UNIVERSITY OF LEEDS

From Shortcut Learning to Explainable Prediction: A Framework for Complex Pulmonary Disease Detection

Author
Rachael Harkness

Supervisors

Dr. Nishant Ravikumar
Technical Supervisor

Dr. Kieran Zucker
Clinical Supervisor

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
School of Computer Science
January, 2025

Intellectual Property and Publication Statements

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contribution and the other authors to this work has been explicitly indicated below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

This work is partially adapted from three publications, with all permissions acquired. These are listed below:

1. Harkness R, Hall G, Frangi AF, Ravikumar N, Zucker K. The Pitfalls of Using Open Data to Develop Deep Learning Solutions for COVID-19 Detection in Chest X-Rays. In Otero P, Scott P, Martin SZ, Huesing E, editors, MEDINFO 2021: One World, One Health - Global Partnership for Digital Innovation - Proceedings of the 18th World Congress on Medical and Health Informatics. IOS Press. 2022. p. 679-683. (Studies in Health Technology and Informatics). doi: 10.3233/SHTI220164

2. Harkness R, Frangi AF, Zucker K, Ravikumar N. Multi-centre benchmarking of deep learning models for COVID-19 detection in chest x-rays. Frontiers in radiology. 2024 May 21;4:1386906. doi: 10.3389/fradi.2024.1386906

3. Harkness R, Frangi AF, Zucker K, Ravikumar N. Learning disentangled representations for explainable chest X-ray classification using Dirichlet VAEs. In Colliot O, Isgum I, editors, Medical Imaging 2023: Image Processing. SPIE. 2023. 1246411. (Progress in Biomedical Optics and Imaging - Proceedings of SPIE). doi: 10.1117/12.2654345

[1] is adapted for Chapter 4, [2] is adapted for Chapter 5, and [3] is adapted for Chapter 6.

While portions of this research have appeared in co-authored publications, I confirm that I was solely responsible for conducting all experiments, analysing the data, and writing the manuscripts. Co-authors provided supervisory guidance and editorial feedback only. I was the primary contributor to all aspects of the research and writing presented in this thesis.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr Nishant Ravikumar and Dr Kieran Zucker, for their invaluable guidance, constructive criticism, and unwavering support throughout my PhD journey. Their expertise and insights have been instrumental in shaping this research. I am grateful to the Center for Computational Imaging & Simulation Technologies in Biomedicine (CISTIB) for providing the funding and resources that made this research possible.

Finally, I would like to extend my heartfelt thanks to my family and friends for their support throughout my PhD.

Abstract

Medical imaging plays a critical role in diagnosing and monitoring various diseases, with chest radiographs (CXRs) being one of the most widely used tools for pulmonary disease detection. However, the interpretation of CXRs is often challenging due to overlapping tissue features, low contrast, and the presence of co-occurring diseases. Traditional deep learning approaches, which often focus on single-disease classification, fail to account for the complexities of multi-pathology presentations and raise concerns about bias and interpretability. This thesis addresses these limitations by advancing explainable, multi-label deep learning frameworks tailored for the detection and explanation of co-occurring pulmonary diseases in CXRs.

I highlight the risks of single-disease approaches, using COVID-19 detection as a case study, and demonstrate the benefits of multi-label classification in capturing disease interdependencies and mitigating model bias. To improve interpretability, I propose sparse prior variational autoencoder (VAE) and hierarchical VAE models, which provide precise visual explanations through gradient-guided latent traversals. These methods outperform traditional deep CNN-based explainability techniques in feature isolation and disease localisation but face challenges with reconstruction quality and predictive accuracy. By advancing explainable, multi-label frameworks, this thesis advances the development of trustworthy, transparent diagnostic tools.

Contents

1 Introduction 1
1.1 Explainability Methods 2
1.2 Thesis Overview 3

2 Background 5
2.1 Radiology & Computer Vision 5
2.2 The Respiratory System 8
2.3 Thesis Aims & Objective 10
  2.3.1 Thesis Structure & Outline 10

3 Methods 13
3.1 Convolutional Neural Networks 13
  3.1.1 Residual Networks 16
  3.1.2 U-Net 17
  3.1.3 Monte Carlo Dropout Uncertainty Estimation 18
3.2 Variational AutoEncoders 18
  3.2.1 Optimisation 20
  3.2.2 Disentanglement & Decomposition 22
    3.2.2.1 β-VAE 23
    3.2.2.2 β-TCVAE 24
    3.2.2.3 FactorVAE 25
    3.2.2.4 Sparsity-driven decomposition 25
3.3 Hierarchical VAEs 27
  3.3.1 Multi-level VAE 30
  3.3.2 Ladder VAE 30
  3.3.3 Bidirectional inference VAE 31
  3.3.4 Nouveau VAE 31
3.4 Data 33
  3.4.1 COVID-19 CXR datasets 33
  3.4.2 Multi-label CXR datasets 35

4 The Pitfalls of using Open Data for COVID-19 Detection in CXRs 37
4.1 Introduction 37
4.2 Experiment I: Model generalisability 38
4.3 Experiment II: Analysis of *Frankenstein* Data 41
  4.3.1 *Frankenstein* Dataset 41

4.4 Results 41
4.4.1 Experiment I: Model generalisability 41
4.4.2 Experiment II: *Frankenstein* data 42
4.5 Discussion 46
4.6 Conclusion 47

5 Multi-centre Benchmarking of Deep Learning Models for COVID-19 Detection in CXRs 49
5.1 Introduction 49
5.2 Materials and Methods 50
5.2.1 Overview of the experimental approach 50
5.2.2 Data 51
5.2.2.1 Label generation 55
5.2.2.2 Models 57
5.2.3 Model training 59
5.2.4 Performance evaluation 62
5.3 Results 63
5.3.1 Evaluation of model performance in national and international hospital populations 64
5.3.2 Model performance under counterfactual conditions 68
5.3.3 Subgroup analysis 69
5.3.4 Model error analysis 73
5.3.5 Model explanations & shortcut features 80
5.3.6 Comparative validation of the impact of lung segmentation on model performance 80
5.4 Discussion 81
5.5 Conclusion 86

6 Multi-task VAEs for the Explainable Prediction of Co-Occurring Pulmonary Diseases 87
6.1 Introduction 87
6.2 VAEs for Explainable Prediction 89
6.2.1 Prior distributions: Gaussian vs Dirichlet 90
6.2.1.1 Gaussian Distribution 91
6.2.1.2 Dirichlet Distribution 91
6.2.2 Generating visual explanations 92
6.3 Application: Explainable Prediction of Multi-label CXRs 93
6.3.1 Results 97
6.3.2 Discussion 102
6.4 Conclusion 106

7 Multi-task Hierarchical VAEs for Disease Localisation 107

7.1 Introduction . . . 107

7.2 Related Work . . . 108

7.3 Bi-directional inference VAE . . . 111

7.3.1 Prior distributions . . . 112

7.3.1.1 Sparse mixture of Gaussian distributions . . . 113

7.3.2 Generating visual explanations with BIVA . . . 113

7.4 Application: Explainable Prediction of Multi-label CXRs . . . 116

7.4.1 Results . . . 121

7.4.1.1 Model fit . . . 121

7.4.1.2 Multi-label prediction . . . 124

7.4.1.3 Visual explanations . . . 129

7.4.2 Discussion . . . 143

7.5 Conclusion . . . 145

8 Conclusion 147

8.1 Key Findings . . . 147

8.2 Limitations & Future Works . . . 148

A Ethics Approval 151

List of Figures

3.1 **The arrangement of layers in a typical deep CNN architecture.** Annotations describe the type and order of each layer i.e. ′conv-1′ is the first convolutional layer. Each layer is also labelled with its corresponding dimensions. The example input image is a sample from the MNIST dataset, which contains 10 classes of handwritten digits. . 14
3.2 **An example of a convolutional operation, with a 7x7 input and 3x3 filter.** 15
3.3 **Residual connection in convolutional neural network.** 17
3.4 **U-Net model architecture.** 18
3.5 **A L=3 layered example of hierarchical Dirichlet-prior VAEs: (a) Multi-level VAE, (b) Ladder VAE, (c) BIVA and (d) NVAE.** Dirichlet posterior distributions are parameterised by α values. Δ indicates the use of residual parameterisation of posterior distributions, as in NVAE. Red arrows show information flow for image generation. Blue arrows show how deterministic parameters are shared between generative and inference models, this describes the additional deterministic top-down pathway applied in BIVA and NVAE. *Abbrvs: Variational AutoEncoder (VAE); Bidirectional inference VAE (BIVA); Nouveau VAE (NVAE).* 29
3.6 **A graphical diagram of residual parametrisation of variational posteriors in a NVAE with 3 stochastic layers.** At each level, neural networks are applied to learn the modifications of each distribution from r, which represents the deterministic variables for each level. Δ represents residual parametrisation. *Abbrvs: Nouveau VAE (NVAE).* . 32

4.1 **Saliency maps of correct DarkCovidNet predictions of COVIDX test CXRs.** Examples of negative, pneumonia and COVID-19 are included side-by-side with the predicted image and saliency maps are generated with GradCAM. *Abbrvs: Chest X-rays (CXRs)* 43
4.2 **Frankenstein data analysis results.** (a) 2D t-SNE projection of hidden features extracted from the trained *Frankenstein* classifier during inference on the held-out test set of the *Frankenstein* data. (b) *Frankenstein* data classifier ROC curves, ‘area’ denotes the area under the curve metric. (c) Frequency of image source in the *Frankenstein* dataset. *Abbrvs: Receiver Operator Characteristic (ROC).* 44

4.3 **Frankenstein prediction saliency maps.** A random sample of Grad-CAM images generated by Deep CNN predictions on *Frankenstein* test data. 45

5.1 **Overall experimental design for multi-centre evaluation of COVID-19 detection models. (A)** ROI-cropped CXRs are generated from semantic segmentations of the left and right lung fields, automated prediction uncertainty-based post-processing is applied to ensure reliable cropping for both classes of CXR. The red box highlights over-segmentation of the lung fields, post-processing removes this structure prior to extracting the region of interest. **(B)** Some models are pre-trained, for these models hyper-parameters are tuned on ImageNet or CheXpert data (domain-specific dataset). After pre-training, model hyper-parameters are refined for the COVID-19 detection task and trained on full CXRs from the NCCID. Models are subsequently evaluated on three independent populations: the unseen NCCID population, the LTHT, and COVIDGR. **(C)** Following primary training and evaluation, the best performing models are selected for *(ii)* training and *(iii)* evaluation on ROI-extracted CXRs. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Region Of Interest (ROI); Chest X-ray (CXR).* 51

5.2 **Exclusion criteria for pre-processing (A) NCCID and (B) LTHT datasets.** CXRs were excluded if missing crucial acquisition data (exam date and submission centre) and if not frontal view (AP or PA). NCCID CXRs were eliminated if submitted from Leeds-based hospitals. CXRs were divided into two cohorts: pre-2019 and post-2019. Pre-2019 CXRs were automatically labelled COVID-19 negative, while post-2019 CXRs were evaluated for COVID-19 outcomes. Post-2019 CXRs were eliminated if missing data essential for determining COVID-19 outcome i.e., CXR acquisition date, RT-PCR swab date or outcome. CXRs were also excluded if exam date fell between diagnostic windows of multiple positive RT-PCR swabs. The COVIDGR dataset is not subject to the same exclusion criteria due to a lack of patient data. As a pre-prepared dataset, some exclusion criteria is already applied i.e., COVIDGR includes only PA CXRs. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Antero-posterior (AP); Postero-anterior (PA); Chest X-ray (CXR).* 53

5.3 **Labelling schema for (A) NCCID and LTHT datasets, which share identical labelling protocol, and (B) COVIDGR data.** For each protocol I present examples of different label outcomes. **(A) CASE 1**: An illustration of CXR acquisition preceding the RT-PCR swab date diagnostic window (-14/+28 days), this case is therefore considered COVID-19 negative. **(A) CASE 2**: An example of CXR acquisition prior to the RT-PCR swab date but within the diagnostic window, as a result this case is labelled COVID-19 positive. **(A) CASE 3**: A scenario involving CXR elimination, where multiple swab tests are documented for a single case. If a CXR is acquired within the time frame between the windows around the swab dates, it is excluded from the dataset. **(B) CASE 1**: A case in which the CXR was acquired within the diagnostic window, specifically within 24 hours of the RT-PCR swab date (-1/+1 days). As a result, this case is designated as COVID-19 positive. **(B) CASE 2**: An example of CXR acquisition occurring after the diagnostic window, leading to the categorisation of this case as COVID-19 negative. *Abbrvs: Chest X-ray (CXR); National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Reverse Transcription Polymerase Chain Reaction (RT-PCR).* 56

5.4 **Post-processing of COVID-19 CXR semantic lung segmentation to generate reliable ROIs.** **(A)** An example of severe COVID-19 in a CXR resulting in an uncertain prediction, with a total prediction uncertainty of 620. The right lung field is under-segmented, the additional structure is highlighted in the semantic segmentation and corresponding uncertainty map. This creates an overly large ROI around the lungs. Pixel-wise prediction uncertainty is used to isolate the extra structure, which is eliminated during post-processing of the semantic segmentation. **(B)** A COVID-19 CXR with under-segmentation of the lung fields (highlighted in the semantic segmentation and uncertainty map). Pixel-wise uncertainty is used to remove this structure from the semantic segmentation before creating a ROI. Total prediction uncertainty of this example is 230. *Abbrvs: Region Of Interest (ROI); Chest X-ray (CXR).* 60

5.5 **Example of unsupervised lung segmentation post-processing algorithm on NCCID data.** The U-Net++ model is used to generate semantic segmentation of the left and right lung field. Monte Carlo dropout is applied to approximate uncertainty of prediction, total uncertainty is calculated and frequency of uncertainty is evaluated for Bimodality with Hartigans' test. In this example, uncertainty is less than the threshold for required manual inspection and prediction uncertainty is bimodal so automatic post-processing is applied. In post-processing, unconnected structures are identified and the density of uncertainty is calculated per structure. Excess structures (more than the two lung fields) are iteratively removed, with the most uncertain structures removed first. A ROI is generated from the post-processed semantic segmentation, the ROI was selected to be the minimum bounding box around the segmented lung fields. *Abbrvs: Region of interest (ROI); National COVID-19 Chest Imaging Database (NCCID).* 61

5.6 **ROI-extractions of post-processed semantic lung segmentations. (A)** Examples of ROI-extracted NCCID control cases. **(B)** ROI-extracted NCCID COVID-19 cases. *Abbrvs: Region Of Interest (ROI); National COVID-19 Chest Imaging Database (NCCID).* 62

5.7 **ROC curves of COVID-19 detection.** Each row presents a different model, the first of each row presents ROC curves for all test data, the following columns present direct comparison between NCCID ROC curves and the dataset of interest, corresponding AUROC values can be found in Table 5.6. Shaded regions correspond to 95% confidence interval. *Abbrvs: Receiver Operating Characteristic (ROC); National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Area Under the Receiver Operating Characteristic (AUROC).* 68

5.8 **Average FPRs and FNRs of top-performing models, (A) XCEPTION NET, (B) SSL-AM, and (C) XVITCOS, and their ROI-trained counterparts in LTHT subgroups**. Subgroup population size is referred to by n. Error bars correspond to standard deviation across cross-validations. *Abbrvs: False Negative Rate (FNR); False Positive Rate (FPR); Region of Interest (ROI); Leeds Teaching Hospital Trust (LTHT).* . . . 74

5.9 **ROC curves of COVID-19 detection in smoker, sex, ethnicity and age subgroups for top-performing models: (A) XCEPTION NET, (B) SSL-AM, and (C) XVITCOS.** The top row presents model performance in the real-world scenario and the bottoms rows present model performance under counterfactual conditions, LTHT (P) and LTHT (NP). Subgroup population size is referred to by *n*. Subgroups that do not exist in LTHT (P) or LTHT (NP) populations are excluded. Shaded regions correspond to 95% confidence intervals. *Abbrvs: Receiver Operating Characteristic (ROC); Area Under Curve (AUC); Leeds Teaching Hospital Trust (LTHT).* 76

5.10 **Analysis of XCEPTION NET prediction error.** Frequency of **(A)** false positive predictions on NCCID TEST data according to proximity of CXR exam date to the diagnostic window, **(B)** false negative predictions on COVIDGR according to RALE-defined CXR severity and **(C)** false positive predictions on LTHT data according to CXR projection, alongside evaluation of ROI-trained XCEPTION NET false positive frequency. *Abbrvs: Radiographic Assessment of Lung Edema (RALE); Region of Interest (ROI); Chest X-ray (CXR); Antero-posterior (AP); Postero-anterior (PA).* 77

5.11 **(A) XCEPTION NET, (B) SSL-AM and (C) XVITCOS AUROCs for LTHT subgroup and LTHT (P) subgroups.** Only subgroups that exist in the LTHT (P) population are included. *n* is the subgroup population size. Error bars correspond to standard deviation across cross validations. *Abbrvs: Area Under Receiver Operator Characteristic (AUROC); Region of Interest (ROI).* 78

5.12 **Saliency maps of correct XCEPTION NET predictions of COVID-19 positive CXRs.** CXRs are sourced from the LTHT dataset and saliency maps are generated with **(A)** Gradient backpropagation and **(B)** GradCAM. *Abbrvs: Leeds Teaching Hospital Trust (LTHT); Chest X-rays (CXRs).* 80

5.13 **Gradient backpropagation feature attribution maps of true positive COVID-19 predictions.** Saliency maps are generated for **(A)** XCEPTION NET predictions and **(B)** XCEPTION NET (ROI) predictions. XCEPTION NET (ROI) is trained and evaluated on CXRs cropped to the region of interest, while XCEPTION NET is trained on full CXRs. *Abbrvs: Leeds Teaching Hospital Trust (LTHT); Chest X-rays (CXRs); Region Of Interest (ROI).* 81

5.14 **ROC curves of (A) SSL-AM, (B) XCEPTION NET, and (C) XVITCOS model evaluations on each dataset and their ROI-trained counterparts.** Shaded regions correspond to the 95% confidence intervals. *Abbrvs: Receiver Operating Characteristic (ROC); Area Under Curve (AUC); Region Of Interest (ROI).* 82

6.1 **3D Dirichlet distributions with different concentrations (α).** 92

6.2 **CheXpert co-occurrence heatmap, raw counts of each label are included.** Co-occurrence is normalised to a 0-1 scale. 94

6.3 **Overview of DirVAE with auxiliary ensemble of logistic regression classifiers.** The DirVAE is trained under a multi-task objective combining a (a) reconstruction task with (b) a multi-label prediction task. 96

6.4 **Dirichlet-prior VAE 'sparseness'.** (a) Frequency of values sampled from the Dirichlet-prior VAE posterior, $z \sim q(z|x)$. (b) Example of classifier weights extracted from the *Pleural Effusion* classifier paired with Dirichlet-prior VAE. 98

6.5 **Saliency maps representing variance over all reconstructions in a gradient-guided latent traversal series, performed for the *Support Devices* class, with comparison between GVAE and DirVAE.** The colour bar shows pixel variance over the traversal, which is rescaled to 0-1 range. 99

6.6 **DirVAE gradient-guided 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. 100

6.7 **Partial latent interpolations (n=10) between label-positive CXRs and a confounding variable matched *No Finding* sample.** n is the number of latent dimensions interpolated. I include *Pleural Effusion* (top row), *Support Devices* (middle row) and *Lung Opacity* (bottom row) examples, and compare the Dirichlet-prior (left) and Gaussian-prior VAE (right). In each row the first column is the input image, columns 2-4 shows the progressive change in the reconstructions over the interpolation series, and column 5 (final) is the residual between the unchanged reconstruction and reconstruction post-interpolation. 101

6.8 **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-wise 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. 102
6.9 **Partial interpolation *(n=10)* from *Lung Opacity* CXR to *No Finding* CXR latent representation.** n **is the number of latent dimensions interpolated. Top**: Panel of image reconstructions resulting from interpolation between *Lung Opacity* CXR and *No Finding* CXR. (A) Original reconstruction of *Lung Opacity* CXR and (B) Reconstruction after interpolation to *No Finding* CXR. **Bottom**: (A) Probabilities predicted from latent representation of *Lung Opacity* CXR. (B) Probabilities predicted from latent representation after partial interpolation towards *No Finding*. 103

7.1 **3 layer BIVA inference model.** Dotted lines indicate shared parameters. Black arrows indicate bottom-up information flow and red arrows indicate top-down information flow. 112
7.2 **Simulated sparse Gaussian mixture.** I sample from four 2D Gaussian distributions and fit a Gaussian kernel to estimate density across the mixture of distributions. To simulate the sparsity constraint I restrict the variance of two out of the Gaussian distributions to 0.01. . . 113
7.3 **A comparison of a (a) L=3 layered BIVA and (b) single-level VAE generative models.** 116
7.4 **Examples of bounded BIVA generative models of** $L = 3$ **layered BIVA.** Red nodes indicate where the conditional prior is used and k defines the bounding level. 117
7.5 **Heatmap of label co-occurrence in CheXpert training data.** 118
7.6 **Heatmap of disease co-occurrence in VinDr-CXR training data, with raw counts of class co-occurrence.** 119
7.7 **Gaussian-prior BIVA model fit.** Model fit is described by (a) Bounded model image generation, where k is the bounding level i.e, $\mathcal{L} > k$, (b) image reconstruction, first row is the input image, second row is the image reconstruction, the third row is the image reconstruction from the mean and, (c) images generated from a random sample of the prior distribution. 122

7.8 **Sparse-prior BIVA model fit.** Model fit is described by (a) bounded model image generation, where k is the bounding level i.e, $\mathcal{L} > k$, (b) image reconstruction, first row is the input image, second row is the image reconstruction, the third row is the image reconstruction from the mean and, (c) images generated from a random sample of the prior distribution. 123

7.9 **(A) Gaussian-prior VAE, (B) Sparse-prior VAE, and (C) Dirichlet-prior VAE model fit**. Top images are image reconstructions: first row is the input image, second row is the image reconstruction, the third row is the image reconstruction from the mean. Bottom images are samples generated from a random sample of the prior distribution. . 124

7.10 **ROC curves of (A) Gaussian-prior VAE, (B) Dirichlet-prior VAE, (C) Sparse-prior VAE, (D) Gaussian-prior BIVA, (E) Sparse-prior BIVA, and (F) Multi-label deep CNN predictions on CheXpert data. *Abbrvs: Receiver Operator Characteristic (ROC).*** 125

7.11 **ROC curves of Gaussian-prior BIVA, Sparse-prior BIVA and multi-label deep CNN on VinDr-CXR data.** *Abbrvs: Receiver Operator Characteristic (ROC).* 128

7.12 **Gaussian-VAE GGLTs for each evaluated label of the CheXlocalise dataset.** *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. Red boxes are predicted disease regions (derived from differences between start traversal and end traversal) and blue boxes are radiologist annotated ground truths. *Abbrvs: Gradient-guided latent traversals (GGLTs).* 131

7.13 **Dirichlet-VAE GGLTs for each evaluated label of the CheXlocalise dataset.** *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. Red boxes are predicted disease regions (derived from differences between start traversal and end traversal) and blue boxes are radiologist annotated ground truths. *Abbrvs: Gradient-guided latent traversals (GGLTs).* 132

7.14 **Sparse-VAE GGLTs for each evaluated label of the CheXlocalise dataset.** *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. Red boxes are predicted disease regions (derived from differences between start traversal and end traversal) and blue boxes are radiologist annotated ground truths. *Abbrvs: Gradient-guided latent traversals (GGLTs).* 133

7.15 **Comparison of GGLT generated by unbounded Sparse-BIVA (k=0) and bounded $\mathcal{L} > 2$ Sparse-BIVA (k=2).** Each case corresponds to a different multi-label CXR. *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. *Expected residual* is a map of pixel variance over repeated reconstructions without traversal. *Residual* is the absolute difference in pixel values between the start traversal and end traversal images. Red arrows point to observable changes in the images generated over the traversal. *Abbrvs: Gradient-guided latent traversal (GGLT)* 135

7.16 **Side-by-side comparison of Gaussian BIVA ($\mathcal{L} > 2$) and Sparse BIVA ($\mathcal{L} > 2$) GGLTs applied to multi-label VinDr-CXR examples.** Each case corresponds to a different multi-label CXR. *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. *Expected residual* is a map of pixel variance over repeated reconstructions without traversal. *Residual* is the absolute difference in pixel values between the start traversal and end traversal images. Red arrows point to areas of significant change. *Abbrvs: Gradient-guided latent traversal (GGLT); Chest X-ray (CXR).* . . . 136

7.17 **Sparse-BIVA ($\mathcal{L} > 2$) GGLTs applied to multi-label VinDr-CXR examples.** Each case corresponds to a different multi-label CXR. *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. *Expected residual* is a map of pixel variance over repeated reconstructions without traversal. *Residual* is the absolute difference in pixel values between the start traversal and end traversal images. Red arrows point to structural changes that relate to the evaluated disease in generated images. Red boxes are predicted disease regions (derived from the residual) and blue boxes are radiologist annotated ground truths. 137

7.18 **Sparse-BIVA ($\mathcal{L} > 2$) OLTs applied to multi-label VinDr-CXR examples.** *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. *Expected residual* is a map of pixel variance over repeated reconstructions without traversal. *Residual* is the absolute difference in pixel values between the start traversal and end traversal images. Red boxes are predicted disease regions (derived from the residual) and blue boxes are radiologist annotated ground truths. *Abbrvs: Optimised Latent Traversals (OLTs).* 138

7.19 **Probability changes over label-specific OLT for multi-label example of VinDr-CXR.** (A) Changes in probabilities over OLT for Pleural Thickening. (B) Changes in probabilities over OLT for Aortic Enlargement. (C) Changes in image generation over OLT for Pleural Thickening. (D) Changes in image generation over OLT for Aortic Enlargement. Red asterisks * point to the evaluated class. *Abbrvs: Optimised Latent Traversals (OLTs)* 140

7.20 **Probability changes over a GGLT for a multi-label example of VinDr-CXR.** (A) Changes in probabilities for GGLT for Pleural Thickening. (B) Changes in probabilities for GGLT for Aortic Enlargement. (C) Changes in image generation over GGLT for Pleural Thickening. (D) Changes in image generation over GGLT for Aortic Enlargement. Red arrows point to areas of significant change in the traversal. The residual image is the absolute pixel-wise difference between the start traversal and end traversal images. Bounding boxes are identified from the binarised residual image. Red asterisks * point to the evaluated class. *Abbrvs: Gradient-guided latent traversals (GGLTs).* 141

7.21 **Comparison of BIVA GGLT visual explanations and multi-label deep CNN GradCAM++ with post-processed localisation.** *Start traversal* is the original image reconstruction prior to GGLT. *End traversal* is the image generated post-GGLT. GGLT overlay displays the GGLT-derived bounding boxes (red) and ground truth radiologist annotations (blue), likewise for GradCAM++ overlay. GradCAM++ was selected as the baseline saliency method due to its improved localisation performance over standard GradCAM and its ability to pass sanity checks that verify dependence on learned model parameters (Chattopadhay et al., 2018; Adebayo et al., 2020). Red arrows point to structural changes that relate to the evaluated disease in generated images. *Abbrvs: Gradient-guided latent traversals (GGLTs).* 142

A.1 Required evidence of ethics approval granted by the Health Research Authority for use of NHS data. 152

List of Tables

4.1 **A summary of public datasets commonly used in DL systems for detecting COVID-19 from CXRs.** *MOONEY is the same as the KERMANY dataset, but hosted on Kaggle. **RICORD provides only COVID-19 images. *Abbrvs: Chest X-ray (CXR); Deep Learning (DL).* . . . 42
4.2 **Model performance metrics across all test datasets, with standard deviation, across cross-validation folds.** 43
4.3 ***Frankenstein* Deep CNN classifier performance metrics.** The Deep CNN classifier is trained to predict image source. RSNA provides only COVID-19 negative images, CHOWDHURY and COHEN give a mix of COVID-19 negative and COVID-19 positive images and are themselves Frankenstein dataset, RICORD gives only COVID-19 positive images. 44

5.1 **Demographic subgroups of training and test data.** Age is presented as mean ± standard deviation. Sex and ethnicity are presented both as absolute counts (n) and as percentages relative to the COVID-19 positive/negative cohort. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Region of Interest (ROI).* 52
5.2 **Inclusion and exclusion criteria of CXR exam from NCCID, LTHT and COVIDGR data.** For all datasets, CXRs are eliminated if not frontal view. NCCID and LTHT CXR exams conducted after 2019 are eliminated if COVID-19 swab or image acquisition data is incomplete. For NCCID data, CXRs are also eliminated if submission centre data is incomplete or if the CXR exam date falls in between two non-overlapping windows of COVID-19 infection. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Antero-posterior (AP); Postero-anterior (PA); Chest X-ray (CXR).* 55

5.3 **Overview of individual dataset use throughout this study, including ROI-extraction, model training and evaluations.** For evaluation of models under counterfactual conditions I used a subset of LTHT with recorded non-COVID-19 pneumonia status. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Region of Interest (ROI); Radiographic Assessment of Lung Edema (RALE).* 57

5.4 **Summary of the evaluated models.** Models are described and presented alongside source reference, pre-training status and deep learning category. Models are referred to by their designated abbreviations. *Abbrvs: Deep Learning (DL); Neural Architecture Search (NAS); Parameters (Params.).* 58

5.5 **Multiple (pair-wise) model performance comparisons with Tukey's Honest Significant Difference (HSD) test.** *Abbrvs: Accuracy (Acc.); Precision (Prec.); Area Under the Receiver Operator Characteristic (AUROC); XCEPTION NET (XCEPTION); Adjusted p-value (p-adj).* 65

5.6 **Average model performance metrics for each test dataset, with standard deviation, across cross-validation folds.** N is the total number of cases in each test population. **Bold** indicates the highest average metric per dataset. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Accuracy (Acc); Precision (Prec); Area Under the Receiver Operator Characteristic (AUROC).* 66

5.7 **One-way analysis of variance (ANOVA) for comparison of mean model performance on population subgroups.** *Abbrvs: Accuracy (Acc.); Precision (Prec.); Area Under the Receiver Operator Characteristic (AUROC).* 71

5.8 **Average model performance metrics for each subgroup in the LTHT population.** Standard deviation of metrics across cross-validation folds is included. **Bold** indicates the highest average metric per dataset. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Accuracy (Acc); Precision (Prec); Area Under Receiver Operator Curve (AUROC); Female (F); Male (M); XCEPTION NET (XCEPTION).* 72

5.9 **Comparison of average full CXR-trained and ROI-trained model performance metrics.** Standard deviation of metrics across cross-validation folds is included. *N* is the total number of cases in each test population. *Abbrvs: Chest X-ray (CXR); National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Accuracy (Acc); Precision (Prec); Area Under the Receiver Operator Characteristic (AUROC).* 79

6.1 **Experimental Settings for Dirichlet-prior and Gaussian-prior VAE training.** . . . 96
6.2 **Label-wise logistic regression classification performance results.** . 97

7.1 **Experimental settings for optimisation of VAE and 3-layer BIVA models. λ is the weighting factor for the classification term.** 118
7.2 **Model fit metrics for VAE and BIVA models.** 124
7.3 **Sparsity measure of VAE and BIVA model posteriors.** Sparsity is measured in the top stochastic level of BIVA models 124
7.4 **Label-wise performance metrics on CheXpert for VAE models, BIVA models, and the multi-label deep CNN.** 126
7.5 **Label-wise performance metrics on CheXlocalise for Dirichlet-prior VAE, Gaussian-prior VAE, and Sparse-prior VAE models.** 126
7.6 **Multi-label prediction performance of BIVA models and multi-label deep CNN on VinDr-CXR test data.** 127
7.7 **VAE localisation of CheXlocalise disease features by OLTs and GGLTs methods across different labels.** *Abbrvs: Optimised Latent Traversal (OLT); Gradient Guided Latent Traversal (GGLT); Hit-or-Miss (HoM).* . . . 129
7.8 **Localisation performance of Sparse-prior BIVA and Gaussian-prior BIVA GGLTs on VinDr-CXR data compared to localisation of multi-label deep CNN GradCAM++.** Classes with low prevalence in VinDr-CXR test data are omitted. *Abbrvs: Gradient-guided latent traversals (GGLTs)* 129

List of Algorithms

1 Gradient-guided latent traversal 93
2 Optimised Latent Traversals 115

List of Abbreviations

AI	**A**rtificial **I**ntelligence
AP	**A**ntero-**P**osterior
AUROC	**A**rea **U**nder **R**eceiver **O**perator **C**haracteristic
BIVA	**Bi**-directional **I**nference **V**ariational **A**uto Encoder
BCE	**B**inary **C**ross **E**ntropy
BU	**B**ottom **U**p
CNN	**C**onvolutional **N**eural **N**etwork
CONSORT	**C**onsolidated **S**tandards **O**f **R**eporting **T**rails
COPD	**C**hronic **O**bstructive **P**ulmonary **D**isease
CXR	**C**hest **X**-**R**ay
DL	**D**eep **L**earning
ELBO	**E**vidence **L**ower **BO**und
FC	**F**ully **C**onnected
FNR	**F**alse **N**egative **R**ate
FPR	**F**alse **P**ositive **R**ate
GGLT	**G**radient **G**uided **L**atent **T**raversal
HVAE	**H**ierarchical **V**ariational **A**uto**E**ncoder
HoM	**H**it **o**r **M**iss
IoU	**I**ntersection **o**ver **U**nion
IW	**I**mportance **W**eighted
KL	**K**ullback **L**eiber
LVAE	**L**adder **V**ariational **A**uto**E**ncoder
LTHT	**L**eeds **T**eaching **H**ospital **T**rust
NAS	**N**eural **A**rchitecture **S**earch
NCCID	**N**ational **C**OVID-19 **C**hest **I**maging **D**atabase
NLL	**N**egative **L**og **L**ikelihood
NVAE	**N**ouveau **V**ariational **A**uto **E**ncoder
OLT	**O**ptimised **L**atent **T**raversal
PA	**P**ostero-**A**nterior
RALE	**R**adiographic **A**ssessment of **L**ung **E**dema
RICORD	**R**SNA **I**nternational **C**OVID-19 **O**pen Annotated **R**adiology **D**atabase
ROC	**R**eceiver **O**perator **C**haracteristic
ROI	**R**egion **O**f **I**nterest
RSNA	**R**adiological **S**ociety of **N**orth **A**merica
RT-PCR	**R**everse **T**ranscription **P**olymerase **C**hain **R**eaction

TD	**T**op **D**own
VAE	**V**ariational **A**uto**E**ncoder

Chapter 1

Introduction

Medical imaging is essential to modern medicine, enabling non-invasive visualization of internal structures and physiological processes. Traditional imaging modalities, including X-rays, computed tomography, magnetic resonance imaging, ultrasound, and positron emission tomography, continue to play a pivotal role in detecting, diagnosing, and monitoring a wide range of diseases and conditions—from fractures and tumours to cardiovascular diseases and neurological disorders. Among various imaging modalities, chest radiographs (CXRs) is one of the most widely used diagnostic tools, particularly for pulmonary diseases (Hussain et al., 2022). However, the interpretation of medical images presents significant challenges, especially when patients present with multiple concurrent conditions, as is likely in hospital populations. The complexity of co-occurring diseases, combined with increasing clinical workloads and the subtle nature of radiographic findings, creates a pressing need for automated assistive systems (Fabbri et al., 2023). In particular, the interpretation of 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. The advent of deep learning, particularly its application in computer vision, has demonstrated remarkable potential in medical image analysis (Hosny et al., 2018). These systems can process vast amounts of imaging data with high accuracy, potentially matching or exceeding human performance in specific diagnostic tasks. However, the transition from promising research to reliable clinical implementation faces several critical challenges. Most notably, current approaches often struggle with model bias and the complexity of co-occurring diseases, where multiple pathologies manifest simultaneously in a single image (DeGrave, Janizek, and Lee, 2021).

Traditional deep learning models for medical image analysis have predominantly focused on single-disease classification, treating each pathology as an independent entity. This approach, while computationally convenient, fails to capture the intricate relationships between different diseases and their combined radiographic presentations. Furthermore, the use of uninterpretable deep learning systems, often described as "black box" models, raises concerns about their reliability and trustworthiness in clinical settings (Hosny et al., 2018).

1.1 Explainability Methods

Deep learning models are "black boxes", referred to as such because they are too complex for humans to understand directly. This creates problems in healthcare, where doctors need to understand why a model made a particular diagnosis. Without this understanding, clinicians cannot trust the system or know when it might fail. Two approaches exist to address this problem, adding explanations to black box models, or using simpler, interpretable models like decision trees or linear regression. While interpretable models are transparent, they often lack the performance needed for complex medical imaging tasks. Explainable AI tries to solve this by creating post-hoc explanations of black box decisions; however existing methods are flawed because if an explanation perfectly captured how the model works, it would be as complex as the original model, defeating the purpose (Rudin, 2019).

A number of post-hoc explainability methods are popularly applied to deep learning computer vision models. Gradient-weighted Class Activation Mapping (Grad-CAM) generates heatmaps by leveraging gradient information flowing into the final convolutional layer, highlighting image regions deemed most important for specific class predictions (Selvaraju et al., 2017a). Local Interpretable Model-agnostic Explanations (LIME) takes a different approach by segmenting images into superpixels and determining which regions contribute positively or negatively to model decision (Ribeiro, Singh, and Guestrin, 2016). Guided backpropagation modifies standard backpropagation to suppress negative influences during the backward pass, theoretically producing cleaner saliency maps (Springenberg et al., 2015).

However, these saliency-based methods suffer from significant limitations that compromise their utility. Saliency maps often provide insufficient detail to understand what the model is actually doing with highlighted image regions—they indicate generally where the model is "looking" but fail to explain the reasoning processes applied to those regions (Rudin, 2019). Furthermore, saliency maps frequently highlight border regions and edges, often providing remarkably similar explanations across different pathological classes, which raises questions about their discriminative value.

Recent research has revealed more fundamental concerns about the trustworthiness of these explanations. Adebayo et al. (2020) introduced randomisation tests that assess whether saliency methods are sensitive to changes in model parameters, a basic requirement for meaningful explanations. Their cascading randomisation through model layers investigates whether saliency depends on learned parameters and the relationship between training instances and labels. Alarmingly, their findings suggest that some widely deployed saliency methods produce explanations that are independent of both the data and the learned model parameters. Of the methods commonly used in medical imaging, Grad-CAM passes these sanity checks, while guided backpropagation and guided Grad-CAM fail by remaining invariant to higher layer parameters.

In medical imaging specifically, saliency maps performed dramatically worse than purpose-built localisation networks when tested on pneumonia detection. This suggests current explanation methods may mislead rather than help clinicians (Arun et al., 2021).

1.2 Thesis Overview

This thesis addresses the challenges outlined above by:

1. Demonstrating the risks associated with taking a standard single-disease classification approach, as was popular in the development of COVID-19 detection models.
2. Developing interpretable, multi-label deep learning frameworks, which I apply for the detection and explanation of co-occurring pulmonary diseases in chest radiographs.

The significance of this research extends beyond technical innovation. The ability to identify model bias and detect and explain multiple co-pathologies enhances diagnostic support and better assists the clinical workflow where an understanding of the interplay between different pathologies is crucial for effective treatment planning. We aim to address the flaws of explainable AI by developing inherently interpretable models from the ground up rather than relying on post-hoc explanations for black box models. By developing interpretable models that can handle disease co-occurrence, this work aims to help bridge the gap between artificial intelligence and clinical practice.

Chapter 2

Background

2.1 Radiology & Computer Vision

In practice, assessment of medical images is performed by expert radiologists and specialist physicians. The role of the radiologist is to visually evaluate medical imaging data in order to detect, characterise or monitor patient disease (Shen, Wu, and Suk, 2017). With increasing workloads faced by radiologists and the growing availability of high-quality medical imaging datasets, deep learning, particularly computer vision, is being increasingly recognised as a transformative tool in clinical radiology (Sogani et al., 2020; Strohm et al., 2020). Recent advancements in deep learning have demonstrated significant success across various radiological applications, with many models achieving diagnostic accuracies comparable to those of expert radiologists. For instance, computer vision algorithms have shown remarkable performance in detecting pulmonary tuberculosis and other lung pathologies from chest radiographs (Lakhani and Sundaram, 2017; Wang et al., 2017), as well as identifying breast masses in mammography scans (Arevalo et al., 2015). These successes highlight the potential of deep learning to assist expert humans, streamline workflows, and improve diagnostic outcomes.

Computer vision, a subfield of deep learning, focuses on enabling machines to interpret, analyse, and derive meaningful information from digital images and other forms of visual data. It encompasses a wide range of machine learning and deep learning techniques that allow computers to *see* and understand visual inputs in ways that simulate human perception (Forsyth and Ponce, 2002). Within computer vision, there are numerous specialised tasks, including image classification (assigning categories to images), object detection (locating specific regions or features within an image), and object tracking (following identified objects across a sequence of images). Each of these tasks plays a critical role in advancing automated medical image analysis, particularly within radiology.

Computer vision algorithms learn hierarchical representations of data through artificial neural networks. These networks capture data at multiple levels of abstraction, from low-level features like edges and textures to higher-level patterns such as anatomical structures or pathological regions (LeCun, Bengio, and Hinton, 2015).

In clinical radiology, deep learning-based computer vision has proven especially effective for tasks involving the analysis of large imaging datasets, such as the detection of lung pathologies in chest X-rays. For example, convolutional neural networks (CNNs)—a class of computer vision algorithm—are commonly employed to identify anomalies, such as nodules, consolidations, and other abnormalities, with high accuracy and efficiency. By training on large-scale annotated datasets, these models can learn to recognise patterns that might be overlooked during routine human interpretation. The aim of deep learning is to learn high-level abstractions of raw input data, where important features of the data are automatically amplified and insignificant features are deemed irrelevant and suppressed (LeCun, Bengio, and Hinton, 2015). To accomplish this, deep learning models are typically built in hierarchical layers, where every layer learns data representations at different levels of abstraction. Using non-linear modules, each stage in the hierarchy transforms its input to a higher level, learning more abstract representations. Moving further up the hierarchy and away from the raw data input compounds these modules, building a complex function that can be learned (LeCun, Bengio, and Hinton, 2015).

For a deep learning model to perform its task successfully, it must undergo a training process. In a typical supervised classification task, where both images and corresponding ground truth labels are provided as inputs, the model is shown an image and generates an output vector that assigns scores to each category. To function as an effective classifier, the model must learn to assign the highest score to the correct class of the input image. This learning process involves adjusting the model's internal parameters, known as "weights" (LeCun, Bengio, and Hinton, 2015). Initially, these weights are randomly initialised with small values. During training, they are refined through an optimisation process called gradient descent, which systematically minimises the difference between the model's output scores and the target scores. This difference is quantified using an objective function (also referred to as the loss function). The objective function directs the weight adjustments to minimise this distance, improving the model's overall performance (LeCun, Bengio, and Hinton, 2015).

The application of deep learning in radiology provides several key advantages. A significant benefit lies in its ability to automatically learn critical features from data without requiring pre-defined visual features extracted by human experts. This allows deep learning models to extract and characterise important phenotypic features of tissues directly from imaging data (Hosny et al., 2018). Without the constraints of human-defined rules, deep learning systems have the capacity to uncover novel features that may not be immediately apparent to human observers. When applied to radiology, this capability enables the identification of disease-specific patterns within medical images, offering unique insights that are especially valuable for detecting newly emerging diseases with unclear or atypical radiological features, such as COVID-19.

Another notable advantage of deep learning in radiology is its quantitative nature. Radiological evaluation has traditionally been viewed as a qualitative and hypothesis-driven process. Radiologists rely on subjective reasoning, which is often influenced by their personal experience, education, and the clinical context of the case provided by non-imaging information (Hosny et al., 2018). As a result, interpretations can vary between practitioners, leading to potential inconsistencies. Deep learning, on the other hand, facilitates a more quantitative assessment of medical images. By automatically identifying and analysing features and patterns within the image, deep learning models can provide an objective, measurable evaluation that supports clinical decision-making (Prior et al., 2020). Quantitative analysis enabled by deep learning not only enhances the reproducibility of findings but also improves the reliability of medical diagnoses. This is particularly important in complex diagnostic cases where the radiological presentation of a pathology is subtle or overlaps with other conditions.

However, the growing promise of computer vision in healthcare brings with it increasing concerns regarding safe and reliable clinical implementation. These concerns primarily revolve around three critical aspects: data quality, model interpretability, and model fairness.

- **Data Quality**: The performance of computer vision models is intrinsically tied to the quality, quantity, and diversity of the training data. Medical imaging datasets often suffer from biases such as imbalances in patient demographics, disease prevalence, and imaging techniques. Poorly curated datasets can lead to models that generalise poorly in real-world clinical settings, particularly when deployed across different healthcare systems or populations. To mitigate this, rigorous data acquisition, cleaning, and labelling processes, as well as diverse and representative datasets, are essential (Mittermaier, Raza, and Kvedar, 2023).

- **Model Fairness**: Bias in computer vision models represents a critical risk, as models trained on non-representative data can inadvertently propagate and exacerbate health disparities. For example, models trained predominantly on images from one demographic group may underperform when applied to others, leading to inaccurate or inequitable outcomes. Ensuring fairness requires ongoing efforts to identify and mitigate bias through diverse, well-balanced datasets and robust evaluation frameworks (Drukker et al., 2023).

- **Model Interpretability**: The "black box" nature of many deep learning models poses significant challenges for clinical adoption. These so-called black box models can produce highly accurate predictions, but the underlying reasoning behind their outputs often remains opaque to users. This lack of interpretability is a major concern in healthcare, where trust, transparency, and explainability are crucial for clinical decision-making. Clinicians need to understand how a model arrives at its conclusions to ensure they align with medical reasoning (Saporta, Gui, Agrawal, et al., 2022).

The use of "black box" models that generate unexplainable predictions compounds these issues, as they increase the risk of model bias, poor generalisability, and silent errors—errors that may go undetected during clinical practice. For instance, a model might perform well under controlled conditions but fail when exposed to variations in imaging protocols, equipment, or patient populations. Such failures can have serious consequences, particularly in high-stakes medical applications where diagnostic errors can impact patient safety and outcomes (Drukker et al., 2023). To address these challenges, a multi-disciplinary approach, involving collaboration between AI researchers and clinicians, is essential. This includes developing rigorous validation protocols and fostering transparency in model development and prediction.

In summary, while deep learning-based computer vision holds immense promise for transforming medical image analysis, its success depends on addressing critical concerns around data quality, model interpretability, and model fairness. By overcoming these challenges, computer vision can assist humans to improve diagnostic accuracy, and ultimately enhance patient care, paving the way for a more efficient and equitable healthcare system.

2.2 The Respiratory System

The respiratory system is a complex network of organs and structures responsible for the vital exchange of oxygen and carbon dioxide between the body and the environment. This system plays a crucial role in maintaining homeostasis, supporting cellular respiration, and facilitating the removal of metabolic waste products. The system includes the nasal cavity, pharynx, larynx, trachea, bronchi, and lungs. Each component contributes to the efficient passage, filtration, and conditioning of inhaled air before it reaches the alveolar surfaces where gas exchange occurs. The structural integrity and functional efficiency of the lungs are paramount for respiration and overall health. Disruption to the lung structure or mechanism of gas exchange gives rise to pulmonary disease (Haddad and Sharma, 2023).

As a critical function, any respiratory system impairment or disease can have a devastating impact on quality of life. Lung diseases, ranging from chronic conditions like asthma and chronic obstructive pulmonary disease (COPD) to acute infections such as pneumonia and tuberculosis, present a substantial global health burden (Collaborators, 2020). Early detection and accurate diagnosis of these conditions are essential for effective treatment and management, ultimately improving patient outcomes. Advancements in medical imaging technologies have revolutionised the field of respiratory medicine, providing non-invasive, detailed visualisation of lung anatomy and pathology. Techniques such as chest radiography, computed tomography, magnetic resonance imaging, and positron emission tomography have become indispensable tools in the detection, diagnosis, and monitoring of lung diseases (Hosny et al., 2018). These imaging modalities offer critical insights into the presence, extent, and nature of pulmonary abnormalities, aiding clinicians in making informed decisions regarding patient care.

