explainable classification of multiple co-occurring labels/class (i.e. multi-label classification).[1] VAEs are able to learn compressed representations of data, in which observed variations of salient

visual features are captured by a number of latent factors. Disentanglement is achieved when each factor corresponds uniquely to some meaningful composite part of the data. Disentangled representations offer a more explainable latent space, unlike their entangled counterparts. The matching of latent factors to interpretable variations in the data manifold, afforded by disentangled representations, makes them advantageous to downstream predictive tasks that utilise the learned latent space (e.g. classification, regression, clustering).[2] Popular approaches to learning disentangled representations with VAEs, includes, β-VAE,[3] Factor-VAE[4] and β-TCVAE.[5]

While these approaches improve disentanglement in Gaussian-prior VAEs, success is limited due to the dense and unimodal nature of a multivariate Gaussian prior. With this in mind, we propose to use a VAE with a Dirichlet prior, referred to as Dirichlet VAE (DirVAE). We hypothesise that a DirVAE will allow a multimodal latent representation of CXRs to be learned through multi-peak sampling and will encourage latent disentanglement due to the sparse nature of the Dirichlet prior.[6]

To evaluate the potential benefits of DirVAE over a conventional VAE with a Gaussian prior (GVAE), we use the CheXpert dataset to tackle a complex multi-label classification problem.[7] We convert the multi-label task to a series of binary classification problems. Logistic regression classifiers are trained jointly within a representation learning-centered framework, to predict each class of interest, with all binary classification tasks tackled simultaneously. Specifically, we jointly train the class-specific logistic regression classifiers with the DirVAE, to encourage disentanglement of class-specific latent factors, and evaluate model capacity for learning diagnostically separable representations of CXRs. Model evaluations investigate classifier performance, latent factorisation and its explainability. We train and evaluate the GVAE identically to the DirVAE for fair comparison, and highlight the benefits of the latter in terms of predictive performance and explainability.

3. METHODOLOGY

3.1 Data

The CheXpert data set, one of the largest publicly available CXR datasets with multi-label outcomes, is used throughout this study. To overcome issues of severe class imbalance in the data, and simplify training of the logistic regression classifiers, we sample 17,000 images of each of the four most represented classes featured in the CheXpert dataset, namely, 'No Finding', 'Lung Opacity', 'Pleural Effusion', and 'Support Devices'. Figure 1 presents the frequency of classes of interest as well as their frequency of co-occurrence. We split the CheXpert dataset into train ($n = 52,943$), validation ($n = 5057$) and test ($n = 10,000$) sets.
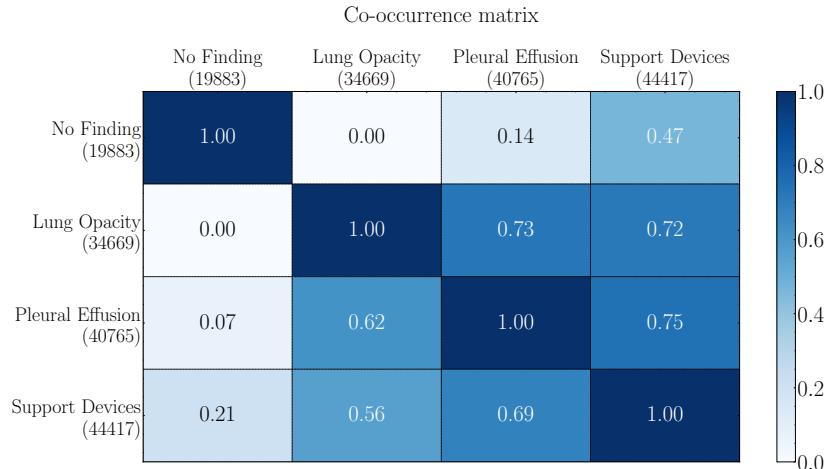

Co-occurrence matrix

	No Finding (19883)	Lung Opacity (34669)	Pleural Effusion (40765)	Support Devices (44417)
No Finding (19883)	1.00	0.00	0.14	0.47
Lung Opacity (34669)	0.00	1.00	0.73	0.72
Pleural Effusion (40765)	0.07	0.62	1.00	0.75
Support Devices (44417)	0.21	0.56	0.69	1.00

Figure 1: Label co-occurrence matrix of multi-label CheXpert data (training data)

3.2 Dirichlet VAE

The goal of this study is to explore the benefits of using the DirVAE to learn a multi-modal latent representation of CXR images, comprised of explainable latent factors. To accomplish this we model the Dirichlet prior explicitly

as in Gawlikowski, J. et al. (2022).[8] Alternative approaches frequently utilize a softmax-Laplace approximation.[9] However, such approaches to modelling the Dirichlet prior prevents true approximation of multi-modal representations as shown in Joo, W. et al. (2020).[10] The Dirichlet distribution is a continuous multivariate probability distribution defined over a set of discrete distributions. It is parameterised by a K-dimensional vector typically referred to as the concentration, where, K corresponds to the set of discrete distributions. Here, the set of discrete distributions represent the latent space in a DirVAE.

We define the Dirichlet probability density function as,

$$\mathrm{Dir}(x; \alpha) = \frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)} \prod x_k^{\alpha_k - 1}, \qquad \left(x_k \geq 0; \ \sum_k x_k = 1 \right) \tag{1}$$

where, $\alpha_k > 0$ and $\Gamma(\cdot)$ denotes the gamma function. The Dirichlet distribution can expressed with a composition of gamma random variables with the Gamma distribution defined as,

$$\mathrm{Gamma}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\beta x}, \qquad (\alpha > 0; \ \beta > 0) \tag{2}$$

where β is a constant, if $X_k \sim \mathrm{Gamma}(\alpha_k)$ are independently distributed, then $(X_1/X_0, \ldots, X_n/X_0)$ follows a Dirichlet distribution parameterised by a vector of $\alpha_1, \ldots, \alpha_n$, with $X_0 = \sum X_k$. Alternatively, we can say where K independent random variables are sampled from the multivariate Gamma distribution $X \sim \mathrm{MultiGamma}(\alpha, \beta \cdot \mathbf{1}_K)$, then we get $\mathbf{Y} \sim \mathrm{Dir}(\alpha)$, where $Y_k = X_k/X_0$.[10]