Many pulmonary diseases are closely interlinked, the presence of a primary disease often gives rise to a secondary disease, with shared pathophysiology increasing the likelihood of co-occurrence. In real-world clinical settings, it is common for patients to present with multiple co-occurring lung conditions simultaneously, which complicates diagnosis, treatment, and prognosis. For example, a patient with COPD may also develop pneumonia, or individuals suffering from COVID-19 may show signs of secondary infections, such as bacterial pneumonia, or long-term lung damage, such as fibrosis. These overlapping conditions often manifest with shared clinical symptoms (e.g., shortness of breath, chest pain) and radiological features, such as opacities, nodules, or consolidations on CXRs. The presence of co-occurring pathologies increases diagnostic complexity and uncertainty for human experts, for computer vision algorithms to assist they must perform reliably where there is co-occurrence (Fabbri et al., 2023; Putcha et al., 2015). However, the co-occurrence of pulmonary diseases makes medical image classifiers vulnerable to learning spuriously correlated features, relying on ‘shortcut’ features to make predictions. The risk of shortcut learning is heightened in cases of co-occurring diseases because of their complex and subtle interplay in radiological images (Ong Ly, Unnikrishnan, Tadic, et al., 2024).

Explainable models can help identify when predictions rely on spurious correlations rather than clinically relevant features, enabling detection of model reliance on shortcut learning. Furthermore, interpretability mechanisms can reveal how different pathologies interact and manifest in radiological images, potentially uncovering new insights into disease co-occurrence patterns. This is particularly valuable in complex cases where multiple conditions may present with overlapping or atypical features, or for newly emerging diseases. By facilitating a transparent decision-making process, explainable AI systems can help clinicians validate model predictions, identify potential biases, and consider model decisions in the context of their own clinical expertise. This capability is critical for accurately detecting diseases in clinical populations, where comorbidities are the norm rather than the exception.

2.3 Thesis Aims & Objective

This thesis explores the applications of computer vision to medical imaging tasks, including, disease detection, classification and segmentation. I aim to develop explainable, trustworthy computer vision models that perform reliably in the presence of data biases and co-occurring pulmonary diseases. Enhancing interpretability not only improves diagnostic support but also bridges the gap between artificial intelligence and clinical decision-making, ensuring models are aligned with expert reasoning and clinical needs.

I aim to develop reliable, generalisable, and interpretable deep learning models capable of detecting co-occurring pulmonary pathologies in CXRs. By evaluating open datasets and their associated limitations, I aim to highlight the risks of data biases and the need for real-world, meta data-rich clinical datasets. Through extensive multi-centre benchmarking of binary COVID-19 classifiers, I aim to evaluate risk of bias in models trained on national hospital datasets. I do this with the larger motivation of demonstrating the importance of mitigating spurious correlations.

To improve both the performance and interpretability of pulmonary disease diagnosis, I propose advanced deep learning frameworks that move beyond single-disease models. Specifically, I introduce novel approaches leveraging variational autoencoders (VAEs) and hierarchical VAEs for multi-label classification. These methods enable the simultaneous prediction of multiple pathologies, providing a more comprehensive understanding of co-occurring diseases. Additionally, I address the critical need for diagnostic transparency by developing explainable AI techniques that generate interpretable visual explanations and facilitate precise disease localisation.

By focusing on multi-pathology detection and interpretability, I aim to contribute to the development of deep learning solutions that are better suited for real-world clinical applications.

2.3.1 Thesis Structure & Outline

I present the methods proposed and work undertaken in this thesis in Chapters 3 - 7. I give concluding remarks and discuss future directions in Chapter 8. In Chapter 3 I describe the key deep learning methods used in my research. I evaluate the risk of using open source CXR data for the development of binary COVID-19 classifiers as it pertains to blind shortcut learning and poor research practice in Chapter 4. I further my investigation of model bias in Chapter 5 by conducting a multi-centre benchmarking of deep learning COVID-19 classifiers trained on real-world, multi-site hospital data.

In Chapter 6 I introduce a novel approach for explainable, multi-label prediction of CXRs, using a sparse prior VAE model. I validate this model to assess if lung pathology features can be localised through visual explanations. Chapter 7 explores the application of hierarchical VAEs (HVAEs) to overcome the limitations identified in earlier chapters. I present a comprehensive evaluation of these models' capacity for precise explanations, quantifying disease localisation against radiologist annotations and comparing results with established methods like Grad-CAM++. I also further my work on sparse prior distributions and their impact on disease feature isolation in variational approaches.

Chapter 3

Methods

In this section, I provide a detailed overview of the deep learning models and core principles applied throughout my research. I begin by introducing key concepts in computer vision, including convolutions, convolutional neural networks, and residual networks. I then explain the theory behind variational autoencoders (VAEs) and hierarchical variational autoencoders (HVAEs). By leveraging these methods, I aim to develop models capable of both explainable prediction and the localisation of lung pathologies. Specifically, I plan to utilise VAEs and HVAEs to isolate disease features within CXRs, with the goal of offering precise visual explanations of medical image predictions. I contextualise my use of VAE approaches by introducing the key concepts and theories behind interpretable VAEs and exemplify these by explaining key methods in this area of research.

3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of deep learning network that have become essential to computer vision. CNNs are designed to process imaging data, where an image is represented by 2D or 3D arrays. To learn visual patterns and spatial hierarchies CNNs require a combination of convolutional and pooling layers. The architecture of a simple CNN classifier is presented in in Figure 3.1. CNNs are structured in stages, which comprise groups of layers. At its core, CNNs require only three main types of layer: the convolutional layer, the pooling layer, and fully connected (FC) layer. Each layer in the architecture serves a purpose (Goodfellow, Bengio, and Courville, 2016). Convolutional layers extract image features by applying a set of filters (or kernels). In this operation, called a convolution, each filter is passed over the input image, and with each step a dot product is calculated between the filter and image patch it covers. As the network is optimised filter values are updated. With optimisation of the filters, I expect the output of this operation to be a set of feature maps that highlight certain image features. Early convolutional layers typically highlight low-level features, such as, edges, textures. By applying pooling layers, which reduce the spatial dimensions of its input, features maps can be abstracted and combined by subsequent convolutional layers. By stacking a series of convolutional and pooling layers the model is able to learn complex, high-level features. The relationship between these high-level feature is learned by fully connected layers. The final fully connected layer produces the class scores, which are then normalised to a probability space with the *softmax* transformation,

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

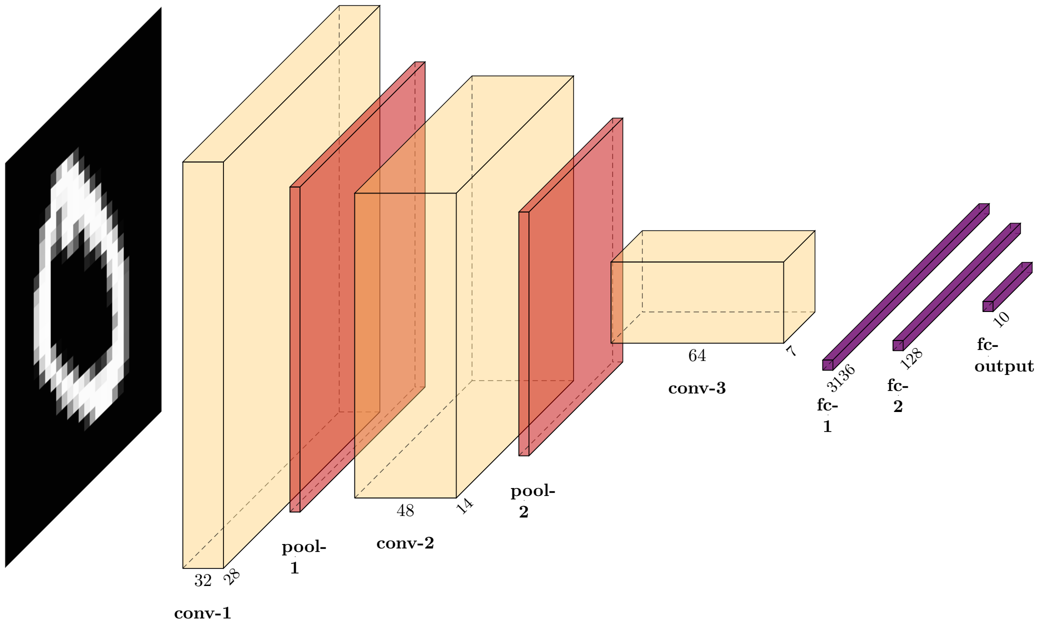


FIGURE 3.1: **The arrangement of layers in a typical deep CNN architecture.** Annotations describe the type and order of each layer i.e. 'conv-1' is the first convolutional layer. Each layer is also labelled with its corresponding dimensions. The example input image is a sample from the MNIST dataset, which contains 10 classes of handwritten digits.

Here I describe the example of a simple CNN shown in Figure 3.1 and the role played by each layer. The input image (28x28x3) holds the raw pixel value of the images, where each image has a resolution of 28x28 and pixel values for each of the three RGB colour channels. The convolutional layer learns features from the previous layer through use of computational operations called convolutions, an example convolution operation is presented in Figure 3.2. The convolution operation itself requires an element called a filter (also referred to as kernel), which is a matrix of weights. The filter is passed over the image until the entire image is traversed. For each stride across the image a matrix multiplication is performed between the region of the image and the filter matrix, producing a convolved feature or feature map. The number of feature maps I generate depends on the number of filters I apply. In this example I apply 32 filters in the first *convolutional* layer, producing 32 feature maps. Convolutional layers are typically paired with an activation function, like ReLU, which is important for introducing non-linearity to the model (Goodfellow, Bengio, and Courville, 2016). The ReLU function is shown below in equation 3.1.

$$ReLU(x) = max(0, x) \tag{3.1}$$

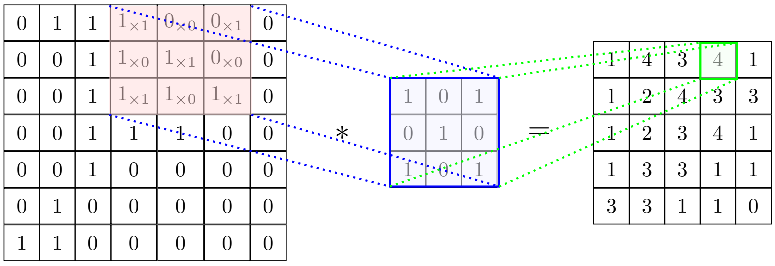


FIGURE 3.2: **An example of a convolutional operation, with a 7x7 input and 3x3 filter.**

A *pooling* layer is applied following the *convolutional* layer. This reduces the spatial size of the convolved feature map, reducing the number of parameters to learn and thus computational cost of model training. In this example (Fig. 3.1), the first pooling layer reduces the size of the feature maps from 28x28 to 14x14. There are two main types of pooling used in CNNs, max pooling and average pooling. Max pooling operations compute the maximum value from the area of the input covered by the pooling kernel, and average pooling returns the average value. Max pooling operations work well in CNNs, they help to suppress noise and learn higher level features (Goodfellow, Bengio, and Courville, 2016).

To complete the CNN classifier, a fully connected layer is required. Neurons in a fully-connected (FC) layer are each connected to all activations in the previous layer. In order to combine fully connected layers with convolutional layers, the feature maps produced by the final convolutional layer are flattened into 1-dimension and passed to the first FC layer. This is an efficient way of learning non-linear combinations of the high-level features extracted by the final convolutional layer. The dimensions of the final FC layer are equivalent to the number of classes in the classification task, this layer produces an output of scalar values that represent class scores. In this example, the final FC layer has the dimensions $1x1x10$, and so will produce 10 scalar values, one for each class. A softmax operation is typically applied to the output to allow us to represent these discrete values as a probability distribution (Goodfellow, Bengio, and Courville, 2016).

Under supervised conditions, CNNs can be trained using an objective function, such as, negative log likelihood (NLL) loss. NLL represents the error of a given classifier, by comparing the predicted probability of each class with the ground truth class of the image. Loss can be minimised during model training, through a process of gradient descent. Not all layers have parameters that can be optimised during training, for example, max pooling is only defined by hyperparameters, and activation operations, such as ReLU, are not parameterised at all. However, importantly, the parameters, or weights, of the convolutional layer filters and fully connected layers are learnable. Meaning, during model training, these are adjusted to optimise the learning of important image features (Goodfellow, Bengio, and Courville, 2016).

The development of CNNs has spurred on significant improvements in computer vision, inspiring the use of deep learning in medical imaging tasks, such as, image segmentation, image fusion, computer-aided diagnosis and prognosis, lesion detection, and many more (Esteva et al., 2021). CNNs, in one form or another, appear essential to deep learning tasks for medical image analysis. Whether used as the sole deep learning technique or incorporated into a larger strategy, CNNs have been applied to a number of important tasks in this field, including, risk stratification of COVID-19 patients, and lesion detection and segmentation from chest X-rays or CTs.

3.1.1 Residual Networks

Residual Networks, or ResNets, were created to address the vanishing gradient problem, a phenomenon in which gradients become exceedingly small as they backpropagate through multiple layers. This phenomenon is commonly observed in CNNs as many layers are required for processing of higher resolution images. By introducing residual connections, or skip connections, these networks can learn more efficiently and go much deeper. During model training, gradients are computed and backpropagated through the many layers of a deep learning model, as they are passed backward each layer introduces a multiplication operation through the network's weight matrices, which can lead to gradients approaching zero. This leads to slow convergence during training and prevents deep networks from capturing hierarchical features and learning meaningful representations. ResNets provide a solution to this problem through residual learning. Residual transformations are applied to create shortcut, or skip, connections which allow the direct flow of information from input to output. Figure 3.3 shows a residual connection in between convolutional layers of a CNN. Mathematically, the residual block can be represented as $y = F(x) + x$, where x is the input, $F(x)$ represents the transformation learned by the layers within the block, and y is the output. This formulation ensures that even if the layers in $F(x)$ contribute little or no change, the identity transformation x can still be directly propagated through the shortcut connection (Fig. 3.3). Consequently, gradients can flow more easily through the network, particularly in the residual path, facilitating the training of very deep architectures. This means medical images can be processed at higher resolutions, revealing subtle details that were previously hard to detect.

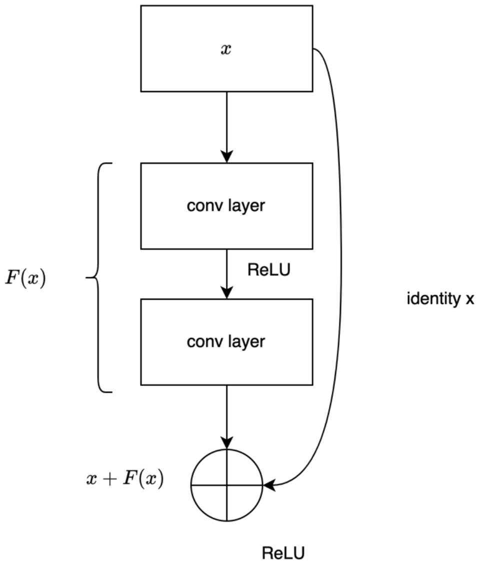


FIGURE 3.3: **Residual connection in convolutional neural network.**

3.1.2 U-Net

U-Net is a type of CNN specifically designed for image segmentation tasks, where image pixels are predicted to belong to different classes (Ronneberger, Fischer, and Brox, 2015). Originally developed for biomedical image segmentation, the U-Net has since been applied to a wide range of segmentation problems. The U-Net consists of an encoder and decoder, which combine with skip connections to give the U-Net architecture it's characteristic "U" shape. In medical image analysis U-Nets have been applied to a host of tasks and have become the standard in image segmentation tasks. The encoder, which consists of repeated convolution and pooling operations, extracts complex image features. The decoder consists of upsampling operations and convolutional layers, it increases the spatial dimensions and combines high-level features with corresponding low-level features from the encoder. Skip connections link corresponding layers of the encoder and decoder through feature map concatenation. Similar to residual connections, these skip connections propagate information from the earlier layers to the deeper ones, combining high-level features with corresponding low-level features, and helping to mitigate the 'vanishing gradients' problem. Figure 3.4 shows the architecture of a typical U-Net model. The U-Net++ extends this approach by adding encoder and decoder subnetworks that are connected through a series of nested, dense skip pathways and can be trained under deep supervision, where each nested U-Net is optimised to its own objective function (Zhou et al., 2018).

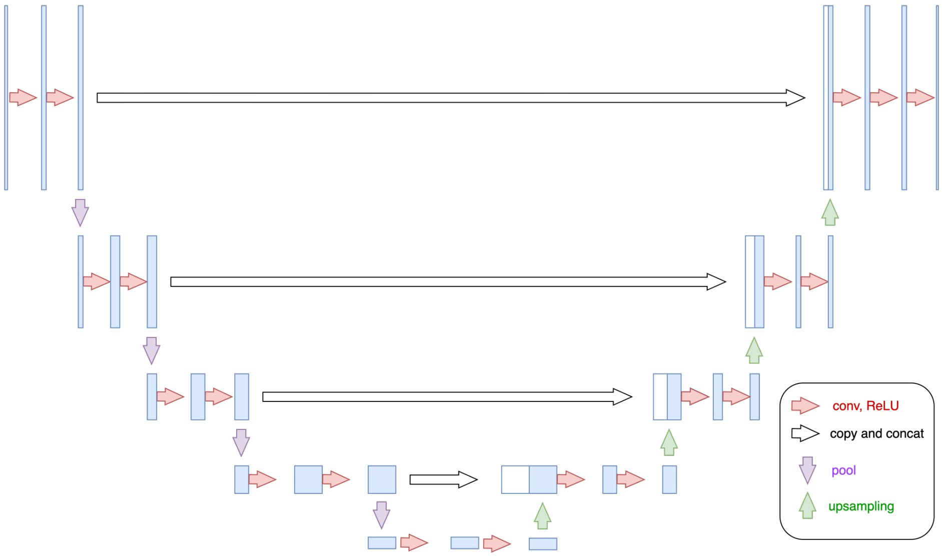


FIGURE 3.4: **U-Net model architecture.**

3.1.3 Monte Carlo Dropout Uncertainty Estimation

Dropout is a regularisation technique where, during training, randomly selected neurons are ignored, or dropped out, which prevents the network from becoming overly reliant on specific neurons and promotes the learning of robust features. Monte Carlo Dropout extends the dropout technique by enabling dropout during inference to obtain multiple stochastic forward passes through the network. Each pass results in a slightly different prediction due to the random dropout of neurons. The uncertainty of prediction can be estimated by computing the variance of the predictions (Gal and Ghahramani, 2016). This approach is reasoned through principles of Bayesian statistics where uncertainties are expressed as distributions and properties of the distribution reflect uncertainty i.e., high variances indicates greater uncertainty.

3.2 Variational AutoEncoders

First introduced by Kingma, Welling, et al. (2019), the Variational AutoEncoder (VAE) is a class of likelihood-based generative model that merges the autoencoder framework with variational inference. Like autoencoders, VAEs consist of an encoder and decoder which are used together for data compression tasks i.e., the encoder maps data into a lower dimensional latent representation, then the decoder projects the latent representation to recreate, or reconstruct, the input. Intuitively, the optimisation of the reconstruction task pushes the model to learn efficient lower-dimensional representations, where the salient features of the data are preserved in the encoding. By compressing the input into a compact latent space, the model filters out irrelevant details or noise while retaining essential information required for accurate reconstruction. This process enables the model to focus on the key patterns and structures in the data, ensuring that the latent representation captures its most meaningful characteristics.

Variational inference is used in Bayesian statistics as a method to approximate complex posterior distributions that arise when applying Bayes' theorem. In many real-world problems, the exact posterior distribution $p(\theta|\mathcal{D})$ is computationally intractable due to high-dimensional integrals or large datasets. Variational inference addresses this challenge by framing the problem as an optimisation task: it approximates the true posterior $p(\theta|\mathcal{D})$ with a simpler, tractable distribution $q(\theta;\phi)$, parametrised by ϕ. The quality of the approximation is measured using the Kullback-Leibler (KL) divergence, which quantifies the difference between $q(\theta;\phi)$ and the true posterior. With its reliance on variational inference, VAEs can be described as latent variable models (see below for further explanation).

By using variational inference to learn to represent data with probability distributions, the VAE gains capacity for novel generation i.e., creating new examples not seen in the training data. This capacity for novel generation is made possible through distributional variance, where the same image is compressed to representations that differ with each sampling. This pushes the model to learn to structure representations efficiently, i.e., similar images are positioned close together in the distribution and different images far apart. However, this also leads to blurriness in the generated images due to a overlapping latent representation between different data samples.

VAEs as probabilistic autoencoders Unlike the 'classic' deterministic autoencoder, the VAE is probabilistic with its optimisation relying on variational inference, hence the name 'variational autoencoder'. The encoder, also called the inference model, learns to map observed, complex $\mathcal{D}$-dimensional data distributions to a lower-dimensional space, the distribution for which is typically much simpler (i.e., multivariate Bernoulli, multivariate Gaussian, or multinomial Dirichlet). The generative process requires that I draw sample $z \in \mathbb{R}^J$ where $J < \mathcal{D}$ from some parametric distribution, such as a multivariate Gaussian distribution,

$$z \sim \mathcal{N}(0, \mathcal{I}) \tag{3.2}$$

I use the decoder to reconstruct the data $\hat{x}$. The decoder is a deterministic function that I use to map sample z to a set of parameters ψ that define another distribution (e.g., data likelihood distribution) that I use to sample $x \in \mathbb{R}^{\mathcal{D}}$. For the task of image generation this distribution is typically either Gaussian or Bernoulli, depending on whether pixel values are continuous or discrete.

VAEs as latent variable models VAEs can be considered deep latent variable models, for which I apply the assumption that there exists some hidden or 'latent' variable z that generates observation x. As only x is observed, I need to infer the characteristic of z, or in other words I need to compute $p(z|x)$. I do this by applying Bayes' theorem,

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \tag{3.3}$$

However, computing $p(x)$ presents a problem, its solution is typically intractable.

$$p(x) = \int p(x|z)p(z)dz \tag{3.4}$$

Instead, I can approximate the posterior distribution $p(z|x)$ with $q(z|x)$ using variational inference i.e., I define tractable distribution $q(z|x)$ and define its parameters in such a way that it is very similar to $p(z|x)$,

$$q_\phi(z|x) \approx p_\theta(z|x) \tag{3.5}$$

where ϕ and θ define the parameters of the distributions. To ensure similarity of the true and approximate posterior distributions, I can minimise the KL divergence, where this term is a measure of difference between two distributions.

$$\min \mathrm{KL}(q(z|x))||p(z|x)) \tag{3.6}$$

The derivation of this term results in the evidence lower bound (ELBO) term (described in detail in Section 3.2.1).

$$\mathcal{L} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - \mathrm{KL}(q(z|x)||p(z)) \tag{3.7}$$

The parameters ϕ for the approximate posterior distribution $q(z|x)$ are approximated through optimisation of encoder parameters, which is used to perform posterior inference over all data points in the dataset i.e., amortized inference. While the likelihood distribution $p(x|z)$ is defined by the decoder function, with parameters θ. Together with a prior distribution $p(z)$, the decoder $p_\theta(x|z)$ becomes a generative function that learns a complex joint distribution over data points and the latent representation $p_\theta(x, z)$.

3.2.1 Optimisation

The VAE optimisation objective is the evidence lower bound (ELBO), also referred to as the variational lower bound (VLB). ELBO comprises two terms, the first term maximises the expected log-likelihood of the data under the decoder distribution, this captures the model's ability to reconstruct input data. The second term minimises the Kullback-Leibler (KL) divergence between the encoder distribution $q_\phi(z|x)$ and a chosen prior distribution, typically a unit Gaussian. The ELBO term can be formulated as:

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

Often, in the absence of a likelihood distribution, the expected log-likelihood of the generated data is represented by a reconstruction loss term, such as: mean-squared error ($\mathcal{L}2$) loss, mean-absolute error ($\mathcal{L}1$) loss, and binary cross entropy (BCE) loss.

Reparameterisation trick VAEs employ a reparameterisation trick to facilitate gradient-based optimization of ELBO, as is required for model training. The reparameterisation trick decouples the stochastic sampling process from the parameterised network. Here, the latent variables **z** are sampled from a simple distribution i.e., Gaussian with parameters μ and σ which are computed by the encoder,

$$\mathbf{z} \sim \mathcal{N}(\mu_{\text{enc}}, \sigma_{\text{enc}})$$

The reparameterisation trick is then applied to re-express the latent representation. This is mathematically expressed as,

$$\mathbf{z} = \mu + \sigma \odot \epsilon$$

where **z** is the sampled latent variable, μ and σ are the mean and variance vectors of the Gaussian distribution as computed by the encoder and ϵ is a random noise vector drawn from the chosen distribution (i.e., Gaussian noise with zero mean and unit variance). Through reparameterisation, the gradient of the sampling operation can be propagated through the network's parameters μ and σ, allowing optimisation via gradient-based techniques.

Importance-weighted Sampling Importance weighting (IW) is used in variational inference to give a tighter lower bound to the data likelihood, with increasing samples the lower bound approaches the true log-likelihood (Burda, Grosse, and Salakhutdinov, 2015).

Here I consider the use of IW in Gaussian-prior VAE optimisation. During training, multiple samples are drawn from the posterior distributions,

$$(z_1, z_2, z_3, \cdots, z_k) \sim \mathcal{N}(\mu, \sigma)$$

For each sample i an importance weighting w_i is computed, these weights are based on the sample likelihoods under the decoder and prior, i.e.,

$$w_i = \frac{p(x|z_i)p(z_i)}{q(z_i)}$$

Data samples are then generated by taking a weighted average of decoder output,

$$x_{\text{samples}} = \Sigma_{i=1}^{N} w_i \cdot D(z_i)$$

where D represents the decoder function. More weight is given to samples that are more likely under the true data distribution i.e., better reconstructions.

3.2.2 Disentanglement & Decomposition

Disentangled representations are known to improve model interpretability, and can be useful in a multitude of tasks, e.g., controllable data generation through the manipulation of latent variables, counterfactual generation, disease decomposition (separation of normal from abnormal), image registration, semantic segmentation, and classification tasks, etc. (Liu et al., 2022).

While a formal definition is yet to be agreed, disentanglement, as it is commonly referred to in the literature, requires explicit independence between latent factors, i.e. no single latent factor may describe more than one generative factor and no two latent factors may describe the same generative factor (Higgins et al., 2017). Disentanglement of the latent space is achieved when individual latent units describe a single generative factor which is largely invariant to other factors i.e., a VAE trained on a dataset of celebrity faces (CelebA) has learnt to describe a single independent salient feature (e.g., eyes, hair, beard) with a single latent unit (Higgins et al., 2017). In line with the agreed definition, disentanglement metrics are commonly derived from measures of independence between latent dimensions. Typically, disentanglement is evaluated on simple synthetic datasets e.g. dSprites, for which the assumption of independence between generative factors holds true (Matthey et al., 2017). However, I propose that achieving disentanglement, by this definition, is unsuitable for complex datasets like medical imaging. Dependencies exist between the true generative factors of complex images, pursuing disentanglement by imposing independence between generative factors results in the loss of true semantic meaning and thus loss of interpretability.

To redirect work on structured representation learning towards more practical approaches that capture the true generative factors that describe the data, Mathieu et al. (2019) introduce *decomposition* of latent representations. In this, the image is decomposed into salient features within the latent representation, but explicit independence between latent factors is not required i.e, two latent factors may describe the same salient feature. In other words, decomposition is a generalisation of disentanglement that permits dependencies between dimensions of the structured latent representation. At a high-level, decomposition arises by successfully imposing a desired structure on the learned representation. Decomposition can therefore give rise to disentanglement, where the desired structure specifically requires independence between latent dimensions i.e., a multivariate Gaussian with a diagonal co-variance matrix (identity co-variance). Decomposition is achieved where there is appropriate overlap between latent representations, i.e., a complete representation of the data is learned, and where the posterior resembles the prior - with the prior imposing the desired structure over latent factors. Successful strategies for inducing decomposition in the latent space includes clustering of the latent space and use of sparsity-inducing priors. Modelling sparsity in the latent space is motivated by the presence of 'non-coding' latent dimensions in large latent spaces, these dimensions do not carry useful information and makes the latent representation less generalisable and less interpretable (Tonolini, Jensen, and Murray-Smith, 2020).

Here I outline well-known strategies for inducing decomposition and/or disentanglement in the latent space, these include the selective use of prior distributions, semi-supervision, and the application of inductive priors. I explore and/or adapt some of these approaches for work in the explainable prediction of CXRs (see Chapters 6 and 7).

3.2.2.1 β-VAE

β-VAE is a well known approach for disentanglement (Higgins et al., 2017). β-VAE is an extension of the original VAE framework that introduces adjustable hyperparameter β which constrains model optimisation. More specifically, the β value controls the weight of the KL term in the ELBO objective. With $\beta > 1$ the optimisation problem is constrained and the model is forced to learn statistically independent latent units as prescribed by the use of a factorised unit Gaussian prior, i.e., more Gaussian noise is added. Greater restriction on the latent bottleneck pushes the model to learn a more efficient representation of the data. Intuitively, I would expect the model to learn to group composite features into generative factors. Where independence between generative factors is observed, unsupervised disentanglement is achieved. Note that when $\beta = 1$, the β-VAE is equivalent to the original VAE.

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

However, as a consequence of the constrained latent bottleneck β-VAE produces poor quality image reconstructions compared to the original VAE, i.e., improved disentanglement is at the expense of good quality reconstructions (Higgins et al., 2017).

An in-depth study of β-VAE considers the effects of the constrained optimisation framework on the learned posterior and how the pressures exerted on the posterior gives rise to structured latent representations and disentanglement (Burgess et al., 2018). Reducing the KL divergence between prior and posterior reduces the spread of posterior means and increases posterior variances (to resemble the unit Gaussian prior). Ultimately, this means posterior distributions are forced to overlap across different samples of the dataset, which creates confusion between sampled points i.e., where $q(z_1|x_1)$ overlaps with $q(z_2|x_2)$, $\hat{x}_{q(z_2|x_2)}$ can be confused with $\hat{x}_{q(z_1|x_1)}$ and vice versa, and thus leads to worse log-likelihoods. This confusion becomes increasingly likely as the posterior overlap increases i.e., as KL divergence is more greatly enforced.

The cost to log-likelihoods can be minimised by arranging points that are nearby in data space close together in the latent space; so when a predicted data point x_1 is more likely under posterior $q(z_2|x_2)$ close proximity between x_1 and x_2 (in the latent space) prevents significant cost to the log-likelihood. In simple terms, it is beneficial for the VAE to learn to arrange representations of similar data points close together in the latent space, then if there is confusion the log-likelihoods are less affected.

The ongoing incentive to preserve 'locality' throughout optimisation gives rise to disentanglement. Specifically, the alignment of generative factors with independent latent dimensions, i.e., disentanglement, emerges due to pressure from the log-likelihood term. During training, the log-likelihood term incentivises the VAE to first learn features that are most significant for the efficient improvement of log-likelihoods. As training continues, more precise information for this feature is learned, until the log-likelihood plateaus i.e., learning of this feature no longer has a significant impact on the log-likelihood. At this point of plateau, the VAE is pushed to learn another factor of variation in the dataset. To preserve the 'locality' property, this new factor of variation is captured in a new latent dimension (Burgess et al., 2018). With this 'locality' property latent traversals, in which a single latent factor is gradually changed and step-wise latent reconstructions are generated, give rise to small changes in a generative factor of the data which gives the effect of a smooth transition.

3.2.2.2 β-TCVAE

Chen et al. (2018) build on the β-VAE by introducing the β- Total Correlation VAE (β-TCVAE). Assuming a factorised prior, Chen et al. (2018) reformulate ELBO to identify a term that measures the total correlation between latent variables, which quantifies the mutual information between latent variables and the data variable.

$$\begin{aligned}\mathbb{E}_{p(n)}[KL(q(z|n)||p(z)] = KL(q(z,n)||q(z)p(n)) + KL(q(z)||\prod q(z_i)) + \\ \sum_i KL(q(z_i)||p(z_i))\end{aligned} \tag{3.8}$$

where z_i is the ith dimension of the latent representation, and $q(z|n) = q(z|x_n)$ and $q(z,n)p(n) = q(z|n)\frac{1}{N}$, and therefore $q(z) = \sum_{n=1}^{N} q(z|n)p(n)$ is the aggregate posterior, which is an approximation of the latent variables under the data distribution.

Here $KL(q(z)||\prod_i q(z_i))$ is the total correlation term, the term of interest. Where all dimensions of z are independent, the joint distribution across $q(z)$ should equal the product of individual z_i distributions, the KL divergence between these terms should therefore be 0. With increasing correlation between z_i variables, $q(z)$ diverges from $\prod_i q(z_i)$, therefore latent dimension independence decreases as the quantifying divergence term increases.

Chen et al. (2018) induce disentanglement by evaluating $q(z)$, which depends on the full dataset, through a Monte Carlo approximation that relies on a mini batch estimator that weights each batch appropriately (inspired by importance sampling). From here, the total correlation term can be minimised to enforce independence between latent dimensions, facilitating disentanglement. This re-formulation of standard ELBO provides another perspective on how β-VAE achieves disentanglement within the latent space.

3.2.2.3 FactorVAE

FactorVAE also achieves disentanglement through inducing independence between latent dimensions (Kim and Mnih, 2018). Motivated by the poor trade-off between disentanglement and reconstruction observed in β-VAE, FactorVAE introduces a reinterpretation of the ELBO that includes an additional term for learning a factorial marginal distribution, i.e., latent variables are made to be independent, with no affect on reconstruction quality. This term is referred to as the total correlation penalty, and is a popular measure of dependence for multiple random variables,

$$\mathrm{TC} = \mathrm{KL}(q(z) \parallel \bar{q}(z)) = \mathbb{E}_q(z)\left[log\frac{q(z)}{\bar{q}(z)}\right] \approx \mathbb{E}_q(z)\left[log\frac{D(z)}{1-D(z)}\right]$$

where $q(z)$ is the marginal posterior and $\bar{q}(z)$ is the product of all marginals in a batch, $\bar{q}(z) := \prod_{j=1}^{d} q(z_j)$. In this approach $\bar{q}(z)$ itself is approximated by sampling a batch from $q(z)$ and permuting across the batch for each latent dimension. The total correlation (TC) penalty is approximated and minimised with the use of a simple discriminator model D, which is tasked with predicting whether a sample z belongs to the a marginal distribution $q(z)$ or the product of all marginal distributions $\bar{q}(z)$. With this task the classifier approximates the density ratio for estimating total correlation, $log\frac{q(z)}{\bar{q}(z)}$

Crucially, minimisation of total correlation alone is not enough to encourage disentanglement. For example, under latent variable collapse i.e., $q(z|x) = p(z)$, $TC = 0$. Hence, it is crucial to correctly balance the TC term with reconstruction error.

3.2.2.4 Sparsity-driven decomposition

Here I discuss sparse representation learning and its relationship to latent space decomposition. Based on the principle of sparse coding, these approaches aim to learn a sparse posterior, in which the encoder is induced to represent the data in as few active latent variables (non-zero) as possible, with a varying number and differing combination of active latent variables for each data sample (Tonolini, Jensen, and Murray-Smith, 2020). Intuitively, sparse representations are well-suited for learning latent representations of images. Most image pixels are redundant, sparse representations are able to capture key elements while ignoring irrelevant details.

As previously discussed in Section 3.2.2, sparsity can give rise to decomposition of the latent space. While the mechanism for sparsity-induced decomposition has not been formally evaluated in the literature, I suggest that sparsity gives rise to latent decomposition by further constraining the information bottleneck. To preserve the required information for image reconstruction, the model is motivated to push low-level features to sparse areas of the posterior and retain high-level, semantic features in dense areas. Moreover, since the model is restricted to using only a small proportion of active variables, each variable is encouraged to learn a specific, isolated features of the image with little overlap or redundancy between them. This strategy for learning interpretable latent spaces may be more desirable for complex imaging, where the assumption of independence between image features is incorrect.

Within the framework of variational inference, representation of data through sparse posteriors can be achieved with the application of sparse-inducing prior distributions, such as, the Spike-and-Slab probability distribution, a 'sparsified' mixture of Gaussian distributions, and the Dirichlet distribution (Tonolini, Jensen, and Murray-Smith, 2020; Mathieu et al., 2019; Joo et al., 2020)

Tonolini, Jensen, and Murray-Smith (2020) propose to model sparsity in the latent representation with a Spike-and-Slab probability distribution prior VAE. In Bayesian statistics, the Spike-and-Slab prior is used to separate relevant variables, or features, from irrelevant variables. This is achieved by defining latent variables with two different distributions. The "Spike" distribution, typically a Dirac delta function (δ_0) is concentrated near zero and has little-to-no capacity for information coding. The "Slab" distribution is broader, typically Gaussian, and allows non-zero values. The Spike-and-Slab prior is expressed as a mixture distribution, the sparsity parameter controls the proportion of values expected to be zero i.e., the proportion of "Spike" distributions within the mixture distribution. With this approach, Tonolini, Jensen, and Murray-Smith (2020) observed decreased tendency towards posterior collapse, and present an analysis of latent variable activation the shows improved axis-alignment of generative factors.

Joo et al. (2020) introduce the Dirichlet-prior VAE for sparse representation learning. Parameterising the Dirichlet prior distribution with α values less than one creates a sparse, multi-modal distribution. Therefore, when regularising the VAE to this prior it is forced to learn a sparse posterior over latent variables. Classification and clustering-based evaluations suggest the Dirichlet-prior VAE learns superior latent representations to the Gaussian-prior VAE. Note that in works modelling on the Dirichlet prior, the Dirichlet distribution is often approximated by applying a softmax function to a unimodal Gaussian. However, this transformation cannot generate a multi-modal distribution and is therefore a poor approximation of a multi-modal Dirichlet distribution.

Moreover, Mathieu et al. (2019) model a sparse posterior using a mixture of

Gaussian distributions, where a narrow Gaussian component pushes latent variables towards zero. This prior is defined as,

$$p(z) = \prod_i (1-\lambda)\mathcal{N}(z_i; 0, 1) + \lambda\mathcal{N}(z_i; 0, \sigma_0^2),$$

with $\sigma_0^2 = 0.05$. A proportion of this mixture of Gaussian distributions is 'turned off', this is controlled by the λ parameter. Where $\lambda \gg 0$, a significant proportion of Gaussians within the mixture are pushed towards zero, creating a sparse prior distribution. Mathieu et al. (2019) use $\lambda = 0.8$ and observe successful decomposition of the latent space without a notable decrease to reconstruction quality.

3.3 Hierarchical VAEs

Hierarchical Variational Autoencoders (HVAE) have been employed as generative models across various domains, including medical imaging, where they excel in capturing complex, multi-scale data distributions. In medical imaging, HVAE have been utilised for high-fidelity image reconstruction, unsupervised anomaly detection, and the generation of realistic synthetic data to address data scarcity in training pipelines (Maaløe et al., 2019; Dorent et al., 2023; Havtorn et al., 2021). By leveraging their hierarchical latent structure, HVAE effectively represent the inherent multi-resolution features and dependencies present in modalities such as CT scans, MRIs, and CXRs, facilitating advanced applications in automated diagnosis, disease progression modelling, and treatment outcome prediction. For example, Dorent et al. (2023) use a HVAE, specifically the Nouveau VAE (NVAE), for the synthesis of missing images from various medical imaging modalities, such as ultrasounds. They extend the principal of multi-modal VAEs with a hierarchical latent structure and apply adversarial learning to generate sharper images. Moreover, Biffi et al. (2020) learn a hierarchy of conditional latent variables that both models anatomical segmentations and enables the classification of distinct clinical conditions. The highest stochastic level of the deep hierarchical VAE is specialised for the classification of clinical conditions, and the generative model is used to facilitate explainable prediction. This works follows the Ladder VAE formulation.

Hierarchical VAEs (HVAEs) extend the basic VAE framework by introducing a hierarchy of L stochastic latent variables $z = z_1, \cdots, z_L$. The hierarchy is ordered such that each stochastic level is conditioned on the level above i.e., $z = \{z_1, \cdots, z_L\}$ and $q_\phi(z_1, \cdots, z_L|x)$, so $L = 1$ is equivalent to the basic VAE framework. This process requires ancestral sampling, where the sampling of z_k is dependent on the prior sampling of its parent stochastic variables i.e., z_{l-1} (Kingma, Welling, et al., 2019). Figure 3.5 shows examples of different types of HVAEs.

Typically, Gaussian-prior VAEs assume latent variable independence and are therefore incapable of capturing dependencies between factors of variation in the data. By introducing a multi-level framework, where dependencies between the lower-level latent variables is permitted, HVAEs are able to learn more flexible posterior distributions that may better approximate the true posterior i.e., they are not restricted to diagonal multivariate Gaussians (Havtorn et al., 2021).

Hierarchical generative models are typically defined in a top-down manner (Havtorn et al., 2021),

$$p_\theta(x|z) = p(x|z_1)p_\theta(z_1|z_2)\cdots p_\theta(z_{L-1}|z_L),$$

where L is the top level.

While the inference model can be defined from either the bottom-up,

$$q_\phi(z|x) = q_\phi(z_1|x)\prod_{i=2}^{L} q_\phi(z_i|z_{i-1})$$

or the top-down,

$$q_\phi(z|x) = q_\phi(z_L|x)\prod_{i=L-1}^{1} q_\phi(z_i|z_{i+1})$$

Choice of inference model dictates information flow through the model. For simple HVAEs (like multi-level VAEs), at inference information flows in the bottom-up direction (BU) and at generation information flows in the top-down (TD) direction (Fig. 3.5a). To draw comparison with the deterministic autoencoder framework, the BU direction encodes data while the TD direction decodes data.

Typically, HVAEs are much larger models than VAEs, requiring many more parameters. Generally, the dimensionality of the lowest level stochastic variable within the hierarchy is similar to the number of pixels of the input image, and the dimensionality of the latent representation exponentially decreases further up through the hierarchy (Dorent et al., 2023). Therefore, low level features are typically captured by bottom levels of the hierarchy. While composite, higher level features are captured by top levels of the hierarchy.

A deep hierarchy of stochastic latent variables allows a much more expressive variational model, and facilitates a better approximation of the true posterior. However, modelling several layers of dependent stochastic variable presents a number of challenges for training, often higher stochastic levels become too ‘noisy’ and are vulnerable to latent variable collapse, or posterior collapse, and therefore become unstable to train[1]. Subsequent research has aimed to increase the number of trainable stochastic variable levels with the view to improve log-likelihoods.

Major advancements in HVAE performance have been achieved through implementing more complex information flow, namely, the addition of top-down inference. Top-down dependencies are already observed in the generative distribution i.e., information is passed from the highest level of abstraction (top-level stochastic layers) down to observable variables.

---

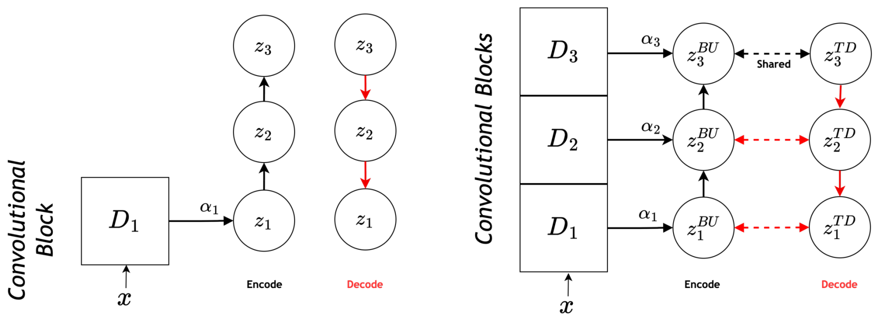
[1] Posterior collapse can lead a latent variable to become inactive. A latent variable can be defined as collapsed/inactive if the KL divergence between the posterior and prior is $\approx$0.

Top-down inference mirrors these dependencies i.e.,

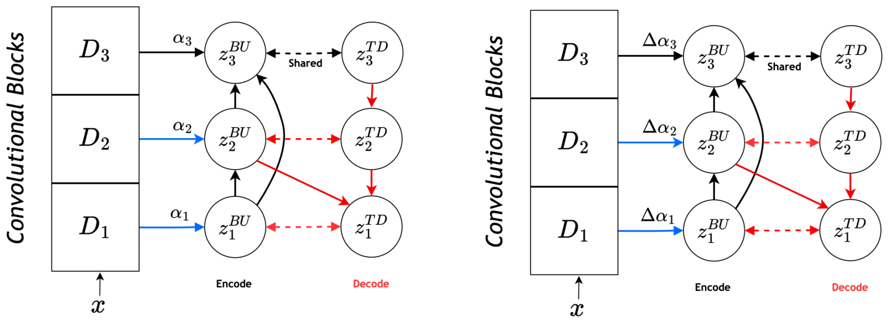
$$Q(z_1, z_2|x) = q(z_1|z_2, x)q(z_2|x)$$

and permits a shared top-down pathway connecting inference and generative models through a shared parametrisation. Alongside this a deterministic path takes information from observable variable x to the top latent variables. This approach supports much larger hierarchies without posterior collapse, and is a key feature of LadderVAEs and Bi-directional inference VAEs (BIVA) (Sønderby et al., 2016b; Maaløe et al., 2019). Further to this, research has explored the use of deep learning engineering techniques and training strategies for optimal generative performance (Vahdat and Kautz, 2020).

In this section, I introduce models that represent these major steps in the advancement of HVAEs, these are: the Multi-level VAE, the Ladder VAE, Bi-directional inference VAE (BIVA), and the Nouveau-VAE (NVAE) (Havtorn et al., 2021).



(a) Multi-level VAE (b) Ladder VAE



(c) BIVA (c) NVAE

FIGURE 3.5: **A L=3 layered example of hierarchical Dirichlet-prior VAEs: (a) Multi-level VAE, (b) Ladder VAE, (c) BIVA and (d) NVAE.** Dirichlet posterior distributions are parameterised by α values. Δ indicates the use of residual parameterisation of posterior distributions, as in NVAE. Red arrows show information flow for image generation. Blue arrows show how deterministic parameters are shared between generative and inference models, this describes the additional deterministic top-down pathway applied in BIVA and NVAE. *Abbrvs: Variational AutoEncoder (VAE); Bidirectional inference VAE (BIVA); Nouveau VAE (NVAE).*

3.3.1 Multi-level VAE

The multi-level VAE is the most simple extension of a VAE. Here, I define a multi-level VAE as a VAE with multiple stochastic levels and separate computation of inference and generative distributions (Fig. 3.5a). In this, top-down information is incorporated indirectly through conditional priors in the generative model only. With only this direct information pathway the multi-level VAE can only support a maximum of two stochastic layers before becoming prone to posterior collapse (Maaløe et al., 2019).

I introduce an example of a two-level VAE, made up of latent variables z_1 and z_2. For this HVAE the generative process is simple, relying solely on ancestral sampling. z_2 is sampled first, then z_1 is sampled given z_2 and finally x is sampled given z_2. Variational inference is used to approximate the posterior, $q(z_1, z_2|x)$ with KL divergence computed per layer. However, previous research shows that with a powerful decoder and random initialisation of latent variables, the model will likely find the optimum of the top-level KL term $KL[q(z_2|z_1)||p(z_2)]$. This means the top-level will collapse to the prior, i.e., $q(z_2|z_1) \approx p(z_2) \approx \mathcal{N}(0, 1)$, and the HVAE is effectively reduced to a single-level. As more levels are added to the hierarchy, this phenomenon becomes more likely.

3.3.2 Ladder VAE

Sønderby et al. (2016b) introduce a new strategy for information flow through the stochastic levels of Hierarchical VAEs. They introduce the Ladder VAE (LVAE) by proposing a new inference model that combines a Gaussian likelihood $p(x|z)$ with the generative model $p(x, z)$ by sharing parameters between inference and generative models and adding a top-down inference method (Fig. 3.5b). In this process, the likelihood distribution is first approximated with a deterministic upward pass i.e., images are deterministically embedded into feature space and distribution parameters are estimated from this. This is followed by a stochastic downward pass, in which stochastic latent variables are ancestrally sampled to recursively compute the approximate posterior,

$$q(z|x) = q(z_L|x) \prod_{i=1}^{L-1} q(z_i|z_{i+1})$$

and generative distribution $p(x, z)$. Note that while the inference model is made more complex, the generative model of LVAE follows the same information flow as a multi-level VAE i.e., simple ancestral sampling.

The Ladder VAE achieves much better log-likelihoods, supporting active latent variables in up to 5 levels of stochastic layers. However, despite significant progress, LVAEs are still challenging to train, top-level stochastic latent variables have a tendency collapse to the prior.. Gradual warm up from deterministic to variational was found to be crucial to prevent posterior collapse in higher layers (Sønderby et al., 2016a).

3.3.3 Bidirectional inference VAE

The Bidirectional inference VAE (BIVA) approach introduces a bi-directional stochastic inference pathway with the goal of further improving the expressivity of the approximate posterior (Sønderby et al., 2016a). BIVA is made up of a deep hierarchy of stochastic variables, using skip-connections to improve information flow and to prevent inactive latent units. As in the Ladder VAE, the bidirectional inference pathway uses stochastic variables in top-down (TD) inference pathways. However, BIVA extends the Ladder VAE by adding a stochastic bottom-up path (Fig. 3.5c). The inference model can therefore be thought of as comprising top-down and bottom-up stochastic latent variables,

$$z_i = \{z_i^{BU}, z_i^{TD}\}$$

where z_i^{BU} belongs to a bottom-up inference path and z_i^{TD} belongs to the top-down path. This factorisation is done at each level of the hierarchy. Information from the BU approximated likelihood $p(x|z)$ is combined with TD information from the generative distributions $p(x,z)$ to give the approximate posterior $q(z|z,x)$. Here, both the generative model and inference model are dependent on top down information flow. In other words, the new inference model recursively corrects the generative distribution $p(x,z)$. First, data flows through the deterministic upward pass to approximate the likelihood distribution $p(x|z)$. Then the stochastic downward pass computes the approximate posterior and generative distribution. The inference model is therefore a combination of BU information an TD information flowing from the prior.

Skip connections between data dependent information and lower stochastic levels are also incorporated, allowing data dependent information to skip stochastic variables lower in the hierarchy. Together, these additional information paths facilitate a highly expressive model that is capable of approximating very complex datasets. BIVA significantly outperforms the Ladder VAE, supporting 20-level hierarchies and improving substantially on Ladder VAE log-likelihoods.

3.3.4 Nouveau VAE

By exploring the best model architecture and training strategies, with Nouveau VAE (NVAE) Vahdat and Kautz (2020) achieve state-of-the-art results among non-auto-regressive likelihood-based generative models (Fig. 3.5d). Through extensive engineering, Vahdat and Kautz (2020) identify three key model design considerations for better log-likelihoods and higher-resolution image generation with a BIVA-based information flow strategy, these are: (1) residual cells, (2) spectral regularisation and (3) residual paramerisation of the approximate posterior for improved optimisation of KL objective.

Residual cells are used in both the inference and generative model to allow the long-range correlations in the data to be learned. Spectral regularisation controls model complexity by directly penalising the spectral norm of model weight matrices, these values typically correspond to weights with high variance. NVAE uses spectral regularisation to prevent the encoder output from changing dramatically with input changes and to stabilise training (Vahdat and Kautz, 2020; Dorent et al., 2023).

Residual parametrisation of posteriors is also key to stabilising HVAE training, Figure 3.6 gives an overview of how this is implemented in NVAE. The deterministic bottom-up path is used to modify the parameters of the top-down path. The deterministic path gives two deterministic variables $r_1 = f_1(x)$ and $r_2 = f_2(r_1)$, where f_1 and f_2 are deep neural networks (DNNs). Additional DNNs are then used to compute modifications to distribution parameters i.e., for a Gaussian NVAE, $[\Delta\mu_i, \Delta\sigma_i^2] = NN_{\Delta i}(r_i)$. Modifications are used to parametrise the variational posteriors for top-down sampling in the generative process. For the top-level of a Gaussian NVAE this can be expressed as, $z_l \sim \mathcal{N}(0 + \Delta\mu_l, 1 \cdot \Delta\sigma_l^2)$, and for subsequent layers as, $z_i \sim \mathcal{N}(\mu_i + \Delta\mu_i, \sigma_i^2 \cdot \Delta\sigma_i^2)$, where $[\mu_i, \sigma_i^2] = DNN_i(z_i + 1)$.

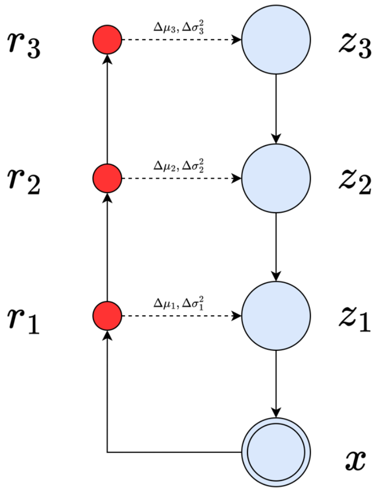

r_3
$\Delta\mu_3, \Delta\sigma_3^2$
z_3
r_2
$\Delta\mu_2, \Delta\sigma_2^2$
z_2
r_1
$\Delta\mu_1, \Delta\sigma_1^2$
z_1
x

FIGURE 3.6: **A graphical diagram of residual parametrisation of variational posteriors in a NVAE with 3 stochastic layers.** At each level, neural networks are applied to learn the modifications of each distribution from r, which represents the deterministic variables for each level. Δ represents residual parametrisation. *Abbrvs: Nouveau VAE (NVAE).*

3.4 Data

I share a number of datasets across multiple chapters, here I introduce these datasets as a reference for subsequent research. I divide these datasets into two categories, COVID-19 CXR datasets and multi-label CXR datasets. The COVID-19 CXR datasets used in this work are exclusively binary or multi-class datasets, meaning each instance can only be assigned to one label.

For each dataset I give detail on labelling protocol (where possible). I consider risk of inaccurate or 'noisy' ground truth with datasets that provide little-to-no information on labelling protocol and datasets which use only one radiologist. In particular, I note that CheXpert labels are extracted from radiologist reports using NLP, adding another level of labelling noise i.e., the report must be entirely accurate and the NLP algorithm must predict accurately. If ground truth labels are provided by expert radiologists, risk of inaccuracy is increased with less experienced radiologists and where multiple classes are considered. Inaccurate labelling can be mitigated by use of expert panels. With multiple labels, majority voting and consensus can be used to derive more accurate labels and provide a measure of label uncertainty.

3.4.1 COVID-19 CXR datasets

The COVID-19 CXR datasets used in this work includes a mix of public datasets and hospital datasets, for which I was granted exclusive access. I consider domain-specific pre-training datasets applied to COVID-19 detection models, namely, CheXpert (as a multi-label dataset CheXpert is described in the relevant section).

COVIDX COVIDX is the largest and most popularly used open-source COVID-19 CXR dataset, comprising 13,975 CXR images across 13,870 patient cases at the time of the use. I use this dataset in Chapter 4 in my evaluation of potential pitfalls in using open-source data to develop medical image classifiers. First introduced in Wang, Lin, and Wong (2020), COVIDX is made up of a number of public data repositions these are, RSNA, CHOWDHURY, COHEN, Figure1, and ActualMed (Cohen, Morrison, and Dao, 2020; Rahman, n.d.; Cohen and Chung, n.d.; IEEE8023, n.d.; Chung, 2020). To form this dataset, the creators combined and modified these five data repositories, leveraging specific patient cases from each of the data repositories. Exclusively COVID-19 CXRs are taken from Figure1, ActualMed and CHOWDHURY repositories. COVID-19 and non-COVID-19 pneumonia CXRs are taken from COHEN. Only non-COVID-19 pneumonia and pneumonia negative CXRs are taken from RSNA. This dataset provides CXRs with corresponding labels for three classes: COVID-19, non-COVID-19 pneumonia and pneumonia negative. Note that data labelling protocols for these sources were not accessible and likely vary across data sources.

RSNA International COVID-19 Open Annotated Radiology Database The RSNA International COVID-19 Open Annotated Radiology Database (RICORD) comprises 998 CXRs from 361 patients located at four international sites (Tsai et al., 2021). All participants are over 18 years old and have received a positive diagnosis for COVID-19. I combine this dataset with CheXpert to create an external test set used to evaluate COVID-19 detection models in Chapter 4.

Leeds Teaching Hospital NHS Trust Data Leeds Teaching Hospital NHS Trust (LTHT), a large teaching hospital based in Leeds, UK, provided a dataset of CXR images of patients alongside PCR test results for COVID-19 diagnosis and non-COVID pneumonia diagnosis. I use data from this source in Chapter 4 and Chapter 5. Access was limited at time of study for Chapter 4, I was provided with a total of 1369 CXRs with COVID-19 labels. COVID-19 label was derived from the outcome of RT-PCR testing, where the swab was taken during a patient's hospital stay. No further patient information was given. For Chapter 5, full data access was granted which provided us with 11,204 CXRs with labelled COVID-19 outcomes. I additionally gained participant demographic data, such as, sex, age and ethnicity data, which I link to CXRs.

National COVID-19 Chest Imaging Database The National COVID-19 Chest Imaging Database (NCCID) is a centralised UK COVID-19 database (Cushnan et al., 2021). Data is collected from 26 hospital centres, totalling 45,635 CXRs from 19,700 patients across the UK in the form of de-identified DICOM image files and header information (at time of access). NCCID provides clinical data associated with imaging, these include the results of RT-PCR tests. Dates for both CXR exams and COVID-19 RT-PCR swabs are provided. COVID-19 outcomes are derived from this information. This dataset is used in Chapter 5 as the training data for COVID-19 detection model benchmarking experiments.

COVID-GR The COVIDGR dataset was developed in collaboration with expert radiologists at Hospital Universitario San Cecilio. It comprises 852 anonymised CXRs, evenly balanced between 426 COVID-19 positive and 426 COVID-19 negative cases. COVID-positive cases were defined by a positive RT-PCR test within 24 hours of the image acquisition. All images were acquired using standardised equipment and protocol, consisting exclusively of postero-anterior chest radiographs (Tabik et al., 2020). This dataset is used in Chapter 5 as a test dataset for benchmarking COVID-19 detection models.

COVID-QU-Ex COVID-QU-Ex was released as a public COVID-19 dataset, the dataset comprises 11,956 COVID-19, 11,263 non-COVID-19 pneumonia, and 10,701 normal CXR images. In addition to CXRs and class labels, ground-truth lung masks were also included. Masks were generated by humans with machine assistance. All

CXRs are either postero-anterior or antero-posterior, i.e., only frontal view CXRs are included. I use this data to train and evaluate a U-Net++ lung segmentation model, described in Chapter 5.

3.4.2 Multi-label CXR datasets

A multi-label dataset refers to a dataset where each instance can be assigned multiple labels simultaneously. This contrasts with single-label classification, where each instance is associated with only one label. The concept of multi-label classification is important for tasks where multiple categories or attributes are relevant for each input. I use a number of multi-label CXR datasets in my work on explainable prediction of co-occurring lung pathologies (Chapters 6 and 7), these include, CheXpert, CheXlocalise, and VinDr-CXR.

Multi-label CXR datasets include classes for common lung pathologies, such as, consolidation, lung opacity or airspace opacity, and pleural effusion. Consolidation refers to the filling of airspaces (like alveoli) in the lung with fluid, inflammatory exudate, or other material, obscuring the margins of vessels and airway walls. As such, consolidation is often observed in the middle-to-lower regions of the lung space (Lee et al., 2013). Pleural effusion is defined as the abnormal accumulation of fluid in the pleural space, which is the thin cavity between the pleural layers surrounding the lungs (Krishna et al., 2024). Lung opacity, or airspace opacity, is a broader descriptor, referring to any area of increased density or whiteness on an x-ray. This term can be used to describe consolidation, pleural effusion, and other conditions like ground-glass opacity (Türk and Kökver, 2023). Support devices are also included as a class in the multi-label CXR datasets considered. Support devices, or medical devices, represent a broad category of therapeutic and monitoring equipment visible on chest radiographs. This term encompasses various implanted or externally placed devices including pacemakers, central venous catheters, endotracheal tubes, chest tubes, and other interventional hardware used for patient care and monitoring Hunter et al., 2004.

CheXpert The CheXpert dataset is a large-scale dataset of CXRs, it contains over 224,000 CXRs from more than 65,000 patients (Irvin et al., 2019). CXRs are labelled for 14 common chest conditions, which are generated using a NLP-based radiology report labelling tool. The CheXpert labels are:

- No Finding
- **Enlarged Cardiomediastinum**: Cardiomegaly
- Lung Lesion
- **Lung Opacity:** Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other

- Fracture
- Support Devices

Labels follow a hierarchical structure, which groups pathologies by radiographic features. For example, observations like "Lung Opacity" encompass several specific findings, such as "Edema" and "Consolidation," which are often interrelated in clinical contexts. Additionally, labels include uncertainty annotations (e.g., positive, negative, uncertain), which reflects the ambiguity of ground truth annotations.

CheXlocalise CheXlocalise is a subset of the CheXpert dataset with radiologist annotations for disease localisation (Saporta, Gui, Agrawal, et al., 2022). The dataset is divided into two annotation sets: (1) ground-truth pixel-level segmentations, which is provided by two board-certified radiologists, and (2) benchmark pixel-level segmentations and most-representative points, which is provided by a separate group of three board-certified radiologists. I use only the ground-truth pixel-level annotations which is annotated on the validation and test set of CheXpert. The validation set contains 234 chest X-rays from 200 patients, while the test set includes 668 chest X-rays from 500 patients. The dataset focuses on 10 pathologies: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Lung Lesion, Airspace/Lung Opacity, Pleural Effusion, Pneumothorax, and Support Devices. As a localisation dataset, the CheXlocalise dataset does not consider the *No Finding* class of CheXpert. I use this dataset in Chapter 7 to quantitatively evaluate visual explanations generated by multi-label prediction models.

VinDr-CXR The VinDr-CXR dataset is a large publicly available dataset of chest radiographs (Nguyen, Lam, Le, et al., 2022). The dataset contains more than 18,000 CXR scans collected from two major hospitals in Vietnam from 2018 to 2020. Each CXR is provided with radiologists' annotations for classification of common thoracic lung diseases and localisation of disease findings. The images were labelled for 28 different radiographic findings, where each scan was annotated by a group of three radiologists, with a total of 17 experienced radiologists contributing annotations. The dataset is divided into a training set of 15,000 and test set of 3,000 CXR. Limited by open access restrictions, I exclusively use the training set and split this into training, validation, and test data. I use this dataset in Chapter 7 to evaluate the predictive performance and visual explanations of lung pathologies.

Chapter 4

The Pitfalls of using Open Data for COVID-19 Detection in CXRs

4.1 Introduction

The unprecedented clinical need arising from the recent severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) or COVID-19 pandemic prompted considerable interest from the AI community, with particular focus on the detection of COVID-19 in CXRs. Supported by shared data repositories, publicly available anonymised datasets, and open-source software, researchers developed deep learning systems with the aim of assisting with COVID-19-related clinical tasks (Cohen, Morrison, and Dao, 2020). A large volume of research was published for the detection of COVID-19 in chest radiographs, with researchers developing deep learning solutions to assist with the triaging of patients to prioritise primary diagnostic resources i.e., polymerase chain reaction with reverse transcription (RT-PCR) assays or more sensitive imaging techniques e.g. CT scans. With further research investigating the potential for automated imaging-based COVID-19 detection systems as a 'second-check' option for cases with suspected false negative RT-PCR results, a consequence of this test's well-documented low sensitivity (Kortela et al., 2021; Watson, Whiting, and Brush, 2020).

Early work into deep learning models for the automated prediction of COVID-19 from CXRs reported exceptional model performance, matching or even surpassing the reported capabilities of deep learning for pneumonia-detection prior to the COVID-19 outbreak (Wang, Lin, and Wong, 2020). However, upon a closer review of the literature and in-depth analysis of the used datasets, I identified serious limitations associated with this area of research. As detailed in this chapter, I perform an in-depth analysis of a popularly used open-source COVID-19 dataset, COVIDX (Wang, Lin, and Wong, 2020). I explore its risk of bias to deep learning models through evaluation of performance on an external hospital dataset, implementation of post-hoc explainability methods, and an in-depth exploration of sources of bias within the dataset. Through in-depth data analysis and model evaluation, I aim to show that the popular open-source data is not representative of the real clinical problem and that model performance results on these datasets are inflated.

With the sudden advent of the COVID-19 virus and limited availability of COVID-19 imaging, the majority of early publications relied on a heterogeneous mix of open-source/public data repositories, with non-COVID-19 CXRs sourced from larger pre-existing repositories and COVID-19 CXRs obtained from the public datasets curated in response to the sudden demand for COVID-19 related data. Four COVID-19 CXR repositories were used most frequently: (1) COVID-19 Image Data Collection (COHEN) (Cohen, Morrison, and Dao, 2020), (2) COVID-19 Chest X-ray Dataset Initiative (Cohen and Chung, n.d.) (FIG1), (3) ActualMed COVID-19 Chest X-ray Dataset Initiative (ACTMED) (Chung, 2020), and (4) COVID-19 Radiography Database (CHOWDHURY) (Rahman, n.d.). These four data sources were combined with established pre-COVID-19 datasets that comprise CXRs of patients with various other lung pathologies e.g., the Radiological Society of North America (RSNA) Pneumonia Detection Challenge dataset (*RSNA Pneumonia Detection Challenge* 2018). These five data repositories have been collated into one large open-source dataset, termed COVIDX (Wang, Lin, and Wong, 2020). I term this aggregate-style of dataset, a *Frankenstein* dataset (Roberts et al., 2021). As a purpose-built dataset, this style of open source COVID-19 data typically includes a maximum of three outcomes: *normal*, *pneumonia*, and *COVID-19*; and gives no additional patient data.

In this work I explore the risks and common pitfalls associated with the use of this style of open data for the detection of COVID-19 in CXRs. Specifically, I conduct an in-depth review of the COVIDX dataset and evaluate the generalisability of COVIDX-trained models to real-world hospital data, with careful consideration of model bias.

4.2 Experiment I: Model generalisability

In this component of my research I aim to assess bias within models trained on COVIDX data. In assessing the generalisability of existing deep learning approaches for the detection of COVID-19 from CXRs, I select three highly cited models. I train these models on a balanced version of COVIDX and evaluate them on two external test datasets. Models are evaluated for the task of classifying CXRs into one of three classes: COVID-19, non-COVID-19 pneumonia, and pneumonia negative. Classification performance is assessed through ROC curves and standard classification metrics, such as, AUROC, F1 score, precision, recall, and accuracy. Post-hoc explainability methods are implemented to explore the image features considered most important to classification.

Models were selected based on publication impact factor from those with open-source data and code availability. From these criteria I selected three models: COVID-Net (Wang, Lin, and Wong, 2020), DarkCovidNet. (Ozturk et al., 2020) and Coro-Net (Khobahi, Agarwal, and Soltanalian, 2020). All three models utilise convolutional neural networks (CNNs) for feature learning. While COVIDNet and DarkCovidNet are standard deep CNN classifiers, CoroNet relies on a two stage process to classify images, comprising convolutional autoencoders in the first stage and a standard CNN classifier in the second stage (Khobahi, Agarwal, and Soltanalian, 2020).

The DarkCovidNet is adapted from the Darknet-19 model, a network that has proven itself in computer vision through success in object detection tasks. Darknet-19 consists of 19 convolutional layers and five pooling layers, DarkCovidNet makes small adjustments to this architecture, removing two convolutional layers. The final result is a straight forward deep CNN that makes efficient use of convolutional and max pooling operations (Ozturk et al., 2020). The COVIDNet architecture is more complex with much greater diversity in design. It leverages residual connections alongside selective long range connectivity and projection-expansion-projection-extension (PEPX) design patterns. It is worth noting that the COVIDNet model itself is inaccessible due to copyright restrictions and so, while training and evaluation is possible, implementation of additional explainability modules is challenging (Wang, Lin, and Wong, 2020).

Unlike COVIDNet and DarkCovidNet, CoroNet employs a strategy of a representation learning, where two separate autoencoders are assigned a class of either 'healthy' or non-COVID pneumonia and are trained independently to learn the latent representation of their assigned class. The learned latent representations are then used to create reconstruction of any class of input image. Calculating pixel-wise intensity differences between the input image and the reconstructed image allows for anomaly detection, I would expect to observe anomalous reconstructions where a given input image falls outside of the autoencoder's data manifold, or in other words if the image does not belong to the learned class of the autoencoder. The differences between image input and reconstruction are calculated and termed residual images, these are passed to a ResNet-18 classifier, which is trained to produce class predictions (Khobahi, Agarwal, and Soltanalian, 2020).

For this work, I define a training dataset and two external test sets. As the largest and most popular open-source COVID-19 CXR dataset, I selected COVIDX as my exemplar open-source training dataset. I approximated the COVIDX data used in previous research using files and code made available at: `https://github.com/lindawangg/COVIDNet`. The COVIDX dataset is comprised of multiple open-source data repositories, including *RSNA*, *CHOWDHURY*, *COHEN*, etc. The COVIDX dataset combines pre-COVID-19 repositories with purpose-built COVID-19 datasets made readily available online, providing *normal*, *pneumonia*, and *COVID-19* CXRs.

In updated versions of COVID-19, the RSNA International COVID-19 Open Radiology Database (RICORD) dataset is also included (Tsai et al., 2021). However, to more closely match the COVIDX data used in early research and to reserve RICORD for use in external testing, I exclude all RICORD CXRs from the training data. I exclude pneumonia cases and normal cases at random to better balance the data classes. Ultimately, the balanced COVIDX data contains 4,638 *normal* cases, 4,347 *pneumonia* CXRs and 3,027 *COVID-19* CXRs.

I also create a bespoke external test set from the well-established open-source CXR repositories, the CheXpert dataset, and the reserved RICORD dataset for external evaluation of the trained models (Irvin et al., 2019; Tsai et al., 2021). I refer to this dataset as the *external open test data* throughout. In my inclusion of CheXpert data, I am careful to restrict pneumonia positive cases to include pneumonia-labelled cases exclusively. Sampling from the CheXpert dataset provides 998 pneumonia negative and 997 non-COVID-19 pneumonia cases. My assignment of CheXpert labels to the negative and pneumonia classes was reviewed by a clinical expert to ensure they were appropriate. I ensured the pneumonia negative class includes a wide variety of other lung pathologies, such as, cardiomegaly, pleural effusion, and lung mass/nodules. While I take care to keep the non-COVID-19 pneumonia class to exclusively pneumonia CXRs (although these may include CXRs with co-occurring pathologies). I exclude any pneumonia case with an uncertain label, as defined by the CheXpert labelling schema, to mitigate potential noise. However, I acknowledge that uncertain labels may correspond to more complex CXRs and therefore the blanket removal of CXRs with uncertain labels may bias the dataset.

Additionally, a large teaching hospital based in Leeds, UK, provided a dataset of CXR images of patients alongside RT-PCR test results for COVID-19 diagnosis and non-COVID pneumonia diagnosis. I randomly sampled a subset of CXRs from the Leeds data, giving 611 COVID-19 cases, 459 pneumonia-negative cases, and 299 non-COVID pneumonia cases. Here, a CXR is considered COVID-19 positive if the CXR has an associated positive RT-PCR swab, i.e., both exams conducted during a patient's hospital stay. Due to data access limitations, I was unable to define a more uniform diagnostic window. I refer to this dataset as LTHT.

I trained models on COVIDX data under 3-fold cross validation. Hyperparameters were selected according to the training protocol described in the original publication. Where possible I implemented GradCAM to identify the features that have the greatest impact on classification[1] (Selvaraju et al., 2017b). I implement a learning rate scheduler and early stopping criteria, where if validation loss does not improve for 10 epochs the learning rate is reduced by 10%, and if validation loss does not improve for 20 epochs, training is stopped.

I conducted three rounds of evaluations using (1) the COVIDX test dataset, (2) the external open test dataset, and (3) the LTHT data. I evaluated model weights from each set of cross validation and recorded all performance metrics, i.e., area

---

[1] It was not possible to perform GradCAM on COVIDNet due to restricted model access.

under the receiver operator characteristic (AUROC), accuracy, recall, precision, and F1 score. For each model, I selected the best cross validation iteration according to F1 score; ROC curves were evaluated based on this selection.

4.3 Experiment II: Analysis of *Frankenstein* Data

To evaluate risk of model bias in using a *Frankenstein* dataset, I conduct further, more data-centric investigations. Specifically, I train and test a deep CNN on the COVIDX dataset to predict the original source of COVIDX images. The capacity for a deep learning model to separate CXR by source is indicative of the presence of source-specific features that COVID-19 predictors are vulnerable to learning to rely on i.e., shortcut features.

4.3.1 *Frankenstein* Dataset

I use COVIDX data as the basis for the exemplar *Frankenstein* data. I include RICORD data but to prevent severe imbalance of target classes, I exclude images from less significant COVIDX contributors, such as ACTMED and FIG1. The end result is a COVIDX dataset that comprises images from only the four main COVIDX contributors, RSNA, COHEN, CHOWDHURY and RICORD. The frequency of image by source is presented in Figure 4.2c. For this task I resize all images in the *Frankenstein* data to 224x224 and apply min-max normalisation. During training I apply data augmentation techniques, such as, image rotation, zoom and flipping.

To predict image source, I use a deep CNN based on the established ResNet-18 classifier (He et al., 2015). During training if the validation loss does not improve for 10 epochs, learning rate is reduced by 10%, and if validation loss does not improve for 20 epochs, training is halted. I evaluate the deep CNN model on a held-out COVIDX test set with standard classification performance metrics, i.e., ROC curves, accuracy, precision, recall,and F1 score.

I generate saliency maps to identify the most predictive features for this task. I assess class separability by extracting the hidden features from the final layer of the deep CNN during inference. I use principal component analysis (PCA) to reduce feature dimensionality from 512 to 20, then use t-SNE to project this into 2D for visualisation (Wold, Esbensen, and Geladi, 1987). I use this t-SNE plot to evaluate the hidden features associated with each image source (Maaten and Hinton, 2008).

4.4 Results

4.4.1 Experiment I: Model generalisability

I present model performance metrics for all three test sets in Table 4.2. All models achieved AUROCs of more than 0.93 across all classes, also exceeding 86% prediction accuracy when evaluated on COVIDX test data. Results indicate that all of

Dataset (abbr.)	Frankenstein	Sub-datasets	Open contribution
COHEN	✓	SIRM, EURORAD, CORONACASES	✓
RSNA	×	N/A	×
KERMANY	×	N/A	×
MOONEY*	×	N/A	×
CHOWDHURY	✓	SIRM, COHEN, MOONEY	✓
FIG1	×	N/A	✓
ACTMED	×	N/A	✓
CHEST-XRAY-8	×	N/A	×
CHEST-XRAY-14	×	N/A	×
COVIDX	✓	COHEN, FIG1, ACTMED, RSNA, CHOWDHURY	✓
CHEXPERT	×	N/A	×
SIRM	×	N/A	×
RADIOPAEDIA	×	N/A	✓
EURORAD	×	N/A	✓
CORONACASES	×	N/A	×
JSRT	×	N/A	×
RICORD**	×	N/A	×

TABLE 4.1: **A summary of public datasets commonly used in DL systems for detecting COVID-19 from CXRs.** *MOONEY is the same as the KERMANY dataset, but hosted on Kaggle. **RICORD provides only COVID-19 images. *Abbrvs: Chest X-ray (CXR); Deep Learning (DL).*

the chosen models were able to reliably separate each of the target classes. However, comparison of model performance on the COVIDX test set with model performance on the external and LTHT datasets shows a steep decline, e.g., COVIDNet prediction accuracy falls from 0.86 on COVIDX test data to 0.44 on LTHT data and 0.38 on External open test data. Moreover, the CoroNet model shows a drop in prediction accuracy of 66% when comparing performance on COVIDX data versus LTHT data.

Figure 4.1 shows GradCAM saliency maps of correct DarkCovidNet predictions for all classes. From this I observe a trend in highlighted features positioned outside the lung field. I observe a pattern of highlighted regions around lettering, markers, and the collarbone, although as these features are small and irregularly shaped it is difficult to be precise. Clinical review of GradCAM saliency maps confirms that, despite strong prediction performance on COVIDX test data, models rely on clinically irrelevant features often highlighting confounding features originating in the COVIDX dataset.

4.4.2 Experiment II: *Frankenstein* data

The *Frankenstein* classifier predicted CXR data source with an overall F1 score of 0.89. Model performance varies with CXR source, achieving AUROC scores of 1.00, 0.99, 0.91, and 1.00 on CHOWDHURY, RSNA, COHEN and RICORD CXRs, respectively (Figure 4.2b).

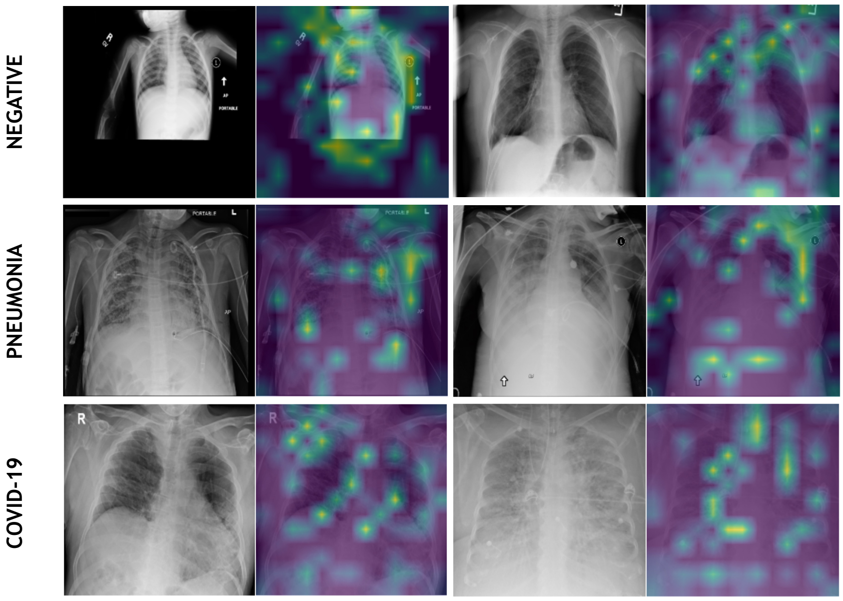


FIGURE 4.1: **Saliency maps of correct DarkCovidNet predictions of COVIDX test CXRs.** Examples of negative, pneumonia and COVID-19 are included side-by-side with the predicted image and saliency maps are generated with GradCAM. *Abbrvs: Chest X-rays (CXRs)*

The t-SNE projected 2D features learned by the deep CNN model shows distinct, separate clustering of the RICORD CXR features, which contributes exclusively COVID-19 positive CXRs (Fig. 4.2a). These features share little-to-no overlap with RSNA CXR features, demonstrating the separability of COVID-19 positive CXRs from COVID-19 negative CXRs by source-specific features alone. CHOWDHURY features are especially clustered, and are far removed from alternative sources. While COHEN CXR features are more dispersed, overlapping with the feature clusters of other CXR sources. I attribute this to the heterogenous nature of the COHEN dataset i.e., its own incorporation of source repositories (Table 4.1). Moreover, CHOWDHURY is the only significant contributor of paediatric images to the

Model	Test set	Precision	Recall	F1 score	Accuracy
DarkCovidNet	COVIDX Data	0.87 ±0.00	0.80 ±0.00	0.82 ±0.00	0.88 ±0.00
	External Data	0.44 ±0.00	0.43 ±0.00	0.41 ±0.00	0.43 ±0.00
	LTHT Data	0.47 ±0.01	0.46 ±0.00	0.44 ±0.01	0.45 ±0.00
CoroNet	COVIDX Data	0.81 ±0.05	0.90 ±0.01	0.84 ±0.05	0.88 ±0.03
	External Data	0.18 ±0.07	0.34 ±0.02	0.19 ±0.03	0.35 ±0.01
	LTHT Data	0.24 ±0.01	0.30 ±0.00	0.15 ±0.01	0.22 ±0.00
COVIDNet	COVIDX Data	0.86 ±0.03	0.69 ±0.05	0.72 ±0.05	0.86 ±0.02
	External Data	0.34 ±0.05	0.36 ±0.01	0.29 ±0.02	0.38 ±0.01
	LTHT Data	0.43 ±0.01	0.39 ±0.00	0.37 ±0.01	0.44 ±0.03

TABLE 4.2: **Model performance metrics across all test datasets, with standard deviation, across cross-validation folds.**

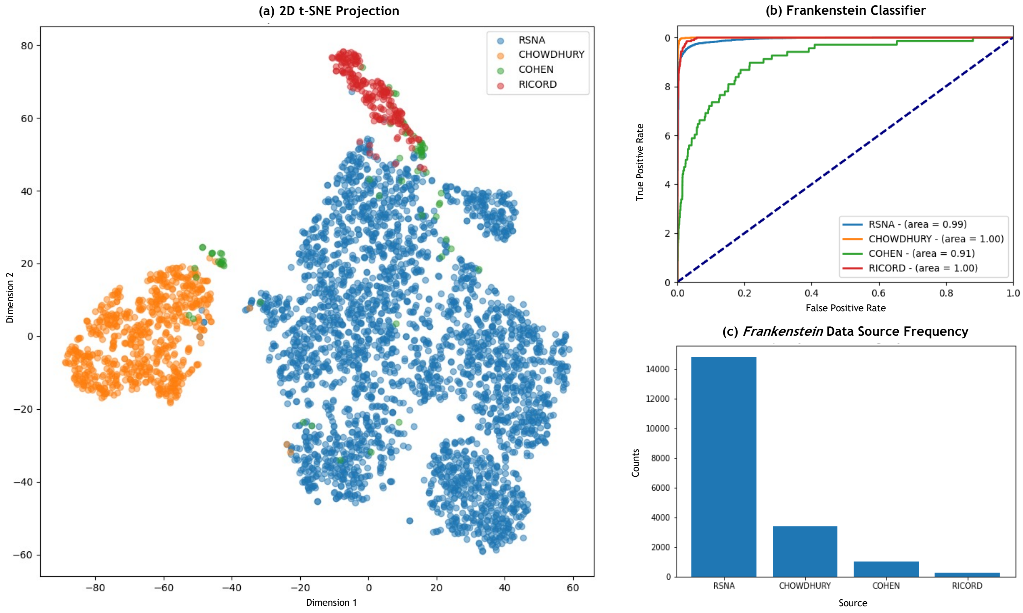


FIGURE 4.2: **Frankenstein data analysis results.** (a) 2D t-SNE projection of hidden features extracted from the trained *Frankenstein* classifier during inference on the held-out test set of the *Frankenstein* data. (b) *Frankenstein* data classifier ROC curves, ‘area’ denotes the area under the curve metric. (c) Frequency of image source in the *Frankenstein* dataset. *Abbrvs: Receiver Operator Characteristic (ROC).*

Frankenstein dataset and so can be separated on this basis. Notably, all CHOWDHURY paediatric images are COVID-19 negative, it therefore follows that age-related features can be used as a ‘shortcut’ for the detection of COVID-19 in CXRs.

GradCAM saliency maps highlight the risk of bias associated with using unverified and untrustworthy public datasets (Fig. 4.3). The saliency maps generated for the CHOWDHURY and RSNA images demonstrate the importance of symbols and annotations for source classification e.g., arrows and lettering. Despite the inclusion of independent COVID-19 source repositories, predictive features do not appear related to COVID-19 disease features. Few saliency maps consistently highlight clinically-relevant areas, instead features of significance are generally localised to the collarbone and shoulders.

Data source	Precision	Recall	F1 score	Accuracy	AUROC
RSNA	1.00	0.88	0.94	0.88	0.99
CHOWDHURY	0.97	0.99	0.98	0.99	1.00
COHEN	0.18	0.66	0.28	0.66	0.91
RICORD	0.58	0.99	0.73	0.99	1.00

TABLE 4.3: ***Frankenstein* Deep CNN classifier performance metrics.** The Deep CNN classifier is trained to predict image source. RSNA provides only COVID-19 negative images, CHOWDHURY and COHEN give a mix of COVID-19 negative and COVID-19 positive images and are themselves Frankenstein dataset, RICORD gives only COVID-19 positive images.

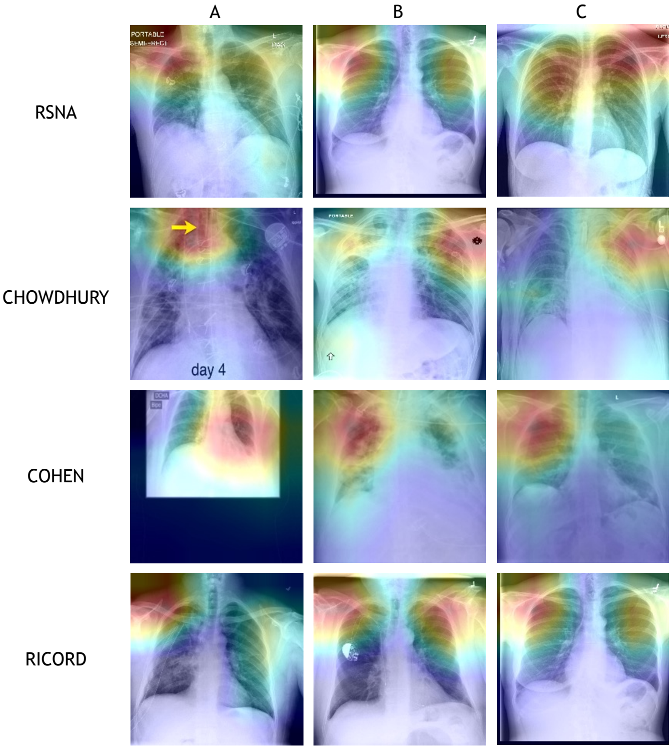


FIGURE 4.3: **Frankenstein prediction saliency maps.** A random sample of GradCAM images generated by Deep CNN predictions on *Frankenstein* test data.

The strong performance of the *Frankenstein* classifier prompted a manual review of each data source. From this, I identified a number of potential sources of bias in the COVIDX data set. CHOWDHURY, a major contributor of COVIDX COVID-19 images, blindly blends paediatric CXR images with adult images and is the only large source of paediatric images within COVIDX, introducing significant feature biases. Manual inspection of ACTMED data revealed the consistent presence of disk-shaped markers in COVID-19 CXRs. For most data sources, images are pre-processed images and vary in size. For example, the images from the RSNA data set are 1024x1024 in resolution, while all CHOWDHURY-provided images are a resolution of 299x299. Most methods resize images to resolutions between these; smaller images must be up-sampled and larger images must be down-sampled. This risks the generation of artefacts that may bias the model. Similarly, additional bias may be introduced by the random application of image pre-processing (i.e., cropping and padding) in COVID-19 repositories, COHEN and CHOWDHURY, and the absence of image pre-processing in pre-established non-COVID-19 data sources.

4.5 Discussion

By reproducing their work and evaluating models on external datasets, I show that models fail to generalise outside COVIDX and perform poorly in hospital populations.

Through the evaluation of COVIDX as a *Frankenstein* dataset I highlight the risk of bias and confounding to models trained on undocumented open data. I identify spurious non-clinical features (e.g., image resolution, symbols and annotations, or age) that correlate with image source within the *Frankenstein* dataset. Risk of bias increases where data sources only provide one class of image. For example, the RSNA dataset provides only COVID-19 negative CXRs and is the largest contributor to COVIDX. RSNA-specific features can therefore be used by models as shortcuts for the prediction of COVID-19 negative images. These shortcuts do not translate to real world hospital data, hence I observe poor model generalisation to the LTHT dataset. Many of the studies that do not employ exact replicas of the COVIDX dataset still use *Frankenstein* datasets, combining CXR repositories from the pre-COVID era with COVID-19 data sources. The broad use of *Frankenstein* data indicates that my findings are of significance to this domain as a whole and may apply to other problem domains that rely on similar approaches for data collation.

Central to the inappropriate use of open-source data in the detection of COVID-19 from CXRs is the lack of patient information. Open-source datasets often lack critical details, such as demographic data (e.g., age, gender, ethnicity), clinical context (e.g., comorbidities, disease severity), and image acquisition information (e.g., type of imaging equipment, protocols). Without this essential metadata, researchers cannot identify or account for key confounding factors that could influence the performance and reliability of deep learning models. In this work, I identify age as a major confounding factor i.e., the inclusion of paediatric chest X-rays in only COVID-19 negative classes. Without age-related metadata, models may inadvertently learn patterns that reflect the distribution of ages in the dataset rather than the underlying disease characteristics. The presence of age-related bias can lead to incorrect predictions, particularly for age groups under-represented or over-represented in the population, raising serious concerns about the generalisability and fairness of these deep learning models.

Clinical review of GradCAM saliency maps confirm that to predict COVID-19 models depend on clinically irrelevant features, and exhibit a pattern of relying on bias and confounding originating from the *Frankenstein* nature of the COVIDX dataset. This highlights the critical need for model interpretability in the clinical domain, particularly where external validation is unavailable or limited.

4.6 Conclusion

My work demonstrates that without external validation, early research using the COVIDX dataset reported overly optimistic results. By reproducing these studies and evaluating the models on external datasets, I show that models trained on COVIDX fail to generalise to real-world hospital populations and perform poorly outside of the training distribution. My evaluation of COVIDX as a *Frankenstein* dataset highlights the significant risk of bias and confounding factors in models trained on undocumented open data that comprise many data repositories. I identified various non-clinical features, or "shortcuts" (such as image resolution, symbols, annotations, and age), that correlated with image sources within this dataset. I found that a lack of metadata prevents researchers from accounting for confounding factors, like age, which I identified as a major source of model bias. To identify model bias interpretability is essential, especially in the absence of external validation where evaluation of generalisability is not possible. Clinical review of GradCAM saliency maps reveals that models are heavily influenced by non-clinical features derived from the biases within the COVIDX dataset. Real-world datasets with detailed patient information are required for the development of automated diagnostic tools. Moreover, thorough model evaluation on hospital populations made up of diverse demographics is essential to properly investigate the efficacy of deep learning in assisting with the detection of COVID-19 from CXRs. Overall, poor performance of deep learning models trained on the COVIDX dataset on both publicly available external test data and LTHT data demonstrates that the exceptional performance reported widely across the problem domain is inflated, that model performance results are misrepresented, and that models do not generalise well to clinically-realistic data.

Chapter 5

Multi-centre Benchmarking of Deep Learning Models for COVID-19 Detection in CXRs

5.1 Introduction

The identification of risk of bias and confounding to COVID-19 classifiers trained on an amalgamation of open data sources in Chapter 4 has motivated further investigation. In this work, I aim to address the key questions raised by my previous research:

- How do deep learning models perform in real world hospital populations?
- Which clinical factors contribute to an increased risk of model errors?
- Does risk of model bias persist when models are trained on a real-world, multi-site hospital dataset?

I conduct a retrospective study to evaluate the performance of existing deep learning models developed for COVID-19 detection in CXRs. I train models on a national dataset of COVID-19 CXRs purposely set up to support the development of deep learning solutions. I validate these models, under clinical guidance, by considering the practical challenges of interpreting chest X-rays in suspected COVID-19 cases. Perhaps foremost of these challenges is that COVID-19 infection often does not develop into COVID-19 pneumonia, in which case diagnostic features of COVID-19 cannot be observed in the CXR. Moreover, where COVID-19 pneumonia can be observed, its heterogeneous presentation mimics a broad spectrum of lung pathologies, making it difficult to identify COVID-19 pneumonia due to confounding conditions. The presence of co-occurring conditions, or comorbidities, can also complicate the detection of COVID-19, especially in cases where the disease is mild and features are subtle. Furthermore, the unpredictable temporal progression of COVID-19 presents a challenge for radiologists, as unexplained rapid advancements in the disease and low resolution in CXRs contribute to ambiguity (Sverzellati et al., 2020). Collectively, these factors lead to substantial diagnostic uncertainty when using medical imaging for the detection of COVID-19. With this understanding I conduct a thorough evaluation of model error, scrutinising these factors to better understand model failures, the demographics of those affected, and potential avenues for model improvement.

Here, I present a comprehensive benchmarking study comparing state-of-the-art DL methods and conducting exhaustive model evaluations on independent, multi-national clinical datasets, with the goal of identifying model strengths and weaknesses while assessing the suitability of automated DL systems as clinical decision support tools in COVID-19 detection.

5.2 Materials and Methods

In this section, I present my methodology, providing detailed descriptions of the evaluated models, my training procedures, my evaluation methods, and a thorough review of the datasets used (Fig. 5.1).

5.2.1 Overview of the experimental approach

I utilise two independent UK-based datasets and a further dataset from outside the UK. I train a diverse set of deep learning models on one of the UK-based datasets (NCCID) and validate national generalisability using data from the other UK-based dataset, which is from an independent hospital site, the Leeds Teaching Hospital Trust (LTHT) (Cushnan et al., 2021). I consider international generalisability using open-source data from a Spanish hospital (COVIDGR) (Tabik et al., 2020).

I investigate model performance variation by patient-level factors e.g., demographic and smoking history. I also evaluate model vulnerability to confounding variables, which requires the use of counterfactual datasets created from a subset of the LTHT population for whom non-COVID-19 pneumonia status was recorded. In this population I modified the definition of the positive and control classes, resulting in two additional counterfactual datasets. The first dataset referred to as *LTHT PNEUMONIA (P)* simulates a pneumonia detection setting where the positive class includes non-COVID-19 pneumonia cases i.e., no distinction is drawn between COVID-19 and other pneumonia types. The second scenario named *LTHT NO PNEUMONIA (NP)* replicates a COVID-19 detection scenario where all instances of non-COVID-19 pneumonia were deliberately excluded.

Following primary evaluation, I identify the top-performing models for further analysis. I train and validate the best models on region-of-interest (ROI)-extracted CXRs to test whether overall performance of COVID-19 detection is improved with the use of ROIs and if, as is commonly assumed, cropping to the ROI helps to mitigate any inherent data biases. Furthermore, I apply explainable AI techniques to examine highlighted features, i.e., features significant to model prediction. Identification of certain features can indicate model reliance on spurious correlations, which can lead to poor generalisation. The presence of these ‘shortcut’ features has been identified in prior work on models trained with open-source data, I evaluate NCCID-trained models for reliance on similar ‘shortcut’ features (DeGrave, Janizek, and Lee, 2021).

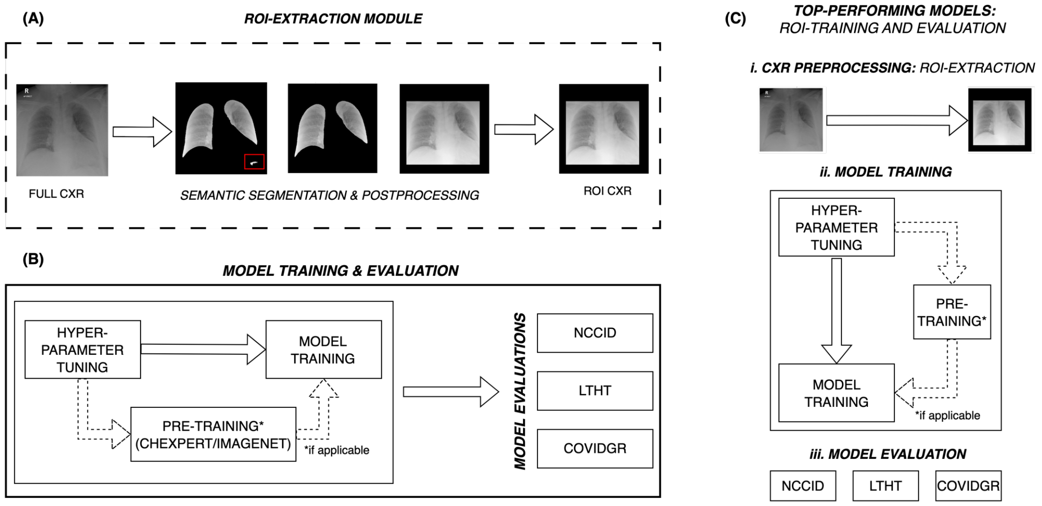


FIGURE 5.1: **Overall experimental design for multi-centre evaluation of COVID-19 detection models. (A)** ROI-cropped CXRs are generated from semantic segmentations of the left and right lung fields, automated prediction uncertainty-based post-processing is applied to ensure reliable cropping for both classes of CXR. The red box highlights over-segmentation of the lung fields, post-processing removes this structure prior to extracting the region of interest. **(B)** Some models are pre-trained, for these models hyper-parameters are tuned on ImageNet or CheXpert data (domain-specific dataset). After pre-training, model hyper-parameters are refined for the COVID-19 detection task and trained on full CXRs from the NCCID. Models are subsequently evaluated on three independent populations: the unseen NCCID population, the LTHT, and COVIDGR. **(C)** Following primary training and evaluation, the best performing models are selected for *(ii)* training and *(iii)* evaluation on ROI-extracted CXRs. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Region Of Interest (ROI); Chest X-ray (CXR).*

5.2.2 Data

This study utilises three independent datasets, NCCID, COVIDGR sourced from a hospital in Spain, and a purpose built single site dataset derived from Leeds Teaching Hospitals NHS Trust (LTHT). The NCCID dataset is available upon request, COVIDGR can be found online and the LTHT dataset is not available publicly, however the hospital has a formal data access process through which researchers may apply.

Uniform exclusion criteria were applied to the NCCID and LTHT datasets. CXRs were excluded if case data was insufficient to confidently assign a COVID-19 label, e.g. missing RT-PCR swab date or missing RT-PCR test result data for CXRs collected post-2019. Inclusion criteria was considered using data collected from Digital Imaging and Communication in Medicine (DICOM) headers and associated radiology reports. Note that the international dataset (COVIDGR) did not include RT-PCR swab date or CXR exam date data. Instead, CXR labels were pre-defined with CXRs considered positive if acquired 24 hours before or after a positive COVID-19 swab. The labelling schema for all datasets are described and I outline this in Figure 5.3. For all datasets, only frontal CXRs, antero-posterior (AP) and postero-anterior (PA), were included and only clinical testing (SARS-CoV-2 RT-PCR) results were used in producing COVID-19 labels, radiological features indicative of COVID-19 infection were not considered. Figure 5.2 presents a CONSORT diagram describing the full exclusion criteria applied during data preparation for this study.