The Dirichlet concentration governs the shape of distribution on the probability simplex and is selected according to an acceptable trade-off between distributional sparsity and distributional smoothness. This can be observed in Figure 2. For enhancing explainability and to capture multi-modal latent distributions a Dirichlet concentration that enforces sparsity is essential. For these reasons we apply a Dirichlet prior parameterised with a concentration of $0.5 \cdot \mathbf{1}_{1024}$, where $\mathbf{1}_{1024}$ represents the K-dimensional latent space with $K = 1024$. This latent dimension was used throughout this study for both the DirVAE and the GVAE.
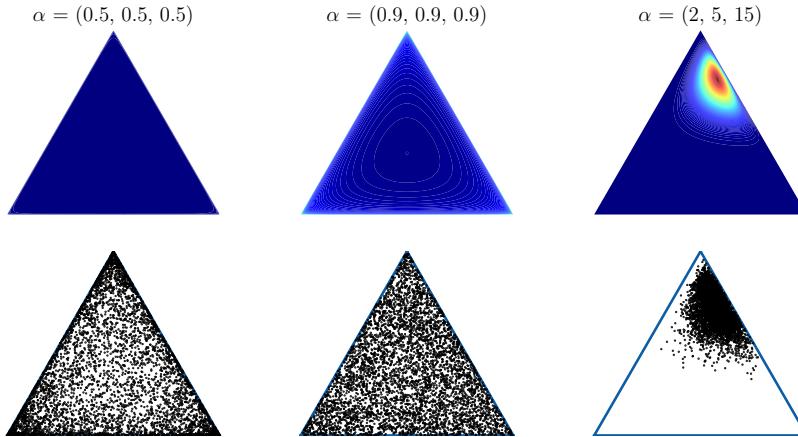

Figure 2: 3-dimensional probability density functions from the Dirichlet distribution, each with different α values. The values of α are set to (0.5, 0.5, 0.5), (0.9, 0.9, 0.9) and (2, 5, 15).

3.3 Experimental Settings

Both the DirVAE and GVAE were trained in four stages categorised as follows: (1) reconstruction, (2) reconstruction and regularisation, (3) classifier initialisation, and (4) joint training. During the first stage, the model is trained using just the reconstruction loss (L1 loss), defined as $L_1 = \sum_{i=1}^n |x_i - f(x_i)|$, where f denotes the encoder-decoder functions. In the second stage, the Kullback-Leibler (KL) divergence term is used to regularise model training by minimising the divergence between the approximated posterior distribution of the latent

factors and the assumed prior distribution (i.e. Dirichlet for DirVAE and centered multivariate Gaussian for GVAE). KL divergence is defined as $L_{KL} = D_{KL}(q_\phi(z|x) \parallel p_\theta(z|x))$, and for DirVAE is derived as in below,

$$KL(Q \parallel P) = \sum log\Gamma(\alpha_k) - \sum log\Gamma(\hat{\alpha}_k) + \sum (\hat{\alpha}_k - \alpha_k)\psi(\hat{\alpha}_k) \tag{3}$$

where P is the prior distribution, equal to MultiGamma($\alpha, \beta \cdot 1_k$), Q represents the learned posterior, $Q = $ MultiGamma($\hat{\alpha}, \beta \cdot 1_k$) and ψ is a digamma function.[10]

In the next stage, the weights of the pre-trained VAE are frozen and each of the four class-specific logistic regression classifiers are independently optimised for their respective binary classification tasks with binary cross entropy loss, using CXR latent representations as inputs. Once each classifier is fully optimised, the ensemble of logistic regression models are trained together, this objective takes the form,

$$L_{BCE} = \frac{1}{4}\sum_{c=1}^{4} -\frac{1}{n}\sum_{i=1}^{n}(y_{c,i} \cdot log(p_{c,i}) + 1 - y_{c,i} \cdot log(1 - p_{c,i})), \tag{4}$$

where y represents the class labels, p is the predicted probabilities and c represents the different classes.

In the final stage of training, the VAEs and logistic regression classifiers are trained jointly, where, the reconstruction loss, KL divergence loss and all four classifier losses are combined into a single training objective and minimised. Based on preliminary experiments, we found that training the multi-label GVAE and DirVAE models in this stage-wise manner helped stabilise the training process, relative to training the models end-to-end from the beginning. Optimising the latent space with the multi-label classification task encourages latent factors to explain class-specific visual features and allows the classifiers to predict target classes from the learned latent space. This enhances explainability of the learned latent space as latent factors with higher importance/discriminative power for predicting a given class correctly, should correspondingly encode visual features in the CXR image that are representative of that class.

Our working hypothesis was that this method of training combined with the properties of a DirVAE should yield a disentangled latent space resulting in improved explainability over the GVAE. Both the DirVAE and GVAE were optimized using the Adam algorithm. Logistic regression models were optimized under SGD.
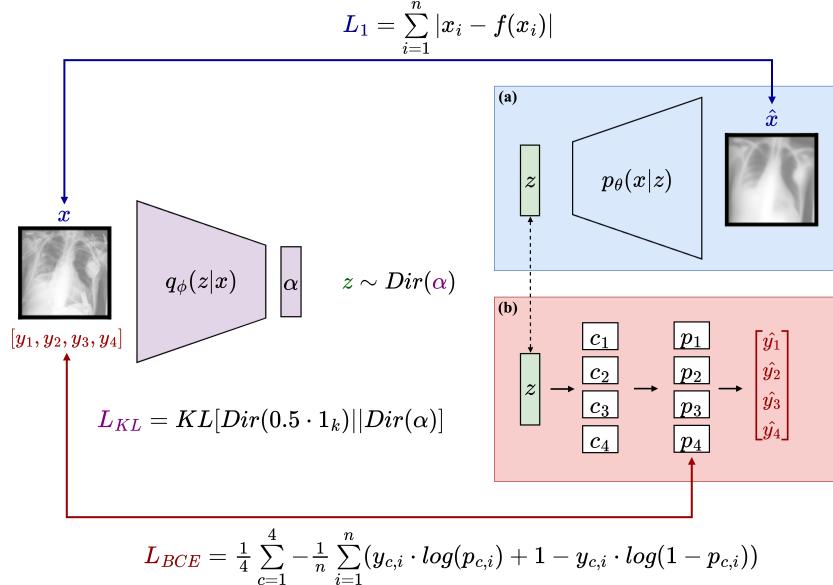

Figure 3: Dirichlet VAE encoder with a (a) decoder which generates image reconstructions from the latent space z and (b) an ensemble of logistic regression classifiers C, each predicting independent class probabilities p from latent space z. Classification thresholds are applied to the probabilities to produce class predictions $\hat{y}$.