I note a methodological concern in COVID-19 chest radiograph analysis that relates to the temporal heterogeneity of imaging data. A substantial proportion of COVID-19 CXR datasets contain images acquired using pre-2019 protocols and scanner configurations, which are likely to change post-2019 due to the pandemic. This temporal mismatch introduces a possible systematic bias into model training and evaluation, as the imaging parameters and image acquisition protocols may differ between pre-2019 and COVID-19 CXRs.

		Sex (n)		**Age**			**Ethnicity (n)**			
	Male	*Female*	*Unknown*		*Asian*	*Black*	*Multiple*	*Other*	*White*	*Unknown*
NCCID										
Positive 41% (8,337)	61% (5113)	38% (3215)	0% (9)	67±17	13% (1120)	8% (739)	1% (118)	4% (379)	67% (5,592)	5% (389)
Negative 59% (12,178) N=20,515	58% (7122)	42% (5054)	0% (2)	70±17	14% (1687)	7% (913)	2% (230)	3% (421)	70% (559)	3% (3680
NCCID TEST										
Positive 33% (268)	71% (190)	29% (78)	0% (0)	66±15	11% (29)	7% (19)	1% (4)	5% (13)	72% (192)	4% (11)
Negative 67% (556) N=824	64% (357)	36% (199)	0% (0)	69±16	12% (67)	8% (42)	3% (14)	4% (20)	71% (392)	4% (21)
LTHT										
Positive 16% (1,752)	17% (1061)	14% (691)	0% (0)	72±16	10% (177)	7% (125)	1% (8)	3% (47)	67% (1171)	13% (224)
Negative 84% (9,452) N=11,204	83% (5034)	86% (4417)	0% (1)	63±26	6% (549)	2% (117)	1% (64)	1% (113)	59% (5,566)	32% (2,983)

TABLE 5.1: **Demographic subgroups of training and test data.** Age is presented as mean ± standard deviation. Sex and ethnicity are presented both as absolute counts (n) and as percentages relative to the COVID-19 positive/negative cohort. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Region of Interest (ROI).*

Pre-training data (ImageNet & CheXpert) Pre-trained models were trained on either ImageNet natural images or domain-specific CheXpert CXRs prior to NCCID training (as dictated by original model implementations). ImageNet is a large-scale image classification dataset comprising 14 million annotated natural images

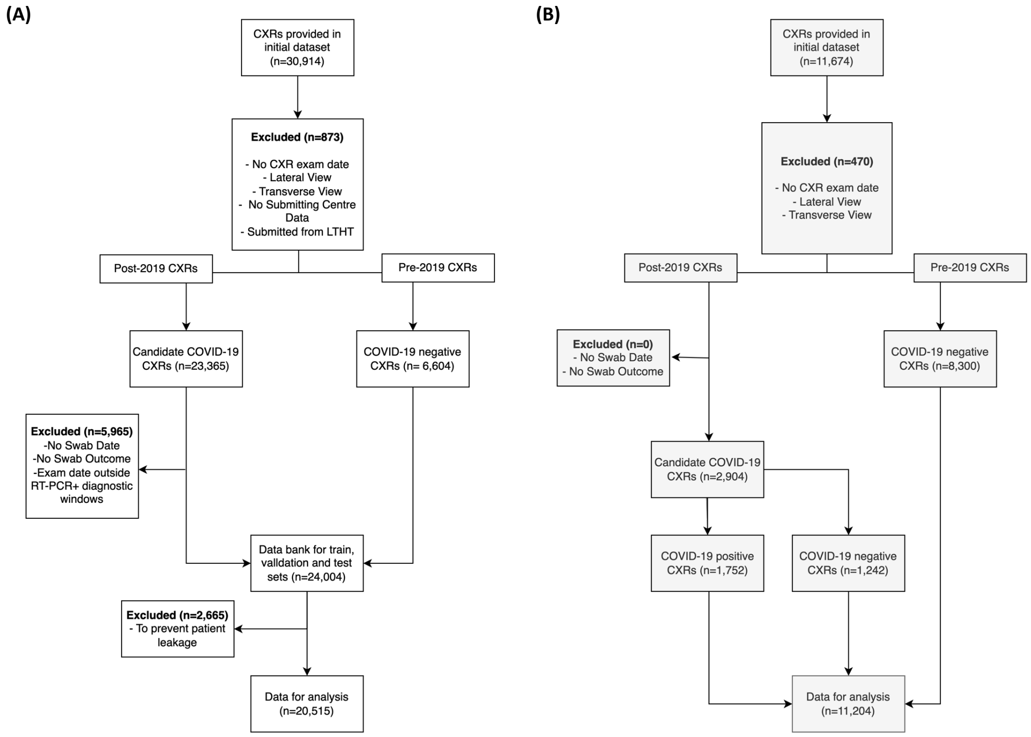

(A)
CXRs provided in initial dataset (n=30,914)
Excluded (n=873)
- No CXR exam date
- Lateral View
- Transverse View
- No Submitting Centre Data
- Submitted from LTHT
Post-2019 CXRs
Pre-2019 CXRs
Candidate COVID-19 CXRs (n=23,365)
COVID-19 negative CXRs (n= 6,604)
Excluded (n=5,965)
-No Swab Date
-No Swab Outcome
-Exam date outside RT-PCR+ diagnostic windows
Data bank for train, valldation and test sets (n=24,004)
Excluded (n=2,665)
- To prevent patient leakage
Data for analysis (n=20,515)
(B)
CXRs provided in initial dataset (n=11,674)
Excluded (n=470)
- No CXR exam date
- Lateral View
- Transverse View
Post-2019 CXRs
Pre-2019 CXRs
Excluded (n=0)
- No Swab Date
- No Swab Outcome
COVID-19 negative CXRs (n=8,300)
Candidate COVID-19 CXRs (n=2,904)
COVID-19 positive CXRs (n=1,752)
COVID-19 negative CXRs (n=1,242)
Data for analysis (n=11,204)

FIGURE 5.2: **Exclusion criteria for pre-processing (A) NCCID and (B) LTHT datasets.** CXRs were excluded if missing crucial acquisition data (exam date and submission centre) and if not frontal view (AP or PA). NCCID CXRs were eliminated if submitted from Leeds-based hospitals. CXRs were divided into two cohorts: pre-2019 and post-2019. Pre-2019 CXRs were automatically labelled COVID-19 negative, while post-2019 CXRs were evaluated for COVID-19 outcomes. Post-2019 CXRs were eliminated if missing data essential for determining COVID-19 outcome i.e., CXR acquisition date, RT-PCR swab date or outcome. CXRs were also excluded if exam date fell between diagnostic windows of multiple positive RT-PCR swabs. The COVIDGR dataset is not subject to the same exclusion criteria due to a lack of patient data. As a pre-prepared dataset, some exclusion criteria is already applied i.e., COVIDGR includes only PA CXRs. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Antero-posterior (AP); Postero-anterior (PA); Chest X-ray (CXR).*

from more than 21,000 classes e.g., hummingbird, hen, lion etc. ImageNet [1] is publicly accessible and available for download. CheXpert is a large dataset containing 224,316 chest X-rays from 65,240 patients, each image has recorded outcomes for 14 observations, such as, pleural effusion, cardiomegaly and consolidation (generated from radiology reports) (Irvin et al., 2019). CheXpert[2] is also publicly accessible and available for download.

Training data (NCCID) The National COVID-19 Chest Imaging Database (NCCID) is a centralised UK database derived from 26 hospital centres, storing 45,635 CXRs from 19,700 patients across the UK in the form of DICOM image files and header information (de-identified) (Cushnan et al., 2021). To preserve the independence of the single-site evaluation dataset (LTHT), I excluded from NCCID all cases originating from the Leeds area, leaving CXRs collected from 25 different hospital centres. The removed CXRs were neither utilised for model training nor model evaluation on NCCID. NCCID CXRs are provided alongside clinical data, including the results of RT-PCR tests. Dates for both CXR exams and RT-PCR swabs are provided. If exam date or RT-PCR dates were unavailable, the CXR was excluded from the study. RT-PCR was used to define ground truth labels for CXRs. As no standard recognised definition exists within the literature, I sought expert opinion from a radiologist, a respiratory physician, and a clinical oncologist to inform my definition of COVID-19 positive CXRs. I treated CXRs with a positive COVID-19 RT-PCR test anywhere from 14 days before to 28 days after image acquisition as COVID-19 positive. I treated images without a positive RT-PCR test within this diagnostic window as COVID-19 negative (Fig. 5.3). After data preparation, the NCCID training dataset consists of 20,515 exams, with 8,337 positive exams and 12,178 control CXRs. Figure 5.2 presents a CONSORT diagram outlining the full exclusion criteria applied to both NCCID and LTHT datasets.

Testing data (LTHT & COVIDGR) External validation data is collected from two independent sources, LTHT, a UK-based hospital in Leeds (nationally-sourced), and COVIDGR, made up of CXRs from San Cecelio University Hospital in Granada, Spain (internationally-sourced) (Tabik et al., 2020). LTHT provides patient CXR images (DICOMs), with RT-PCR test results for COVID-19 diagnosis. In LTHT, RT-PCR date is provided relative to CXR exam date to allow precise classification of COVID-19 status according to my chosen diagnostic window (Fig. 5.3). The exclusion criteria for LTHT and COVIDGR datasets is summarised in Table 5.2).

Additionally, for a subset of LTHT patients, non-COVID-19 pneumonia diagnostic status was available, from this subset of the LTHT population the counterfactual datasets LTHT (P) and LTHT (NP) were created. To create LTHT (NP) all participants with recorded non-COVID-19 pneumonia were removed from the

---

[1] https://www.image-net.org/

[2] https://stanfordaimi.azurewebsites.net/datasets

		Data					
		NCCID		**LTHT**		**COVIDGR**	
Criteria	Principle	*Pre-2019*	*Post-2019*	*Pre-2019*	*Post-2019*	*Pre-2019*	*Post-2019*
Inclusion	AP/PA CXRs	✓	✓	✓	✓	✓	✓
	Submission centre data	✓	✓				
	Complete swab data (data & outcome)		✓		✓		
	Complete image acquisition data (date & ID)	✓	✓	✓	✓		
Exclusion	Lateral or transverse CXRs	✓	✓	✓	✓	✓	✓
	LTHT submission centre	✓	✓				
	Exam data outside RT-PCR+ swab window		✓				

TABLE 5.2: **Inclusion and exclusion criteria of CXR exam from NCCID, LTHT and COVIDGR data.** For all datasets, CXRs are eliminated if not frontal view. NCCID and LTHT CXR exams conducted after 2019 are eliminated if COVID-19 swab or image acquisition data is incomplete. For NCCID data, CXRs are also eliminated if submission centre data is incomplete or if the CXR exam date falls in between two non-overlapping windows of COVID-19 infection. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Antero-posterior (AP); Postero-anterior (PA); Chest X-ray (CXR).*

LTHT dataset. To construct the LTHT (P) dataset the image labelling criteria was changed such that all CXRs positive for pneumonia (COVID-19 or non-COVID-19 pneumonia) were labelled positive. I do this to evaluate the models' capacity to separate COVID-19 from non-COVID-19 pneumonia cases, a major confounding pathology. For both LTHT (P) and LTHT (NP) populations, participants without non-COVID-19 labels were not considered.

The COVIDGR dataset provides a total of 852 CXRs sourced from the San Cecelio University Hospital in Granada, Spain. The dataset is balanced, containing 426 positive and 426 negative CXRs. In the creation of the COVIDGR dataset CXRs were chosen through manual selection. COVID-19 CXRs are defined by a positive RT-PCR test, conducted within 24 hours of the CXR exam. All CXRs were pre-cropped prior to being compiled into COVIDGR. COVIDGR includes only postero-anterior (PA) views which were acquired with the same scanner type. In addition, RALE severity scores are provided for all positive cases, as well as 76 CXRs in which COVID-19 is not observed (NORMAL-PCR+), 100 mild (MILD), 171 moderate (MODERATE) and 79 serious (SEVERE) cases.

5.2.2.1 Label generation

For NCCID and LTHT data, CXR labels were generated according to a pre-defined diagnostic window. Under clinical guidance, I defined the COVID-19 diagnostic window as 14 days before and 28 days after the acquisition data of a positive RT-PCR test swab. CXR exam date was evaluated relative to the nearest positive RT-PCR COVID-19 swab date, CXRs that fell inside this window (-14/+28 days) were

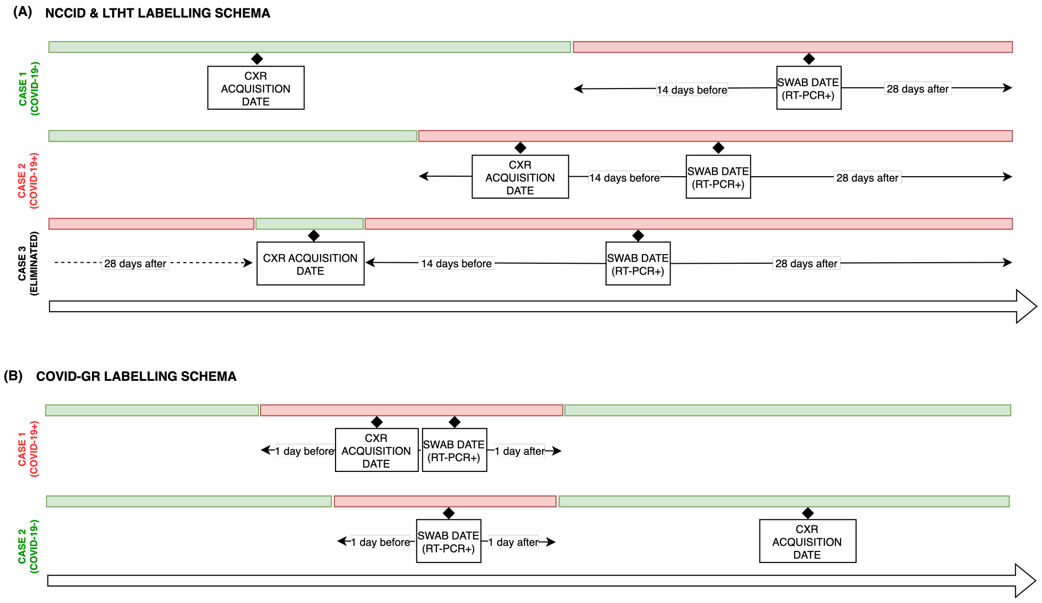


FIGURE 5.3: **Labelling schema for (A) NCCID and LTHT datasets, which share identical labelling protocol, and (B) COVIDGR data.** For each protocol I present examples of different label outcomes. **(A) CASE 1**: An illustration of CXR acquisition preceding the RT-PCR swab date diagnostic window (-14/+28 days), this case is therefore considered COVID-19 negative. **(A) CASE 2**: An example of CXR acquisition prior to the RT-PCR swab date but within the diagnostic window, as a result this case is labelled COVID-19 positive. **(A) CASE 3**: A scenario involving CXR elimination, where multiple swab tests are documented for a single case. If a CXR is acquired within the time frame between the windows around the swab dates, it is excluded from the dataset. **(B) CASE 1**: A case in which the CXR was acquired within the diagnostic window, specifically within 24 hours of the RT-PCR swab date (-1/+1 days). As a result, this case is designated as COVID-19 positive. **(B) CASE 2**: An example of CXR acquisition occurring after the diagnostic window, leading to the categorisation of this case as COVID-19 negative. *Abbrvs: Chest X-ray (CXR); National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Reverse Transcription Polymerase Chain Reaction (RT-PCR).*

	NCCID **(TRAIN)**	NCCID **(TEST)**	LTHT	COVIDGR	COVID-QU-Ex
Lung segmentation model training & testing					✓
Model training	✓				
Model performance evaluation	✓	✓	✓	✓	
Sub-population analysis			✓		
Counterfactual evaluation			✓(subset w/ known pneumonia outcomes)		
RALE-dependent performance evaluation				✓	

TABLE 5.3: **Overview of individual dataset use throughout this study, including ROI-extraction, model training and evaluations.** For evaluation of models under counterfactual conditions I used a subset of LTHT with recorded non-COVID-19 pneumonia status. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Region of Interest (ROI); Radiographic Assessment of Lung Edema (RALE).*

labelled COVID-19 positive. In some cases, evaluation of serial patient swab dates created multiple non-overlapping diagnostic windows, I treated these as separate instances of COVID-19 infection and CXRs that fell between these windows were removed from the dataset. For COVIDGR, CXRs are provided with COVID-19 labels, COVID-19 CXRs are defined by a positive RT-PCR swab within 24 hours of CXR acquisition. I provide an illustration of the labelling schema through case by case examples (Fig. 5.3).

5.2.2.2 Models

The models I selected for this benchmarking study are diverse in design and leverage different learning paradigms i.e., supervised, transfer, semi-supervised and self-supervised learning (Table 5.4). I found that the majority of proposed DL methods for COVID-19 detection in CXRs rely on supervised or transfer learning. Here I define supervised models as models trained for COVID-19 detection from randomly initialised weights. All transfer learning approaches used weights pre-trained on either ImageNet or CheXpert and were later fine-tuned in a fully supervised manner on the training dataset (NCCID) for the task of COVID-19 detection.

Within the supervised learning category, I explored the use of various well-established deep convolutional neural network (CNN) backbones. Of these, I identified XCEPTION NET (Khan, Shah, and Bhat, 2020) and ECOVNET (Chowdhury et al., 2021) from highly cited publications as influential models of interest. Similar approaches place emphasis on domain-specific tuning, rather than applying a pre-defined deep CNN backbone. For example, COVIDNET (Wang, Lin, and Wong, 2020) is defined by a generative neural architecture search (NAS) for optimal COVID-19 detection performance. Other deep CNN approaches employ unique designs to encourage the recognition of domain-specific features, such as RES. ATTN. (Sharma and Dyreson, 2020) which incorporates attention mechanisms, and MAG-SD (Li et al., 2021) which uses hierarchical feature learning. Although all of these approaches share some similarities, training strategies differ. For each model, I reproduce the pre-training strategies outlined in their respective studies in order to maintain consistency with first reported implementations. Under the supervised learning category, I also include a vision transformer (XVITCOS), a deep CNN-ensemble network (FUSENET) (Abdar et al., 2023), and a capsule network (CAPSNET) (Afshar et al., 2020).

I select CORONET (Khobahi, Agarwal, and Soltanalian, 2020) as an example of semi-supervised learning. First introduced in Chapter 4, CORONET relies on a two stage process to classify images, comprising a convolutional autoencoder in the first stage and a standard CNN classifier in the second stage. First, the convolutional autoencoder is trained to reconstruct COVID-19 negative CXRs from learned low-dimensional latent representations. The classifier is then trained to predict CXR outcomes taking images comprising the pixel-wise differences between original CXRs and autoencoder reconstructions (residual images) as inputs. The intuition is that the reconstructions of CXRs from the unseen class (COVID-19 positive) will fail to preserve radiographic features of COVID-19 infection, which will appear in residual images. Other approaches like SSL-AM (Park, Kwak, and Lim, 2021), follow a self-supervised pre-training strategy. In SSL-AM, representations learned during pre-training are enhanced through 2D transformations, such as, distortion, in-painting and perspective transformations. During pre-training, SSL-AM is comprised of a U-Net-style network architecture which learns domain-specific features independent of the disease class. Following pre-training, the decoder portion of the U-Net is subsequently discarded, while, the encoder and its pre-trained, domain-specific weights are incorporated into a COVID-19 classifier.

Model	**Reference**	**Abbrvs.**	**DL type**	**Pre-trained [Y/N] *(Data)***	**Params. (#)**
Deep CNN generated by NAS	(Wang, Lin, and Wong, 2020)	COVIDNET	Supervised	Y (CheXpert)	50,150,485
Multiscale attention guided network with soft distance regularisation	(Li et al., 2021)	MAG-SD	Supervised	Y (ImageNet)	23,835,968
Vision transformer	(Mondal et al., 2022)	XVITCOS	Supervised	Y (CheXpert)	86,537,477
Ensemble of deep CNNs	(Abdar et al., 2023)	FUSENET	Supervised	Y (ImageNet)	17,245,921
Deep CNN with Xception backbone	(Khan, Shah, and Bhat, 2020)	XCEPTION NET	Supervised	Y (ImageNet)	21,331,753
Deep CNN with residual connections and attention component	(Sharma and Dyreson, 2020)	RES. ATTN.	Supervised	N	5,476,673
Deep CNN with EfficientNet backbone	(Chowdhury et al., 2021)	ECOVNET	Supervised	Y (ImageNet)	7,304,737
Convolutional capsule network	(Afshar et al., 2020)	CAPSNET	Supervised	Y (CheXpert)	523,072
Convolutional autoencoder with classifier	(Khobahi, Agarwal, and Soltanalian, 2020)	CORONET	Semi-supervised	Y (ImageNet)	11,230,978
Deep CNN with attention mechanism, pre-trained under self-supervised conditions	(Park, Kwak, and Lim, 2021)	SSL-AM	Self-supervised	Y (CheXpert)	6,753,905

TABLE 5.4: **Summary of the evaluated models.** Models are described and presented alongside source reference, pre-training status and deep learning category. Models are referred to by their designated abbreviations. *Abbrvs: Deep Learning (DL); Neural Architecture Search (NAS); Parameters (Params.).*

5.2.3 Model training

I apply a pre-defined training protocol designed to facilitate uniform comparison in model performance. I train models on NCCID training data across 5-fold cross-validation experiments, each of which comprises a balanced number of COVID-19 negative and COVID-19 positive cases. Prior to training I, where necessary, adapt the original models for the task of binary classification of CXRs, i.e., to accommodate larger image resolution (than was used in the original implementation of any of the selected models) or to predict two classes instead of three.

I used the CheXpert dataset for models that required pre-training on domain-specific datasets. Specifically, model weights were optimised for the task of predicting lung pathologies in CXRs.

Models that required pre-training on natural images were pre-trained on ImageNet. The choice of dataset, and if pre-training is even required, is dictated by the original model implementation. For all training stages, images were resized to 480x480 and standard image transformations were applied. I also tuned the learning rates for each model, at each stage of training, using Optuna which is an open source hyperparameter optimisation framework. As models are identified from pre-existing, published works I accepted model architecture hyper-parameters as fixed and do not tune these to the training datasets.

Lung segmentation (ROI) Automatic segmentation of lung fields is often applied to mitigate the influence of confounding variables and background artefacts/noise. To test this, the top three performing models are also trained using CXRs cropped to the lung fields, which have been separated from background tissue using semantic segmentation. To generate these segmentations I trained a U-Net++ model on the open-source dataset COVID-QU-Ex, containing 33,920 COVID-19, pneumonia, and normal CXRs, all with ground truth segmentation masks (Tahir et al., 2021).

To improve segmentation robustness and reduce the risk of introducing a segmentation quality bias to the downstream classification task, I applied a novel epistemic uncertainty-based post-processing algorithm to revise predictions or flag predictions for manual inspection where necessary (Stone et al., 2022). For the task of lung field segmentation, correct segmentations are expected to comprise two connected components, each component corresponding to the left or right lung field. Additionally, successful segmentations are assumed to have corresponding pixel-wise prediction uncertainty maps that are unimodal with uncertainty predominantly concentrated along the borders of the lungs. This is what would be expected if a panel of radiologists were tasked with outlining lung fields in CXRs (and inter-rater variability/uncertainty was quantified). Thus, I also assume that a bimodal uncertainty frequency is evidence of erroneous segmentation outside normal inter-rater variability.

If predicted segmentation masks were found to have more than two unconnected components, their corresponding uncertainty maps were then assessed for

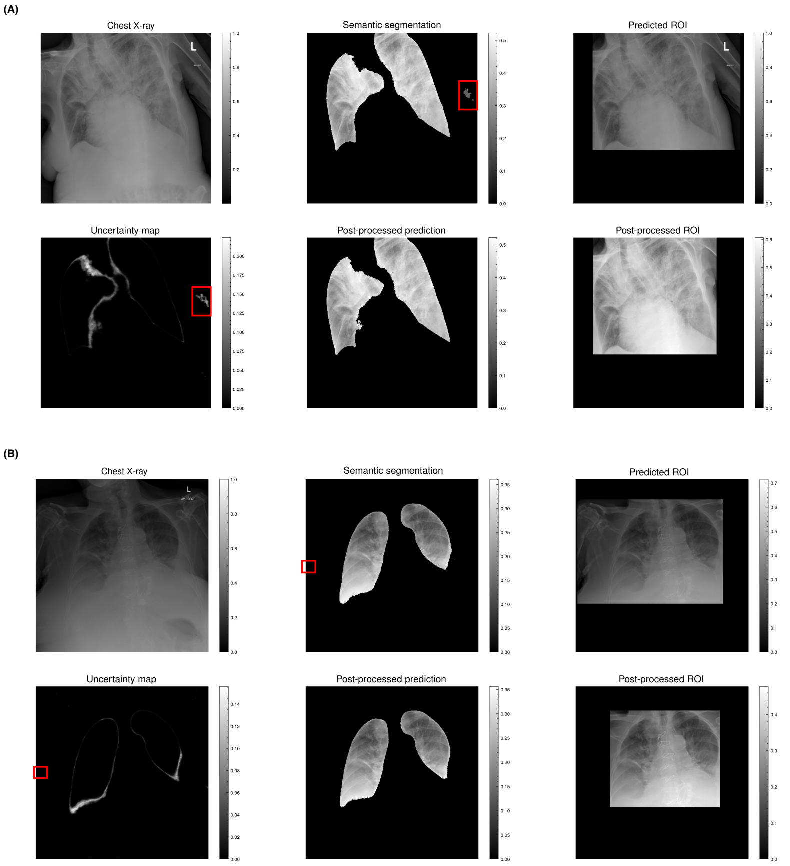


FIGURE 5.4: **Post-processing of COVID-19 CXR semantic lung segmentation to generate reliable ROIs.** **(A)** An example of severe COVID-19 in a CXR resulting in an uncertain prediction, with a total prediction uncertainty of 620. The right lung field is under-segmented, the additional structure is highlighted in the semantic segmentation and corresponding uncertainty map. This creates an overly large ROI around the lungs. Pixel-wise prediction uncertainty is used to isolate the extra structure, which is eliminated during post-processing of the semantic segmentation. **(B)** A COVID-19 CXR with under-segmentation of the lung fields (highlighted in the semantic segmentation and uncertainty map). Pixel-wise uncertainty is used to remove this structure from the semantic segmentation before creating a ROI. Total prediction uncertainty of this example is 230. *Abbrvs: Region Of Interest (ROI); Chest X-ray (CXR).*

Bimodality using Hartigans' dip test. Predictions that produce bimodal pixel-wise uncertainty frequency distributions, and give a total uncertainty below an empirically defined threshold, are highlighted as likely erroneous predictions and excess structures are iteratively eliminated according to greatest total uncertainty per segmented area i.e., structures with the greatest density of uncertainty are removed first. Figure 5.5 provides an overview of this post-processing algorithm and Figure 5.4 gives a visual example of its application. Predictions that exceed the total uncertainty limit are put forward for manual inspection. As a result of preliminary experiments, I applied a total uncertainty limit of 800, which I found facilitated selection of the best candidates for post-processing. Once this process is applied, I crop CXRs to the remaining segmented areas, this produces the region of interest (ROI). I use ROI instead of semantic segmentation for added robustness and to ensure that all clinically significant thoracic structures are included e.g., the mediastinum.

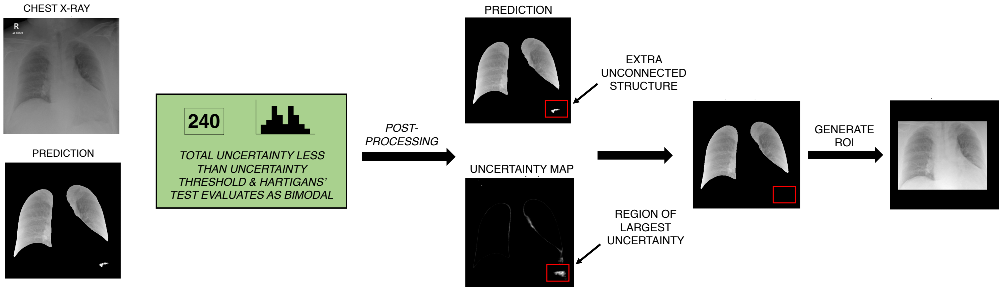

CHEST X-RAY
PREDICTION
240
TOTAL UNCERTAINTY LESS THAN UNCERTAINTY THRESHOLD & HARTIGANS' TEST EVALUATES AS BIMODAL
POST-PROCESSING
PREDICTION
EXTRA UNCONNECTED STRUCTURE
UNCERTAINTY MAP
REGION OF LARGEST UNCERTAINTY
GENERATE ROI

FIGURE 5.5: **Example of unsupervised lung segmentation post-processing algorithm on NCCID data.** The U-Net++ model is used to generate semantic segmentation of the left and right lung field. Monte Carlo dropout is applied to approximate uncertainty of prediction, total uncertainty is calculated and frequency of uncertainty is evaluated for Bimodality with Hartigans' test. In this example, uncertainty is less than the threshold for required manual inspection and prediction uncertainty is bimodal so automatic post-processing is applied. In post-processing, unconnected structures are identified and the density of uncertainty is calculated per structure. Excess structures (more than the two lung fields) are iteratively removed, with the most uncertain structures removed first. A ROI is generated from the post-processed semantic segmentation, the ROI was selected to be the minimum bounding box around the segmented lung fields. *Abbrvs: Region of interest (ROI); National COVID-19 Chest Imaging Database (NCCID).*

With a total uncertainty threshold of 800, region of interest (ROI) prediction Dice scores improved from 0.96 to 0.98. While improvements in scores on data from the same training distribution are modest, it is expected that applying the proposed uncertainty-based post-processing algorithm will help improve overall ROI-extraction accuracy for CXR data from unseen domains. Qualitative evaluation of segmentations performed on the NCCID training data showed that applying the post-processing algorithm improved the accuracy and robustness of predicted ROIs (Fig. 5.6).

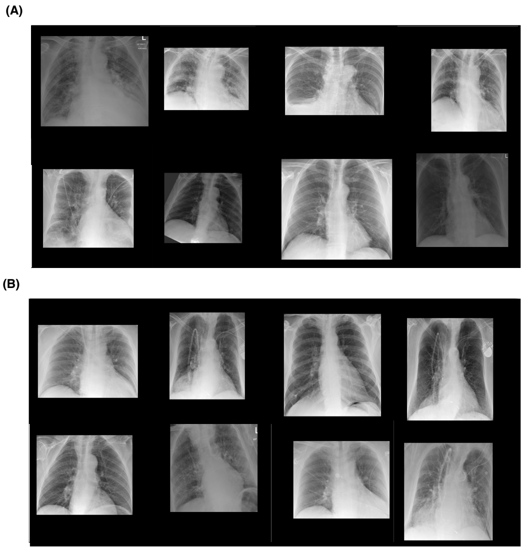


FIGURE 5.6: **ROI-extractions of post-processed semantic lung segmentations. (A)** Examples of ROI-extracted NCCID control cases. **(B)** ROI-extracted NCCID COVID-19 cases. *Abbrvs: Region Of Interest (ROI); National COVID-19 Chest Imaging Database (NCCID).*

5.2.4 Performance evaluation

I evaluated predictive performance on multiple independent test populations, with classification thresholds set to 0.5 for ease of comparison. However, this may be a limitation for performance accuracy metrics that rely on a probability threshold. To compare the classification performance of all models, I evaluated performance metrics, such as, accuracy, precision, recall, F1, and AUROC. To consider average performance over all iterations of the 5-fold cross-validation, I calculated confidence intervals for all ROC curves and mean ± standard deviations for classification metrics. Models were ranked according to their individual performance metrics and all metric rankings were considered equally to give an overall model ranking. I performed Tukey's honestly significant difference (HSD) statistical test to compare model performance. I investigated model explanation techniques, including Grad-CAM and guided backpropagation visualisation methods. Additionally, I trained the top-performing models on ROI-extracted CXRs, allowing us to directly compare these ROI-trained models with their counterparts trained on the entire CXR.

Model evaluation in national and international hospital populations Model capacity to generalise to national populations was evaluated using external NHS hospital data from LTHT. With this evaluation I estimate how the models perform in an unseen hospital trust, in which patient demographics and clinical practices may vary.

Furthermore, I conducted an assessment of model generalisability to international hospital populations, utilising data from the Grenada Hospital in Spain (referred to as COVIDGR). Note that the evaluation of international generalisability is limited due to an uncontrollable label shift across patient populations (a consequence of different labelling strategies).

Model performance under counterfactual conditions I also created counterfactual datasets from a subset of LTHT data for which non-COVID-19 pneumonia labels are also available. I adjusted the definition of the positive and control classes in LTHT data, resulting in the creation of two alternative scenarios for comparison of model performance under counterfactual conditions. The first scenario, referred to as LTHT (P), encompasses a general pneumonia detection scenario where the positive class includes non-COVID-19 pneumonia CXRs. The second scenario, named LTHT (NP), represents a COVID-19 detection scenario where all instances of non-COVID-19 pneumonia were excluded. I evaluate models according to standard performance metrics and perform sub-population analysis under counterfactual conditions.

Model performance variation by patient-level factors Sub-population analysis was performed on LTHT data. I assessed model performance across different patient sub-populations, grouped according to ethnicity, age, sex, smoking status, and the presence of comorbidities within the CXR. To create CXR-observable comorbidity subgroups I convert patient recorded comorbidities into a binary label that describes whether the comorbidity is likely observable in the CXR. Moreover, ethnic subgroups are defined according to NHS ethnic categories, which I in turn group into five larger populations: Black, White, Asian, Multiple and Other. 'Other' describes any ethnicity that does not fall under the aforementioned ethnic categories, cases with unknown ethnicity are not considered in my analysis.

I perform one-way analysis of variance (ANOVA) tests to evaluate the statistical significance of differences in model performance across different subgroups. I also assessed model error rate and its correlations with various clinical and demographic factors. I examined the effects of CXR projection, RALE-defined CXR severity, and proximity to the COVID-19 diagnostic window on the rates of false positive and false negative predictions. Refer to Table 5.3 for a summary of which dataset is used for each specific task.

5.3 Results

In this section I present the results of my study, providing a comprehensive description of the key findings and observations drawn from the analysis of the considered models and datasets.

5.3.1 Evaluation of model performance in national and international hospital populations

All models generalised well to a national-level but perform poorly when applied to international datasets. Table 5.6 shows that there is a marginal decrease in model performance when applied beyond the training domain (NCCID) to the unseen NHS trust population (LTHT).

During LTHT population testing AUROC scores ranged from 0.65 to 0.78. XCEPTION NET (Khan, Shah, and Bhat, 2020), XVITCOS (Mondal et al., 2022) and SSL-AM (Park, Kwak, and Lim, 2021) emerged as top-performing models, with AUROCs between 0.74 and 0.78. I identify RES. ATTN. (Sharma and Dyreson, 2020), CAPSNET (Afshar et al., 2020), and FUSENET (Abdar et al., 2023) as the poorest performing models. Table 5.6 shows that RES. ATTN., the only model without a pre-training strategy, gives the lowest performance across all evaluated metrics. RES. ATTN., FUSENET and CAPSNET AUROCs scores are lower than all other models, this is statistically significant to a confidence interval of 95% (Tukey multiple comparisons tests, $p < 0.05$; Table 5.5).

Even top-performing models are susceptible to returning high rates of false positives, as evidenced by universally low precision scores (Table 5.6). However, even without classification threshold tuning, the top-performing models detect COVID-19 in LTHT populations similarly to radiologist performance in a variety of performance metrics. Model accuracy scores ranged from 0.69 to 0.75 and one study reports the average accuracy scores of radiologist groups as between 0.76 to 0.84, depending on professional experience (Cozzi et al., 2020). Comparison with another study shows that the best performing model AUROCs exceed radiologist performance, with scores of 0.78 compared to 0.71 (Albiol et al., 2022).

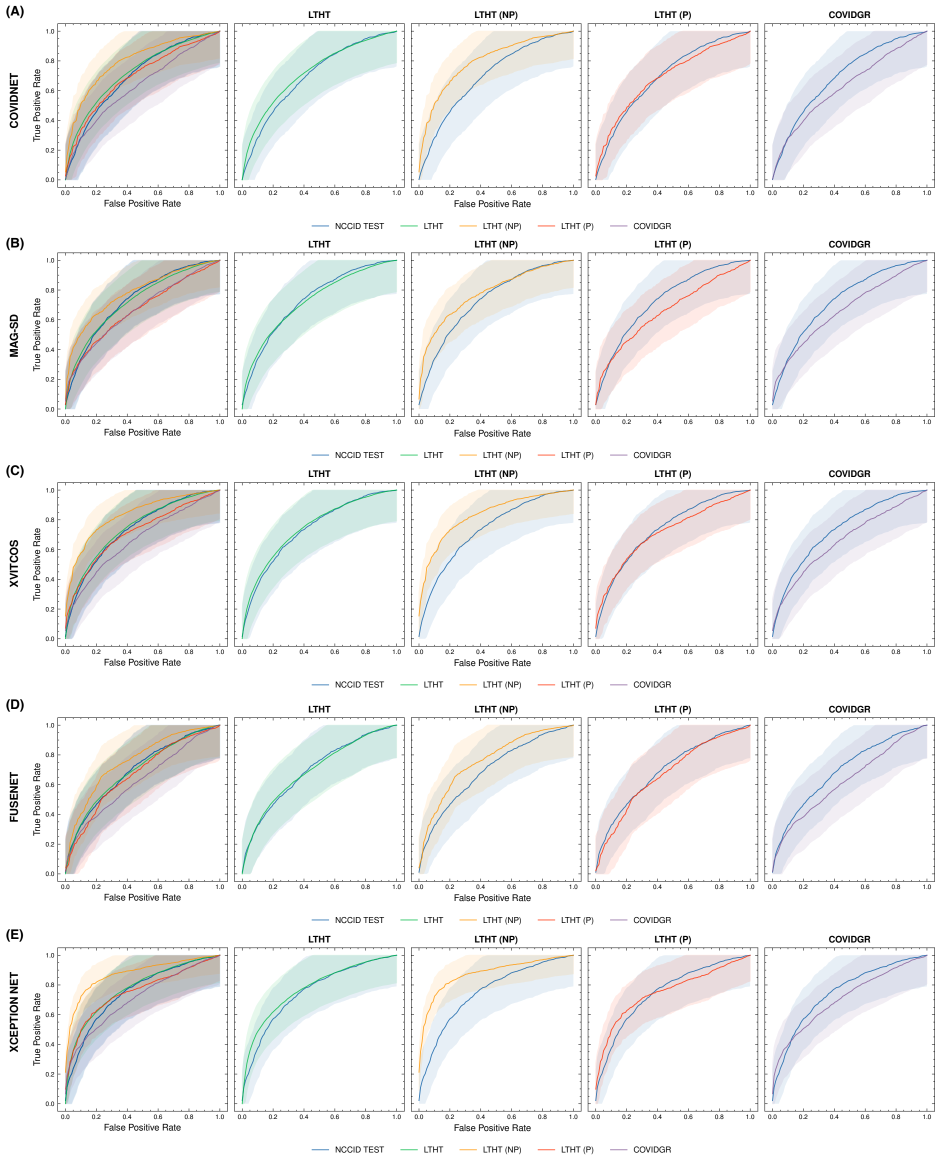
Figure 5.7 shows a significant drop in performance when models are applied to an international dataset (COVIDGR). I found CORONET gives the most substantial decrease in performance, with model recall halving from LTHT (0.52) to COVIDGR (0.26). This decline in performance is further evidenced by a large drop in AUROC values from 0.70 in the national population (LTHT) to 0.60 in the international population (COVIDGR) (Table 5.6).

		Acc.		**AUROC**		**F1**		**Prec.**		**Recall**	
Group 1	*Group 2*	*mean diff.*	*p-adj*	*mean diff.*	*p-adj*	*mean diff.*	*p-adj*	*mean diff.*	*p-adj*	*mean diff.*	*p-adj*
CAPSNET	CORONET	0.0550	0.6572	0.0571	0.0308	0.0314	0.8277	0.0416	0.4585	-0.0005	1.0000
CAPSNET	COVIDNET	0.0165	0.9999	0.0779	0.0008	0.0607	0.0814	0.039	0.5505	0.1243	0.3164
CAPSNET	ECOVNET	0.0862	0.1107	0.0748	0.0014	0.0684	0.0307	0.0767	0.0067	0.0046	1.0000
CAPSNET	FUSENET	-0.0127	1.0000	0.0499	0.0913	0.0342	0.7488	0.0195	0.9873	0.1039	0.5636
CAPSNET	MAG-SD	0.0223	0.9985	0.0818	0.0004	0.0630	0.0616	0.0426	0.4272	0.1174	0.3940
CAPSNET	RES. ATTN.	-0.0101	1.0000	-0.0000	1.0000	-0.0006	1.0000	-0.0051	1.0	0.0176	1.0000
CAPSNET	SSL-AM	0.0533	0.6936	0.0961	0.0000	0.0816	0.0047	0.0643	0.0405	0.1019	0.5889
CAPSNET	XCEPTION	0.0895	0.0856	0.1301	0.0000	0.125	0.0000	0.1056	0.0001	0.1316	0.2456
CAPSNET	XVITCOS	0.0360	0.9576	0.1019	0.0000	0.0822	0.0043	0.0565	0.1085	0.1426	0.1613
CORONET	COVIDNET	-0.0385	0.9366	0.0208	0.9503	0.0293	0.8772	-0.0027	1.0000	0.1248	0.3117
CORONET	ECOVNET	0.0312	0.9832	0.0177	0.9821	0.0370	0.6591	0.0351	0.6843	0.0050	1.0000
CORONET	FUSENET	-0.0677	0.3741	-0.0072	1.0000	0.0028	1.0000	-0.0221	0.9715	0.1043	0.5576
CORONET	MAG-SD	-0.0327	0.9770	0.0247	0.8705	0.0316	0.8226	0.0009	1.0000	0.1178	0.3886
CORONET	RES. ATTN.	-0.0650	0.4296	-0.0571	0.0307	-0.0321	0.8107	-0.0468	0.3006	0.0180	1.0000
CORONET	SSL-AM	-0.0016	1.0000	0.0390	0.3431	0.0502	0.2530	0.0226	0.9665	0.1024	0.5830
CORONET	XCEPTION	0.0345	0.9676	0.0730	0.0019	0.0935	0.0007	0.0640	0.0421	0.1321	0.2416
CORONET	XVITCOS	-0.0190	0.9996	0.0448	0.1793	0.0508	0.2382	0.0148	0.9983	0.1430	0.1583
COVIDNET	ECOVNET	0.0697	0.3349	-0.0031	1.0000	0.0077	1.0000	0.0378	0.5921	-0.1197	0.3663
COVIDNET	FUSENET	-0.0292	0.9893	-0.0280	0.7680	-0.0265	0.9285	-0.0194	0.9879	-0.0204	1.0000
COVIDNET	MAG-SD	0.0058	1.0000	0.0039	1.0000	0.0023	1.0000	0.0036	1.0000	-0.0070	1.0000
COVIDNET	RES. ATTN.	-0.0265	0.9946	-0.0779	0.0008	-0.0614	0.0753	-0.0441	0.3784	-0.1067	0.5264
COVIDNET	SSL-AM	0.0369	0.9509	0.0182	0.9786	0.0208	0.9842	0.0253	0.9345	-0.0224	1.0000
COVIDNET	XCEPTION	0.0730	0.2759	0.0523	0.0648	0.0642	0.0531	0.0666	0.0293	0.0073	1.0000
COVIDNET	XVITCOS	0.0195	0.9995	0.0240	0.8893	0.0215	0.9806	0.0175	0.9942	0.0183	1.0000
ECOVNET	FUSENET	-0.0989	0.0387	-0.0249	0.8649	-0.0342	0.7487	-0.0572	0.0994	0.0993	0.6232
ECOVNET	MAG-SD	-0.0639	0.4552	0.0070	1.0000	-0.0054	1.0000	-0.0342	0.7157	0.1128	0.4493
ECOVNET	RES. ATTN.	-0.0962	0.0488	-0.0748	0.0014	-0.0691	0.0282	-0.0819	0.0030	0.0130	1.0000
ECOVNET	SSL-AM	-0.0328	0.9764	0.0213	0.9432	0.0131	0.9995	-0.0125	0.9996	0.0974	0.6482
ECOVNET	XCEPTION	0.0033	1.0000	0.0553	0.0406	0.0565	0.1320	0.0289	0.8669	0.1271	0.2885
ECOVNET	XVITCOS	-0.0502	0.7606	0.0271	0.8007	0.0138	0.9993	-0.0203	0.9837	0.1380	0.1933
FUSENET	MAG-SD	0.0350	0.9642	0.0319	0.6163	0.0288	0.8877	0.0230	0.9629	0.0135	1.0000
FUSENET	RES. ATTN.	0.0027	1.0000	-0.0499	0.0910	-0.0349	0.7289	-0.0247	0.9435	-0.0863	0.7820
FUSENET	SSL-AM	0.0661	0.4080	0.0462	0.1498	0.0474	0.3249	0.0447	0.3591	-0.0019	1.0000
FUSENET	XCEPTION	0.1022	0.0289	0.0803	0.0005	0.0907	0.0012	0.0861	0.0016	0.0277	0.9999
FUSENET	XVITCOS	0.0487	0.7887	0.0520	0.0673	0.0480	0.3075	0.0369	0.6224	0.0387	0.9987
MAG-SD	RES. ATTN.	-0.0324	0.9785	-0.0818	0.0004	-0.0637	0.0568	-0.0477	0.2756	-0.0998	0.6172
MAG-SD	SSL-AM	0.0310	0.9837	0.0143	0.9962	0.0185	0.9930	0.0217	0.9744	-0.0154	1.0000
MAG-SD	XCEPTION	0.0672	0.3854	0.0483	0.1131	0.0619	0.0705	0.0630	0.0477	0.0143	1.0000
MAG-SD	XVITCOS	0.0137	1.0000	0.0200	0.9602	0.0192	0.9911	0.0139	0.999	0.0252	1.0000
RES. ATTN.	SSL-AM	0.0634	0.4654	0.0961	0.0000	0.0822	0.0043	0.0694	0.0199	0.0844	0.8029
RES. ATTN.	XCEPTION	0.0995	0.0366	0.1302	0.0000	0.1256	0.0000	0.1107	0.0000	0.114	0.4338
RES. ATTN.	XVITCOS	0.0461	0.8367	0.1019	0.0000	0.0829	0.0039	0.0616	0.0576	0.1250	0.3093
SSL-AM	XCEPTION	0.0361	0.9568	0.0341	0.5297	0.0434	0.4448	0.0413	0.4683	0.0297	0.9998
SSL-AM	XVITCOS	-0.0173	0.9998	0.0058	1.0000	0.0006	1.0000	-0.0078	1.0000	0.0406	0.9981
XCEPTION	XVITCOS	-0.0534	0.6913	-0.0283	0.7586	-0.0428	0.4655	-0.0491	0.2398	0.0110	1.0000

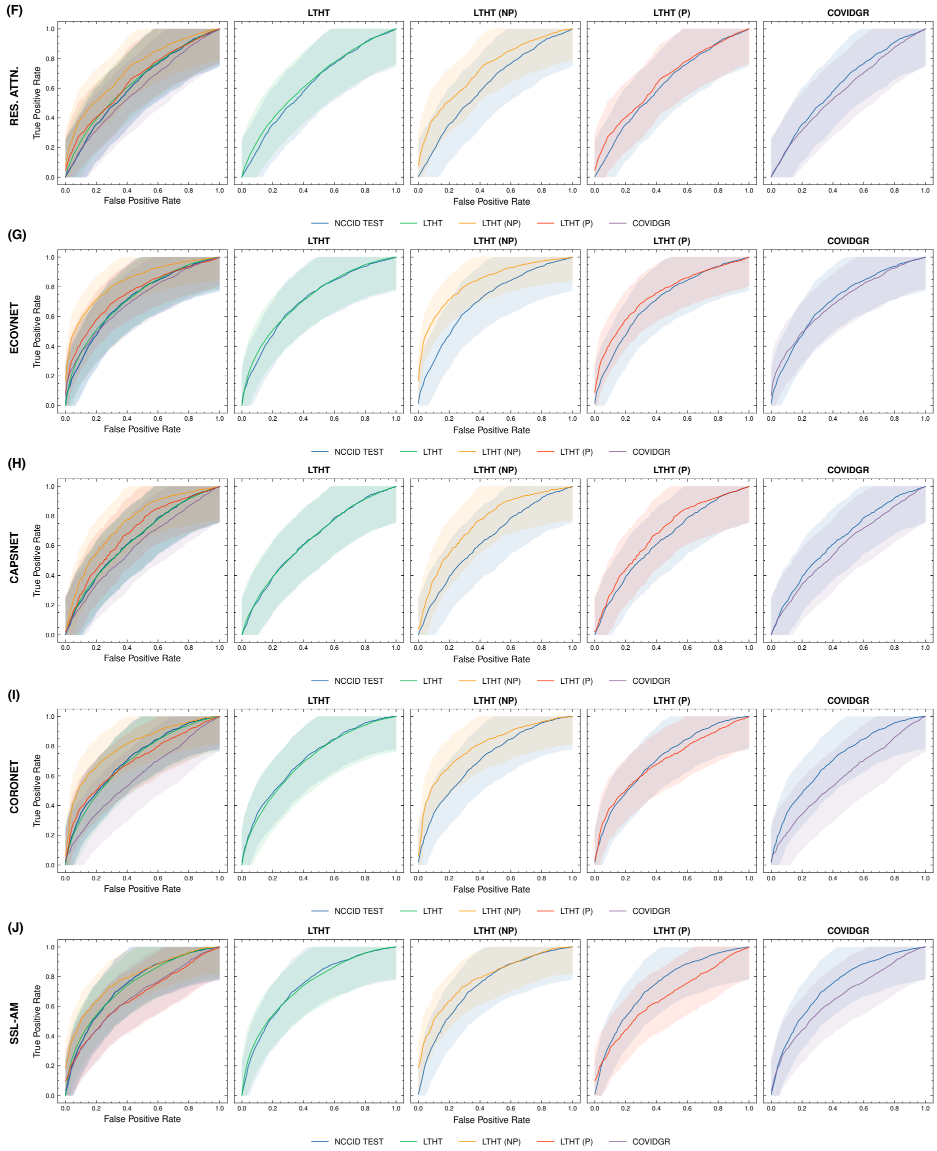
TABLE 5.5: **Multiple (pair-wise) model performance comparisons with Tukey's Honest Significant Difference (HSD) test.** *Abbrvs: Accuracy (Acc.); Precision (Prec.); Area Under the Receiver Operator Characteristic (AUROC); XCEPTION NET (XCEPTION); Adjusted p-value (p-adj).*

		MODELS									
		COVIDNET	*MAG-SD*	*XVITCOS*	*FUSENET*	*XCEPTION NET*	*RES. ATTN.*	*ECOVNET*	*COVID-CAPS*	*CORONET*	*SSL-AM*
NCCID TEST	*Acc.*	0.66±0.04	0.63±0.01	0.67±0.03	0.65±0.02	**0.70±0.02**	0.59±0.04	0.67±0.02	0.56±0.03	0.67±0.03	0.69±0.03
	Prec.	0.50±0.04	0.46±0.01	0.49±0.03	0.47±0.02	**0.53±0.02**	0.42±0.03	0.50±0.02	0.40±0.01	0.50±0.04	0.52±0.03
	Recall	0.63±0.12	**0.78±0.01**	0.73±0.07	0.64±0.07	0.69±0.04	0.61±0.11	0.66±0.03	0.69±0.05	0.64±0.10	0.70±0.05
	F1	0.54±0.01	0.58±0.00	0.59±0.01	0.54±0.02	**0.60±0.01**	0.49±0.02	0.57±0.02	0.51±0.01	0.56±0.02	0.59±0.01
	AUROC	0.70±0.01	0.73±0.01	0.74±0.01	0.70±0.01	**0.75±0.01**	0.63±0.01	0.71±0.01	0.65±0.01	0.72±0.01	0.74±0.02
N=824	*Ranking*	7	3	6	8	2	10	5	9	4	1
LTHT	*Acc.*	0.68±0.04	0.68±0.04	0.70±0.03	0.65±0.08	**0.75±0.01**	0.65±0.04	0.75±0.01	0.66±0.04	0.72±0.04	0.72±0.03
	Prec.	0.28±0.03	0.28±0.03	0.30±0.02	0.26±0.03	**0.34±0.01**	0.23±0.02	0.32±0.02	0.24±0.02	0.28±0.04	0.30±0.03
	Recall	0.64±0.05	0.63±0.06	**0.66±0.04**	0.62±0.10	0.65±0.02	0.53±0.07	0.52±0.03	0.52±0.05	0.52±0.16	0.62±0.03
	F1	0.38±0.01	0.39±0.03	0.41±0.01	0.36± 0.02	**0.45±0.01**	0.32±0.02	0.39±0.02	0.32±0.01	0.36±0.06	0.41±0.03
	AUROC	0.72±0.01	0.73±0.03	0.75±0.00	0.70±0.01	**0.78±0.01**	0.65±0.03	0.72±0.02	0.65±0.01	0.70±0.05	0.74±0.03
N=11,204	*Ranking*	6	5	2	8	1	10	4	9	7	3
LTHT (NP)	*Acc.*	0.71±0.04	0.70±0.05	0.74±0.03	0.68±0.08	**0.75±0.02**	0.59±0.07	0.62±0.04	0.58±0.04	0.60±0.15	0.69±0.03
	Prec.	0.96±0.01	0.96±0.01	0.97±0.01	0.96±0.01	**0.98±0.00**	0.95±0.01	0.98±0.01	0.96±0.01	0.97±0.01	0.96±0.01
	Recall	0.71±0.05	0.69±0.06	0.73±0.04	0.68±0.10	**0.73±0.02**	0.57±0.07	0.59±0.04	0.56±0.05	0.57±0.17	0.68±0.03
	F1	0.81±0.03	0.80±0.04	0.83±0.02	0.79±0.07	**0.84±0.01**	0.71±0.06	0.73±0.03	0.71±0.04	0.70±0.15	0.80±0.03
	AUROC	0.80±0.02	0.78±0.05	0.83±0.01	0.76±0.02	**0.88±0.02**	0.73±0.06	0.83±0.02	0.75±0.03	0.80±0.06	0.79±0.05
N=3,948	*Ranking*	3	6	2	8	1	9	4	10	7	5
LTHT (P)	*Acc.*	0.54±0.05	0.53±0.06	**0.55±0.04**	0.55±0.10	0.52±0.02	0.47±0.06	0.45±0.02	0.46±0.05	0.43±0.12	0.49±0.02
	Prec.	0.98±0.00	0.98±0.01	0.99±0.00	0.98±0.00	**0.99±0.00**	0.98±0.01	0.99±0.00	0.98±0.00	0.99±0.01	0.98±0.00
	Recall	0.53±0.06	0.52±0.06	**0.54±0.04**	0.54±0.11	0.50±0.02	0.46±0.07	0.43±0.02	0.45±0.06	0.41±0.13	0.48±0.03
	F1	0.69±0.05	0.68±0.05	**0.70±0.04**	0.69±0.09	0.67±0.02	0.62±0.06	0.60±0.02	0.61±0.05	0.57±0.13	0.65±0.02
	AUROC	0.69±0.03	0.67±0.04	0.72±0.02	0.67±0.02	**0.75±0.03**	0.66±0.06	0.75±0.02	0.69±0.03	0.70±0.05	0.66±0.04
N=1,451	*Ranking*	3	6	1	4	2	10	5	9	7	8
COVIDGR	*Acc.*	0.58±0.01	0.62±0.01	0.61±0.02	0.58±0.02	**0.63±0.01**	0.56±0.04	0.58±0.04	0.55±0.02	0.56±0.03	0.59±0.03
	Prec.	0.73±0.06	0.71±0.06	0.73±0.04	0.80±0.08	0.86±0.05	0.59±0.06	**0.90±0.07**	0.62±0.04	0.66±0.07	0.78±0.05
	Recall	0.31±0.11	**0.42±0.10**	0.34±0.04	0.23±0.09	0.30±0.02	0.38±0.11	0.18±0.09	0.27±0.07	0.26±0.11	0.26±0.10
	F1	0.42±0.08	**0.51±0.06**	0.46±0.04	0.34±0.11	0.44±0.02	0.46±0.10	0.29±0.13	0.37±0.07	0.36±0.12	0.38±0.10
	AUROC	0.63±0.03	0.67±0.02	0.67±0.03	0.63±0.01	**0.71±0.02**	0.59±0.10	0.70±0.03	0.60±0.03	0.60±0.05	0.67±0.01
N=852	*Ranking*	6	2	3	7	1	8	5	9	10	4

TABLE 5.6: **Average model performance metrics for each test dataset, with standard deviation, across cross-validation folds.** N is the total number of cases in each test population. **Bold** indicates the highest average metric per dataset. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Accuracy (Acc); Precision (Prec); Area Under the Receiver Operator Characteristic (AUROC).*



(A) **(A-E)** ROC curves of COVID-19 detection.



(B) **(F-J)** ROC curves of COVID-19 detection.

FIGURE 5.7: **ROC curves of COVID-19 detection.** Each row presents a different model, the first of each row presents ROC curves for all test data, the following columns present direct comparison between NCCID ROC curves and the dataset of interest, corresponding AUROC values can be found in Table 5.6. Shaded regions correspond to 95% confidence interval. *Abbrvs: Receiver Operating Characteristic (ROC); National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Area Under the Receiver Operating Characteristic (AUROC).*

5.3.2 Model performance under counterfactual conditions

The exploration of model performance in counterfactual scenarios gives insight into the impact of confounding disease features on COVID-19 detection i.e., non-COVID-19 pneumonia. In LTHT (NP) which removes other pneumonia types from the population, I observed clear improvements. All models achieve near perfect precision scores, ranging from 0.95 to 0.98, as well as substantially improved AUROC and recall scores (Fig. 5.7 and Table 5.6). XCEPTION NET, the best-performing model in real-world LTHT cohorts, further improved with increases in AUROC scores from 0.78 to 0.88, recall scores from 0.65 to 0.73, and precision scores from 0.34 to 0.98 (Table 5.6).

When evaluating models under the alternative counterfactual scenario, using LTHT (P) where both COVID-19 and non-COVID-19 pneumonia are treated as the positive class i.e., models become general pneumonia classifiers, model performances diverge relative to performance on LTHT. Top-performing models in real-world data (LTHT) decrease in performance, as evidenced by especially reduced recall scores (Table 5.6). Of the best performing models I observe the greatest decline in AUROC in SSL-AM, from 0.74 (LTHT) to 0.66, suggesting that SSL-AM is best able to isolate features of COVID-19 pneumonia from non-COVID-19 pneumonia. The worst performing models (RES. ATTN., ECOVNET, CAPSNET and CORONET) all demonstrate improved performance on LTHT (P), suggesting that these were unable to learn to separate features of COVID-19 pneumonia from other pneumonia types.

When comparing top-performing models with their ROI-trained counterparts, the reduction in their performance under this counterfactual is greater. For example, XCEPTION NET (ROI) AUROC falls from 0.77 to 0.71, while the decrease in XCEPTION NET performance is less substantial. This indicates that ROI-trained models, compared to their full CXR trained counterparts, may have improved capacity for separating COVID-19 from non-COVID-19 pneumonia (Fig. 5.14b).

5.3.3 Subgroup analysis

During sub-population analysis with independent NHS hospital data (LTHT), I observe disparities in model performance across demographic subgroups. I consider sex, ethnicity, age, smoking, and subgroups with comorbidities that are likely observable in a CXR. I report smoking and comorbidity analysis together due to overlap in clinical interest. Subgroups are described in detail in Section 5.2.

Performance variations across demographic and clinical populations raise FAIR principle concerns regarding fairness, accountability, interpretability, and reliability. While overall model performance was satisfactory, certain subgroups demonstrated notable disparities in performance metrics, indicating potential equity issues in deployment across diverse patient populations. To mitigate these FAIR concerns I recommend establishing clearer indications for use that explicitly define optimal performance populations and contexts where alternative approaches may be more appropriate. Additionally, developing specialised models tailored to specific subgroups could address observed performance disparities through more representative training datasets. Future iterations should incorporate enhanced data collection strategies for balanced subgroup representation and implement ongoing monitoring systems to detect real-world performance degradation.

Sex I found that, according to AUROC values, models perform better in male populations compared to female populations (Fig 5.9 and Table 5.8). There is a consistent pattern of increased false negative predictions in the female population i.e., a greater proportion female COVID-19 cases are missed (Fig. 5.8). Statistical significance in model AUROC disparities is confirmed in 5 out of 10 models (One-way ANOVA, $p < 0.05$; Table 5.7).

Sex bias persists even under counterfactual conditions; with the exception of XVITCOS, I observe this bias in models when applied to populations without alternative pneumonia types, LTHT(NP). This suggests that this bias cannot be due to differences in the prevalence of the non-COVID-19 pneumonia across the sexes. Upon further examination using real-world data (LTHT), I see that sex bias is not mitigated by ROI-extraction, the ROI-trained version of XVITCOS returns a higher rate of false negatives in the female population compared to the male population which is reflected in a larger recall scores in males (0.75) compared to females (0.62)(Table 5.8 and Fig. 5.8).

Ethnicity I also see statistically significant model performance disparities across ethnic subgroups (One-way ANOVA, $p < 0.05$; Table 5.7). XCEPTION NET AUROCs vary from 0.72 to 0.91 (Table 5.8). All models appear to perform better when applied to Black and Asian groups, with significantly fewer false negatives. Table 5.8 shows all models return higher precision score when applied to Black and Asian groups compared to White groups. XCEPTION NET precision falls from 0.73 (Black) and 0.52 (Asian) to 0.37 (White). This disparity of performance is also observed in the counterfactual without other pneumonia types, LTHT (NP). The performance gap between White, and Black and Asian groups, is unchanged with the use of ROI-trained models i.e., training on ROI CXRs has no effect.

Age I can observe similar statistically significant disparities in model performance across different age groups (One-way ANOVA, $p < 0.05$; Table 5.7). Generally, models perform best in the 50-74 age group, which is in line with COVID-19 prevalence by age group in the training data (Table 5.1). However, I observe that top-performing models appear more likely to return false negatives for the 75-99 age group, an age group at greater risk of adverse COVID-19 outcomes (Fig. 5.8). Under counterfactual conditions, where all pneumonia types are included in the positive class, I see XCEPTION NET and XVITCOS models improve in performance in the 75-99 age group, indicating a reduced ability to separate COVID-19 pneumonia from non-COVID-19 pneumonia in older age groups (Fig. 5.8). These comparisons should be interpreted cautiously, considering that the prevalence of non-COVID-19 pneumonia differs among subgroups.

Figure 5.8 shows that models return the lowest false positive rate in the youngest age group i.e., 0-24 years, although this may be due to low prevalence of COVID-19

		SSL-AM		MAG-SD		COVIDNET		XVITCOS		FUSENET		CORONET		RES. ATTN.		ECOVNET		CAPSNET		XCEPTION NET	
		F-statistic	*P-value*	*F-statistic*	*P-value*	*F-statistic*	*P-value*	*F-statistic*	*P-value*	*F-statistic*	*P-value*	*F-statistic*	*P-value*	*F-statistic*	*P-value*	*F-statistic*	*P-value*	*F-statistic*	*P-value*	*F-statistic*	*P-value*
Age	Acc.	6.4300	0.0000	1.4200	0.2600	2.0600	0.1300	4.9600	0.0100	0.3500	0.8400	1.3100	0.3000	1.6400	0.2000	102.3600	0.0000	6.1800	0.0000	49.0200	0.0000
	AUROC	4.8300	0.0100	11.5200	0.0000	46.8300	0.0000	13.0300	0.0000	28.6700	0.0000	2.3200	0.0900	3.3500	0.0300	17.8500	0.0000	29.0300	0.0000	17.5500	0.0000
	F1	199.8700	0.0000	78.9200	0.0000	183.3100	0.0000	178.3900	0.0000	101.2100	0.0000	21.3900	0.0000	43.8200	0.0000	125.2400	0.0000	121.2800	0.0000	149.3500	0.0000
	Prec.	175.5400	0.0000	102.4400	0.0000	118.8600	0.0000	144.4000	0.0000	68.9700	0.0000	45.2500	0.0000	53.6800	0.0000	223.9200	0.0000	110.5500	0.0000	169.0900	0.0000
	Recall	8.5800	0.0000	4.8000	0.0100	7.3400	0.0000	7.3700	0.0000	2.4500	0.0800	0.9200	0.4700	1.0700	0.4000	9.5900	0.0000	7.3700	0.0000	18.1300	0.0000
Comorbidity	Acc.	7.6700	0.0200	3.9500	0.0800	17.7100	0.0000	24.2000	0.0000	6.5300	0.0300	0.2800	0.6100	2.4600	0.1600	2.7000	0.1400	2.2800	0.1700	118.0900	0.0000
	AUROC	6.4400	0.0300	5.5100	0.0500	102.1400	0.0000	25.4400	0.0000	21.1600	0.0000	1.0600	0.3300	9.3300	0.0200	9.6400	0.0100	0.0300	0.8700	132.0600	0.0000
	F1	26.9400	0.0000	6.1400	0.0400	11.6400	0.0100	25.0100	0.0000	3.5200	0.1000	0.4500	0.5200	3.1800	0.1100	10.9300	0.0100	0.1900	0.6700	109.8800	0.0000
	Prec.	73.9700	0.0000	65.4300	0.0000	187.5000	0.0000	87.0800	0.0000	43.3800	0.0000	11.3500	0.0100	59.2800	0.0000	157.6400	0.0000	59.9000	0.0000	276.5400	0.0000
	Recall	4.2400	0.0700	0.5400	0.4800	0.6500	0.4400	0.1200	0.7400	0.3000	0.6000	0.0900	0.7700	0.1700	0.6900	1.7300	0.2300	0.7800	0.4000	8.5900	0.0200
Ethnicity	Acc.	3.2700	0.0200	0.7400	0.6000	2.1900	0.0900	2.9100	0.0300	1.3100	0.2900	0.4000	0.8400	5.1300	0.0000	51.2000	0.0000	0.2600	0.9300	52.3700	0.0000
	AUROC	19.1600	0.0000	9.5900	0.0000	24.1500	0.0000	22.5600	0.0000	96.8300	0.0000	1.1200	0.3800	3.8500	0.0100	28.2900	0.0000	8.9300	0.0000	104.8200	0.0000
	F1	74.1200	0.0000	33.6400	0.0000	117.8800	0.0000	107.6800	0.0000	142.6400	0.0000	3.6800	0.0100	24.1000	0.0000	41.7800	0.0000	55.4400	0.0000	56.0500	0.0000
	Prec.	81.7900	0.0000	27.1300	0.0000	61.0000	0.0000	81.0100	0.0000	34.4500	0.0000	13.3000	0.0000	82.8800	0.0000	8.6400	0.0000	55.8700	0.0000	51.8300	0.0000
	Recall	31.0900	0.0000	28.7100	0.0000	13.2500	0.0000	10.9200	0.0000	4.7500	0.0000	0.3900	0.8500	1.4100	0.2600	52.7400	0.0000	13.0000	0.0000	42.8200	0.0000
Gender	Acc.	0.2300	0.6400	8.9000	0.0200	10.2800	0.0100	1.6800	0.2300	3.8900	0.0800	0.1600	0.7000	0.3000	0.6000	9.0400	0.0200	0.6600	0.4400	34.9800	0.0000
	AUROC	3.1400	0.1100	3.7300	0.0900	23.2500	0.0000	23.1000	0.0000	1.8100	0.2200	0.2500	0.6300	0.2500	0.6300	7.0800	0.0300	13.4500	0.0100	26.7800	0.0000
	F1	1.0200	0.3400	18.2600	0.0000	18.2700	0.0000	3.0000	0.1200	5.6800	0.0400	1.0400	0.3400	0.9700	0.3500	38.4300	0.0000	6.6800	0.0300	46.8600	0.0000
	Prec.	4.1000	0.0800	44.6300	0.0000	106.6500	0.0000	16.0900	0.0000	30.0400	0.0000	6.8500	0.0300	30.8700	0.0000	176.7200	0.0000	137.2700	0.0000	12.6300	0.0100
	Recall	0.6700	0.4400	7.7300	0.0200	5.7400	0.0400	2.2700	0.1700	1.7800	0.2200	0.6700	0.4400	0.0100	0.9400	23.9600	0.0000	0.6000	0.4600	34.4600	0.0000
Smoker	Acc.	1.4600	0.2700	0.3000	0.7400	0.2700	0.7700	0.2800	0.7600	0.0300	0.9700	0.3000	0.7400	0.5200	0.6100	32.1700	0.0000	0.1800	0.8400	4.0400	0.0500
	AUROC	6.9800	0.0100	12.1800	0.0000	31.4000	0.0000	19.8900	0.0000	20.6300	0.0000	2.0500	0.1700	8.1700	0.0100	15.9800	0.0000	23.7800	0.0000	42.3900	0.0000
	F1	56.9300	0.0000	111.7300	0.0000	212.9700	0.0000	189.8600	0.0000	99.3200	0.0000	12.7900	0.0000	104.2500	0.0000	64.5700	0.0000	144.0400	0.0000	244.0800	0.0000
	Prec.	57.9300	0.0000	85.2600	0.0000	87.5600	0.0000	115.0600	0.0000	36.4900	0.0000	24.7700	0.0000	138.2100	0.0000	91.7100	0.0000	79.4300	0.0000	279.6400	0.0000
	Recall	10.3100	0.0000	7.3400	0.0100	2.4400	0.1300	1.4100	0.2800	0.9500	0.4100	0.5300	0.6000	3.9500	0.0500	16.7600	0.0000	12.8400	0.0000	10.0200	0.0000

TABLE 5.7: **One-way analysis of variance (ANOVA) for comparison of mean model performance on population subgroups.** *Abbrvs: Accuracy (Acc.); Precision (Prec.); Area Under the Receiver Operator Characteristic (AUROC).*

		Ethnicity					**Age Group**					**CXR Comorbidity**		**Smoking Status**			**Sex**	
Subgroups		*White*	*Black*	*Asian*	*Multiple*	*Other*	*0-24*	*25-49*	*50-74*	*75-99*	*100+*	*Y*	*N*	*Never*	*Former*	*Current*	*M*	*F*
COVIDNET	*Acc.*	0.68±0.04	0.74±0.01	0.71±0.04	0.68±0.07	0.73±0.05	0.75±0.08	0.68±0.04	0.70±0.03	0.65±0.04	0.68±0.05	0.59±0.00	0.63±0.02	0.68±0.04	0.66±0.04	0.68±0.04	0.63±0.02	0.59±0.01
	Prec.	0.30±0.03	0.66±0.05	0.46±0.05	0.17±0.04	0.54±0.06	0.03±0.01	0.17±0.02	0.39±0.03	0.25±0.02	0.20±0.04	0.62±0.01	0.72±0.01	0.31±0.03	0.18±0.02	0.13±0.01	0.72±0.01	0.64±0.01
	Recall	0.60±0.05	0.81±0.06	0.76±0.05	0.48±0.15	0.84±0.05	0.80±0.10	0.68±0.06	0.77±0.04	0.51±0.06	0.63±0.14	0.62±0.01	0.65±0.06	0.68±0.05	0.57±0.05	0.59±0.11	0.68±0.05	0.58±0.05
	F1	0.40±0.01	0.72±0.12	0.57±0.01	0.25±0.06	0.65±0.04	0.05±0.01	0.28±0.01	0.51±0.01	0.33±0.00	0.30±0.05	0.62±0.02	0.68±0.03	0.42±0.02	0.28±0.01	0.21±0.02	0.70±0.03	0.61±0.03
	AUROC	0.71±0.01	0.85±0.01	0.80±0.01	0.58±0.10	0.86±0.02	0.84±0.03	0.74±0.01	0.80±0.00	0.63±0.01	0.71±0.04	0.61±0.01	0.66±0.01	0.75±0.00	0.65±0.01	0.69±0.03	0.65±0.00	0.62±0.01
MAG-SD	*Acc.*	0.69±0.04	0.72±0.03	0.71±0.04	0.65±0.17	0.75±0.07	0.69±0.16	0.68±0.06	0.73±0.03	0.65±0.04	0.58±0.09	0.59±0.02	0.62±0.02	0.70±0.03	0.67±0.04	0.68±0.06	0.63±0.02	0.58±0.03
	Prec.	0.30±0.03	0.63±0.03	0.46±0.04	0.20±0.09	0.56±0.09	0.03±0.01	0.18±0.03	0.41±0.03	0.26±0.03	0.22±0.06	0.63±0.02	0.72±0.01	0.32±0.03	0.18±0.02	0.13±0.02	0.73±0.02	0.63±0.02
	Recall	0.59±0.07	0.80±0.03	0.75±0.04	0.53±0.05	0.83±0.04	0.83±0.06	0.68±0.06	0.73±0.03	0.54±0.09	0.55±0.21	0.61±0.06	0.64±0.06	0.67±0.05	0.54±0.04	0.59±0.05	0.68±0.06	0.56±0.06
	F1	0.40±0.02	0.71±0.02	0.57±0.03	0.28±0.10	0.67±0.07	0.05±0.03	0.28±0.04	0.52±0.02	0.35±0.03	0.22±0.06	0.62±0.04	0.68±0.03	0.43±0.02	0.27±0.02	0.21±0.02	0.70±0.03	0.59±0.04
	AUROC	0.71±0.02	0.81±0.02	0.79±0.03	0.62±0.12	0.86±0.05	0.86±0.09	0.74±0.04	0.80±0.01	0.66±0.04	0.61±0.08	0.63±0.02	0.66±0.02	0.76±0.02	0.67±0.02	0.69±0.03	0.65±0.02	0.62±0.03
XVITCOS	*Acc.*	0.70±0.03	0.77±0.02	0.70±0.03	0.71±0.08	0.76±0.03	0.78±0.06	0.69±0.04	0.72±0.03	0.67±0.03	0.64±0.07	0.60±0.01	0.63±0.01	0.70±0.03	0.68±0.04	0.69±0.03	0.64±0.00	0.59±0.01
	Prec.	0.32±0.02	0.69±0.03	0.44±0.02	0.22±0.06	0.56±0.04	0.02±0.01	0.18±0.01	0.40±0.03	0.27±0.01	0.19±0.04	0.63±0.01	0.71±0.01	0.32±0.02	0.21±0.01	0.15±0.01	0.73±0.01	0.64±0.01
	Recall	0.62±0.04	0.81±0.03	0.77±0.02	0.60±0.15	0.90±0.04	0.63±0.08	0.66±0.05	0.77±0.03	0.55±0.05	0.63±0.07	0.66±0.03	0.67±0.04	0.68±0.04	0.62±0.06	0.69±0.08	0.70±0.03	0.60±0.04
	F1	0.42±0.01	0.75±0.02	0.56±0.01	0.32±0.06	0.69±0.03	0.04±0.01	0.28±0.02	0.53±0.02	0.29±0.05	0.29±0.05	0.64±0.01	0.69±0.02	0.43±0.02	0.31±0.01	0.24±0.01	0.71±0.01	0.62±0.02
	AUROC	0.74±0.01	0.85±0.01	0.80±0.01	0.70±0.06	0.87±0.01	0.73±0.03	0.73±0.01	0.81±0.01	0.68±0.01	0.70±0.05	0.63±0.01	0.66±0.01	0.77±0.01	0.71±0.01	0.74±0.02	0.65±0.01	0.64±0.01
XVITCOS (ROI)	*Acc.*	0.67±0.01	0.71±0.04	0.68±0.03	0.65±0.03	0.70±0.03	0.64±0.03	0.63±0.02	0.69±0.02	0.65±0.02	0.59±0.04	0.61±0.01	0.64±0.01	0.67±0.02	0.63±0.02	0.62±0.02	0.64±0.01	0.59±0.01
	Prec.	0.30±0.01	0.62±0.04	0.42±0.03	0.18±0.02	0.49±0.03	0.01±0.00	0.15±0.01	0.38±0.01	0.27±0.01	0.17±0.02	0.62±0.01	0.71±0.01	0.30±0.01	0.18±0.01	0.12±0.01	0.71±0.01	0.64±0.01
	Recall	0.67±0.03	0.82±0.02	0.81±0.03	0.63±0.08	0.88±0.04	0.70±0.06	0.68±0.03	0.80±0.02	0.59±0.03	0.73±0.08	0.68±0.02	0.70±0.02	0.71±0.03	0.64±0.04	0.66±0.01	0.75±0.02	0.62±0.02
	F1	0.41±0.00	0.70±0.03	0.55±0.02	0.28±0.03	0.63±0.03	0.03±0.00	0.25±0.01	0.52±0.01	0.37±0.01	0.28±0.04	0.65±0.01	0.71±0.01	0.42±0.01	0.28±0.01	0.20±0.01	0.73±0.01	0.63±0.01
	AUROC	0.73±0.01	0.82±0.02	0.80±0.02	0.68±0.05	0.85±0.03	0.79±0.04	0.72±0.01	0.81±0.01	0.68±0.01	0.68±0.05	0.64±0.01	0.66±0.02	0.76±0.01	0.70±0.02	0.69±0.01	0.64±0.01	0.63±0.01
FUSENET	*Acc.*	0.65±0.07	0.70±0.04	0.68±0.05	0.75±0.01	0.71±0.04	0.69±0.09	0.65±0.10	0.67±0.07	0.62±0.07	0.65±0.09	0.57±0.01	0.61±0.03	0.65±0.07	0.64±0.01	0.65±0.10	0.60±0.03	0.57±0.02
	Prec.	0.28±0.03	0.62±0.06	0.42±0.09	0.25±0.06	0.51±0.05	0.02±0.00	0.15±0.02	0.37±0.05	0.24±0.02	0.16±0.04	0.60±0.01	0.70±0.03	0.28±0.04	0.17±0.02	0.12±0.02	0.70±0.03	0.63±0.02
	Recall	0.59±0.10	0.77±0.07	0.73±0.02	0.50±0.16	0.80±0.08	0.68±0.06	0.60±0.12	0.71±0.09	0.50±0.10	0.50±0.17	0.60±0.09	0.64±0.10	0.65±0.09	0.55±0.09	0.59±0.11	0.66±0.10	0.56±0.10
	F1	0.37±0.01	0.31±0.04	0.53±0.02	0.31±0.04	0.62±0.03	0.03±0.01	0.24±0.01	0.48±0.03	0.33±0.02	0.23±0.06	0.60±0.05	0.66±0.05	0.39±0.02	0.26±0.02	0.20±0.02	0.67±0.05	0.59±0.05
	AUROC	0.68±0.01	0.79±0.02	0.77±0.04	0.64±0.01	0.82±0.02	0.75±0.12	0.68±0.01	0.76±0.01	0.63±0.01	0.60±0.05	0.59±0.01	0.63±0.02	0.71±0.01	0.64±0.01	0.67±0.03	0.61±0.01	0.60±0.01
XCEPTION	*Acc.*	0.75±0.01	0.80±0.01	0.77±0.01	0.87±0.02	0.84±0.02	0.93±0.02	0.75±0.02	0.76±0.01	0.71±0.02	0.69±0.05	0.61±0.01	0.65±0.01	0.74±0.01	0.72±0.01	0.75±0.02	0.64±0.01	0.62±0.00
	Prec.	0.37±0.01	0.73±0.02	0.52±0.01	0.43±0.09	0.68±0.01	0.06±0.01	0.20±0.01	0.46±0.01	0.30±0.00	0.23±0.04	0.65±0.01	0.74±0.01	0.36±0.01	0.22±0.01	0.16±0.01	0.73±0.01	0.68±0.01
	Recall	0.60±0.03	0.80±0.02	0.80±0.00	0.53±0.09	0.89±0.01	0.63±0.00	0.60±0.02	0.78±0.01	0.53±0.03	0.78±0.11	0.62±0.02	0.67±0.02	0.68±0.02	0.59±0.03	0.55±0.06	0.69±0.02	0.58±0.02
	F1	0.46±0.01	0.76±0.01	0.63±0.01	0.47±0.08	0.77±0.02	0.12±0.02	0.30±0.01	0.57±0.01	0.38±0.01	0.36±0.05	0.63±0.01	0.70±0.01	0.47±0.01	0.32±0.01	0.24±0.02	0.71±0.01	0.63±0.01
	AUROC	0.77±0.01	0.86±0.00	0.84±0.00	0.72±0.03	0.91±0.01	0.81±0.01	0.73±0.01	0.84±0.01	0.70±0.00	0.77±0.04	0.65±0.01	0.69±0.01	0.79±0.01	0.73±0.01	0.51±0.02	0.66±0.00	0.66±0.01
XCEPTION (ROI)	*Acc.*	0.67±0.04	0.74±0.03	0.67±0.06	0.70±0.11	0.73±0.07	0.68±0.15	0.69±0.04	0.70±0.03	0.61±0.03	0.52±0.06	0.62±0.01	0.67±0.01	0.65±0.03	0.63±0.03	0.68±0.02	0.66±0.01	0.63±0.01
	Prec.	0.31±0.03	0.64±0.04	0.42±0.05	0.24±0.10	0.54±0.08	0.02±0.01	0.19±0.02	0.40±0.03	0.26±0.01	0.17±0.02	0.63±0.01	0.72±0.01	0.30±0.02	0.19±0.01	0.13±0.01	0.72±0.01	0.66±0.01
	Recall	0.72±0.03	0.87±0.03	0.83±0.01	0.65±0.15	0.88±0.03	0.73±0.05	0.68±0.03	0.83±0.02	0.66±0.04	0.87±0.04	0.72±0.03	0.76±0.01	0.77±0.03	0.68±0.03	0.62±0.05	0.76±0.03	0.71±0.03
	F1	0.43±0.03	0.74±0..02	0.56±0.04	0.35±0.12	0.66±0.06	0.04±0.02	0.29±0.02	0.54±0.02	0.37±0.01	0.29±0.03	0.67±0.01	0.73±0.01	0.43±0.02	0.29±0.02	0.22±0.02	0.74±0.01	0.68±0.01
	AUROC	0.76±0.03	0.86±0.03	0.81±0.04	0.74±0.08	0.89±0.03	0.77±0.09	0.76±0.01	0.83±0.02	0.69±0.02	0.76±0.04	0.66±0.01	0.68±0.01	0.78±0.02	0.71±0.03	0.71±0.03	0.66±0.01	0.67±0.01
RES. ATTN.	*Acc.*	0.66±0.04	0.68±0.02	0.65±0.03	0.73±0.05	0.64±0.01	0.72±0.07	0.67±0.03	0.66±0.03	0.63±0.05	0.63±0.09	0.55±0.02	0.57±0.03	0.65±0.04	0.66±0.05	0.68±0.04	0.55±0.02	0.56±0.03
	Prec.	0.26±0.02	0.60±0.01	0.37±0.03	0.60±0.01	0.42±0.02	0.02±0.01	0.15±0.02	0.33±0.02	0.22±0.02	0.17±0.06	0.60±0.02	0.70±0.02	0.26±0.02	0.16±0.01	0.10±0.01	0.70±0.02	0.62±0.02
	Recall	0.50±0.06	0.67±0.10	0.64±0.11	0.60±0.05	0.60±0.05	0.58±0.20	0.55±0.09	0.62±0.08	0.45±0.07	0.63±0.07	0.52±0.07	0.54±0.07	0.57±0.08	0.45±0.08	0.46±0.04	0.53±0.06	0.53±0.09
	F1	0.34±0.02	0.63±0.05	0.47±0.04	0.34±0.04	0.49±0.01	0.03±0.01	0.23±0.03	0.42±0.02	0.29±0.02	0.26±0.07	0.55±0.04	0.61±0.04	0.35±0.02	0.23±0.02	0.17±0.01	0.60±0.04	0.57±0.05
	AUROC	0.64±0.02	0.75±0.05	0.70±0.03	0.65±0.04	0.69±0.03	0.70±0.08	0.66±0.06	0.70±0.02	0.58±0.02	0.63±0.07	0.57±0.02	0.61±0.02	0.66±0.03	0.61±0.02	0.61±0.02	0.58±0.01	0.59±0.02
ECOVNET	*Acc.*	0.75±0.01	0.78±0.02	0.76±0.02	0.90±0.03	0.84±0.01	0.92±0.02	0.81±0.00	0.76±0.01	0.68±0.02	0.61±0.05	0.57±0.01	0.59±0.02	0.73±0.03	0.73±0.01	0.80±0.01	0.60±0.02	0.56±0.01
	Prec.	0.34±0.02	0.73±0.02	0.50±0.03	0.62±0.24	0.69±0.01	0.04±0.01	0.23±0.02	0.45±0.02	0.24±0.01	0.18±0.02	0.64±0.00	0.74±0.02	0.33±0.02	0.18±0.02	0.16±0.02	0.74±0.01	0.65±0.01
	Recall	0.47±0.03	0.73±0.05	0.65±0.06	0.35±0.05	0.83±0.03	0.40±0.15	0.45±0.07	0.65±0.05	0.41±0.03	0.68±0.10	0.51±0.03	0.54±0.04	0.56±0.04	0.40±0.05	0.42±0.04	0.57±0.04	0.45±0.03
	F1	0.39±0.02	0.73±0.03	0.56±0.04	0.44±0.09	0.75±0.01	0.07±0.03	0.30±0.03	0.53±0.03	0.31±0.02	0.28±0.03	0.56±0.02	0.62±0.03	0.41±0.03	0.25±0.03	0.23±0.02	0.64±0.03	0.53±0.03
	AUROC	0.71±0.02	0.84±0.02	0.77±0.03	0.67±0.05	0.88±0.01	0.77±0.05	0.71±0.02	0.79±0.03	0.62±0.02	0.69±0.04	0.61±0.01	0.65±0.02	0.73±0.03	0.63±0.03	0.68±0.02	0.64±0.01	0.61±0.02
CAPSNET	*Acc.*	0.67±0.04	0.69±0.03	0.66±0.04	0.66±0.09	0.66±0.04	0.80±0.06	0.64±0.07	0.67±0.03	0.63±0.04	0.63±0.07	0.56±0.01	0.55±0.02	0.65±0.05	0.67±0.04	0.66±0.07	0.56±0.02	0.55±0.01
	Prec.	0.26±0.02	0.62±0.05	0.39±0.03	0.15±0.07	0.45±0.04	0.02±0.00	0.14±0.03	0.34±0.02	0.21±0.02	0.15±0.03	0.62±0.01	0.69±0.02	0.26±0.02	0.15±0.01	0.09±0.02	0.71±0.01	0.62±0.01
	Recall	0.48±0.05	0.66±0.03	0.65±0.04	0.40±0.09	0.60±0.05	0.65±0.09	0.54±0.02	0.62±0.04	0.41±0.06	0.48±0.11	0.53±0.04	0.50±0.05	0.55±0.04	0.43±0.05	0.41±0.04	0.53±0.05	0.50±0.04
	F1	0.33±0.01	0.64±0.02	0.48±0.02	0.22±0.08	0.51±0.03	0.05±0.01	0.22±0.03	0.44±0.01	0.28±0.01	0.22±0.04	0.57±0.02	0.58±0.04	0.35±0.02	0.22±0.01	0.15±0.02	0.61±0.03	0.55±0.02
	AUROC	0.64±0.01	0.75±0.04	0.71±0.01	0.61±0.06	0.72±0.03	0.79±0.03	0.64±0.04	0.71±0.01	0.56±0.01	0.61±0.06	0.59±0.00	0.59±0.01	0.65±0.01	0.59±0.02	0.58±0.01	0.60±0.01	0.57±0.01
CORONET	*Acc.*	0.72±0.04	0.68±0.06	0.71±0.03	0.74±0.12	0.73±0.06	0.79±0.14	0.73±0.05	0.72±0.03	0.69±0.05	0.68±0.08	0.55±0.04	0.58±0.08	0.72±0.04	0.72±0.05	0.74±0.05	0.57±0.08	0.56±0.05
	Prec.	0.30±0.04	0.60±0.06	0.44±0.05	0.24±0.08	0.51±0.12	0.04±0.02	0.17±0.04	0.39±0.05	0.26±0.04	0.19±0.10	0.61±0.03	0.71±0.05	0.31±0.06	0.18±0.03	0.13±0.02	0.71±0.03	0.63±0.05
	Recall	0.48±0.15	0.66±0.21	0.62±0.17	0.53±0.25	0.59±0.25	0.60±0.12	0.53±0.16	0.61±0.18	0.42±0.14	0.55±0.19	0.49±0.14	0.53±0.17	0.55±0.15	0.44±0.16	0.47±0.14	0.55±0.17	0.46±0.14
	F1	0.37±0.06	0.62±0.14	0.50±0.08	0.31±0.12	0.54±0.18	0.07±0.04	0.26±0.06	0.47±0.09	0.31±0.06	0.27±0.05	0.53±0.10	0.59±0.13	0.39±0.07	0.25±0.05	0.20±0.04	0.61±0.12	0.52±0.11
	AUROC	0.70±0.04	0.76±0.07	0.76±0.05	0.69±0.08	0.77±0.09	0.76±0.05	0.70±0.05	0.76±0.06	0.64±0.04	0.70±0.10	0.60±0.03	0.64±0.06	0.72±0.06	0.66±0.05	0.66±0.05	0.62±0.04	0.60±0.05
SSL-AM	*Acc.*	0.72±0.03	0.77±0.03	0.74±0.03	0.69±0.06	0.78±0.04	0.75±0.07	0.75±0.02	0.74±0.02	0.68±0.03	0.61±0.07	0.60±0.01	0.63±0.02	0.72±0.03	0.72±0.03	0.75±0.03	0.63±0.02	0.59±0.01
	Prec.	0.32±0.03	0.70±0.03	0.48±0.04	0.18±0.04	0.59±0.06	0.02±0.01	0.20±0.02	0.43±0.03	0.28±0.03	0.16±0.02	0.64±0.01	0.73±0.02	0.34±0.03	0.21±0.02	0.15±0.02	0.74±0.01	0.65±0.02
	Recall	0.57±0.03	0.81±0.04	0.75±0.03	0.48±0.05	0.80±0.08	0.73±0.09	0.59±0.06	0.72±0.03	0.53±0.02	0.60±0.06	0.60±0.02	0.63±0.03	0.66±0.03	0.55±0.05	0.54±0.05	0.65±0.03	0.57±0.02
	F1	0.41±0.03	0.75±0.03	0.58±0.03	0.26±0.05	0.68±0.06	0.04±0.01	0.30±0.03	0.54±0.03	0.36±0.02	0.26±0.02	0.62±0.01	0.68±0.02	0.44±0.03	0.30±0.02	0.24±0.03	0.69±0.02	0.61±0.02
	AUROC	0.73±0.03	0.85±0.03	0.80±0.03	0.64±0.05	0.85±0.04	0.78±0.07	0.74±0.04	0.81±0.03	0.67±0.03	0.69±0.08	0.64±0.01	0.67±0.02	0.77±0.03	0.70±0.03	0.72±0.03	0.66±0.02	0.64±0.02
SSL-AM (ROI)	*Acc.*	0.68±0.02	0.73±0.01	0.69±0.01	0.65±0.07	0.72±0.03	0.53±0.06	0.73±0.02	0.73±0.01	0.67±0.03	0.63±0.06	0.60±0.02	0.61±0.04	0.70±0.2	0.70±0.02	0.72±0.03	0.60±0.03	0.59±0.05
	Prec.	0.28±0.01	0.65±0.02	0.42±0.01	0.18±0.05	0.52±0.03	0.01±0.00	0.18±0.01	0.41±0.03	0.26±0.01	0.21±0.02	0.64±0.01	0.72±0.01	0.32±0.01	0.18±0.02	0.13±0.01	0.72±0.01	0.66±0.01
	Recall	0.56±0.09	0.77±0.06	0.72±0.08	0.55±0.13	0.78±0.03	0.50±0.14	0.54±0.06	0.71±0.06	0.50±0.10	0.82±0.12	0.60±0.08	0.61±0.08	0.63±0.07	0.49±0.13	0.54±0.10	0.63±0.08	0.56±0.08
	F1	0.38±0.03	0.70±0.02	0.53±0.03	0.26±0.07	0.63±0.02	0.02±0.00	0.27±0.01	0.52±0.02	0.34±0.03	0.33±0.02	0.62±0.04	0.66±0.05	0.42±0.02	0.27±0.04	0.21±0.01	0.67±0.05	0.60±0.05
	AUROC	0.70±0.04	0.82±0.03	0.77±0.04	0.62±0.07	0.84±0.01	0.56±0.11	0.71±0.02	0.79±0.03	0.66±0.03	0.73±0.04	0.63±0.02	0.66±0.03	0.75±0.03	0.66±0.04	0.70±0.03	0.63±0.03	0.64±0.03

TABLE 5.8: **Average model performance metrics for each subgroup in the LTHT population.** Standard deviation of metrics across cross-validation folds is included. **Bold** indicates the highest average metric per dataset. *Abbrvs: National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Accuracy (Acc); Precision (Prec); Area Under Receiver Operator Curve (AUROC); Female (F); Male (M); XCEPTION NET (XCEPTION).*

in this age group as indicated by combined low false positive rates and low precision scores (Table 5.8). XVITCOS (ROI) gives improved performance in this age group compared to XVITCOS, with AUROC scores rising from 0.73 to 0.79. However, other ROI-trained models show a trend in reduced performance in this age group when compared to full CXR-trained counterparts, more so than other age groups (Fig. 5.8). Comparison of SSL-AM and SSL-AM (ROI) shows a drastic fall in AUROC scores from 0.78 to 0.56 (Table 5.8). Moreover, under the counterfactual condition in which models become general pneumonia classifiers, LTHT (P), I find that models perform particularly poorly in this age group (Fig 5.9); which could be interpreted as evidence of better separation of COVID-19 from non-COVID-19 pneumonia. I also observe that under LTHT (NP) conditions, with non-COVID-19 pneumonia removed, models perform better than in real-world populations of this subgroup (LTHT) e.g., XVITCOS AUROC scores increase from 0.73 to 0.96 (Fig 5.9).

Smoking status & comorbidities I observe a universal decline in model performance in the subgroup with CXR observable comorbidities. For 9 out of 10 models I evaluate statistically significant differences in model performance (AUROC) in these groups (One-way ANOVA, $p < 0.05$; Table 5.7). However, model performance on these subgroups could not be evaluated under counterfactual scenarios due to a lack of data.

In addition, I find that all models perform worse when applied to subgroups with a history of smoking (both former and current smokers) e.g., XCEPTION NET AUROC falls from 0.79 in subgroups without any smoking history to 0.73 and 0.51 in former and current smoker subgroups, respectively (Table 5.8). I see increased false negative rates in these groups compared to non-smokers (Fig. 5.8). Under counterfactual conditions where non-COVID-19 pneumonia is removed from the population, I see that model performance disparities between groups of different smoking status do not decrease. Models still perform best when applied to subgroups without any history of smoking, and performance disparities between former smoker and current smoker groups is sustained (Fig. 5.9).