3.4 Evaluations

The goal of this study is to evaluate the capacity of the DirVAE to learn explainable, disentangled latent representations of CXR images, for multi-label classification. Accordingly, we assess the performance of the DirVAE by evaluating the performance of their logistic regression classifiers on the multi-label classification problem, and qualitatively assessing the explainability of their respective learned latent spaces. Classifier performance is evaluated through ROC curves and standard classification metrics (evaluated per-class). All results are compared directly with GVAE performance. We perform a qualitative assessment of the explainability of the latent space through a series of 'latent factor traversals'. Once the GVAE and DirVAE models were trained, given a CXR image from the test set as input, we first identify a single latent factor with the largest gradient activation from a positive classifier (i.e. where the positive class of interest to a classifier was predicted correctly for the input image). The identified latent factor is adjusted incrementally or 'traversed', while all other factors are preserved. We visualise the influence of changing the single factor of interest on the decoder reconstructions. Intuitively, if changing only a single latent factor results in class-specific structural changes in the CXR reconstructions, we can assume that the latent space has been successfully disentangled into visual features relating to the class of interest. Pixel-wise variance is created to summarise changes across a traversal and identify features controlled by a specific latent factor. In these evaluations, cases with co-occurring labels are not considered, to simplify interpretation (as co-occurring labels would require exploring multiple relevant latent factors).

4. RESULTS

Table 1 summarises the classification performance of the logistic regression models, for both the DirVAE and the GVAE. The DirVAE classifiers marginally outperform GVAE classifiers with higher values across most classification metrics (see tab. 1) and higher AUC scores as shown in Figure 4. While individual label prediction metrics are good, multi-label metrics show that the DirVAE classifiers offer only modest performance in the prediction of all four labels, achieving an exact match rate (EMR) of 0.38 ± 0.01 and Hamming loss of 0.25 ± 0.02. GVAE achieved an EMR of only 0.37 ± 0.01 and Hamming loss of 0.26 ± 0.01.

		Dirichlet VAE	**Gaussian VAE**
Accuracy	*No Finding*	0.82 ± 0.01	0.82 ± 0.01
	Lung Opacity	0.72 ± 0.01	0.72 ± 0.01
	Pleural Effusion	0.70 ± 0.01	0.70 ± 0.01
	Support Devices	0.74 ± 0.02	0.72 ± 0.01
Precision	*No Finding*	0.61 ± 0.06	0.62 ± 0.03
	Lung Opacity	0.82 ± 0.06	0.82 ± 0.02
	Pleural Effusion	0.81 ± 0.01	0.82 ± 0.02
	Support Devices	0.87 ± 0.01	0.89 ± 0.01
Recall	*No Finding*	0.74 ± 0.01	0.72 ± 0.01
	Lung Opacity	0.69 ± 0.02	0.68 ± 0.02
	Pleural Effusion	0.72 ± 0.01	0.72 ± 0.01
	Support Devices	0.76 ± 0.02	0.74 ± 0.01
AUC	*No Finding*	0.87 ± 0.02	0.87 ± 0.01
	Lung Opacity	0.78 ± 0.02	0.77 ± 0.01
	Pleural Effusion	0.75 ± 0.01	0.74 ± 0.01
	Support Devices	0.78 ± 0.03	0.76 ± 0.02

Table 1: Label-wise table of logistic regression classification performance results
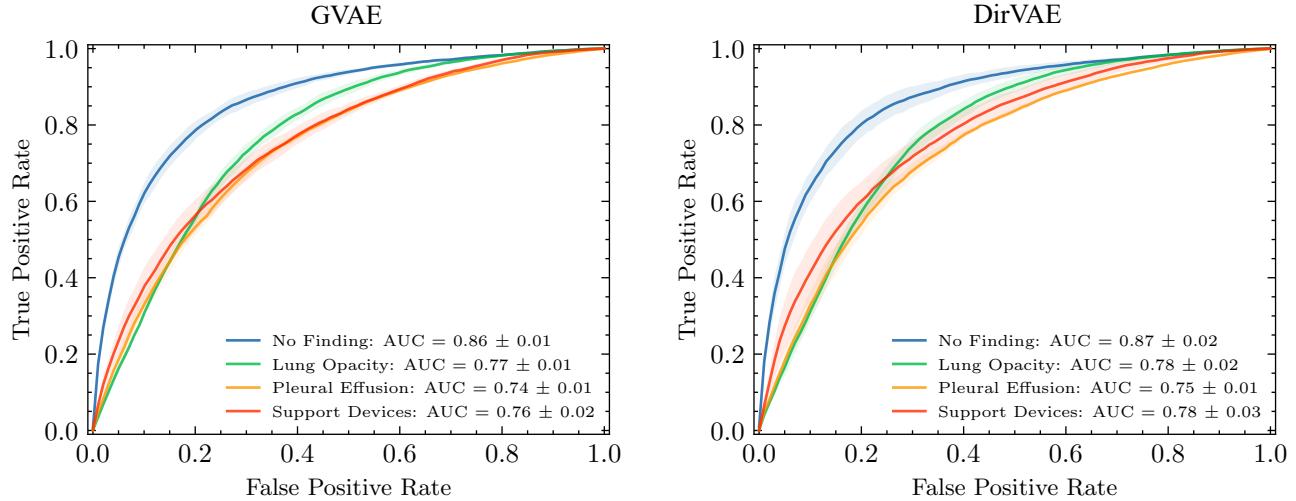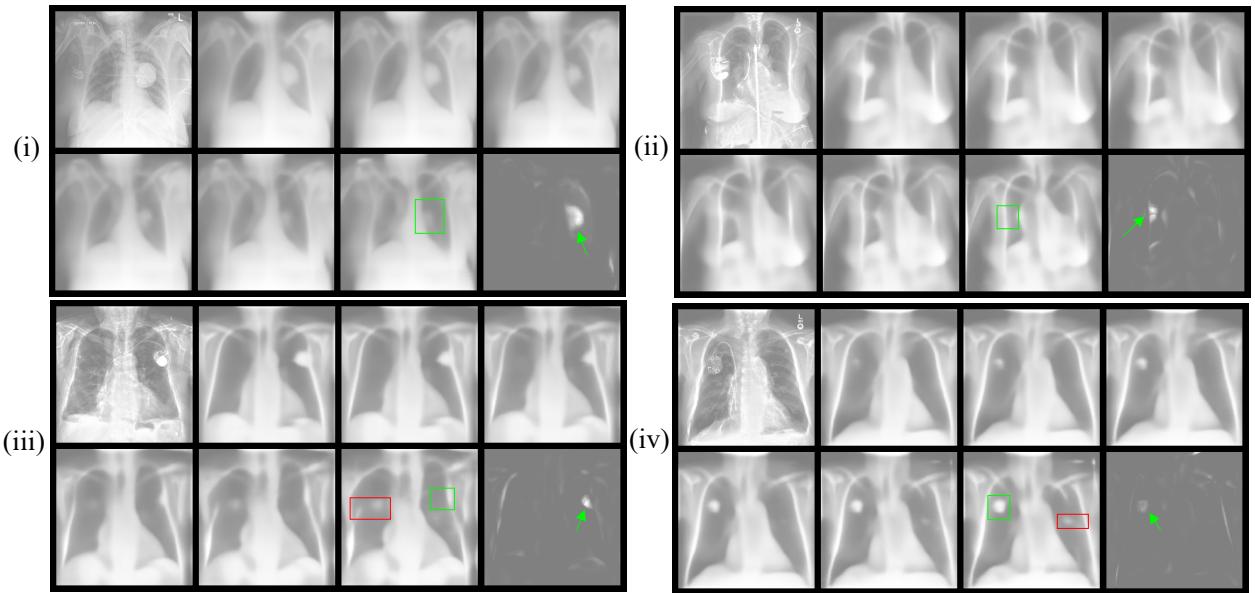
Figure 4: Label-wise ROC curves for GVAE and DirVAE logistic regression classifier performance.