5.3.4 Model error analysis

I explore the influence of clinical and experimental factors on model error rate. As the best performing model, I used XCEPTION NET predictions for this analysis. During NCCID test population evaluation, I examined the relationship between the frequency of false positives and the recorded distance from CXR exam date to swab date-derived diagnostic window. According to this analysis, COVID-19 negative CXRs acquired in close proximity to the COVID-19 diagnostic window are more frequently predicted as COVID-19 compared to those obtained further away, with XCEPTION NET delivering the most false positives for CXRs 1-5 days before or after the diagnostic window (Fig. 5.10A). I examine the correlation of incorrect COVIDGR predictions with radiologist-defined CXR severity. For this evaluation I

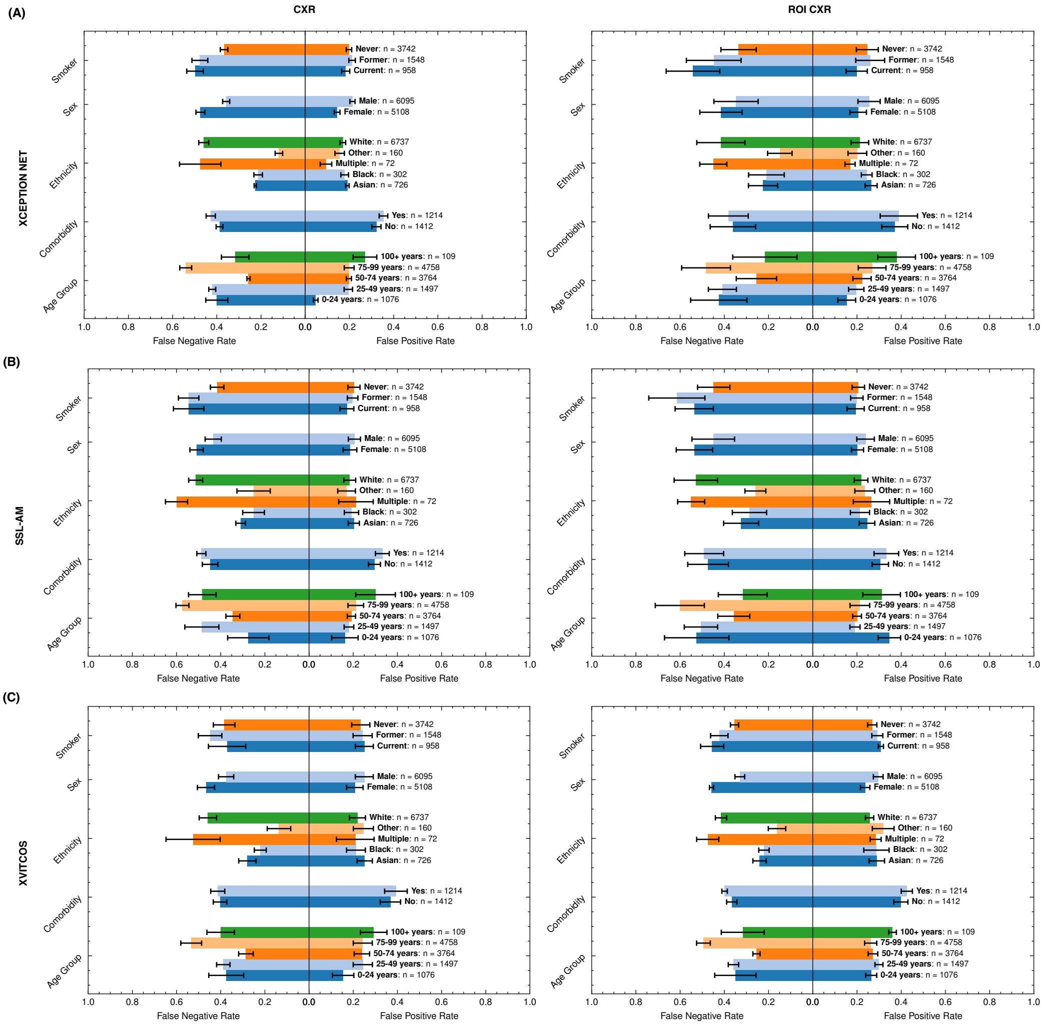
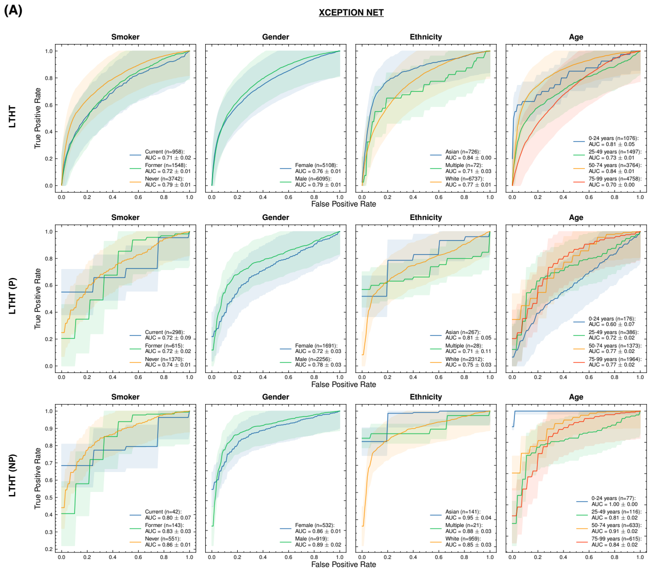
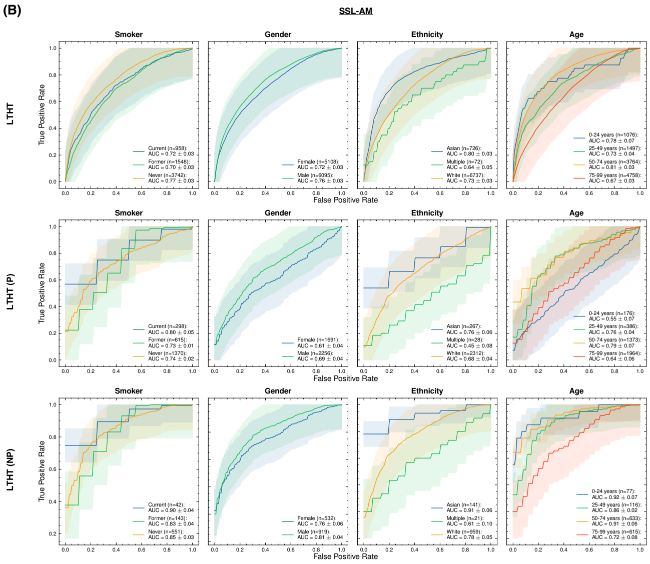


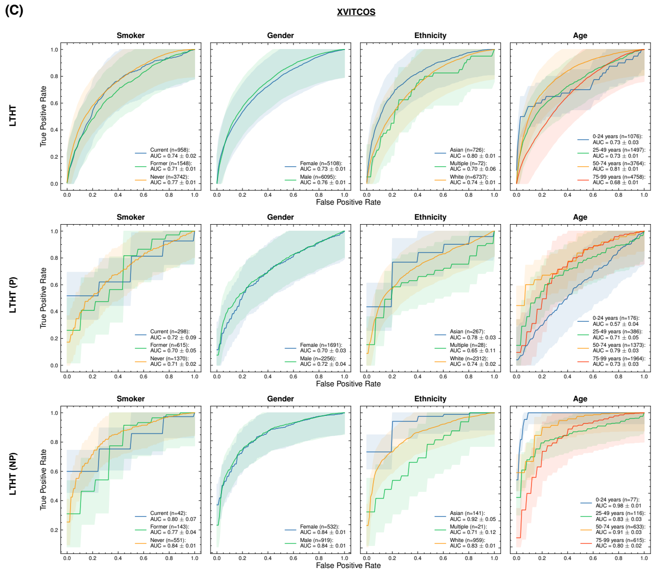
FIGURE 5.8: **Average FPRs and FNRs of top-performing models, (A) XCEPTION NET, (B) SSL-AM, and (C) XVITCOS, and their ROI-trained counterparts in LTHT subgroups**. Subgroup population size is referred to by *n*. Error bars correspond to standard deviation across cross-validations. *Abbrvs: False Negative Rate (FNR); False Positive Rate (FPR); Region of Interest (ROI); Leeds Teaching Hospital Trust (LTHT).*



(A) **ROC curves of XCEPTION NET performance in smoker, sex, ethnicity and age subgroups.**



(B) **ROC curves of SSL-AM performance in smoker, sex, ethnicity and age subgroups.**



(C) **ROC curves of XVITCOS performance in smoker, sex, ethnicity and age subgroups.**

FIGURE 5.9: **ROC curves of COVID-19 detection in smoker, sex, ethnicity and age subgroups for top-performing models: (A) XCEPTION NET, (B) SSL-AM, and (C) XVITCOS.** The top row presents model performance in the real-world scenario and the bottoms rows present model performance under counterfactual conditions, LTHT (P) and LTHT (NP). Subgroup population size is referred to by n. Subgroups that do not exist in LTHT (P) or LTHT (NP) populations are excluded. Shaded regions correspond to 95% confidence intervals. *Abbrvs: Receiver Operating Characteristic (ROC); Area Under Curve (AUC); Leeds Teaching Hospital Trust (LTHT).*

characterised each CXR according to RALE criteria (Warren et al., 2018), labelling as 'NORMAL-PCR+', 'MILD', 'MODERATE' and 'SEVERE'. I observe a strong pattern of increasing frequency of false negative predictions for CXRs with milder features of COVID-19 disease i.e., MILD CXRs are more frequently missed. As expected, I find that 99% NORMAL-PCR+ CXRs are missed i.e., cases in which radiologists were unable to identify COVID-19 features. Even MILD and MODERATE COVID-19 CXRs exhibit high rates of false negatives, with COVID-19 being missed ≈94% and ≈70% of the time (Fig. 5.10B). I find that CXRs categorised as SEVERE are less frequently missed, yet 32.9% are still falsely classed as COVID-19 negative.

With XCEPTION NET predictions, I evaluate the relationship between the projection view of the CXR and the frequency of false positive predictions. I found AP projected CXRs are more commonly misidentified as COVID-19 compared to PA projected CXRs (Fig. 5.10C). Figure 5.10C also shows that ROI-trained XCEPTION NET makes more false positive predictions in AP projected CXRs than the full CXR trained XCEPTION NET.

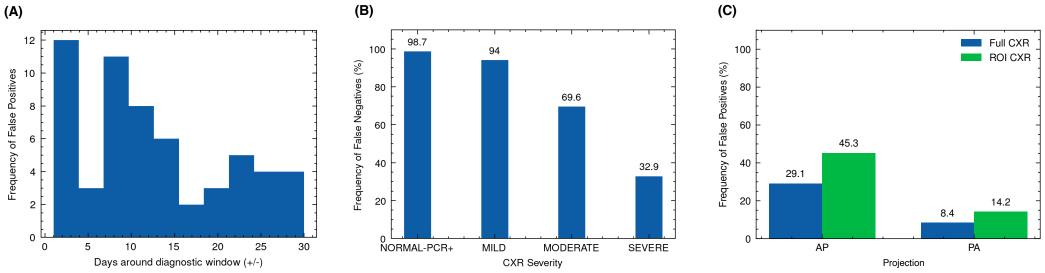


FIGURE 5.10: **Analysis of XCEPTION NET prediction error.** Frequency of **(A)** false positive predictions on NCCID TEST data according to proximity of CXR exam date to the diagnostic window, **(B)** false negative predictions on COVIDGR according to RALE-defined CXR severity and **(C)** false positive predictions on LTHT data according to CXR projection, alongside evaluation of ROI-trained XCEPTION NET false positive frequency. *Abbrvs: Radiographic Assessment of Lung Edema (RALE); Region of Interest (ROI); Chest X-ray (CXR); Antero-posterior (AP); Postero-anterior (PA).*

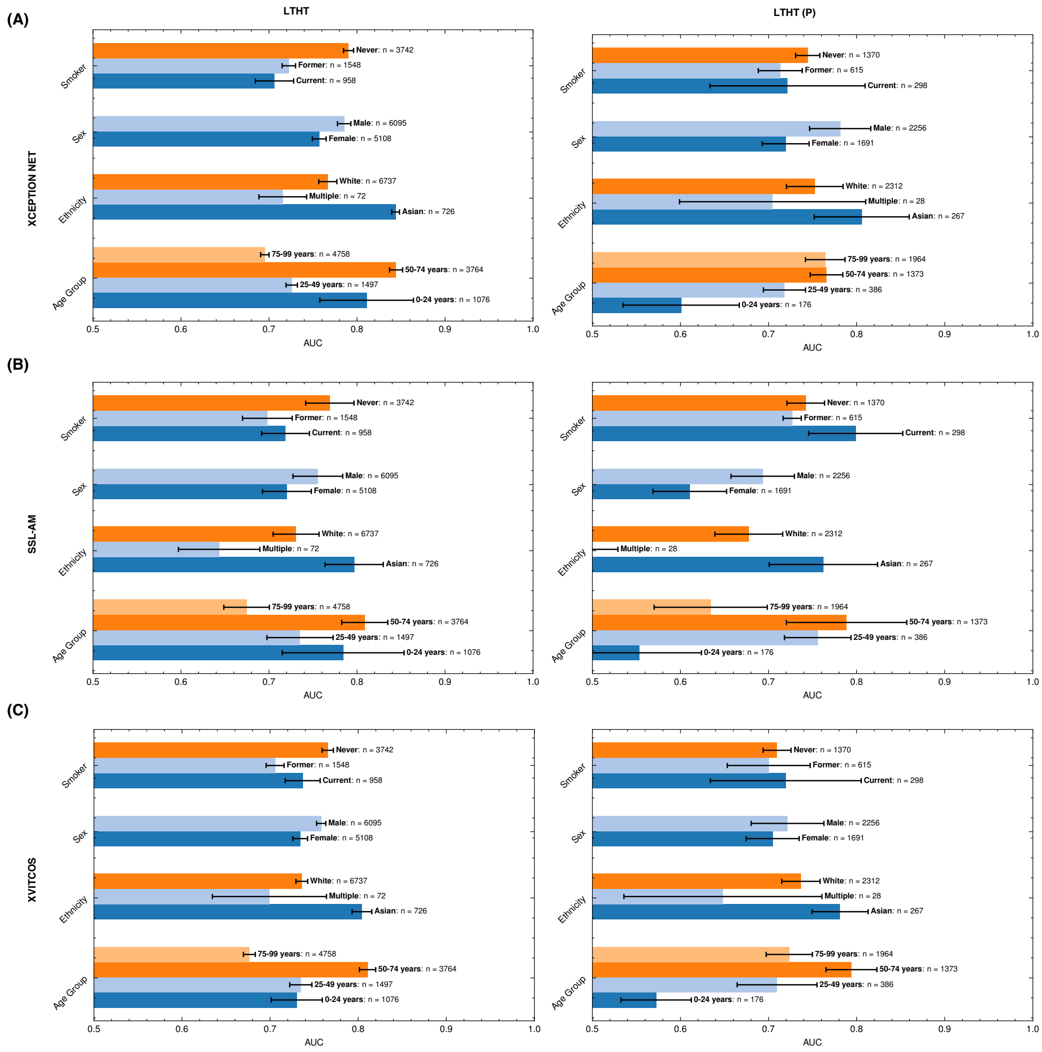


FIGURE 5.11: **(A) XCEPTION NET, (B) SSL-AM and (C) XVITCOS AUROCs for LTHT subgroup and LTHT (P) subgroups.** Only subgroups that exist in the LTHT (P) population are included. n is the subgroup population size. Error bars correspond to standard deviation across cross validations. *Abbrvs: Area Under Receiver Operator Characteristic (AUROC); Region of Interest (ROI).*

		XCEPTION NET		**SSL-AM**		**XVITCOS**	
		FULL	*ROI*	*FULL*	*ROI*	*FULL*	*ROI*
	Acc.	0.65 ± 0.01	0.60 ± 0.04	0.69 ± 0.03	0.61 ± 0.02	0.61 ± 0.03	0.61 ± 0.01
	Prec.	0.47 ± 0.01	0.44 ± 0.03	0.52 ± 0.03	0.44 ± 0.02	0.45 ± 0.02	0.45 ± 0.01
NCCID TEST	Recall	0.79 ± 0.01	0.76 ± 0.03	0.70 ± 0.05	0.75 ± 0.06	0.77 ± 0.02	0.77 ± 0.02
	F1	0.59 ± 0.01	0.55 ± 0.02	0.59 ± 0.01	0.56 ± 0.02	0.57 ± 0.01	0.56 ± 0.01
	AUROC	0.75 ± 0.01	0.71 ± 0.03	0.74 ± 0.02	0.71 ± 0.03	0.74 ± 0.01	0.73 ± 0.01
N=824							
	Acc.	0.75 ± 0.01	0.66 ± 0.04	0.72 ± 0.03	0.68 ± 0.02	0.70 ± 0.03	0.66 ± 0.02
	Prec.	0.34 ± 0.01	0.28 ± 0.02	0.30 ± 0.03	0.27 ± 0.01	0.30 ± 0.02	0.27 ± 0.01
LTHT	Recall	0.65 ± 0.02	0.74 ± 0.02	0.62 ± 0.03	0.60 ± 0.08	0.66 ± 0.04	0.69 ± 0.02
	F1	0.45 ± 0.01	0.41 ± 0.03	0.41 ± 0.03	0.37 ± 0.02	0.41 ± 0.01	0.39 ± 0.01
	AUROC	0.78 ± 0.01	0.77 ± 0.03	0.74 ± 0.03	0.71 ± 0.04	0.75 ± 0.01	0.74 ± 0.01
N=11,204							
	Acc.	0.75 ± 0.02	0.87 ± 0.01	0.69 ± 0.03	0.73 ± 0.05	0.74 ± 0.03	0.84 ± 0.02
	Prec.	0.98 ± 0.00	0.96 ± 0.02	0.96 ± 0.01	0.95 ± 0.01	0.97 ± 0.01	0.95 ± 0.00
LTHT (NP)	Recall	0.73 ± 0.02	0.90 ± 0.02	0.68 ± 0.03	0.75 ± 0.06	0.73 ± 0.04	0.87 ± 0.03
	F1	0.84 ± 0.01	0.93 ± 0.01	0.80 ± 0.03	0.83 ± 0.04	0.83 ± 0.02	0.91 ± 0.01
	AUROC	0.88 ± 0.02	0.87 ± 0.05	0.79 ± 0.05	0.74 ± 0.06	0.83 ± 0.01	0.85 ± 0.01
N=3,948							
	Acc.	0.52 ± 0.02	0.64 ± 0.03	0.49 ± 0.02	0.50 ± 0.06	0.55 ± 0.04	0.58 ± 0.02
	Prec.	0.99 ±. 0.00	0.98 ± 0.01	0.98 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	0.98 ± 0.00
LTHT (P)	Recall	0.50 ± 0.02	0.64 ± 0.03	0.48 ± 0.03	0.50 ± 0.07	0.54 ± 0.04	0.58 ± 0.03
	F1	0.67 ± 0.02	0.77 ± 0.02	0.65 ± 0.02	0.66 ± 0.05	0.70 ± 0.04	0.73 ± 0.02
	AUROC	0.75 ± 0.03	0.71 ± 0.06	0.66 ± 0.04	0.61 ± 0.02	0.72 ± 0.02	0.67 ± 0.03
N=1,451							
	Acc.	0.62 ± 0.01	0.65 ± 0.02	0.59 ± 0.03	0.58 ± 0.03	0.61 ± 0.02	0.59 ± 0.01
	Prec.	0.86 ± 0.05	0.91 ± 0.01	0.78 ± 0.05	0.77 ± 0.05	0.73 ± 0.04	0.69 ± 0.02
COVIDGR	Recall	0.30 ± 0.02	0.33 ± 0.04	0.26 ± 0.10	0.23 ± 0.05	0.34 ± 0.04	0.33 ± 0.03
	F1	0.44 ± 0.02	0.49 ± 0.04	0.38 ± 0.10	0.35 ± 0.07	0.46 ± 0.04	0.44 ± 0.02
	AUROC	0.71 ± 0.02	0.74 ± 0.02	0.67 ± 0.01	0.66 ± 0.03	0.67 ± 0.03	0.64 ± 0.02
N=852							

TABLE 5.9: **Comparison of average full CXR-trained and ROI-trained model performance metrics.** Standard deviation of metrics across cross-validation folds is included. *N* is the total number of cases in each test population. *Abbrvs: Chest X-ray (CXR); National COVID-19 Chest Imaging Database (NCCID); Leeds Teaching Hospital Trust (LTHT); Accuracy (Acc); Precision (Prec); Area Under the Receiver Operator Characteristic (AUROC).*

5.3.5 Model explanations & shortcut features

To investigate model reliance on spurious features, I create saliency maps using GradCAM and gradient backpropagation and examine the features that have the most influence on model predictions for both full CXR and ROI-trained models. I explore true positive COVID-19 predictions made by XCEPTION NET (Fig. 5.12). Review of GradCAM saliency maps shows model reliance on both COVID-19 pathology and undesirable 'shortcut' features. I found that clinically-relevant regions were consistently highlighted, with regions of significance often localised to the lower lung areas, as well as the heart margins of the cardiac silhouette. I observed similar activations in gradient backpropagation saliency maps. However, with this improved granularity I also observe highlighted support devices (i.e., heart monitor wiring, portacaths or endotracheal tubing) and radiograph annotations, possibly representing reliance on spurious 'shortcut' features (Fig. 5.12A).

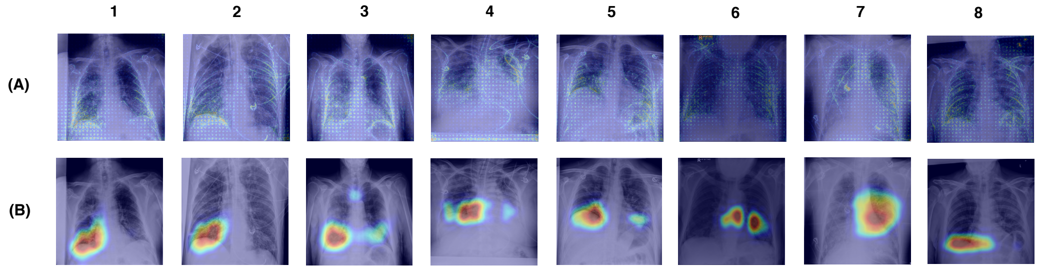


FIGURE 5.12: **Saliency maps of correct XCEPTION NET predictions of COVID-19 positive CXRs.** CXRs are sourced from the LTHT dataset and saliency maps are generated with **(A)** Gradient backpropagation and **(B)** GradCAM. *Abbrvs: Leeds Teaching Hospital Trust (LTHT); Chest X-rays (CXRs).*

5.3.6 Comparative validation of the impact of lung segmentation on model performance

I find that training models on ROI-extracted CXRs does not improve model performance. Against expectation, ROI-trained model performance is marginally worse compared to full CXR trained models when testing in NHS centre populations (LTHT). Notably, for XCEPTION NET and XVITCOS models, I find that training on ROI CXRs does not worsen performance in international population, instead performance marginally improves (Fig. 5.14).

Manual inspection of the ROI-extracted CXRs showed that, with the use of a post-processing algorithm, lung regions were preserved (Fig. 5.6). Without loss of clinical features, I propose that the decrease in model performance is linked to the exclusion of non-clinical 'shortcut' features, such as radiograph annotations, which have been identified in saliency maps of full CXR-trained models (Fig. 5.12). These 'shortcut' features are typically located outside the thoracic area, and are cropped out during ROI extraction (Fig. 5.6). However, while non-clinical features that

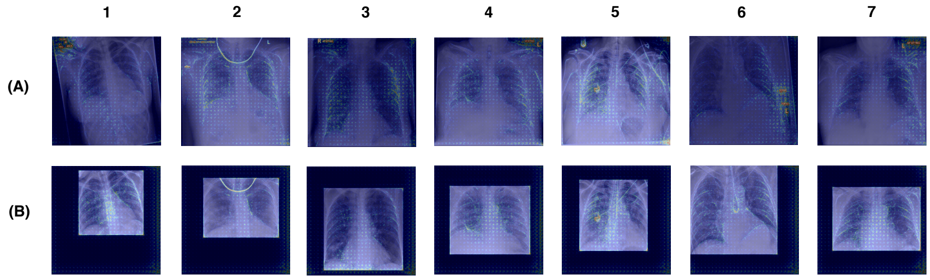


FIGURE 5.13: **Gradient backpropagation feature attribution maps of true positive COVID-19 predictions.** Saliency maps are generated for **(A)** XCEPTION NET predictions and **(B)** XCEPTION NET (ROI) predictions. XCEPTION NET (ROI) is trained and evaluated on CXRs cropped to the region of interest, while XCEPTION NET is trained on full CXRs. *Abbrvs: Leeds Teaching Hospital Trust (LTHT); Chest X-rays (CXRs); Region Of Interest (ROI).*

exist outside the lung fields can be removed by cropping to the ROI, general health indicators e.g., presence of support devices, bone density, etc., remain in view and are highlighted in saliency maps, suggesting influence over model predictions (Fig. 5.13).

Moreover, I find that ROI-trained models are worse at distinguishing COVID-19 from other pneumonia types, as evidenced by the evaluation of ROI-trained models in a subset of the LTHT populations with known pneumonia outcomes: LTHT (P) and LTHT (NP). ROI-trained models produce a much higher error rate in non-COVID-19 pneumonia populations. I observe that the ROI-trained version of XCEPTION NET performs much worse than its full CXR trained counterpart, with error rates of 0.51 compared to 0.40.

5.4 Discussion

The main goal of this research is to evaluate the use of deep learning approaches for the detection of COVID-19. Specifically, to evaluate model performance in real world clinical populations and identify factors contributing to model bias or error.

When comparing the main findings with the existing literature I found that the model performances evaluated in this study contradict initial reports of model performance in many of the source publications, many of which report AUROC scores as exceeding 0.90. I identify several methodological flaws in this literature. Roberts et. al. (2021) considered 62 studies, including many of the studies selected for this benchmarking, and identified substantial limitations that placed the majority of models at high risk of bias (Roberts et al., 2021). The main limitations considered by Roberts et. al. (2021) were the use of inappropriate training data, inadequate external validation and lack of subgroup evaluation (Roberts et al., 2021). An additional critical analysis identifies that the data used in source publications put deep learning models at high risk of learning spurious 'shortcuts' (DeGrave, Janizek,

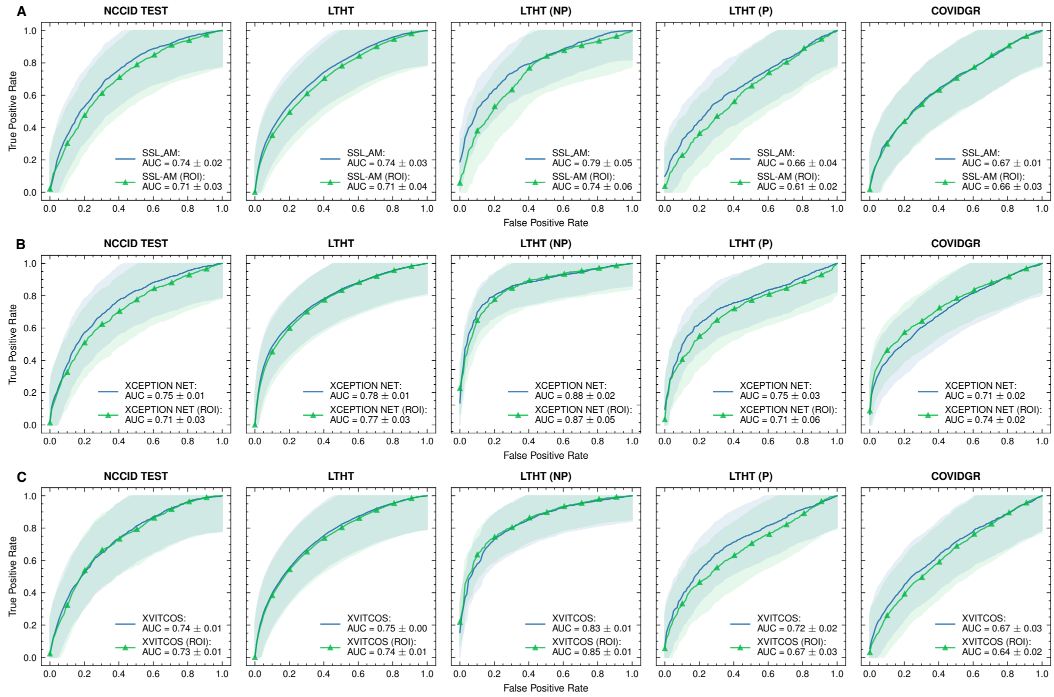


FIGURE 5.14: **ROC curves of (A) SSL-AM, (B) XCEPTION NET, and (C) XVITCOS model evaluations on each dataset and their ROI-trained counterparts.** Shaded regions correspond to the 95% confidence intervals. *Abbrvs: Receiver Operating Characteristic (ROC); Area Under Curve (AUC); Region Of Interest (ROI).*

and Lee, 2021). This retrospective study corrects these issues, with the use of multi-centre hospital data and extensive model validation on independent datasets. I report new findings in deep learning model performance and consider the major pitfalls in the development of deep learning models for clinical application.

Radiologists achieve performances of 0.78 AUROCs, as reported in Albiol et al. (2022). Although direct comparison between radiologist performance and model performance is inappropriate due to differences in test populations, at face-value deep learning models show promise as an assistive tool for use in future pandemics.

A comparison of model performance shows that the best performances on the real-world LTHT population are achieved by supervised deep CNN models that employ transfer learning. Pre-trained with ImageNet weights, top-performer XCEPTION NET can be characterised by the application of depth-wise separable convolutional operations for more efficient use of model parameters. The SSL-AM and XVITCOS models give the next strongest performances. SSL-AM is pre-trained with CheXpert data under self-supervised conditions to fully leverage the underlying data structures of a common domain. XVITCOS is a vision transformer, pre-trained on CheXpert, this approach uses positional embedding and self-attention to learn efficient CXR representations that incorporate both local features and global dependencies. In contrast, RES. ATTN. is the only model that does not apply pre-trained weights, and records the lowest performance metrics out of all evaluated models. From this I can speculate that transfer learning, domain-specific or otherwise, is needed to achieve strong model performance.

While classification metrics indicate adequate model performance, this study gives conclusive evidence that DL models perform poorly on clinically complex cases i.e., where comorbidities/confounding features are present, and frequently fail at separating COVID-19 from non-COVID-19 pneumonia. For example, decreased model performance is observed in populations with an increased incidence of clinically relevant underlying conditions e.g., the 74-99 age group, active/previous smokers, etc., these complex cases are disproportionately represented in hospital populations. Therefore, my findings suggest that existing COVID-19 detection models have limited value as an assistive tool for frontline radiologists, who are tasked with making challenging diagnoses for high risk patients that require urgent treatment.

These model failings can, in part, be attributed to inadequate data. In the absence of labels for alternative pathologies/classes supervised DL models are not equipped to learn to separate the similar features of different pathologies, e.g., non-COVID-19 pneumonia, emphysema, lung cancer, etc. Where pathologies co-occur more frequently with the class of interest than the negative class, models are vulnerable to blindly learning these features as ‘shortcuts’ (DeGrave, Janizek, and Lee, 2021). Scarcity of multi-label datasets and widespread inadequacies in model validation highlight the need for clinicians and deep learning researchers to address the current shortfalls in data collection and to define criteria for clinically-oriented DL model development. My future work centres on multi-labelled prediction to develop models that accurately classify co-occurring pathologies. Using multi-label datasets, I aim to address the complexities of overlapping diseases in medical imaging, reducing reliance on shortcuts and improving generalisability across clinical settings.

I tested models on the COVIDGR dataset to evaluate performance in international populations (Spain) where typical NHS clinical pathways and data acquisition protocols do not apply. Evaluation in COVIDGR shows that models generalise poorly outside NHS populations. However, in addition to changes in population characteristics, there are critical differences in how COVID-19 was defined. The diagnostic window I use to define COVID-19 cases in the NHS populations (-14/+28 days around RT-PCR+ swab) was decided under clinical guidance taking into consideration: the importance of early detection; poor RT-PCR sensitivity, particularly with low viral load as is observed in early stages of infection; and, typical time for CXR resolution post-infection. This is in stark contrast to COVIDGR which was pre-defined with a much shorter diagnostic window of 24 hours before or after a positive RT-PCR swab. This raises the issue that without a standardised COVID-19 labelling protocol, which should balance technical feasibility with clinical utility, detection models are vulnerable to poor generalisability as a consequence of label shift. The likely use of open-source datasets in emergent pandemic situations, which can be compiled with inconsistent labelling as seen with COVID-19, highlights the importance of pro-active collaboration. For example, without clinical input deep learning researchers may unintentionally adopt a disease definition that maximises quantitative metrics i.e., accuracy, precision, recall etc., at the expense of clinical utility.

Moreover, I observe increased rates of false positive predictions in negative COVID-19 CXRs acquired close to the diagnostic window. With this evidence of increased diagnostic uncertainty and the understanding that, for a large portion of COVID-19 CXRs, disease features persist for a long time after infection, I suggest that current labelling strategies result in a noisy ground truth. A portion of post-COVID-19 resolved CXRs are either incorrectly considered COVID-19 or persistent disease features enter the control population. To reconcile this source of label noise, I propose the use of an additional category of COVID-19 disease which would separate chronic changes, i.e., persistent disease features post-infection, from active COVID-19 infection.

Additionally, the detection of COVID-19 through RT-PCR is flawed with low sensitivity and high rates of false negatives. Therefore, deriving COVID-19 status from RT-PCR testing alone adds further noise to the ground truth labels. In practice, the clinical diagnosis of COVID-19 takes into account more than just RT-PCR outcomes, e.g., clinical signs and symptoms, recent exposures, comorbidities, and patient medical history. In fact, 20% of symptomatic patients receive a clinical diagnosis of COVID-19 despite negative RT-PCR testing (Middleton et al., 2021). I further propose a multi-modal labelling approach that would incorporate all relevant patient data, this would drastically reduce ground truth noise and benefit deep learning models, particularly supervised models.

In evaluating model performance awareness of bias and fairness is critical. Inadequate evaluation can allow biased deep learning models to amplify systemic healthcare disparities in under-served communities. My evaluation of the models shows varied performance across different sub-populations, with top-performing models exhibiting obvious demographic biases, including unequal performance depending on ethnicity, sex and age. However, clinical evidence suggests that observed model performance disparities may be a consequence of varied disease severity between demographics. Generally, models perform better when applied to demographics which experience COVID-19 more severely, e.g., ethnic minorities, males, and older age groups (Sun et al., 2022; Joseph et al., 2020). The clinical factors affecting the severity of COVID-19 infection are still not fully understood. Before clinical implementation, a greater understanding is required to determine if these disparities in model performance might result in greater health inequity.

Crucially, my findings show that ROI-extraction was insufficient to prevent these disparities. Therefore, if bias is identified researchers should be cautioned against applying segmentation techniques with the assumption that the removal of background noise will fully mitigate the bias. Additionally, I find that cropping CXRs to the ROI prior to training does not improve overall model performance. This contradicts previous studies in which ROI-trained models performed better (Nafisah et al., 2023). A key difference between their approaches and mine is that I undertake a more rigorous methodology in which the segmentation model is trained on an external dataset. Whereas conflicting studies typically use the same data for both segmentation training and classification training, an approach that is not supportable in a clinical setting (Nafisah et al., 2023).

I evaluated the impact of CXR projection on model predictions, as recommended by Roberts et al. (2021). AP projected CXRs are used when the patient is not able to get into the correct position for the standard PA projection, for example, if the patient is too ill or is in isolation (Rubin et al., 2020). As a result, algorithms are at risk of learning to associate COVID-19 with projection rather than the clinically-relevant CXR features. I observe over-representation of AP CXRs in the disease class of the training data, 83% of positive COVID-19 images were AP projected, whilst only 65% of negative COVID-19 images were AP projected. Saliency maps provide evidence that projection may have been a spurious shortcut features, as they consistently highlight features around heart borders, a region of the CXR that varies greatly depending on projection. Moreover, AP CXR predictions are more commonly false positive. While GradCAM saliency maps are useful here for highlighting vague areas of interest in the CXR, I find evaluation challenging due to a lack of precision. Similarly, I find the widespread pixel activation from gradient activation maps challenging to interpret. In future work I will consider alternative approaches to identify salient features important for image classification, with the goal of improving model transparency in clinical settings.

Evaluations of a wide range of models suggests CXRs alone may not be sufficient to detect COVID-19. In a head-to-head comparison, performance metrics indicate that the top models are unable to compete with the gold standard clinical test, RT-PCR. Models often fail to separate COVID-19 from other pneumonia types and are unable to detect COVID-19 in RALE-defined NORMAL-PCR+ cases, in which 99% of COVID-19 positive CXRs are missed. Here, it is important to note that not all COVID-19 infections develop into COVID-19 pneumonia, in which case diagnostic features of COVID-19 cannot be observed in the CXR and even the very best DL models would be unable to detect COVID-19 infection. In practice, it is rare for a disease diagnosis to be wholly determined by a single test. In fact, reducing the source of diagnostic information to a single modality risks losing diagnostic features of a disease. Where imaging is incongruous with patient health, clinicians often rely on additional sources of information. The incorporation of multi-modal information e.g., exposure data, symptoms, medical history, etc. during data curation should be more widely adopted to facilitate the development of improved DL models for the detection of COVID-19.

5.5 Conclusion

This benchmarking of COVID-19 detection models trained on multi-centre hospital data highlights the need for clinical guidance in developing reliable predictive models, for disease diagnosis and medical image interpretation. In particular, I highlight the need for early and consistent disease definition, in order to ensure model generalisablity across international and jurisdictional populations. Disease definitions should also be continually reviewed for clinical utility, for instance, I suggest COVID-19 detection models could be improved by the separation of CXRs that exhibit long-term changes as a result of prior infection from CXRs of patients with active infection. To the extent that comparison is possible, the deep learning models evaluated detect COVID-19 with apparent similar performance to radiologists. However, both fall short of the gold standard clinical test, RT-PCR. I suggest that a multi-modal approach under clinical guidance, where additional clinical factors are incorporated, can be used to improve model performance; with the aim of developing a reliable assistive tool, on par with the existing gold standard.

Moreover, COVID-19 detection models have extreme difficulty identifying COVID-19 in complex clinical cases, as demonstrated by my evaluation of models in subgroups with higher incidences of confounding pathologies and comorbidities. Models are also vulnerable to learning ‘shortcut’ features. Neither of these issues are mitigated by the use of lung segmentation. Ultimately, I suggest that to accurately predict disease in real clinical populations, where patients have comorbidities, it is essential to apply multi-label training objectives where possible. Multi-label classification requires the model to learn a more complete understanding of the data, preventing excessive reliance on ‘shortcut’ features and improving model generalisability across clinical settings.

Chapter 6

Multi-task VAEs for the Explainable Prediction of Co-Occurring Pulmonary Diseases

6.1 Introduction

The presence of co-occurring diseases in medical imaging datasets makes deep learning models vulnerable to learning shortcut features. Shortcut learning refers to the phenomenon where a model relies on spurious correlations or superficial patterns in the data, rather than the fundamental features indicative of the disease pathology (Ong Ly, Unnikrishnan, Tadic, et al., 2024). Pathologies that frequently co-occur with the disease of interest become 'shortcuts' to the prediction of this disease, particularly where co-occurring pathologies are larger or more salient than the disease of interest, e.g., emphysema as a shortcut to lung nodules. Reliance on 'shortcuts' can lead to poor performance outside the training distribution, harm model robustness and increase risk of model bias, which may go undetected if shortcuts persist outside the training distribution and model explanation is insufficient[1] (Ong Ly, Unnikrishnan, Tadic, et al., 2024). Shortcut learning can be mitigated through careful data curation, adversarial approaches that penalise model reliance on confounding features, the removal of confounding image features from the training dataset, or with the use of multi-label classification objectives. Removal of confounding image features can be achieved through data curation strategies and counterfactual image generation, where data curation is performed to 'rebalance' disease feature occurrence and counterfactual image generation is used to 'edit' images directly to remove secondary disease features. Following a data curation strategy may lead to an unacceptable loss of training data, particularly where over-representation of co-occurrence is severe, as is often observed in medical datasets (Banerjee et al., 2023).

Weng et al. (2023) apply diffusion-based counterfactual image generation to synthetically remove or add shortcut features to samples of a CXR data. Key to this approach is the preservation of the remaining image features in the generated

---

[1] More details on shortcut learning are covered in Chapter 5

shortcut-free counterfactuals, despite best efforts to prevent this, this approach to mitigating shortcut learning risks removal of salient clinically-relevant features.

A multitude of adversarial approaches have been developed to prevent shortcut learning, these approaches create an adversarial optimisation task to preventing shortcut learning. By maximising the primary task (i.e., disease classification) and minimising the performance of spuriously correlated factor prediction, they can enforce shortcut invariance. Another popular solution to this is 'subspace' disentanglement. This term encapsulates a class of approaches that rely on the division of an embedding space into subspaces, where $z1$ encodes the primary classification task and $z2$ encodes the spuriously correlated variable. With this disentanglement regularisation objectives can be applied to enforce independence between the subspaces (Müller et al., 2024).

For example, Fay et al. (2023) use a deep CNN to embed brain MRIs into a feature vector, which is split into two parts. With the first part they train a classifier to predict the disease status of the image, and with the other part they predict spuriously correlated factors, such as, age and sex. To make these two feature sub-vectors independent from each other, they apply an objective to reduce the mutual information between them. Similar approaches have applied a distance correlation objective, which measures the linear and non-linear dependence between two random vectors (in this case sub-vectors), to the same effect (Müller et al., 2024). However, subspace disentanglement and adversarial approaches both require prior knowledge of shortcut features and annotation of the datasets, which is highly impractical. Moreover, these approaches are largely applied to prevent the learning of demographic data, such as sex and age, as shortcuts. Prevention of shortcuts arising from co-occurring pathologies is largely unexplored.

Of the aforementioned strategies, multi-label prediction is preferred as an intuitive regularisation strategy to prevent shortcut learning. By applying a multi-label prediction framework, where the prediction of each label is independent, the model learns a more complex, robust understanding of the data. Framing image classification as a multi-label prediction problem acts as an inherent form of regularisation. The model is required to generalise across multiple labels, reducing the risk of overfitting and preventing the model from narrowing focus to any single aspect of the data and developing a reliance on features that spuriously correlate with this aspect. Consider, an image classification task where the goal is to predict lung cancer in CXRs. In a multi-class framework, the model can learn to associate this label with features that evidence a history of smoking (e.g., emphysema or fibrosis) as smoking is a known risk factor for lung cancer. While in a multi-label setting, the model must learn to identify multiple pathologies within the same CXR (e.g., fibrosis, emphysema, lung nodules, pneumonia etc.), to predict each of these labels the model must learn to focus on the pathological features themselves as how these pathologies present and co-occur will vary.

Moreover, where shortcut learning persists, explainable prediction is essential for the identification of model biases. Clear and precise localisation of predictive features is crucial for providing radiologists with confidence where interpretation and model prediction is uncertain, as is often the case for complex CXRs with co-occurring pathologies. Common methods for explainable prediction of images, include, GradCAM, GradCAM++, and LIME, etc. (Selvaraju et al., 2017a; Chattopadhay et al., 2018; Ribeiro, Singh, and Guestrin, 2016). These are post-hoc approaches that deliver saliency maps or attribution maps that describe the features of the images that are most important to prediction. Crucially, while these approaches offer spatial localisation of significant features, they do not offer clear pixel-level explanation of the relationship between input features and model predictions. The generated saliency maps are imprecise and so produce especially poor localisations of irregularly shaped, small features (Saporta, Gui, Agrawal, et al., 2022). This is particularly problematic for explanation of co-occurring pathologies where features of different pathologies may exist close together. To mitigate the risk of shortcut learning and overcome the limitations of popular post-hoc explainability methods, I propose to use a multi-task generative approach to provide pixel-level explanations of multi-label predictions. This chapter addresses the existing shortcomings and discusses the use of VAEs for the explainable prediction of pulmonary disease in CXRs with co-occurring pathologies.

6.2 VAEs for Explainable Prediction

Variational Autoencoders (VAEs) are a type of generative model that applies principles of variational learning to an encoder-decoder model architecture to learn a distribution over data, typically Gaussian, in which factors of variation are captured in a lower-dimensional, latent representation[2] (Kingma, Welling, et al., 2019).

Achieving explainable prediction with VAEs requires that they learn interpretable latent spaces. Popular methods for learning interpretable latent spaces require disentanglement of the latent space. Disentanglement is achieved when each latent dimension describes some salient feature of the data. While a precise definition of disentanglement has not yet been agreed, a disentangled representation is generally understood to require statistical independence between latent dimensions (Burgess et al., 2018). This is popularly achieved by enforcing independence between the dimensions of the learned multivariate Gaussian posterior e.g., β-VAE (Higgins et al., 2017), Factor-VAE (Kim and Mnih, 2018), etc. (described in Section 3.2.2), and has been demonstrated on simple, toy datasets in which factors of variation are strictly independent. However, disentanglement - which requires independence between latent dimensions - is impractical for complex medical imaging data in which dependencies between salient features is assumed. For example, consider a healthy chest X-ray in which key features such as clear lung areas, normal heart size and shape, and symmetric breathing muscles correlate. These dependencies

---

[2] Described in more detail in chapter 3

capture how different parts of healthy anatomy support each other and are considered together. With this in mind *decomposition* of the latent space is more suitable. In this, data is decomposed into salient features, but independence between latent dimensions is not required [3].

For my application the VAE is required to learn structured representations of clinically complex medical images, where anatomical, pathological, and image acquisition-related features are separated from relevant disease features. Complex medical imaging describes a wide variety of non-independent visual and clinical concepts. Therefore it follows that learning a dense, unimodal Gaussian distribution over complex data, where independence between latent dimensions is enforced, is restrictive and counter-intuitive. To overcome the existing limitation of the Gaussian VAE, I propose to learn a sparse, multi-modal distribution over medical imaging data (Tonolini, Jensen, and Murray-Smith, 2020). Specifically, I introduce the Dirichlet-prior VAE (DirVAE) for decomposed representation learning of medical images. To encourage further separation of disease features I train models under the influence of an auxiliary multi-label classification task. I hypothesise that training classifiers on randomly sampled image representations complements multi-label prediction in regularising against shortcut learning. For explainable prediction, I hypothesise that with a prior distribution that facilitates distributional sparsity and multi-peak sampling, i.e., Dirichlet, the VAE will learn a sparse, multi-modal posterior that can be influenced by the auxiliary classifiers to separate disease-related features that are important for label prediction from non-clinical features that are not significant.

6.2.1 Prior distributions: Gaussian vs Dirichlet

In this work I compare the use of a dense Gaussian-prior with the use of a sparse Dirichlet-prior. Prior distributions serve as a regularisation tool during training and as a means to impose structure and constraints on the latent space. They can therefore be selected to enforce desirable properties in the latent representation (Kingma, Welling, et al., 2019). The preference between dense and sparse priors depends on the purpose of the generative model. A sparse prior distribution is characterised by having many zero or near-zero probabilities for most of the possible values, only a small subset of possible values has a high probability of being selected (Tonolini, Jensen, and Murray-Smith, 2020). While the probabilities of a dense distribution are more evenly spread out across the possible values. This implies that every possible value has a non-negligible probability of being selected. Given these properties, dense priors are generally preferred when the goal is to generate a wide variety of high-quality samples and when the underlying data distribution is complex. While sparse priors are preferred when the goal is to identify

---

[3] Disentanglement and decomposition are discussed in more detail in section 3.2.2

key features, sparsity limits the complexity of the latent distribution and therefore is less suited for the generation of high quality samples.

6.2.1.1 Gaussian Distribution

The latent space of a VAE is typically assumed to follow a multivariate Gaussian distribution, which means that each element of the latent vector follows an independent Gaussian distribution (Kingma and Welling, 2013). The Gaussian distribution is therefore parametrised by two vectors: the mean vector (μ) and the diagonal covariance matrix (σ^2).

The Gaussian probability density function is defined as,

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.1}$$

6.2.1.2 Dirichlet Distribution

The Dirichlet distribution is a continuous multivariate probability distribution, it is defined over a set of discrete distributions and can be considered a multivariate case of the continuous Beta distribution (Joo et al., 2020). The Dirichlet distribution is parametrised by a K-dimensional vector typically referred to as the concentration, where, K corresponds to the number of discrete distributions and $K \geq 2$. This value shapes the distribution and controls how concentrated the distribution is around the mean vector. Here the Dirichlet probability density function is defined as,

$$p(x_1, \cdots, x_K; \alpha_1, \cdots, \alpha_K) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \tag{6.2}$$

Where B is the Beta function and α is the concentration parameter.

Alternatively, the Dirichlet distribution can be defined through Gamma functions,

$$p(x|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \tag{6.3}$$

where, $\alpha_i > 0$ and $\Gamma(\cdot)$ is the gamma function. Note that each x_i is a positive real number and together must sum to 1.

The Dirichlet concentration parameter determines the shape of the distribution on the probability simplex, with its value controlling the balance between distributional sparsity and smoothness. In order to capture latent distributions with multiple modes, which I theorise encourages an explainable latent space, I use a Dirichlet concentration that promotes sparsity.

As a continuous distribution, optimisation of the Dirichlet posterior can be achieved through pathwise derivation (aka the reparameterisation trick), as shown below.

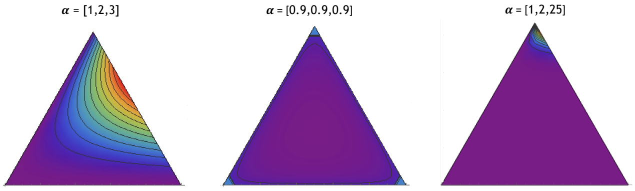


FIGURE 6.1: **3D Dirichlet distributions with different concentrations (α).**

$$z \sim q_{\phi}(z|x), \quad z = g(\phi, x, \epsilon), \quad \epsilon \sim Dir(\alpha) \tag{6.4}$$

where z is continuous, i.e., sampled from a continuous distribution, and z is cast as a function of fixed noise (Jankowiak and Obermeyer, 2018).

6.2.2 Generating visual explanations

I qualitatively assess latent space explainability using two core techniques: latent traversals and latent interpolations. Latent traversals involve varying one or few latent dimensions at a time while fixing all others, allowing observation of how specific features (e.g., size, shape, style) evolve and identification of semantically meaningful attributes. While latent interpolations blend two latent vectors, typically belonging to different classes of data, showing how transitions occur between different data concepts.

I adapt the classical latent traversal approach and introduce a gradient-guided latent traversal (GGLT) method to be used at inference to explain predictions. I first identify a single latent factor with the largest gradient activation from a positive classifier i.e. where the class of interest to a classifier was correctly detected for the input image. The identified latent factor is adjusted incrementally or 'traversed', while all other factors are preserved. I 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, I can assume that the latent space has been successfully structured in a way that isolates visual features relating to the class of interest. Pixel-wise variance is used to summarise changes across a traversal and identify features controlled by a specific latent factor. Algorithm 1 describes this process in pseudo-code.

Additionally, I explore latent interpolation as a method for explainable prediction. To isolate label-specific features I perform partial interpolations between label-positive CXR latent representations and a confounding-variable matched 'No Finding' CXR [4]. The most 'active' latent dimensions of the label-positive CXR is assumed to hold disease specific features and there are expected to be 'inactive' in CXRs without recorded pathologies. I describe interpolations as partial because I interpolate between only the n most different latent dimensions. In the presented

---

[4] To account for possible confounding visual features, the destination 'No Finding' CXR is matched to the label-positive CXRs matching age group and sex.

evaluations, cases with co-occurring labels are not considered, to simplify interpretation (as co-occurring labels would require exploring multiple relevant latent factors simultaneously and ground truth disease features were unavailable).

Algorithm 1 Gradient-guided latent traversal

```
Input: Latent representation, Logistic regression classifiers
Output: Saliency maps for each label prediction
for each label $i$ do
    Predict the probability of label $i$ from latent representation
        $z : Prob_i = \text{Classifier}_i(z)$
    Compute and identify largest gradients in the classifier:
        $Gradients_i = \nabla_{\text{Classifier}_i} \text{Loss}(Prob_i)$
    Extract most predictive latent factors from gradients:
        $LatentFactors_i = z[\text{argmax}(Gradients_i)]$
    Perform latent traversal using the selected factors:
        $LatentTraversal_i = \text{Traverse}(LatentFactors_i)$
    Generate saliency map from pixel value variance across the traversal:
        $\text{Saliency map}_i = \text{Variance}(\text{Pixel values along } LatentTraversal_i)$
end for
```

6.3 Application: Explainable Prediction of Multi-label CXRs

My approach uses a VAE with an attached classification module comprising an ensemble of logistic regression classifiers (a classifier for each label). The encoder is made up of 4 blocks, each of which comprises a convolutional layer, batch normalisation layer, activation layer and pooling layer. With each block the convolutions extract increasing numbers of kernels, from 64 to 512. The decoder is made up of 4 blocks, mirroring the encoder. Transposed convolutions are used to upscale the kernel dimensions. I apply a Gaussian prior distribution of diagonal covariance to define the GVAE and apply a Dirichlet prior distribution for the DirVAE. I define this distribution for extreme sparsity. The Dirichlet prior is defined as, $\boldsymbol{Dir}(0.5 \cdot \mathbf{1}_k)$, where k is the dimensionality of the distribution and the concentration value is set to 0.5 in all dimensions k.

I apply this approach to the CheXpert data set, one of the largest publicly available CXR datasets with multi-label outcomes (described in Chapter 3). To overcome

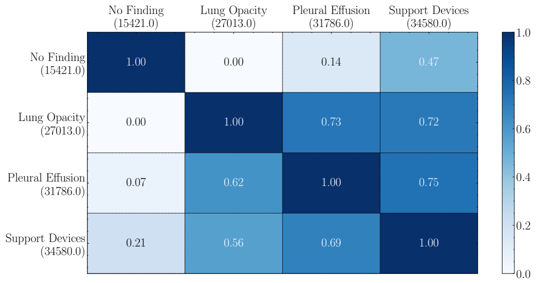


FIGURE 6.2: **CheXpert co-occurrence heatmap, raw counts of each label are included.** Co-occurrence is normalised to a 0-1 scale.

issues of severe class imbalance in the data, and simplify training of the logistic regression classifiers, I randomly 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*, allowing for co-occurrence between classes. Figure 6.2 presents the frequency of classes of interest as well as their frequency of co-occurrence. I split the CheXpert dataset into train (n = 52,943), validation (n = 5057) and test (n = 10,000) sets.

To encourage separation of latent factors according to clinically-significant features I apply an ensemble of simple independent logistic regression models, where each unit is optimised to predict one of four CheXpert labels from the learned latent space and outputs are combined to produce a set of independent label probabilities i.e., multi-label probabilities. For the DirVAE I apply the log transformation to map Dirichlet-sampled latent vectors from the simplex to a Euclidean space. This transformation removes the sum-to-one constraint and enables the use of linear classifiers.

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,

$$KL(Q \parallel P) = \sum log\Gamma(\alpha_k) - \sum log\Gamma(\hat{\alpha}_k) + \sum(\hat{\alpha}_k - \alpha_k)\psi(\hat{\alpha}_k) \tag{6.5}$$

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 (Joo et al., 2020).

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, taking the log of the CXR latent representations as inputs (for the DirVAE only). Once each classifier is fully optimised, the ensemble of logistic regression models are trained together, this objective takes the form,

$$L_{BCE} = \frac{1}{4}\sum_{c=1}^{4} -\frac{1}{n}\sum_{i=1}^{n}(y_{c,i} \cdot log(p_{c,i}) + 1 - y_{c,i} \cdot log(1 - p_{c,i})), \tag{6.6}$$

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, I 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. Figure 6.3 gives an overview of the DirVAE and classifier framework.

In this stage VAE optimisation is influenced by the classification task, and the ELBO term can be re-expressed to reflect this,

$$\text{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - \lambda \text{KL}[q(z|x)||p(z)] + \mathbb{E}_{q(z|x)}[\log q(y|x,z)]$$

where y represents the class label in the auxiliary classification task, and λ is a weighting factor that controls the importance of the KL divergence term relative to the reconstruction and regularisation terms. As the focus of this study is explainable prediction, not image generation, I down-weight the KL divergence term in the multi-task objective and set $\lambda = 0.01$. Preliminary experiments demonstrate that this improved classification performance and reconstruction quality. Both the DirVAE and GVAE were optimised using the Adam algorithm. Logistic regression models were optimised under stochastic gradient descent (SGD). Table 6.1 present the experimental setting used for Dirichlet-prior and Gaussian-prior VAE training.

I pursue joint optimisation on the assumption that optimising the latent space with the multi-label classification task will re-structure the latent space such that 'active' latent dimensions explain class-specific visual features. This allows simple linear classifiers (logistic regression) to predict the CXR target classes from the learned latent space. I theorise that this will also enhance latent space interpretability. Latent factors with greater discriminative power for predicting a given class correctly, should correspondingly encode visual features in the CXR image that are representative of that class. Intuitively, if only a few factors have significant impact on the predicted probabilities of a logistic regression model, the latent space can be assumed to be decomposed into clinical features relating to the pathology of interest. With this in mind, I explore my proposed strategy for explainable multi-label

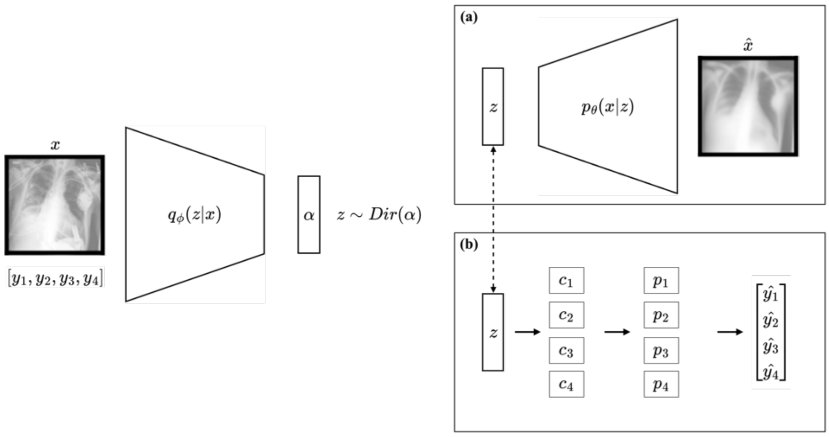


FIGURE 6.3: **Overview of DirVAE with auxiliary ensemble of logistic regression classifiers.** The DirVAE is trained under a multi-task objective combining a (a) reconstruction task with (b) a multi-label prediction task.

prediction, gradient-guided traversals: for each label, I use classifier gradients to identify the most important latent dimensions for classifier prediction and traverse these dimensions to isolate features that are significant for prediction (Algorithm 1). I evaluate the Dirichlet-prior VAE and the conventional Gaussian-prior VAE for their capacity to learn class-separable, explainable representations of medical imaging (Kingma, Welling, et al., 2019).

Parameter	**Setting**
Deterministic WU epochs	1000
Training epochs	1500
KL weighting	0.001
Learning rate	0.0001
Optimiser	Adam

TABLE 6.1: **Experimental Settings for Dirichlet-prior and Gaussian-prior VAE training.**

The goal of this study is to evaluate the capacity of the DirVAE to learn explainable, decomposed latent representations of CXR images, for multi-label classification. Accordingly, I 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 standard classification metrics (evaluated per-class) and multi-label metrics, Hamming score and Exact Match Ratio (EMR). All results are compared directly with GVAE performance.

The Hamming score assesses the accuracy of predicted labels against true labels on a per-sample basis. It is calculated as the proportion of correctly predicted labels, accounting for both matches (correct positives and negatives) and mismatches (false positives and negatives). The score ranges from 0 to 1, with 1 indicating perfect predictions.

In mathematical form,

$$\text{Hamming Score} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\text{True}_i \cap \text{Pred}_i|}{|\text{True}_i \cup \text{Pred}_i|}$$

where True_i and Pred_i are the sets of true and predicted labels for the i-th sample, and n is the total number of samples.

The Exact Match Ratio (EMR) measures the percentage of samples where the predicted labels exactly match the true labels across all classes.

The formula for EMR is,

$$\text{Exact Match Ratio} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(\text{True}_i = \text{Pred}_i)$$

where $\mathbb{1}$ is a function that returns 1 if the true label set True_i exactly matches the predicted label set Pred_i for sample i, and 0 otherwise, and n is the total number of samples.

6.3.1 Results

Table 6.2 shows logistic regression model performances, for both the DirVAE and Gaussian VAE. The DirVAE classifiers perform marginally better than the GVAE classifiers, with performance metrics marginally higher across most metrics. Multi-label prediction metrics show that the DirVAE classifiers perform modestly when considered together, with an exact match rate (EMR) of 0.38±0.01 and Hamming loss of 0.25±0.02. Similarly, the Gaussian VAE achieved an EMR of 0.37±0.01 and Hamming loss of 0.26±0.01.

	Label	**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
AUROC	*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 6.2: **Label-wise logistic regression classification performance results.**

I informally investigate the sparsity of the DirVAE posterior by visualising the frequency of values in a sampled latent representation given a randomly selected

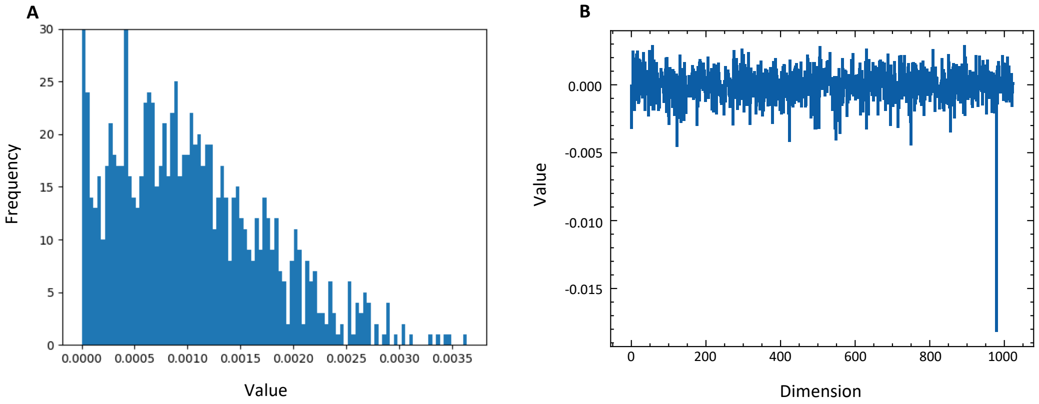


FIGURE 6.4: **Dirichlet-prior VAE 'sparseness'.** (a) Frequency of values sampled from the Dirichlet-prior VAE posterior, $z \sim q(z|x)$. (b) Example of classifier weights extracted from the *Pleural Effusion* classifier paired with Dirichlet-prior VAE.

CXR. Correspondingly a kind of 'sparsity' in the classifier weights is observed. Figure 6.4 shows the *Pleural Effusion* logistic regression classifier weights, the vast majority of values are near zero, while the weight indexed at 987 has a much larger absolute value. This suggests that regularisation to the sparse Dirichlet-prior successfully induced sparsity over the posterior distribution, and that training linear classifiers on samples from this sparse distribution results in disproportionately large weight assigned to a single dimension. This large weight is indicative of feature dominance, where a particular feature (or the corresponding input dimension) has a disproportionately large influence on classifier predictions.

Figure 6.6 presents the results of traversing the latent spaces in the direction of the most predictive latent dimension, where the most predictive latent dimension is identified from classifier gradients i.e., GGLT. For the DirVAE, clear structural changes are observed for each class. Typically, latent traversals gradually remove or intensify disease-related features, in a manner specific to the input CXR image. I consider both types of change as evidence of disease-/class-specific and image-specific isolation, and so generate pixel intensity variance maps (presented alongside each traversal, bottom right corner of each panel) in order to capture both types of feature change.

I 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 costophrenicangles and hemidiaphragm region, two key areas of diagnostic relevance for this pathology (Fig. 6.6). For *Lung Opacity* cases, DirVAE traversals consistently highlight areas of suspected consolidation (verified by clinical expert). Clinically, the radiographic presentation of lung opacities varies greatly, this is reflected in the diversity of feature changes observed during DirVAE traversals. Traversals capture features

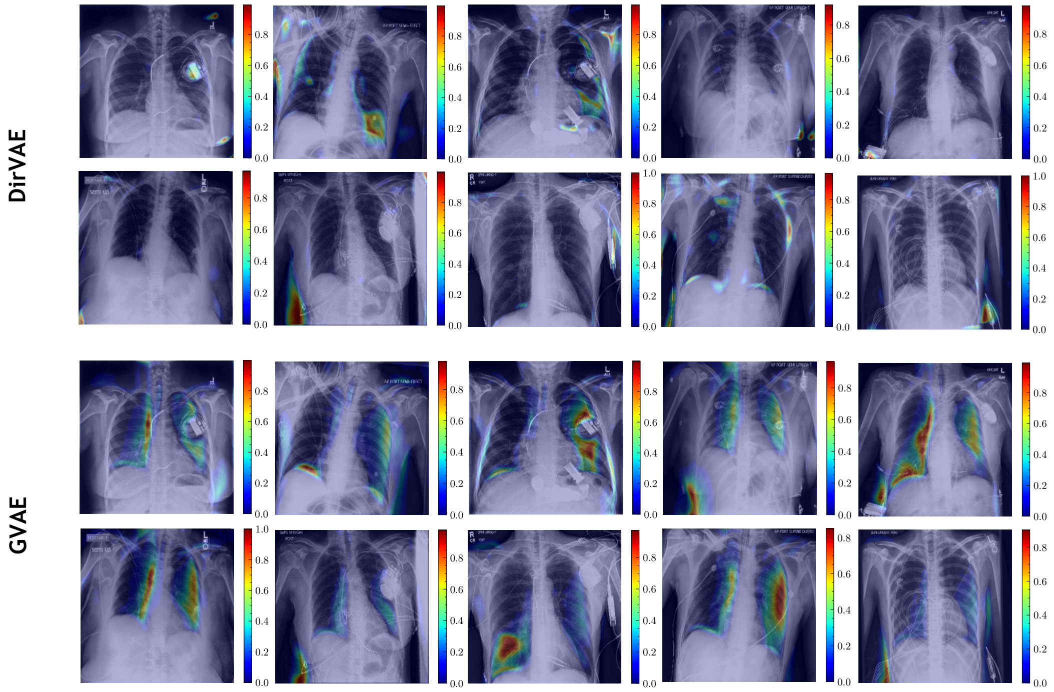


FIGURE 6.5: **Saliency maps representing variance over all reconstructions in a gradient-guided latent traversal series, performed for the *Support Devices* class, with comparison between GVAE and DirVAE.** The colour bar shows pixel variance over the traversal, which is rescaled to 0-1 range.

of consolidation that vary in shape, size and position, matching the appearance of diagnostic features in the example CXRs (Fig. 6.6).

Similarly, in evaluating traversals of the *Support Devices* class, I 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 (Fig. 6.6). 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. I consider the use of gradient-guided traversals for the identification of classifier reliance on 'shortcut' and confounding features. With successful isolation of salient features in the latent space 'spurious' and confounding features can be observed in some cases (indicated by red boxes overlaying CXR images in Fig. 6.6). Here, I also see that latent traversal causes feature changes relating to radiology annotation, co-occurring classes (particularly support devices) and shoulder position. In Figure 6.6 Support Devices C I observe specific changes that resemble wiring begin to appear, this is a feature that spuriously correlates with the support device class but is not itself deemed a support device. This indicates that the decomposed 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. 6.5).

For GVAE GGLTs, traversal variance appears diffuse across the reconstructed

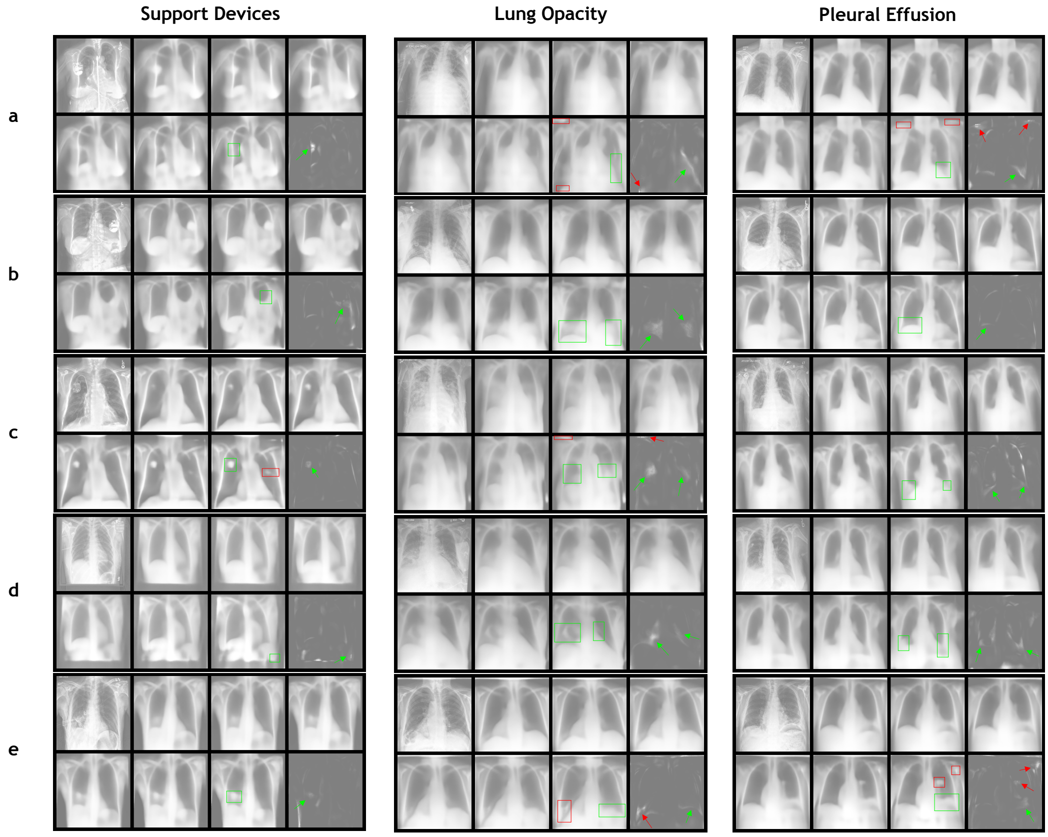


FIGURE 6.6: **DirVAE gradient-guided latent traversals, with examples from each CXR class.** Each grid presents a different example, the top left image in the grid is the original reconstruction, the entire traversal is presented sequentially in the grid. The bottom right image presents a pixel-wise 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.

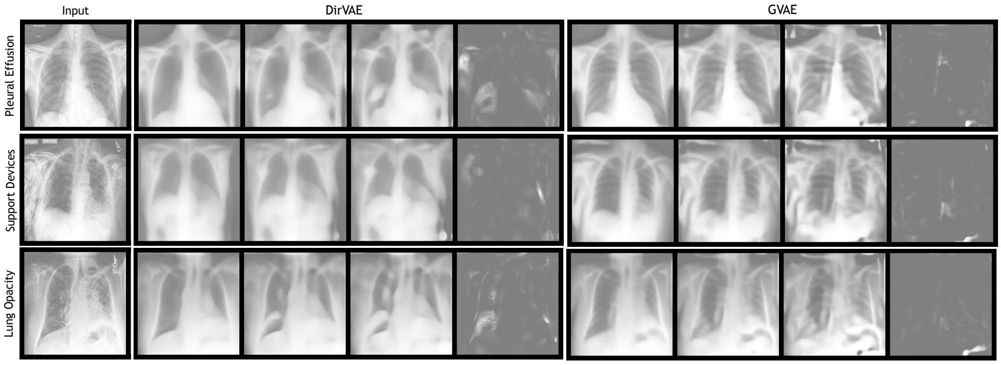


FIGURE 6.7: **Partial latent interpolations (n=10) between label-positive CXRs and a confounding variable matched *No Finding* sample.** *n* is the number of latent dimensions interpolated. I include *Pleural Effusion* (top row), *Support Devices* (middle row) and *Lung Opacity* (bottom row) examples, and compare the Dirichlet-prior (left) and Gaussian-prior VAE (right). In each row the first column is the input image, columns 2-4 shows the progressive change in the reconstructions over the interpolation series, and column 5 (final) is the residual between the unchanged reconstruction and reconstruction post-interpolation.

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 air space. 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 6.5, which presents a direct comparisons between DirVAE and GVAE traversal saliency maps (for the same input CXR images), with *Support Devices* as the class of interest.

Partial interpolations offer another method for generating visual explanation from DirVAE-based predictions. Figure 6.7 demonstrates that DirVAE is able to isolate label-specific features, while the visual explanations from GVAE partial interpolations appear largely unrelated to label features. Figure 6.9 shows image changes and corresponding changes in predicted probability over partial latent interpolation. As the *Pleural Effusion* latent representation is adjusted towards the *No Finding* latent representation, there are clear changes in the lower lung field. Correspondingly, I observe increasing classifier probabilities for *No Finding* and decreasing probabilities for *Lung Opacity* probabilities. I also see small decreases in the predicted probabilities for *Pleural Effusion*, a pathology that shares features with *Lung Opacity*.

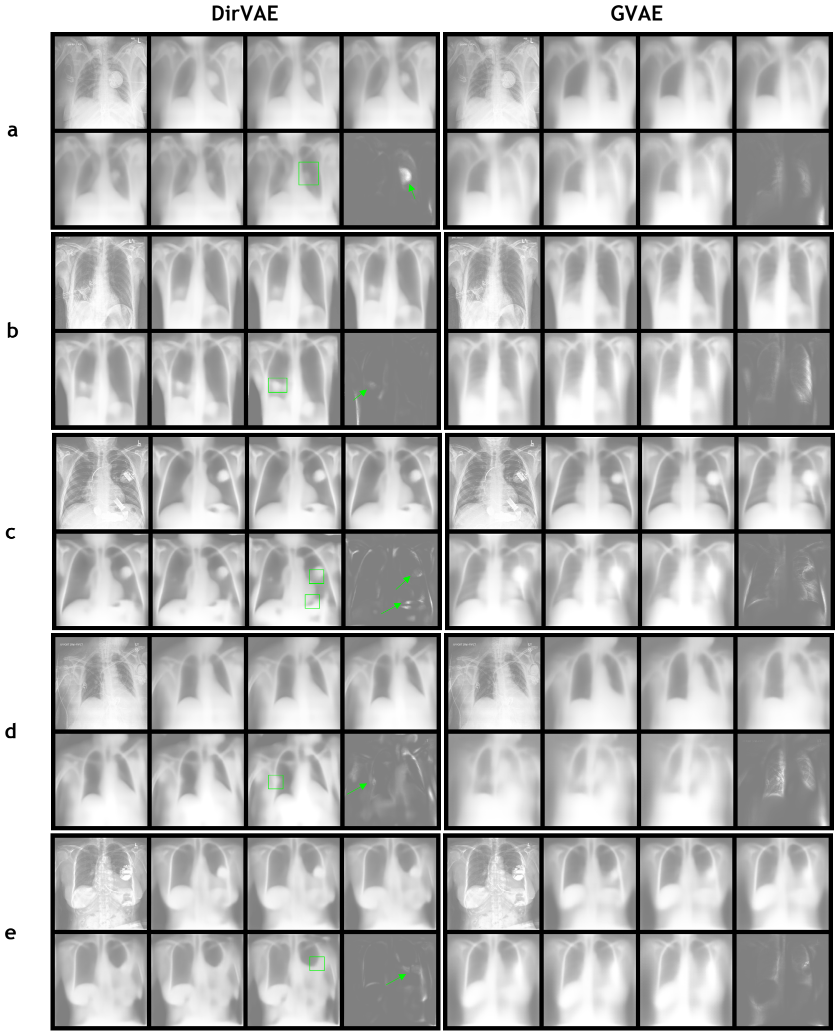


FIGURE 6.8: **Comparison of GVAE and DirVAE traversals, with examples from the *Support Devices* class.** Each row presents a different example, the top left image in the grid is the input CXR, the next image is the original reconstruction, the entire traversal is then presented sequentially in the grid. The bottom right image presents a pixel-wise 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.

6.3.2 Discussion

Analysis of latent traversals suggests that the DirVAE learns a more explainable 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, DirVAE classifiers perform marginally better than GVAE. The improvement in latent structure 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 the Dirichlet prior is enforced by setting its concentration parameter $\alpha = 0.5$ across all experiments. This

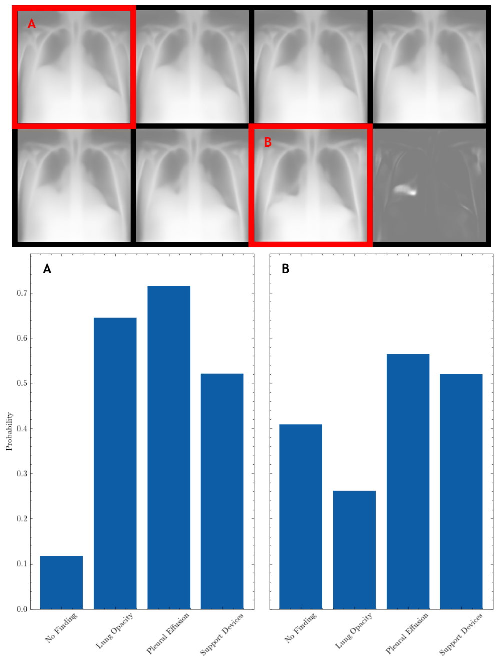


FIGURE 6.9: **Partial interpolation *(n=10)* from *Lung Opacity* CXR to *No Finding* CXR latent representation. n is the number of latent dimensions interpolated. Top**: Panel of image reconstructions resulting from interpolation between *Lung Opacity* CXR and *No Finding* CXR. (A) Original reconstruction of *Lung Opacity* CXR and (B) Reconstruction after interpolation to *No Finding* CXR. **Bottom**: (A) Probabilities predicted from latent representation of *Lung Opacity* CXR. (B) Probabilities predicted from latent representation after partial interpolation towards *No Finding*.

pushes areas of the posterior distribution towards zero, i.e. the inner areas of the simplex (Fig. 6.1) and creates a narrower information bottleneck.

My approach to encouraging interpretability in the latent space follows a similar principle to β-VAE, but with a different strategy for narrowing the information bottleneck (Higgins et al., 2017). In order to generate quality reconstructions under strict sparsity constraints the latent factors have to capture more global features. Consider the impact of a narrower bottleneck on the reconstruction objective as well as the classification objective, which influences the model to prioritise the learning of disease-related features, this strategy encourages the learned latent variables to capture 'global disease features' and explain away noise i.e., in the posterior distribution high-level disease features are pushed to dense regions while non-predictive features are pushed to sparse areas.

This principle can be observed in latent traversal evaluations, by traversing latent dimensions important for lung opacity classification. There are no diffuse changes in unrelated CXR features, but clear changes in homologous features of the image. Moreover, there is evidence of DirVAE capacity to decompose 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 (irregularly shaped and 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. Isolated feature changes is also observed with partial, confounding-controlled interpolation method. Overall, the proposed use of latent representation manipulation for identification of the salient visual features underlying classifier predictions facilitates improved model explainability.

Within my explainability approach, successful decomposition of DirVAE latent representations allow us to more clearly observe the appearance of 'spurious correlations' along side 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) (Jabbour et al., 2020). 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 (Gaube et al., 2023). 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, have been criticised for their inability to highlight smaller pathologies or structures with complex shapes, for example, mechanical wiring (Selvaraju et al., 2017a; Saporta, Gui, Agrawal, et al., 2022). They are similarly poor at highlighting important featured that are distributed far apart in the image.

My 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, my 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 (Saporta, Gui, Agrawal, et al., 2022). This is well-illustrated by an example of confounding highlighted in a *Lung Opacity* traversal, 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) (Fig. 6.6e). 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 multi-label classification. Due to the dense and unimodal nature of Gaussian distributions, the learned latent space appears unable to separate disease-/class-specific features from other image-specific features. This is particularly evident in the latent traversals presented for CXR images from the *Support Device*s class (see Fig. 6.8). From this, I can conclude that in this framework the GVAE has limited functionality as a method for prediction explainability.

While I am able to demonstrate success in explainable prediction and its utility in identifying shortcut features, further evaluation is needed to determine the extent of shortcut feature learning. Moreover, without knowledge of true generative factors of variation in dataset (i.e., dSprites). It is difficult to quantify the 'decomposition' or 'disentanglement' of the latent representations. Some work proposes a linear classifier-based approach to quantifying - on the reasoning that linear separability suggests some degree of axis-aligned semantic feature learning (Carbonneau et al., 2022). However, the influence of the classification objective on representation learning disqualifies this as a suitable approach. Further work should go to assessing and quantifying the decomposability of the latent representations learned by DirVAE under the influence of the auxiliary classification task.

In addition, I observe some limitations compared to existing work, such as counterfactual generation, namely that my current application of latent traversals offers no directional control, i.e., I can only generate feature changes not remove or intensify relevant features. Moreover, I identify that generally CXR reconstructions are of too low quality to be used directly as counterfactual explanations, I find the clearest means of interpretation is via the traversal variance saliency map.

I propose that with improved reconstruction quality and directional control GGLT explanations can rival counterfactual generation. Further work will concentrate on this.

6.4 Conclusion

In this work I demonstrated that decomposition of class-/disease-specific (and potentially, clinically-relevant) features can be achieved using my DirVAE model and that its capacity for class-/disease-specific disentanglement is superior to a GVAE. I introduced a promising new approach for explainable multi-label classification, where I applied an ensemble of simple logistic regression classifiers and explored latent dimensions of significance to class predictions through so-called 'gradient-guided latent traversals'. With this I provided visual explanations that highlighted regions in CXR images clinically relevant to the class(es) of interest and additionally, I was able to identify cases where classification was biased and relied on spurious feature correlations.

Future work will explore the use of metrics to quantify explainability in order to formalise my qualitative assessment and improve image reconstruction/generation quality, with a view to improve the clarity of feature changes during latent traversals. With better reconstruction quality the developed approach would have applications in both explainable medical AI and synthetic data.

Chapter 7

Multi-task Hierarchical VAEs for Disease Localisation

7.1 Introduction

As demonstrated in Chapter 6, through the use of sparse-priors and multi-task training, VAEs are able to learn a posterior distribution in which factors of variation are captured in a structured lower-dimensional, latent representation. I aim to further this work by applying the same principal to Hierarchical VAEs (HVAEs). HVAEs extend the basic VAE framework by introducing a hierarchy of stochastic latent variables. With this hierarchy, HVAEs can better approximate complex distributions by allowing for dependencies between latent variables at different levels (Maaløe et al., 2019). By increasing the expressivity of the learned posterior HVAEs should be able to improve on the capacity for VAEs to predict CXR labels and generate higher resolution images for more explainable prediction, a limitation of single VAEs that I have already observed in Chapter 6. With this in mind I hypothesise that representations learned by HVAEs can give better classification performance and more precise localisation of disease features within the visual explanations.

In this work I introduce multi-task sparse-prior HVAEs for decomposed representation learning of complex medical images under the influence of an auxiliary multi-label classification task. I hypothesise that, like the Dirichlet-prior VAE, the sparse-prior HVAE will learn a sparse, multi-modal posterior that can be influenced by the auxiliary classifier to separate class-related features that are important for label prediction from irrelevant features. I intend to use this property to facilitate the explainable prediction of complex, multi-label CXRs. I explore the use of a sparse mixture of Gaussian distributions as a prior distribution (Sparse-prior) to the HVAE model, specifically the bi-directional inference model (BIVA), and compare the performance of this sparse-prior HVAE with that of the standard dense Gaussian-prior HVAE[1]. HVAEs are evaluated for the predictive performance of their latent representations as well as for their capacity to isolate class-specific features in the latent

---

[1] I initially explore the use of Dirichlet-priors for this but find that the extreme sparsity imposed by the Dirichlet posterior makes HVAE-training unstable, instead I pursue Gaussian-based sparse priors.

representation, which I quantitatively evaluate via a disease localisation task. Localisation is visualised through classifier-informed latent traversals, where classifier gradients are used to identify the most predictive latent factors, which are then traversed to generate a corresponding change in the generated images i.e., GGLT.

For baseline comparison I explore the application of Dirichlet-prior, Sparse mixture of Gaussians-prior, and Gaussian-prior VAEs in this multi-task framework. I evaluate these VAEs for their predictive performance as well as their capacity for explainable prediction. I compare all variational models with a deep CNN trained under a multi-label objective. This avenue of investigation extends my work on Dirichlet-prior VAEs (Chapter 6) with the quantitative evaluation of disease localisation as a measure of explainability and additional comparison with alternative sparse-prior distributions (e.g., Sparse-prior).

In this Chapter my research aims are as follows:

- Evaluate the performance of multi-task sparse-prior HVAEs for the multi-label prediction of co-occurring pathologies in CXRs, with comparison to multi-task VAE performance as the baseline method for variational approaches, and comparison to the Gaussian-prior HVAE as the standard choice of prior distribution for HVAEs.
- Explore the use of HVAEs for explainable prediction, proposing novel methods for visual explanations through latent traversal, and quantifying this through comparison with radiologist annotated disease features.
- Compare variational approaches with a deep CNN trained under a multi-label classification objective, and compare traversal-generated explanations with GradCAM++, a popular post-hoc method for generating visual explanation.

7.2 Related Work

My work using sparse-prior HVAEs draws on principles of sparse coding, generative modelling, and explainable prediction. For context, I introduce the concept of sparse representation learning and its application to variational inference, and I discuss existing methods for the explainable prediction of medical images, with a focus on HVAEs.

A number of published methods have combined principles of sparse coding with variational inference. Based on the core principle of sparse coding, these approaches aim to learn a sparse posterior, in which the encoder is induced to represent the data in as few active latent variables (non-zero) as possible, with a varying number and differing combination of active latent variables for each data sample (Tonolini, Jensen, and Murray-Smith, 2020). Intuitively, sparse representations are well-suited for learning efficient latent representations of images. Complex imaging datasets contain a vast number of features, but only a small subset is present in any given image. Sparse representations can capture the factors of variation that describe these datasets by activating only a few latent dimensions for each image. Different images activate different combinations of these dimensions, which allows for a compact and efficient way to represent diverse visual information with minimal redundancy. The properties of sparse representations naturally facilitate interpretability, and is often demonstrated through isolation of salient visual features in latent traversals. Sparse-inducing prior distributions can be truly sparse or near-sparse. Here I introduce existing methods that apply sparsity to VAEs. Gyawali et al. (2019) use the Indian buffet process as a truly sparse prior distribution in a VAE framework. The Indian buffet process is defined by binary sampling: each observation i draws discrete $0/1$ values for each existing feature k with probability $\frac{mk}{i}$ (where mk is previous occurrences), followed by sampling Poisson($\frac{\alpha}{i}$) new features. This inherently generates sparse matrices since most entries are zero and all entries are discrete binary values. Gyawali et al. (2019) have used this process to demonstrate improved isolation of disease features in complex imaging data, including a skin lesion image dataset.

Similar to the Indian buffet process, the stick-breaking process is a truly sparse distribution that works by recursively breaking off portions of a unit interval i.e., a "stick" of length 1, where each break point is drawn from a Beta distribution. The length of each broken piece determines the probability of selecting that feature, naturally leading to a sparse set of active features as the remaining stick length diminishes exponentially. This creates a sequence of decreasing probabilities that sum to 1, effectively prioritising a small subset of features while allowing for theoretically infinite dimensions. Nalisnick and Smyth (2017) use this process as a prior in a VAE framework and demonstrate improved discriminative power in semi-supervised classification tasks compared to Gaussian-prior VAEs.

Alternatively, near-sparse distributions constrain areas of the distribution to near-zero, through this sparsity can be represented despite distributions not being truly sparse. When applied to variational inference, I observe that the latent representation of data through sparse posteriors can be achieved with the application of near-sparse prior distributions, such as, the Spike-and-Slab probability distribution, the Dirichlet distribution, and a *'sparsified'* mixture of Gaussian distributions.

Tonolini, Jensen, and Murray-Smith (2020) propose to model sparsity in the latent representation with a Spike-and-Slab probability distribution prior VAE. In Bayesian statistics, the Spike-and-Slab prior is used to separate relevant variables, or features, from irrelevant variables. I describe this method in more detail in Chapter 3. Fallah and Rozell (2022) propose a new approach to sparse coding that uses learned thresholding. To facilitate variational sparse coding they apply a soft-threshold function which sets some samples from the latent representation to zero. They demonstrate that the thresholded samples are identically distributed to the

Spike-and-Slab distribution but their method offers greater control over the degree of sparsity imposed. With this method they show efficient feature decomposition of the CelebA dataset.

The Dirichlet-prior VAE was first introduced by Joo et al. (2020). In this work, they consider the sparse multimodal properties of the Dirichlet distribution, by setting $\alpha = 0.99 \cdot 1_{dim}$, where *dim* is the number of latent dimensions. Parametrising the Dirichlet prior distribution with α values less than one creates a sparse, multi-modal distribution. They demonstrate successful factorisation through latent traversals over the MNIST dataset. Similarly, Xu, Fan, and Liu (2023) demonstrate unsupervised *disentanglement* via the Dirichlet VAE through latent traversal over dSprites and 3dShapes datasets. Although this work does rely on the softmax Laplace approximation of the Dirichlet distribution, which has been shown to be a restrictive, inaccurate approximation of a multi-modal Dirichlet.

Moreover, Mathieu et al. (2019) model a sparse posterior using a mixture of Gaussian distributions, where a narrow Gaussian component pushes latent variables towards zero. Of the sparse distributions discussed here I select the sparse mixture of Gaussians-prior as the HVAE sparse prior for two key properties, an easily controllable degree of sparsity and stable Gaussian optimisation. I describe this probability distribution in more detail in Section 7.3.1.

Research on the use of HVAEs for explainable prediction remains relatively limited, although a handful of recent studies across different domains have begun to explore this area. Li et al. (2023) implement a multimodal hierarchical conditional VAE for salient object detection. This method incorporates RGB imaging alongside auxiliary modalities, including, thermal data, depth data, and image captions. This method integrates HVAEs into a mixture of products of experts framework to aggregate the multimodal latent variables, which are also concatenated with deterministic features to be delivered to the decoder for salient object detection. Vercheval and Pižurica (2021) use HVAEs to learn a posterior distribution over latent variables that is conditioned on ground truth data. They visualise salient objects by varying the conditional variable, decoding the latent variables and capturing changes in image reconstruction.

Hierarchical VAEs have also been used for visual counterfactual generation. In this work, the HVAE learns latent representations of images, specifically celebrity faces (CelebA), conditioned on predicted probabilities from an independent classifier trained to predict sex (Vercheval and Pižurica, 2021). Counterfactuals are then generated by varying the conditioning variable i.e., probability value. Both approaches require a conditioning variable to generate explanations and do not present evidence of decomposition of generative factors within an unconditional latent representation.

Recent work indicates it is possible to achieve explainable HVAEs without the need for a conditioning variable. Vafaii, Yates, and Butts (2024) encourage disentanglement through the β-VAE mechanism of up-weighting the KL divergence term of the ELBO objective. In this work an NVAE, a type of HVAE described in Chapter 3, was applied to predict neuron response in the motion processing pathway of primates. With the β-VAE objective they were able to identify generative factors through varying a single latent factor in top level NVAE latent representations. Additionally, in the medical imaging domain Biffi et al. (2020) apply the LVAE (see Chapter 3) to learn 3D anatomical shapes in cardiac and brain MRs. They apply a MLP classifier to the top level of the LVAE for disease prediction, training LVAE+MLP end-to-end to encourage learning of class discriminative features. With this framework they demonstrate explainable prediction of disease classes from anatomical shape segmentations. Explanations are generated through latent traversals in the 2-dimensional top level of the LVAE, these showed clear, clinically-relevant changes in anatomical shape. However, the datasets used here to describe anatomical shape are binary segmentation maps, with only a few generative factors this data is simple compared to the complex array of features comprising CXRs. To my knowledge the explainable prediction of CXRs through HVAEs has yet to be explored.

7.3 Bi-directional inference VAE

For this work I use the bi-directional inference VAE (BIVA). The BIVA model is a type of hierarchical VAE, which extends the classical single-level VAE.

BIVA inference combines the likelihood model $p(x|z)$ with the generative model $p(x,z)$ by sharing parameters between inference and generative models as part of a top-down inference method. In this process, the likelihood distribution is first approximated with a stochastic upward pass i.e., distribution parameters are estimated from the deterministic feature space and distributions are ancestrally sampled.

This is followed by a stochastic downward pass, in which stochastic latent variables are ancestrally sampled i.e., $p(z_i|z_{i+1})$, to recursively compute the approximate posterior

$$q(z|x) = q(z_K|x) \prod_{i=1}^{L-1} q(z_i|z_{i+1}),$$

where L is the number of layers, and the generative distribution $p(x,z)$. The inference model can therefore be thought of as comprising top-down and bottom-up stochastic latent variables,

$$z_i = \{z_i^{BU}, z_i^{TD}\}$$

where z_i^{BU} belongs to a bottom-up inference path and z_i^{TD} belongs to the top-down path. Figure 7.1 shows information flow through the BIVA inference model. I define deterministic blocks as D_i, where the deterministic function D_{i+1} outputs a lower dimensional feature space than D_i. BIVA uses a top-down deterministic path in which $(z_{i+1}^{BU}, z_{i+1}^{TD})$ and D_{i+1} features are combined and passed to z_i. Information from the BU approximated likelihood $p(x|z)$ is combined with TD information from the generative distributions $p(x,z)$ to give the approximate posterior $q(z|z,x)$. Here, both the generative model and inference model are dependent on top down information flow. In other words, the inference model recursively corrects the generative distribution $p(x,z)$. Step-by-step, the data flows through the deterministic upward pass to approximate the likelihood distribution $p(x|z)$. Then the stochastic downward pass computes the approximate posterior and generative distribution. The inference model is therefore a combination of BU information and TD information flowing from the prior. Together, these information paths facilitate highly expressive models that are capable of approximating very complex datasets.

Moreover, in this work I add a classification model $q_\phi(y|x, z_{<L}^{BU})$ to the inference network i.e., I include an auxiliary multi-label prediction task and train an ensemble of logistic regression classifiers with the top latent level as input. Note that for Dirichlet-prior VAEs, a log transform is applied to the classifier input in order to project this vector from the simplex into euclidean space. The number of logistic regression classifiers depends on the number of classes, which varies depending on the prediction task.

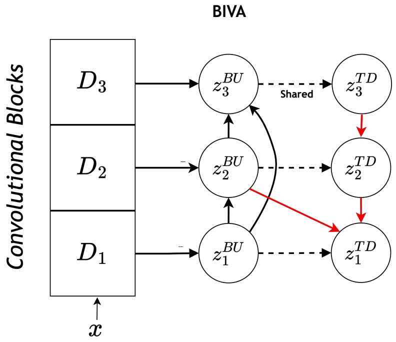


FIGURE 7.1: **3 layer BIVA inference model.** Dotted lines indicate shared parameters. Black arrows indicate bottom-up information flow and red arrows indicate top-down information flow.

7.3.1 Prior distributions

In this work I explore the use of the Gaussian distribution and a Sparse mixture of Gaussian distribution as prior distributions for VAE and BIVA models, I also further my use of the Dirichlet prior for VAEs. I described the Dirichlet and Gaussian distributions in Chapter 6 and introduce the Sparse mixture of Gaussian distributions, which I refer to as the *Sparse* distribution for brevity, here.

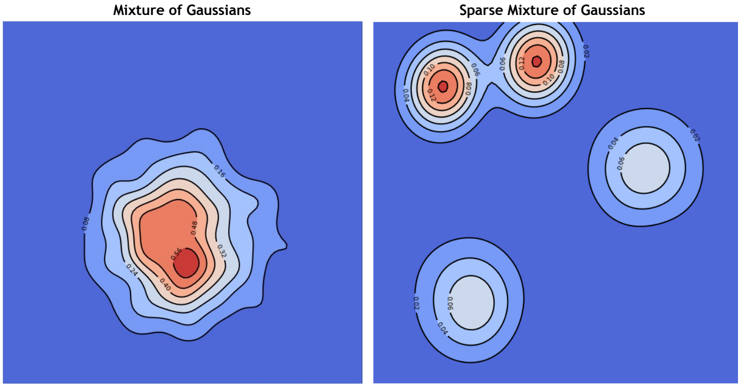


FIGURE 7.2: **Simulated sparse Gaussian mixture.** I sample from four 2D Gaussian distributions and fit a Gaussian kernel to estimate density across the mixture of distributions. To simulate the sparsity constraint I restrict the variance of two out of the Gaussian distributions to 0.01.

7.3.1.1 Sparse mixture of Gaussian distributions

As in Mathieu et al. (2019), I define the sparse Gaussian mixture prior distribution as,

$$p(z) = \prod_{d} (1-\gamma)\mathcal{N}(z_d; 0, 1) + \gamma\mathcal{N}(z_d; 0, \sigma_0^2)$$

where $\sigma_0^2 = 0.01$, d defines the number of distributions within the mixture, and γ defines the proportion of sample distributions within the mixture restricted to near 0, i.e., are bounded to a mean of 0 and variance of σ_0^2. Intuitively, distributions restricted in this way are limited in their capacity to encode information and so are 'switched off'. As γ defines the proportion of distributions that are restricted, this can be thought of as the 'sparsifying' parameter. Figure 7.2 shows a sparse Gaussian mixture alongside an unconstrained Gaussian mixture. Side-by-side comparison demonstrates the effects of 'sparsifying' the mixture of Gaussians, a multimodal distribution with vast areas of sparsity between modes is created.

7.3.2 Generating visual explanations with BIVA

As in Chapter 6, I use latent traversals to generate visual explanations of model predictions. I explore the use of two different algorithms for explanation generation, which I have termed gradient-guided latent traversals (GGLTs)[2] and optimised latent traversals (OLTs).

To perform a GGLT for a given prediction, the largest classifier gradients are identified, the aligning latent dimensions are identified as the most predictive factors and latent traversals are performed over these factors. During traversal iterative changes are made to the selected factors and images are generated with each iteration. Pixel variances across the traversal are calculated to generate a saliency map. Conversely, OLTs require direct optimisation of the latent representation used

---

[2] Introduced and described in more detail in Chapter 6

as input to the classifiers. All models weights are frozen and classifier gradients are used to update the latent representation w.r.t a BCE loss classification objective for which I construct a new target vector with the predicted label set to 0. The latent representation is then updated to minimise the probability of observing the class of interest for any given positive prediction. Algorithm 2 describes this process in pseudo-code. For the OLT experiments, I use a SGD optimiser and optimise over 2000 iterations. The size of the updates is controlled by the optimiser learning rate, which I vary depending on the prior distribution.

Counterfactual generation A key difference between GGLT and OLT, is that OLT explanations are directional. The OLT method updates the latent representation to explicitly minimise label probability, while GGLT makes non-directional changes to the latent representation. OLT can therefore also give directional feature change i.e., features change to explicitly minimise label probability, which may be observed as disappearing disease features. This motivates my use of OLTs despite previous proven success with GGLTs.