Figure 5 presents the results of traversing the latent spaces in the direction of the most predictive latent dimension. For the DirVAE, clear structural changes are observed for each class. Typically, latent traversals will gradually remove or intensify disease-related features, in a manner specific to the input CXR image. We consider both types of change as evidence of disease-/class-specific and image-specific disentanglement and generate pixel intensity variance maps (presented along side each traversal, bottom right corner of each panel) in order to capture both types of feature change. We find that the features highlighted by DirVAE traversals are not only disease- and image-specific, but also clinically relevant. The feature changes observed during latent traversals for the 'Pleural Effusion' class are largely isolated to the lower regions of the lungs, typically affecting the appearance of the costophrenic angles and hemidiaphragm region, two keys areas of diagnostic relevance for pleural effusion (fig. 5c).[11]
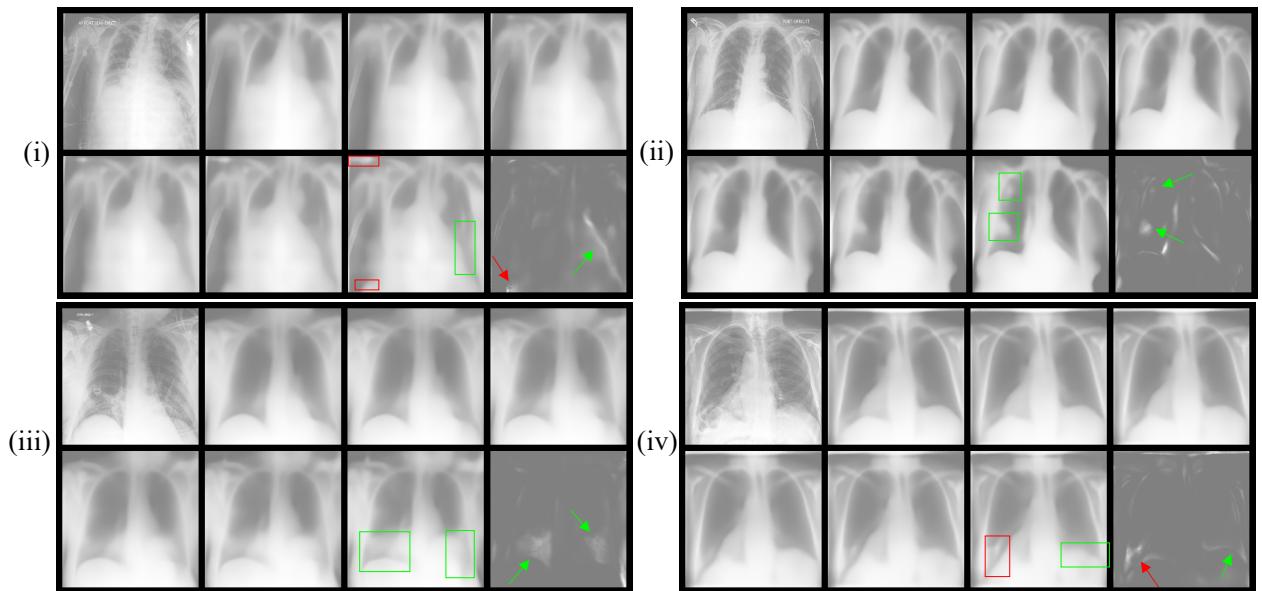
For lung opacity cases, DirVAE traversals consistently highlight areas of suspected consolidation. Clinically, the radiographic presentation of lung opacities can vary greatly, this is reflected in the diversity of feature changes observed during DirVAE traversals. Traversals capture features of consolidation that vary in shape, size and position, matching the appearance of diagnostic features in the example CXRs (fig. 5b). Similarly, in evaluating traversals of the 'Support Devices' class, we observe a diverse set of well-defined structural changes in all areas of the CXR, with significant changes mirroring the location and shape of support devices in the evaluated CXR (see fig. 5a and fig. 8). These results indicate that the DirVAE has the capacity to learn disease/class-specific latent factors that are representative of all modes within the class of interest.

With successful isolation of salient features in the latent space, potential reliance on 'spurious' features for class prediction can be observed in some cases (indicated by red boxes overlaid on CXR images in fig. 5). Here, we see that latent traversal causes feature changes relating to radiology annotation, co-occurring classes (particularly support devices) and shoulder position. This indicates that the disentangled latent space learned by the DirVAE could also be used to identify spurious features resulting from biases in the data, which is a requisite for mitigating the same.
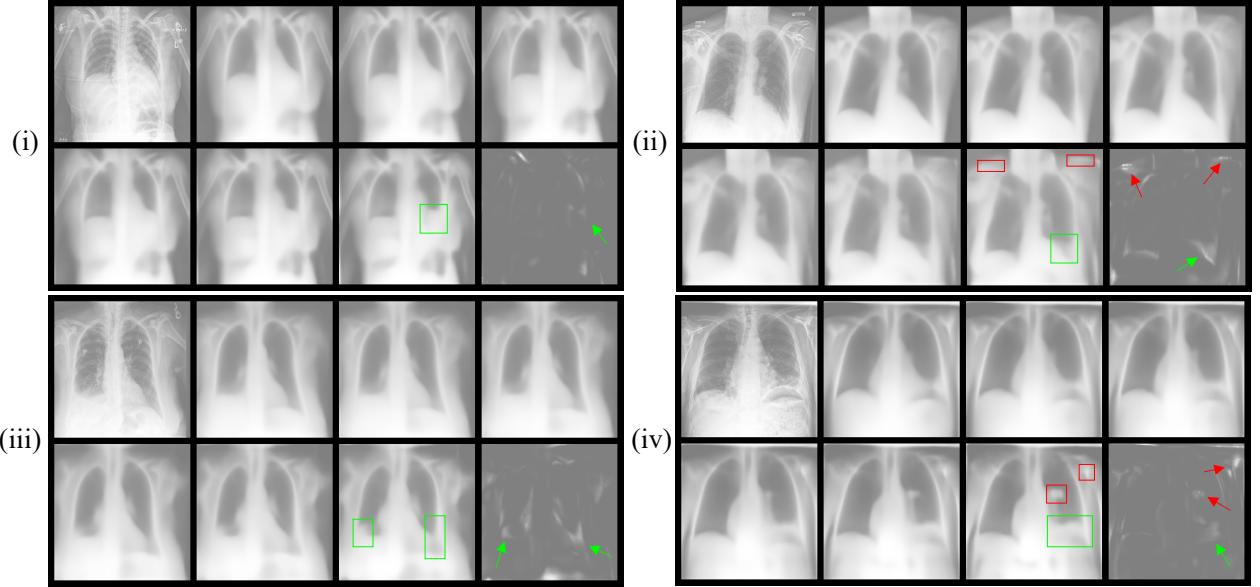
Crucially, no such class-specific changes are observed during evaluation of GVAE latent traversals, for any class (fig. 6). Traversal variance appears diffuse across the reconstructed image, showing only subtle changes in anatomical areas and little to no change in areas of diagnostic relevance. Where disease-related feature changes are observed they are not localised to regions relevant to the class of interest, but are accompanied by changes to other features across the images. Changes in clavicle, shoulder position, mediastinum and lung width are often observed together, causing non-specific changes in the size or shape of the lung airspace. With non-specific changes GVAE traversals appear similar for all examples and localisation of CXR features of significance is near impossible. This is especially apparent in Figure 7 and 8, which present direct comparisons between DirVAE and GVAE traversals (for the same input CXR images), with 'Support Devices' as the class of interest.

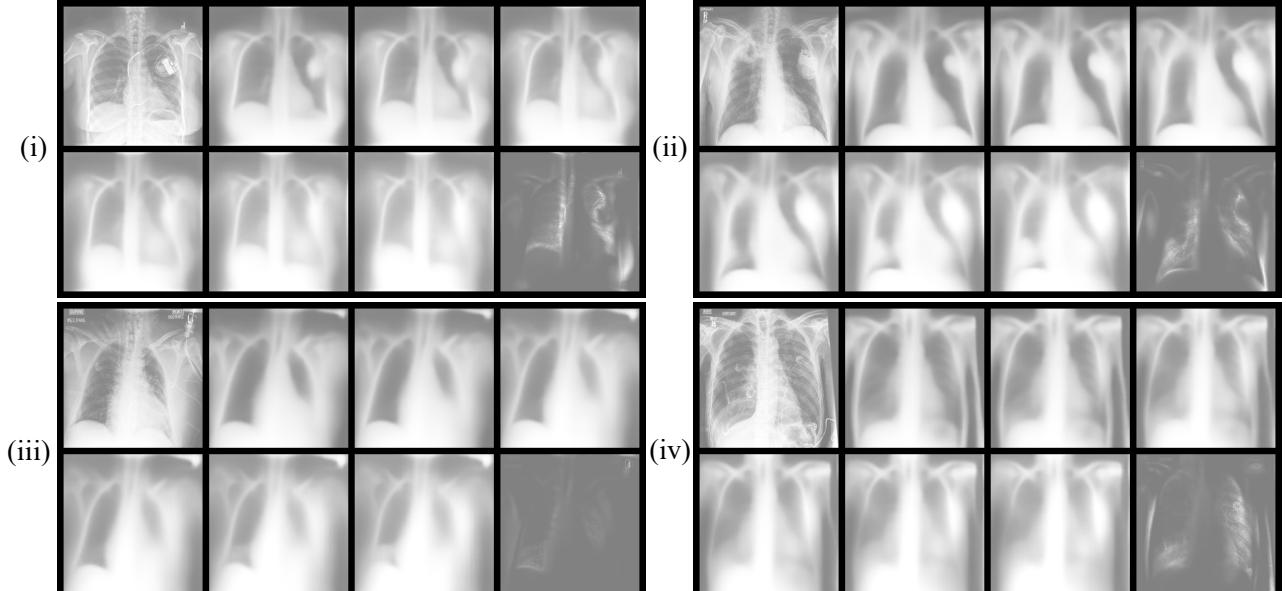(a) Examples of DirVAE latent traversal results of the 'Support Devices' CXR class

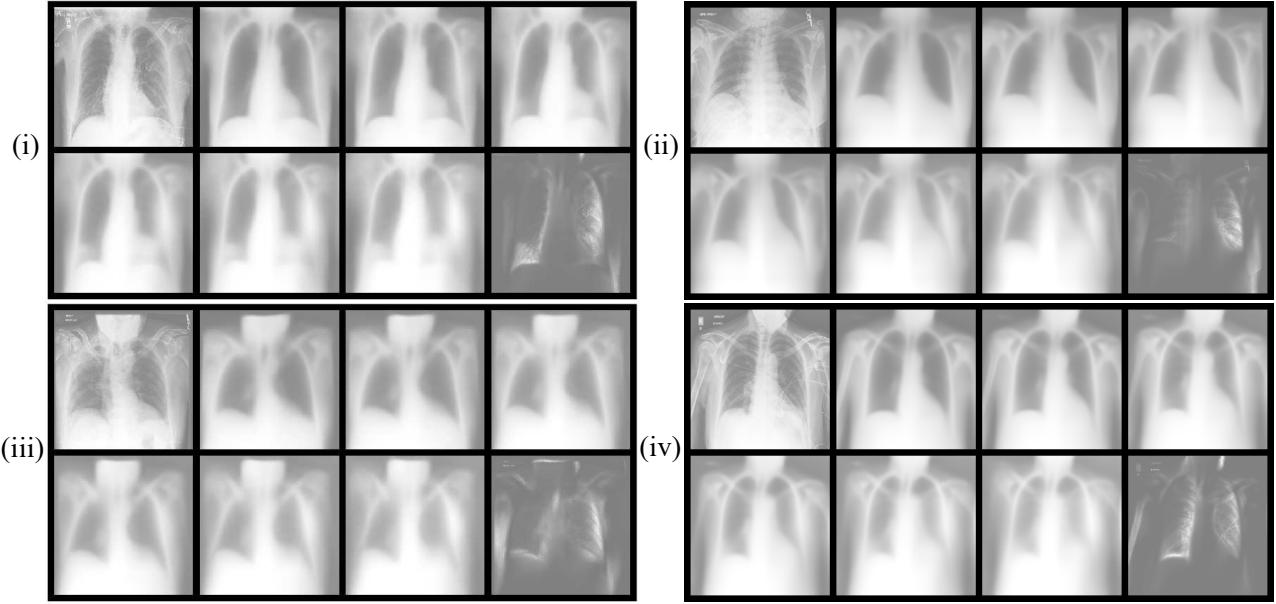(b) Examples of DirVAE latent traversal results of the 'Lung Opacity' CXR class

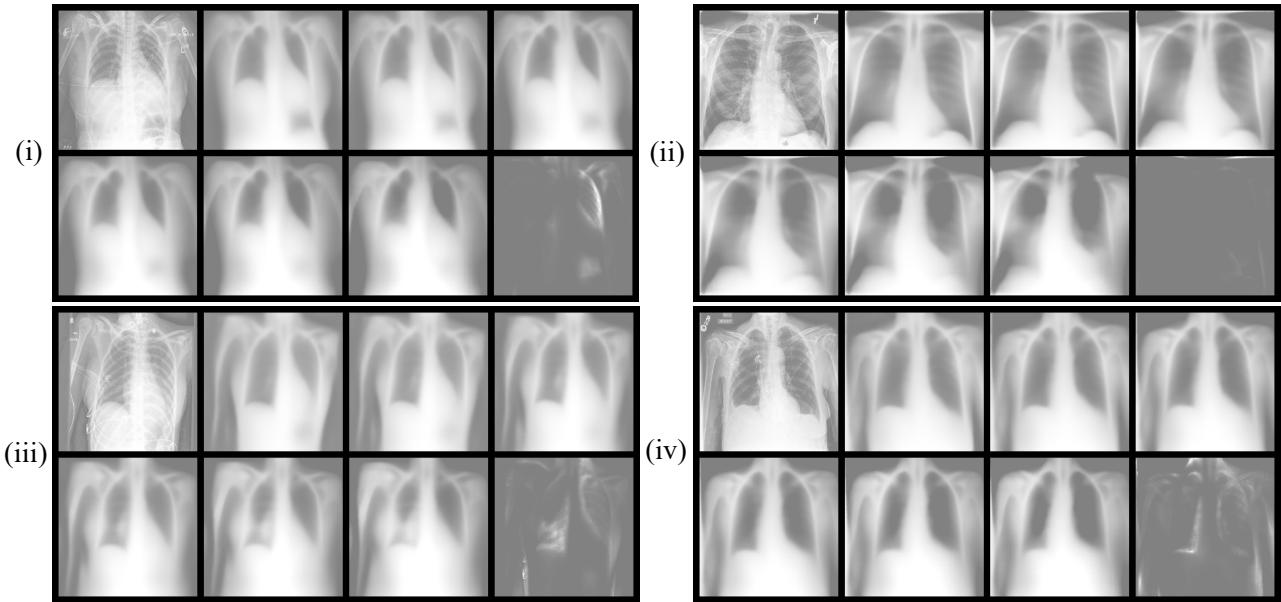(c) Examples of DirVAE latent traversal results of the 'Pleural Effusion' class

Figure 5: DirVAE latent traversals, with examples from each CXR class. Each grid presents a different example, the top left image in the grid is the original reconstruction, the entire traversal is presented sequentially in the grid. The bottom right image presents a pixel-wise variance map summarising the entire traversal. Green arrows (➡) in the variance map point to disease-related feature changes while red arrows (➡) point to likely confounding features. Similarly, red and green boxes are used to highlight key features changes in the final reconstruction.

(a) Examples of GVAE latent traversal results of the 'Support Devices' class

(b) Examples of GVAE latent traversal results of the 'Lung Opacity' class

(c) Examples of GVAE latent traversal results of the 'Pleural Effusion' class

Figure 6: GVAE latent traversals, with examples from each CXR class. Each grid presents a different example, the top left image in the grid is the input CXR, the next image is the original reconstruction, the entire traversal is then presented sequentially in the grid. The bottom right image presents a pixel-wise variance map summarising the entire traversal.

Figure 7: A direct comparison of GVAE and DirVAE traversals, with examples from the 'Support Devices' class. Each row presents a different example, the top left image in the grid is the input CXR, the next image is the original reconstruction, the entire traversal is then presented sequentially in the grid. The bottom right image presents a pixel-w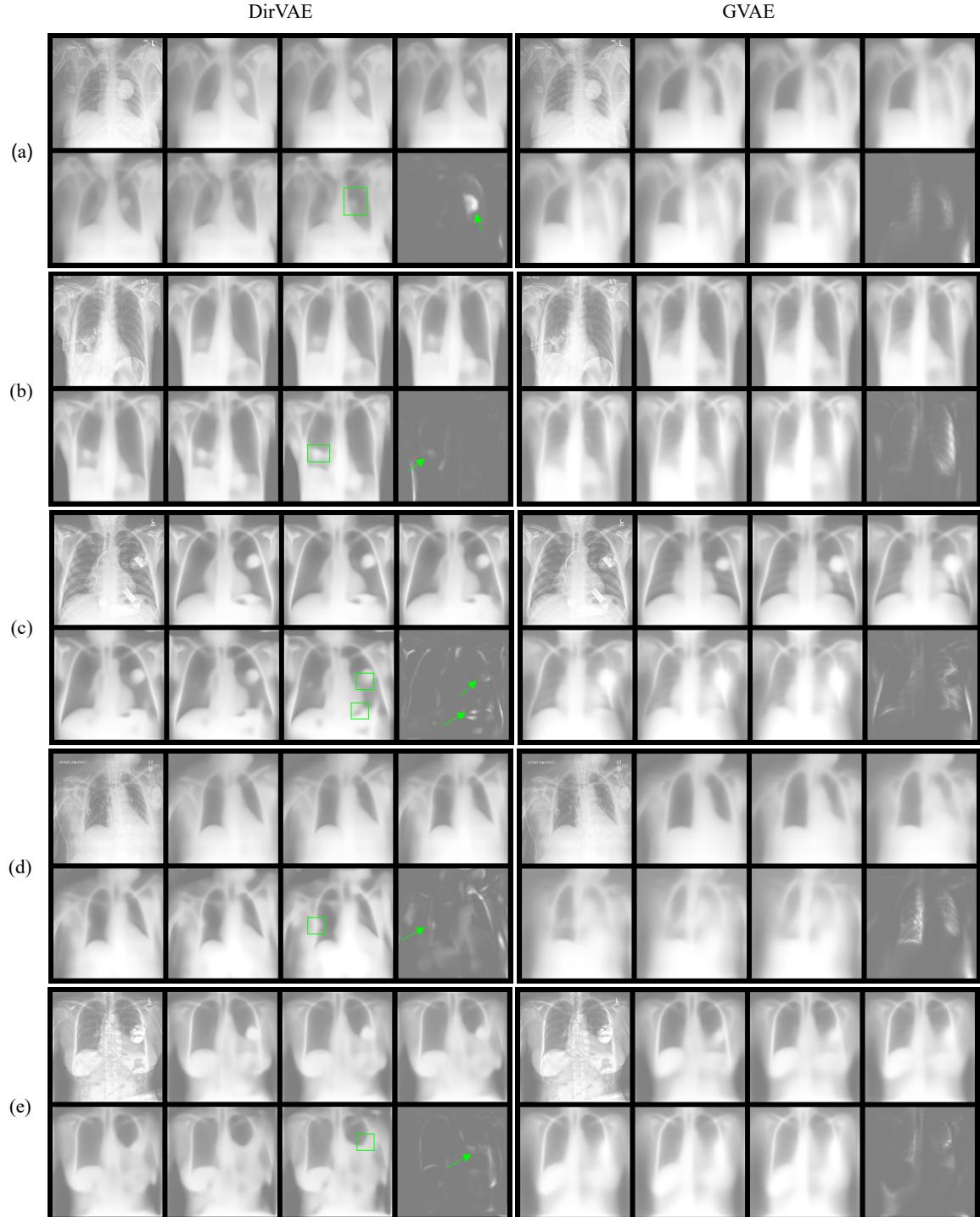ise variance map summarising the entire traversal. Green arrows (→) in the variance map point to class-related feature changes. Similarly, green boxes are used to highlight key features changes in the final reconstruction.
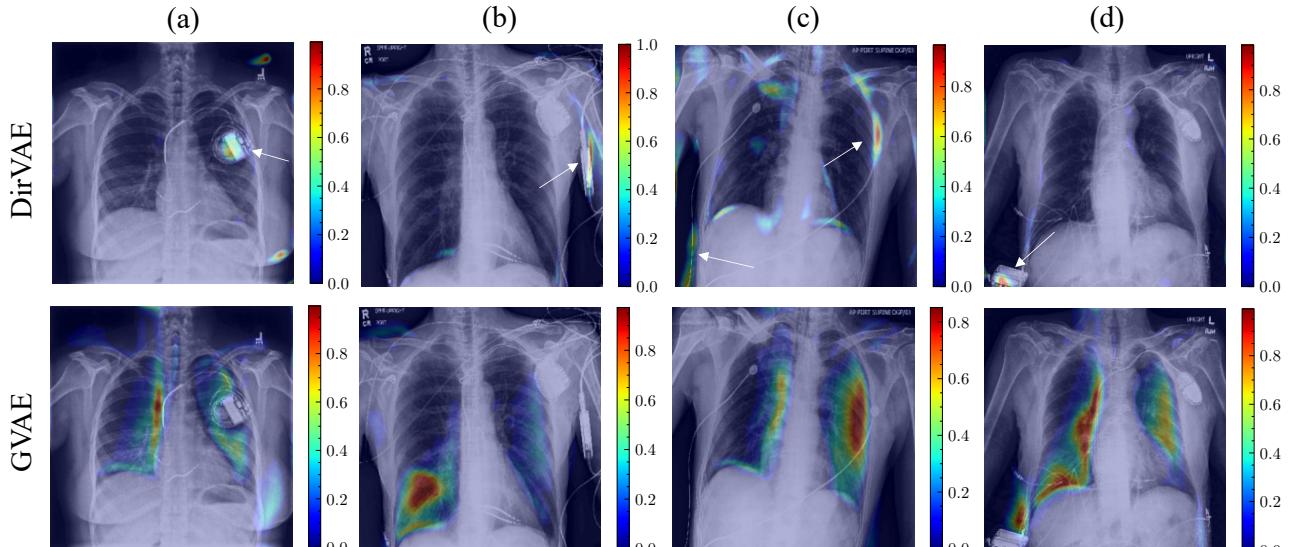
Figure 8: Saliency maps representing variance over all reconstructions in a latent traversal series, performed for the 'Support Devices' class, with comparison between GVAE and DirVAE. We include arrows in DirVAE examples to show areas of class-related feature changes.

5. DISCUSSION

Analysis of latent traversals suggests that the DirVAE learns a more disentangled representation (than its GVAE counterpart), despite the challenges presented by a multi-label task, where the co-occurrence of disease features complicates representation learning. In addition to its superiority in terms of disentanglement and therefore, explainability, DirVAE classifiers perform marginally better than GVAE. The improvement in disentanglement afforded by the DirVAE, relative to the GVAE, is attributed to the sparsity and multi-modal characteristics of the latent space learned by the former. Specifically, the sparsity of our Dirichlet prior is enforced by setting its concentration parameter $\alpha = 0.5$ across all experiments. This pushes areas of the posterior distribution towards zero, i.e. the inner areas of the simplex, and creates a narrower information bottleneck.

In order to generate quality reconstructions under strict sparsity constraints the latent factors have to capture more global features.[12] Our approach to encouraging disentanglement follows a similar principle to β-VAE, but with a different strategy for narrowing the information bottleneck.[13] If we consider this with our classification objective, which influences the model to prioritise learning of disease-related features, we could say that our strategy encourages the learned latent variables to capture 'global disease features' and explain away noise. We can observe this principle in latent traversal evaluations, by traversing latent dimensions important for lung opacity classification, we do not observe diffuse changes in unrelated CXR features, but clear changes in homologous features of the image. Moreover, we observe the capacity of the DirVAE to disentangle global patterns of disease features, with localised regions of consolidation changing across both lungs and in upper and lower lung regions. Similarly, global features are observed in 'Support Device' classification, where all support structures, including connecting wiring (distributed spatially across the image) are altered during traversal; during traversal coordinated feature changes can be observed across the entire image, for all images classes.

Our proposed use of latent embedding manipulation for identification of the salient visual features underlying classifier predictions facilitates improved model explainability. Within our explainability approach, superior disentanglement in DirVAE latent representations allow us to more clearly observe the appearance of 'spurious correlations' alongside disease features, across traversals. Examples of observed 'spurious correlation' includes, sex changes, the appearance of various support devices, as well as changing body position (relating to switches from AP to PA projection).[14]

Highlighting the influence of confounders and biases in predictions made by learning-based systems is essential for building safer and fairer predictive models. This is especially relevant for translating learning-based computer-

aided diagnostic/screening systems to routine clinical care.[15] Particularly, multi-label classification tasks, where the co-occurrence of disease makes deep learning models more vulnerable to reliance on confounding factors. Typical approaches to multi-label image classification explainability, including GradCAM,[16] have been criticised for their inability to highlight smaller pathologies or structures with complex shapes, for example, mechanical wiring.[17] Our work opens doors to a novel approach for this task. By interrogating the direct influence of important/identified latent factors on the generated images and the classification of independent outcomes (or indeed the fluctuation in the predicted class-specific probabilities) for the generated images, our approach provides a systematic framework to identify relevant class-specific features and confounders and/or biases in image data. Interrogating the effects of the most predictive latent factor demonstrates that this approach is able to clearly highlight small and complex features as well as important correlations with, potentially, confounding image features.[17] This is well-illustrated by an example of confounding highlighted in a 'Lung Opacity' traversal (fig. 5b (iv)), where the classifier appears to mistake the gas in the digestive system for opacities in the lower lung regions (highlighted by red box and arrows).

Strong classification results alongside unclear traversals suggest that GVAE classifiers rely on latent factors that combine to explain shared variation and describe the CXR image (and its corresponding class(es)) as a whole, rather than rely on a sparse set of disentangled latent factors that describe the presence/absence of disease-/class-specific features (as with the DirVAE). While some disease-/class-specific features are visible in the latent traversals visualised for GVAE, these are attributed to the effects of multi-task representation learning, i.e. to the joint training of the GVAE with the logistic regression classifiers for the multi-label classification problem. Due to the dense and unimodal nature of Gaussian distributions, the learned latent space appears unable to separate disease-/class-specific features other image-specific features. This is particularly evident in the latent traversals presented for CXR images from the 'Support Devices' class (see fig. 7 and fig. 8). From this, we can conclude that GVAE has limited functionality as a method for prediction explainability.

6. CONCLUSION

In this work we demonstrate that disentanglement of class-/disease-specific (and potentially, clinically-relevant) features can be achieved using our DirVAE model and that its capacity for class-/disease-specific disentanglement is superior to a GVAE. We introduce a promising new approach for explainable multi-label classification, where we apply an ensemble of simple logistic regression classifiers and explore latent dimensions of significance to class predictions through so-called 'latent traversals'. With this we provide visual examples of how our explainability method can highlight regions in CXR images clinically relevant to the class(es) of interest and additionally, identify cases where classification is biased and relies on spurious feature correlations.

7. FUTURE WORK

Future work will explore the use of metrics to quantify disentanglement in order to formalise our qualitative assessment. We will improve image reconstruction/generation quality, with a view to improve the clarity of feature changes during latent traversals. We will explore latent space interpolations between CXR images of different classes and observe corresponding changes to classifier predictions during the transitions between disease representations, for further explainability. With improved reconstruction quality the developed approach would have applications in both explainable medical AI and synthetic data.

ACKNOWLEDGEMENTS

REFERENCES

[1] Kingma, D. P. and Welling, M., "Auto-encoding variational bayes," *CoRR* **abs/1312.6114** (2013).

[2] Tschannen, M., Bachem, O., and Lucic, M., "Recent advances in autoencoder-based representation learning," (2018).

[3] Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A., "beta-vae: Learning basic visual concepts with a constrained variational framework," in [*5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*], OpenReview.net (2017).

[4] Kim, H. and Mnih, A., "Disentangling by factorising," in [*International Conference on Machine Learning (ICML)*], (2018).

[5] Chen, R. T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K., "Isolating sources of disentanglement in variational autoencoders," in [*Advances in Neural Information Processing Systems*], Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., eds., **31**, Curran Associates, Inc. (2018).

[6] Tonolini, F., Jensen, B. S., and Murray-Smith, R., "Variational sparse coding," in [*Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*], Adams, R. P. and Gogate, V., eds., *Proceedings of Machine Learning Research* **115**, 690–700, PMLR (22–25 Jul 2020).

[7] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K. S., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *CoRR* **abs/1901.07031** (2019).

[8] Gawlikowski, J., Saha, S., Kruspe, A., and Zhu, X. X., "An advanced dirichlet prior network for out-of-distribution detection in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–19 (2022).

[9] Srivastava, A. and Sutton, C., "Autoencoding variational inference for topic models," in [*Proceedings for the 5th International Conference on Learning Representations*], (Apr. 2017). 5th International Conference on Learning Representations, ICLR 2017 ; Conference date: 24-04-2017 Through 26-04-2017.

[10] Joo, W., Lee, W., Park, S., and Moon, I.-C., "Dirichlet variational autoencoder," *Pattern Recognition* **107**, 107514 (2020).

[11] Ruskin, J., Gurney, J., Thorsen, M., and Goodman, L., "Detection of pleural effusions on supine chest radiographs," *American Journal of Roentgenology* **148**(4), 681–683 (1987). PMID: 3493648.

[12] Makhzani, A. and Frey, B. J., "k-sparse autoencoders," in [*2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*], Bengio, Y. and LeCun, Y., eds. (2014).

[13] Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A., "Understanding disentangling in β-vae," *CoRR* **abs/1804.03599** (2018).

[14] Jabbour, S., Fouhey, D., Kazerooni, E., Sjoding, M. W., and Wiens, J., "Deep learning applied to chest x-rays: Exploiting and preventing shortcuts," in [*Proceedings of the 5th Machine Learning for Healthcare Conference*], Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., and Wiens, J., eds., *Proceedings of Machine Learning Research* **126**, 750–782, PMLR (07–08 Aug 2020).

[15] Gaube, S., Suresh, H., Raue, M., Lermer, E., Koch, T., Hudecek, M., Ackery, A., Grover, S., Coughlin, J., Frey, D., Kitamura, F., Ghassemi, M., and Colak, E., "Non-task expert physicians benefit from correct explainable ai advice when reviewing x-rays," *Sci Rep* **13**(1), 1383 (2023).

[16] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in [*2017 IEEE International Conference on Computer Vision (ICCV)*], 618–626 (2017).

[17] Saporta, A., Gui, X., Agrawal, A., et al., "Benchmarking saliency methods for chest x-ray interpretation," *Nature Machine Intelligence* **4**, 867–878 (2022).