Moreover, from this perspective OLTs are akin to counterfactual generation, which has been popularly applied to explainable medical image prediction. Counterfactual generation describes the process in which generative models are used to change specific features of an input while keeping others fixed, in order to generate an alternative scenario that would result in a different prediction. By generating these alternative scenarios, it is possible to better understand which features are most influential in the model's decision-making process.

Sun et al. (2023) present a counterfactual generation method specifically for explainable prediction in a multi-label setting. In their framework, they use a generative adversarial network (GAN) to first generate class-specific attribution maps based on counterfactuals, and then simple logistic regression classifier is used to make predictions based solely on these attribution maps. They present convincing qualitative examples, but do not quantitatively evaluate against ground truth radiologist annotations. Due to their adversarial nature, GANs can be difficult and time consuming to train. Cohen et al. (2021) instead take a simple autoencoder approach to generating counterfactual explanations for CXRs. They train an autoencoder and classifier independently, they generate a latent representation, which is then perturbed to cause a change in the classifier predictions. Image samples are produced from the perturbed representations and optical flow is computed on a sequence of generated images to visualise feature changes. They summarise counterfactual changes and compare highlighted regions against ground truth radiologist annotation masks. While, counterfactual image generation in medical imaging is well established, existing approaches do not have a capacity for protecting against shortcut learning. Through my evaluations I aim to demonstrate the natural advantages of OLTs for identifying 'shortcuts' in this framework.

For GGLTs and OLTs, I summarise traversal changes by finding the absolute pixel-wise differences between the initial reconstruction and the post-traversal reconstruction. I consider explanations of binary predictions, where only a single class is evaluated, as well as explanations of multi-label predictions, where the predictive features of all positive predictions in a CXR are interrogated. I use these methods to explain model predictions and compare resulting saliency maps to identify which methods, VAEs or HVAEs, and which priors, Sparse prior or Gaussian prior, offer the clearest explanations. I compare model prediction performance against that of a multi-label deep CNN. Traversal-based saliency maps (derived from OLTs or GGLTs) are also compared against conventional post-hoc explainability modules such as, GradCAM++, which is applied to a deep CNN classifier. To generate visual explanations of multi-label deep CNN predictions, I binarise GradCAM++ saliency maps and quantitively evaluate all disease feature localisation through comparison of resulting bounding boxes with ground truth radiologist annotations.

Algorithm 2 Optimised Latent Traversals

```
Require: Learned posterior distribution $q(z|x)$, set of independent classifiers $\{C_1, C_2, \dots, C_K\}$ (one per class), classification loss function $\mathcal{L}_{\text{class}}$, learning rate $\eta$, input $x$, number of optimisation steps $N$
Ensure: Optimised latent representations $\{z_1^*, z_2^*, \dots, z_K^*\}$ for each class
Initialise: Sample initial latent representation $z \sim q(z|x)$
for each class $k$ in $\{1, 2, \dots, K\}$ do
    Set $z_k \leftarrow z$ (initialise latent representation for class $k$)
    for $t = 1$ to $N$ do
        Compute prediction for class $k$: $\hat{y}_k \leftarrow C_k(z_k)$
        Evaluate classification loss: $\mathcal{L}_k \leftarrow \mathcal{L}_{\text{class}}(\hat{y}_k, y_{\text{target},k})$
        Compute gradient of the loss w.r.t. $z_k$: $\nabla_{z_k} \mathcal{L}_k$
        Update latent representation for class $k$: $z_k \leftarrow z_k - \eta \cdot \nabla_{z_k} \mathcal{L}_k$
    end for
end for
Return: Optimised latent representations $\{z_1^*, z_2^*, \dots, z_K^*\}$
```

Bounded generative model For HVAEs, to visualise the meaningful changes that are observed by the classifier modules, I use bounded variation on the standard ELBO which removes the influence of lower level latent variables (Maaløe et al., 2019). Initially devised to remove the influence of data in the lower-levels of the HVAE for improved out-of-distribution detection, the bounded ELBO term restricts data-dependent information flow through the TD information path of HVAEs (Maaløe et al., 2019). I adapt the use of a bounded ELBO term for improved localisation of disease features in multi-task HVAEs.

The flow of information through the hierarchy of latent variables in the generative model of BIVA makes evaluation through latent traversal more complex. The lower layers, which receive deterministic bottom up information, dilute the impact

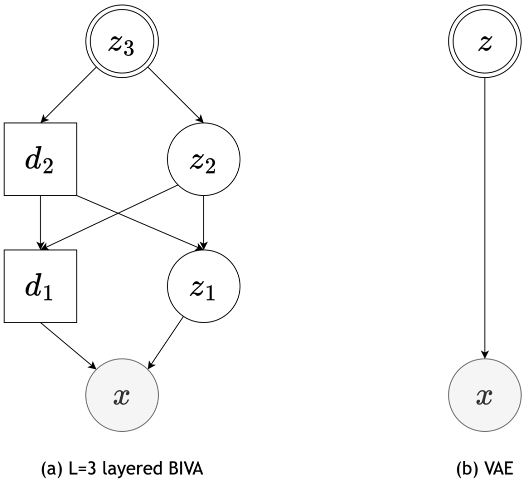


FIGURE 7.3: **A comparison of a (a) L=3 layered BIVA and (b) single-level VAE generative models.**

of top level changes on the reconstructed image (Fig. 7.3). To minimise the impact of BU information at image generation I place a bound on the generative model.

Consider the generative model of a $L = 3$ layered BIVA,

$$p_\theta(x, z) = p_\theta(x|z)p_\theta(z_L)\prod_{i=1}^{L-1} p_\theta(z_i^{BU}|z_{>i})p_\theta(z_i^{TD}|z_>i)$$

where θ are the parameters of the generative model. The likelihood $p_\theta(x|z)$ depends on z_1 and $z_{>1}$ through the deterministic TD path. As shown in Figure 7.3a, Information flows from z_L to z_2 and z_1, which additionally received information from deterministic variable d_2.

By placing a bound on the generative model I sample the k lowest latent variables from the conditional prior $z_1, ..., z_L \sim p_\theta(z_{\leq k}|z_{>k})$ and only the $L > k$ highest level from the approximate posterior $z_{k+1}, ..., z_L \sim q_\theta(z_{>k}|x)$. With this I can evaluate reconstructions x from each latent variable. To maximise the effect of changes in z_L (the traversed latent variable) on the image reconstruction, I set $k = L - 1$. Figure 7.4 shows the impact of bounding the generative model on information flow to the data likelihood $p(x|z)$ i.e., the reconstruction.

The logistic regression classifiers are trained on only the top-level latent variables. Similarly, these are the only variables changed during latent traversal. By eliminating the influence of the lower level stochastic variables, the image reconstructions offer a more direct visual representation of the changes observed by the classifier.

7.4 Application: Explainable Prediction of Multi-label CXRs

I apply models to two examples of populations of multi-label CXRs, CheXpert and VinDr-CXR. Both datasets are described in Chapter 3. When applying VAE and

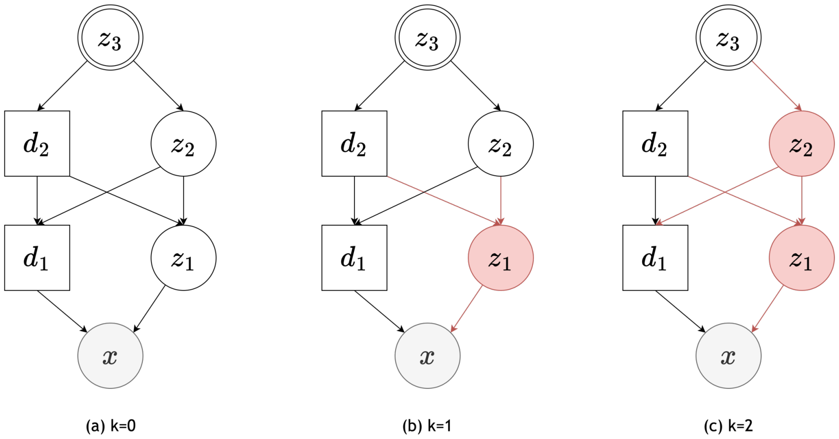


FIGURE 7.4: **Examples of bounded BIVA generative models of $L = 3$ layered BIVA.** Red nodes indicate where the conditional prior is used and k defines the bounding level.

HVAE methods to CheXpert I consider only 4 of the 14 classes labelled in the dataset these are, *No Finding*, *Lung Opacity*, *Pleural Effusion*, and *Support Devices*, which I select for their salient features and prevalence in the dataset. Figure 7.5 presents the co-occurrence matrix of the considered labels in the training set. I observe that the highest rate of co-occurrence is between the *Lung Opacity* class and Support Devices class. Models trained on CheXpert are also evaluated on the CheXlocalise dataset for their explainable prediction performance. CheXlocalise is a subset of the CheXpert dataset with radiologist annotations in the form of pixel-wise class annotation. As a localisation dataset, the CheXlocalise dataset does not include *No Finding* CXRs. I therefore evaluate this dataset for only *Airspace Opacity* (which corresponds to the Lung Opacity class in CheXpert), *Pleural Effusion*, and *Support Devices* labels.

When applying methods to VinDr-CXR, I consider the following seven labels: *Pleural Thickening*, *Lung Opacity*, *No Finding*, *Other Lesion*, *Pleural Effusion*, *Cardiomegaly*, *Aortic Enlargement*. Figure 7.6 presents the co-occurrence of classes in the training dataset, showing the highest frequency co-occurrence is between *Cardiomegaly* and *Aortic Enlargement* classes, I also observe high frequency co-occurrence between *Pleural Thickening* and *Aortic Enlargement* classes.

For my experiments BIVA models have $L = 3$ stochastic layers (depicted in Figure 7.1). For each stochastic level l_i I define a deterministic level D_i which connects the stochastic layers. Deterministic levels are made up of 3 ResNet blocks, which produce the feature maps for the corresponding stochastic level. Each block comprises convolutional layers, ReLU activation functions, and weight normalisation, with residual connections.

Deterministic blocks D_1 and D_2 are defined by 64x5x5 (number of kernels x kernel width x kernel height) convolutional layers and an overall stride of 2. D_3 is defined by 64x3x3 convolutional layers with an overall stride of 2. Stochastic latent variables in L_1 and L_2 are convolutionally connected layers of dimension

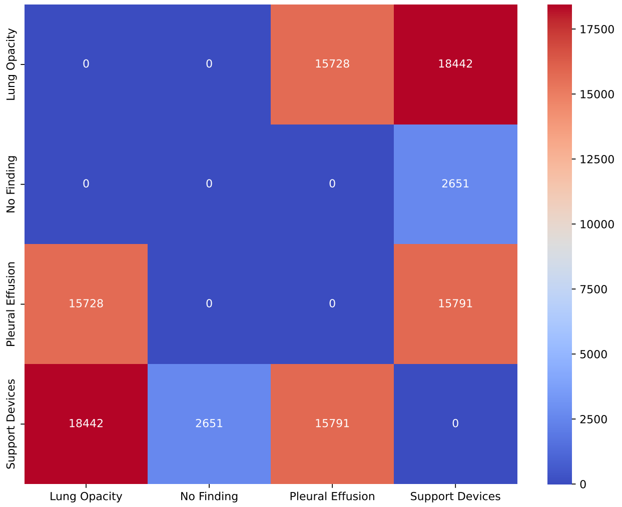


FIGURE 7.5: **Heatmap of label co-occurrence in CheXpert training data.**

32x120x120 and 16x30x30. Stochastic latent variables in l_3 are densely connected with dimensions of 1024.

For my experiments VAE models have a single stochastic layer with a dimensionality of 1024, which is densely connected to the deterministic encoder and decoder functions. The encoder function comprises 3 blocks of convolutional layers, ReLU activation functions, and weight normalisations, with residual connections. All convolutional layers are defined by 64x5x5 with strides of 2.

Hyperparameter	**VAE**	**BIVA**
Optimizer	Adam	Adamax
Learning rate	1e-3	5e-4
Epochs	1500	3000
WU epochs	500	1500
λ	0.01	0.01

TABLE 7.1: **Experimental settings for optimisation of VAE and 3-layer BIVA models. λ is the weighting factor for the classification term.**

Each of the label-specific logistic regression classifiers are optimised for their respective binary classification tasks with minimisation of a multi-label BCE loss, which computes independent losses for each label based on their respective ground-truth values. VAE models and attached classifiers are trained with a batch size of 128 using the Adam optimisation algorithm.

For BIVA models, only the top-level latent representation is used as input to the classifiers. For training on CheXpert data, I use 4 logistic regression classifiers and for the VinDr-CXR task I use 7 logistic regression classifiers.

With the same multi-task optimisation task as in Chapter 6, I apply a similar training strategy. The reconstruction loss, KL divergence loss and classifier losses

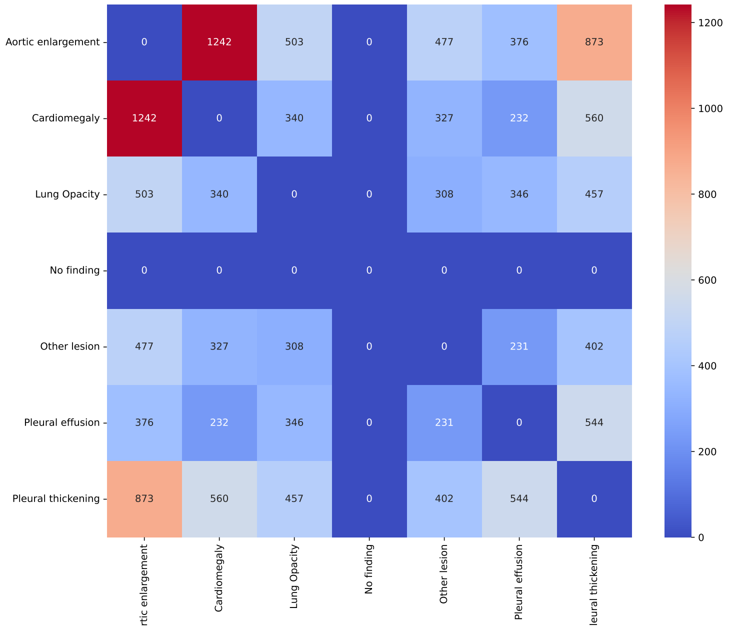


FIGURE 7.6: **Heatmap of disease co-occurrence in VinDr-CXR training data, with raw counts of class co-occurrence.**

are combined into a single training objective and minimised. Instead of applying a multi-stage optimisation strategy (as in Chapter 6), the BIVA and logistic regression classifiers are trained jointly from the start. The classifier losses and KL divergence loss are slowly annealed, this prevents latent variable collapse. KL divergence loss annealing is a commonly used strategy in HVAEs, and is referred to as deterministic warm up (WU) (Maaløe et al., 2019).

Through combined losses, model optimisation is influenced by the classification task, and I can re-express the ELBO term to reflect this,

$$\text{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - \text{KL}[q(z|x)||p(z)] + \lambda \mathbb{E}_{q(z|x)}[\log q(y|x,z)]$$

where y represents the class labels in the auxiliary classification task, and λ is a weighting factor that controls the importance of the classification term relative to the reconstruction and regularisation terms.

BIVA models and attached classifiers are trained with a batch size of 28 and optimised using Adamax as in Maaløe et al. (2019). I train BIVA models for 3000 epochs, with deterministic warm up over 1500 epochs i.e., increasing the weighting of the KL divergence term. The classification loss is included in this warm up. Additional details of BIVA and VAE model training settings are included in Table 7.1.

As seen in Chapter 6, optimising the latent space with the multi-label classification task encourages the latent factors to explain class-specific visual features and allows the classifiers to predict target classes from the learned latent space. This enhances the 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. To this end I select simple linear logistic regression functions as the auxiliary classifiers. I theorise that by restricting the classification model, I reduce the risk of overfitting and the posterior distribution will be greater influenced to minimise the classification objective.

For both GGLT and OLT implementations, hyperparameters were selected through qualitative evaluation of visual explanations. For OLTs I use an Adam optimizer with learning rates of 0.0001 for Gaussian and Sparse-prior models, and 0.00001 for Dirichlet-prior models, optimizing over 2000 iterations. In GGLTs I traverse the single most significant latent dimension (determined by classifier gradient) with step sizes of 0.0001 for Gaussian/Sparse models and 0.00001 for Dirichlet models, over 2000 steps. These hyperparameters were applied consistently across both VAE and BIVA architectures.

For comparison with deep CNN methods I adapt a ResNet-18 architecture, which is characterised by its residual learning framework (He et al., 2015). The network comprises 18 layers with a combination of convolutional, batch normalization, and ReLU activation layers interspersed with residual blocks. I use a pretrained ResNet-18 model initialised with ImageNet weights. The fully connected layer of the original model was replaced to align with the multi-label classification objective i.e., to give the correct number of output nodes for the classification task. I apply a sigmoid activation function to each output. I include a dropout layer ($p = 0.2$) for model regularisation purposes. I trained this model using a multi-label BCE loss function until validation loss converged. For training I used the Adam optimiser with a learning rate of 0.0001.

Under the hypothesis that "sparseness" leads to decomposed latent representations, I evaluate the sparsity of the posterior distribution over latent variables by applying the Hoyer metric and Gini index to the sampled latent representation. The Hoyer metric is a normalised measure of sparsity and is defined as:

$$H(x) = \frac{\sqrt{n} - \frac{\|x\|_1}{\|x\|_2}}{\sqrt{n} - 1},$$

where x is the sampled latent representation under evaluation, n is its dimensionality. This metric ranges from 0, which indicates a uniform distribution of values i.e., no sparsity, to 1, where only one element is non-zero i.e., maximum sparsity.

The Gini index, a standard measure of inequality, is computed as:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n \sum_{i=1}^{n} x_i},$$

where x_i and x_j are individual elements of the posterior distribution and n is the total number of elements. The Gini index ranges from 0 (perfect equality where all values are identical) to 1 (maximum inequality where all values are concentrated in a single element). For my experiments, the Gini index and Hoyer metric were applied to the sampled latent space of VAE and BIVA models.

I train and evaluate single-level VAEs on CheXpert and CheXlocalise data. I do the same for BIVA models, with the addition of training and evaluating on the VinDr-CXR dataset. I evaluate per-label binary model prediction performance through standard classification metrics e.g., AUROC, precision, recall, and F1 score. I also evaluate prediction performance with multi-label metrics such as, Hamming score and Exact Match Ratio (EMR). To assess quality of visual explanations, I manually review latent traversal-derived explanations as well as GradCAM++ saliency maps (applied to deep CNN predictions). I also evaluate visual explanations by applying a disease localisation task in which I compare explanations against radiologist annotations. To measure this I use object detection and segmentation metrics, intersect-over-union (IoU), hit-or-miss (HoM), and Dice scores to quantitatively evaluate how well visual explanations isolate disease features. The HoM metric is calculated as the proportion of 'hits' among all predicted bounding boxes. I define a 'hit' as where the predicted bounding box and ground truth bounding box overlap.

7.4.1 Results

7.4.1.1 Model fit

Comparison of Gaussian-prior VAE and Gaussian-prior BIVA reconstructions (Fig. 7.7a and Fig. 7.9b) shows that, as expected, BIVA models give much better image reconstruction quality. The same is observed for Sparse-prior models (Fig. 7.8a and Fig. 7.9b).

Figure 7.7, 7.8, and 7.9 show that BIVA models also give higher quality image generation, which is quantified by the likelihoods $\log p(x)$ reported in Table 7.4. I evaluate images generated by bounded BIVA models to verify that top-level latent representations have not collapsed to the prior. Figure 7.8b ($k = 2$) and Figure 7.7b ($k = 2$) shows that Gaussian-prior and Sparse-prior BIVA learn meaningful top-level posterior distributions i.e., $k = 2$ bounded reconstructions resemble the input data. This demonstrates that the latent representations seen by the classifiers carry meaningful data for the multi-label prediction task.

Table 7.3 presents measures of the degree of sparsity of VAE and BIVA models latent representations. As expected, according to Hoyer and Gini Index metrics, the Sparse-prior VAE and Sparse-prior BIVA models learn a more sparse posterior distribution than their Gaussian-prior model counterparts. Against expectation, the latent representation of the Dirichlet-prior VAE (parametrised for extreme sparsity) is less sparse than the Sparse-prior VAE (according to these metrics).

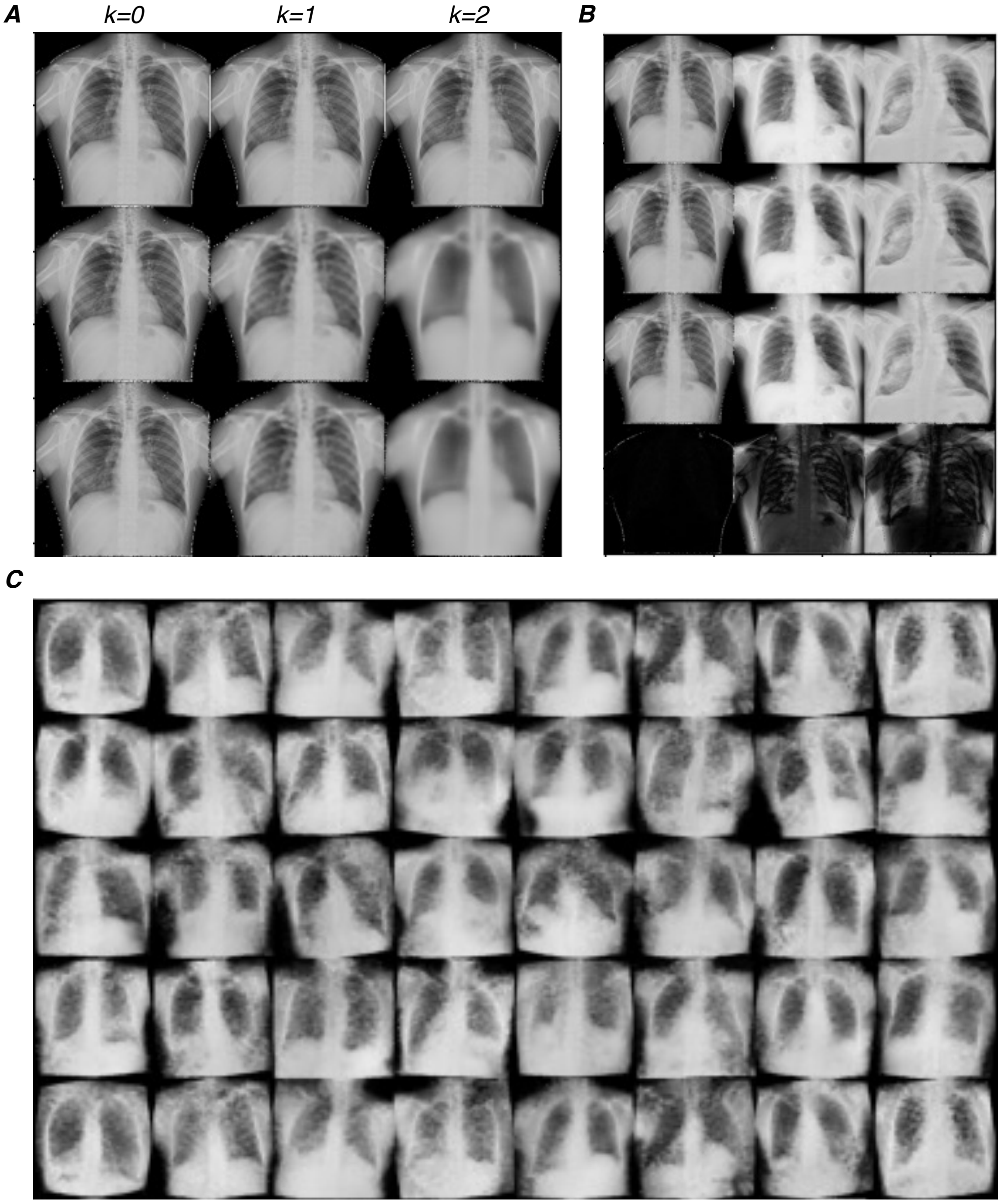


FIGURE 7.7: **Gaussian-prior BIVA model fit.** Model fit is described by (a) Bounded model image generation, where k is the bounding level i.e, $\mathcal{L} > k$, (b) image reconstruction, first row is the input image, second row is the image reconstruction, the third row is the image reconstruction from the mean and, (c) images generated from a random sample of the prior distribution.

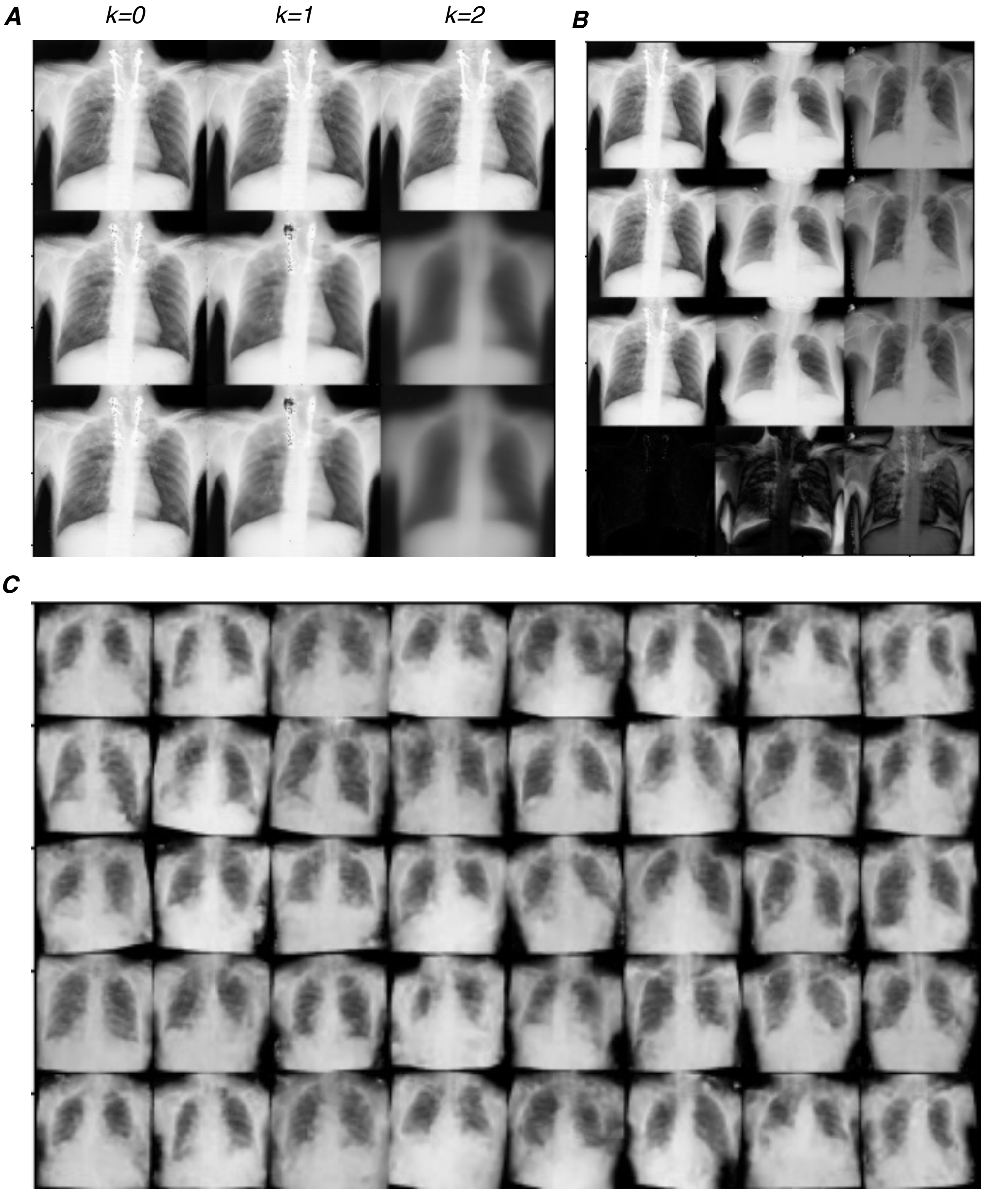


FIGURE 7.8: **Sparse-prior BIVA model fit.** Model fit is described by (a) bounded model image generation, where k is the bounding level i.e, $\mathcal{L} > k$, (b) image reconstruction, first row is the input image, second row is the image reconstruction, the third row is the image reconstruction from the mean and, (c) images generated from a random sample of the prior distribution.

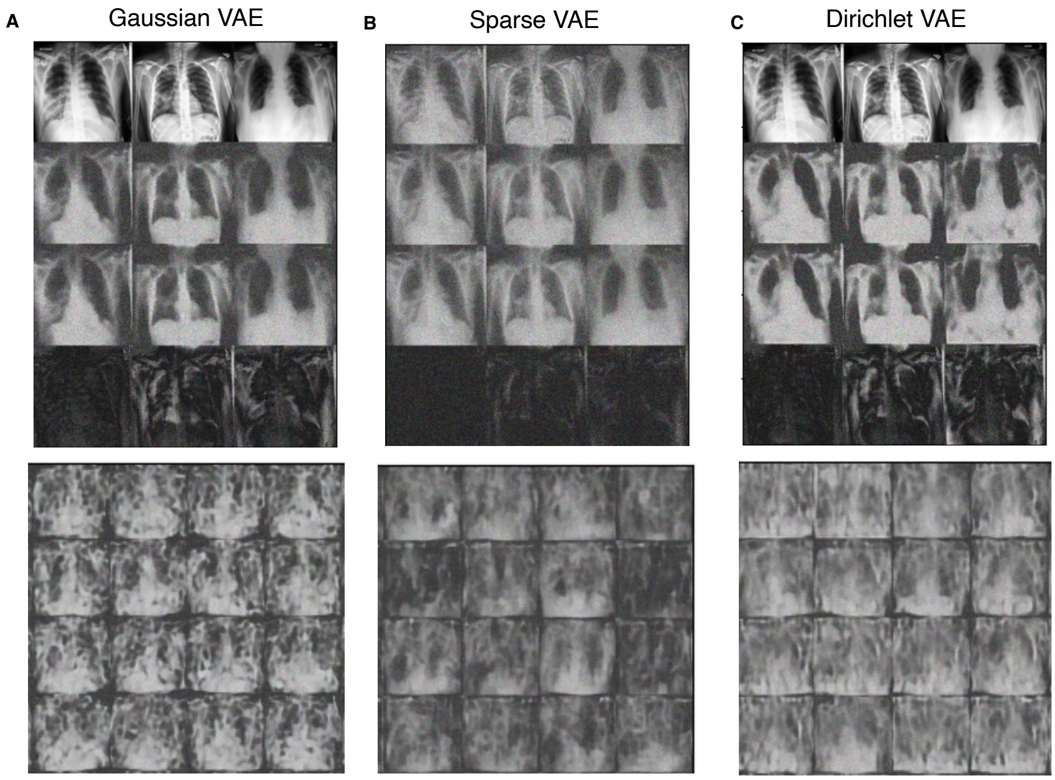


FIGURE 7.9: **(A) Gaussian-prior VAE, (B) Sparse-prior VAE, and (C) Dirichlet-prior VAE model fit**. Top images are image reconstructions: first row is the input image, second row is the image reconstruction, the third row is the image reconstruction from the mean. Bottom images are samples generated from a random sample of the prior distribution.

Metric	**Sparse BIVA**	**Gaussian BIVA**	**Dirichlet VAE**	**Gaussian VAE**	**Sparse VAE**
KL $q(z\|x), p(z)$	15964	31032	4569	695	1102
$\log p(x)$	-194261	-181961	-866960	-871128	-863280
$\log p(x\|z)$	-178296	-150929	-866960	-870432	-862177

TABLE 7.2: **Model fit metrics for VAE and BIVA models.**

Model	**Prior Distribution**	**Hoyer Metric**	**Gini Index**
VAE	Sparse-prior	0.30	0.54
	Gaussian-prior	0.25	0.45
	Dirichlet-prior	0.26	0.48
BIVA	Sparse-prior	0.28	0.53
	Gaussian-prior	0.21	0.42

TABLE 7.3: **Sparsity measure of VAE and BIVA model posteriors.** Sparsity is measured in the top stochastic level of BIVA models

.

7.4.1.2 Multi-label prediction

I compare the predictive performance of the Gaussian-prior, Sparse-prior, and Dirichlet-prior VAE on CheXpert data. Label-wise classification metrics show that VAE models perform similarly on CheXpert data regardless of prior distribution (Table 7.4).

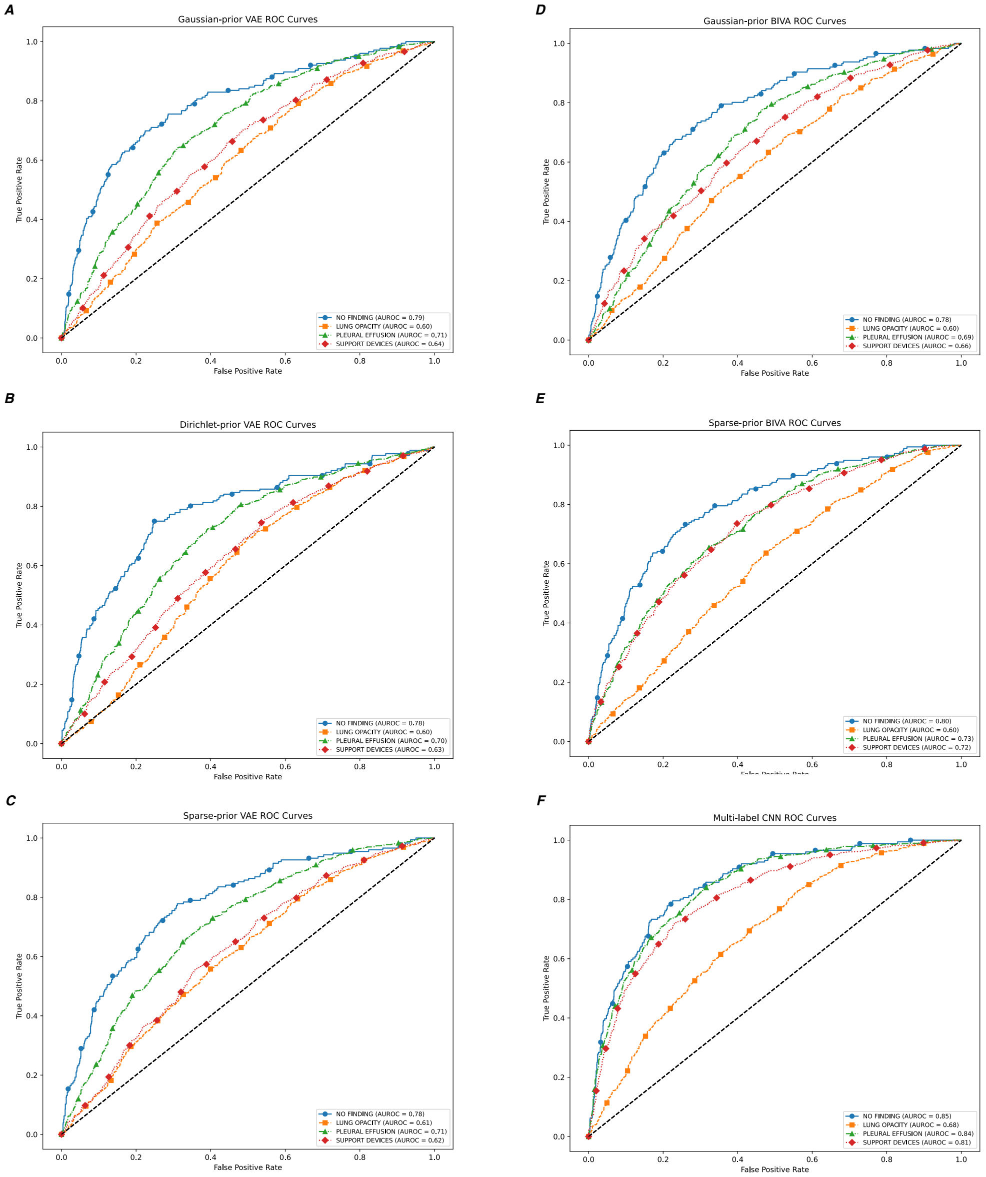


FIGURE 7.10: **ROC curves of (A) Gaussian-prior VAE, (B) Dirichlet-prior VAE, (C) Sparse-prior VAE, (D) Gaussian-prior BIVA, (E) Sparse-prior BIVA, and (F) Multi-label deep CNN predictions on CheXpert data.** ***Abbrvs: Receiver Operator Characteristic (ROC).***

Best performance is recorded for *No Finding* CXRs, with AUROCs ranging from 0.78 to 0.79. VAEs perform the worst at predicting *Lung Opacity* CXRs, with AUROCs between 0.60 and 0.61. Evaluation of multi-label performance metrics shows that the Sparse-prior VAE achieved best performance according to multi-label performance metrics, with an EMR of 0.18 and a Hamming loss of 0.54. While the Gaussian-prior VAE and Dirichlet-prior VAE performed marginally worse, both giving EMR scores of 0.17 and Hamming losses of 0.48.

Surprisingly, the predictive performance of Sparse-prior and Gaussian-prior BIVA models does not far exceed their VAE counterparts. Gaussian-prior BIVA

Model	Label	AUROC	Precision	Recall	F1-Score	F1-Threshold
Sparse-prior VAE	No Finding *(n=176)*	0.78	0.32	0.45	0.37	0.20
	Lung Opacity *(n=934)*	0.61	0.52	0.93	0.67	0.33
	Pleural Effusion *(n=761)*	0.71	0.52	0.80	0.63	0.36
	Support Devices *(n=1089)*	0.62	0.60	0.95	0.73	0.32
Gaussian-prior VAE	No Finding *(n=176)*	0.79	0.31	0.53	0.39	0.17
	Lung Opacity *(n=934)*	0.61	0.51	0.96	0.67	0.28
	Pleural Effusion *(n=761)*	0.72	0.50	0.86	0.63	0.27
	Support Devices *(n=1089)*	0.64	0.58	0.98	0.73	0.27
Dirichlet-prior VAE	No Finding *(n=176)*	0.78	0.29	0.53	0.38	0.21
	Lung Opacity *(n=934)*	0.60	0.52	0.93	0.66	0.28
	Pleural Effusion *(n=761)*	0.70	0.52	0.83	0.64	0.16
	Support Devices *(n=1089)*	0.64	0.58	0.97	0.73	0.15
Sparse-prior BIVA	No Finding*(n=176)*	0.80	0.31	0.52	0.39	0.27
	Lung Opacity *(n=934)*	0.60	0.51	0.97	0.67	0.16
	Pleural Effusion *(n=761)*	0.73	0.51	0.86	0.64	0.26
	Support Devices *(n=1089)*	0.72	0.64	0.91	0.75	0.39
Gaussian-prior BIVA	No Finding *(n=176)*	0.78	0.28	0.55	0.38	0.73
	Lung Opacity *(n=934)*	0.59	0.50	0.96	0.66	0.25
	Pleural Effusion *(n=761)*	0.69	0.51	0.82	0.63	0.73
	Support Devices *(n=1089)*	0.66	0.60	0.94	0.73	0.39
Multi-label deep CNN	No Finding *(n=176)*	0.85	0.42	0.49	0.45	0.21
	Lung Opacity *(n=934)*	0.68	0.56	0.92	0.70	0.31
	Pleural Effusion *(n=761)*	0.84	0.64	0.84	0.73	0.33
	Support Devices *(n=1089)*	0.81	0.71	0.89	0.79	0.29

TABLE 7.4: **Label-wise performance metrics on CheXpert for VAE models, BIVA models, and the multi-label deep CNN.**

Model	Label	AUROC	Precision	Recall	F1-Score	F1-Threshold
Dirichlet-prior VAE	Airspace Opacity *(n=116)*	0.63	0.68	1.00	0.81	0.01
	Pleural Effusion *(n=64)*	0.69	0.53	0.75	0.62	0.33
	Support Devices *(n=99)*	0.49	0.58	1.00	0.73	0.00
Gaussian-prior VAE	Airspace Opacity *(n=116)*	0.62	0.68	1.00	0.81	0.00
	Pleural Effusion *(n=64)*	0.65	0.56	0.66	0.60	0.34
	Support Devices *(n=99)*	0.54	0.60	0.99	0.99	0.75
Sparse-prior VAE	Airspace Opacity *(n=116)*	0.64	0.70	0.98	0.81	0.13
	Pleural Effusion *(n=64)*	0.69	0.50	0.84	0.44	0.40
	Support Devices *(n=99)*	0.57	0.58	1.00	0.73	0.00
Gaussian-prior BIVA	Airspace Opacity *(n=116)*	0.72	0.77	0.87	0.82	0.38
	Pleural Effusion (n=64)	0.68	0.49	0.80	0.60	0.62
	Support Devices *(n=99)*	0.59	0.59	0.99	0.74	0.18
Sparse-prior BIVA	Airspace Opacity *(n=116)*	0.74	0.693	0.99	0.82	0.13
	Pleural Effusion *(n=64)*	0.72	0.52	0.89	0.66	0.28
	Support Devices *(n=99)*	0.66	0.63	0.90	0.74	0.44
Multi-label deep CNN	Airspace Opacity *(n=116)*	0.815	0.831	0.888	0.858	0.29
	Pleural Effusion *(n=64)*	0.883	0.807	0.719	0.760	0.38
	Support Devices *(n=99)*	0.778	0.721	0.889	0.796	0.31

TABLE 7.5: **Label-wise performance metrics on CheXlocalise for Dirichlet-prior VAE, Gaussian-prior VAE, and Sparse-prior VAE models.**

	Label	**AUROC**	**Precision**	**Recall**	**F1-Score**	**F1-Threshold**
Gaussian-prior BIVA	Pleural thickening *(n=43)*	0.84	0.26	0.47	0.33	0.50
	Lung Opacity *(n=18)*	0.81	0.11	0.44	0.17	0.36
	No finding *(n=753)*	0.90	0.87	0.96	0.92	0.09
	Other lesion *(n=11)*	0.74	0.10	0.27	0.15	0.41
	Pleural effusion *(n=10)*	0.91	0.20	0.40	0.27	0.45
	Cardiomegaly *(n=133)*	0.91	0.72	0.61	0.66	0.56
	Aortic enlargement *(n=192)*	0.92	0.78	0.65	0.71	0.61
Sparse-prior BIVA	Pleural thickening *(n=43)*	0.80	0.22	0.44	0.29	0.64
	Lung Opacity *(n=18)*	0.81	0.14	0.17	0.15	0.71
	No finding (n=753)	0.91	0.86	0.97	0.91	0.04
	Other lesion (n=11)	0.85	0.17	0.18	0.17	0.73
	Pleural effusion (n=10)	0.94	0.14	0.60	0.23	0.55
	Cardiomegaly (n=133)	0.92	0.67	0.67	0.67	0.79
	Aortic enlargement (n=192)	0.92	0.64	0.78	0.70	0.65
Multi-label deep CNN	Pleural thickening *(n=43)*	0.86	0.34	0.26	0.31	0.64
	Lung Opacity *(n=18)*	0.86	0.40	0.22	0.29	0.72
	No finding *(n=753)*	0.96	0.95	0.95	0.95	0.11
	Other lesion *(n=11)*	0.84	0.12	0.18	0.14	0.56
	Pleural effusion *(n=10)*	0.99	0.91	1.00	0.95	0.59
	Cardiomegaly *(n=133)*	0.98	0.84	0.77	0.80	0.86
	Aortic enlargement *(n=192)*	0.98	0.81	0.82	0.81	0.76

TABLE 7.6: **Multi-label prediction performance of BIVA models and multi-label deep CNN on VinDr-CXR test data.**

multi-label prediction metrics are indeed the same as the Gaussian-prior VAE, predicting CheXpert multi-label CXRs with an EMR of 0.17 and Hamming score of 0.48. Label-wise metrics show that, like the Gaussian VAE, Gaussian BIVA performs best at predicting *No Finding* CXRs (0.78 AUROC) and gives worst performance on *Lung Opacity* CXRs (0.59 AUROC). Sparse-prior BIVA marginally improves on this predicting *No Finding* with an AUROC of 0.80 and performs much better than the Gaussian BIVA at predicting *Pleural Effusion* and *Support Devices* classes. *Pleural Effusion* AUROC improves from 0.69 to 0.73 and *Support Devices* AUROC increases from 0.66 to 0.72. Both models are significantly outperformed by the multi-label deep CNN (Fig. 7.10F). The multi-label CNN demonstrates improved performance across all CheXpert labels, achieving an AUROC of 0.85 for *No Finding* CXRs and 0.68 for *Lung Opacity* CXRs (Table 7.4).

I compare the predictive performance of the Sparse-prior BIVA and Gaussian-prior BIVA on VinDr-CXR. Both models perform similarly in VinDr-CXR data (Fig. 7.11), achieving strong label-wise prediction performance. Gaussian-prior BIVA predicts *Pleural Effusion*, *Cardiomegaly* and *Aortic Enlargement* CXRs the best, with AUROCs of 0.91, 0.91 and 0.92, respectively. Sparse-prior BIVA performance exceeds that of Gaussian BIVA, with AUROCs of 0.94, 0.92 and 0.92 for the compared classes. Moreover, Sparse-prior BIVA performs notably better at predicting *Other Lesion* CXRs (Gaussian-prior BIVA AUROC of 0.74; Sparse-prior BIVA AUROC of 0.85). However, again, the deep CNN performs better, with near perfect performance in *Pleural Effusion*, *Cardiomegaly* and *Aortic Enlargement* classes, achieving AUROCs ranging from 0.98 to 0.99 (Table 7.6).

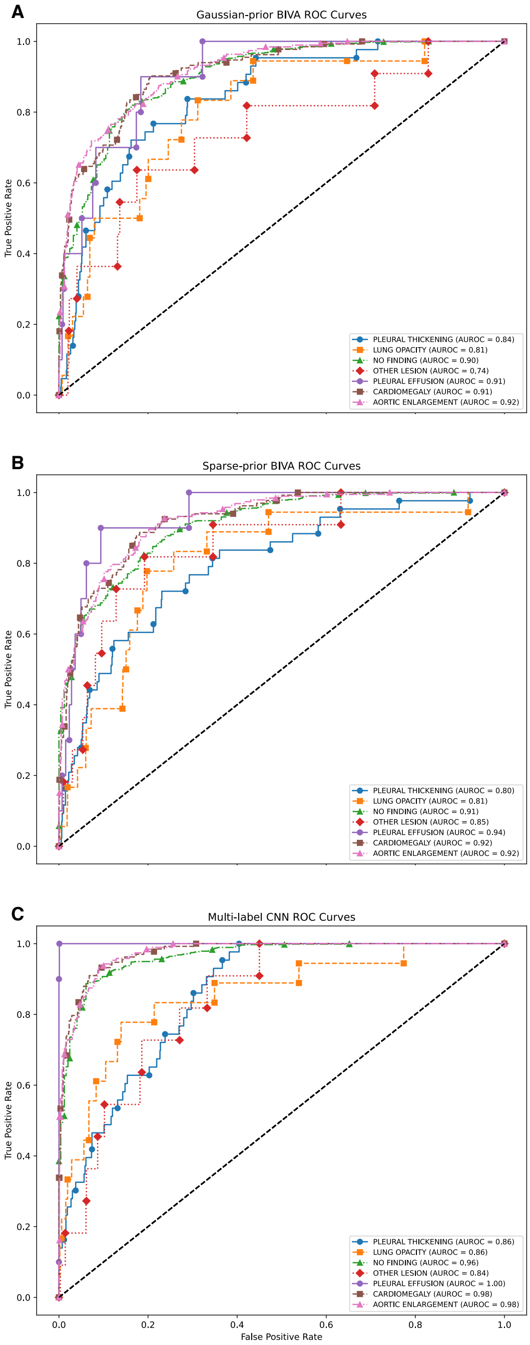


FIGURE 7.11: **ROC curves of Gaussian-prior BIVA, Sparse-prior BIVA and multi-label deep CNN on VinDr-CXR data.** *Abbrvs: Receiver Operator Characteristic (ROC).*

7.4.1.3 Visual explanations

Here I report on the quality of visual explanations generated by the two proposed methods for latent traversals, GGLT and OLT (as applied to VAE and BIVA models). I evaluate the capacity of these methods for explaining multi-label prediction and compare against the popular post-hoc visual explainer, GradCAM++, applied to the deep CNN. I perform qualitative and quantitative evaluation on radiologist annotated datasets, reporting hit-or-miss localisation metrics and IoU metrics for VinDr-CXR and CheXlocalise evaluations. For CheXlocalise evaluations, which provides pixel-wise ground truth masks, I additionally evaluate Dice score.

	Prior	**Metric**	**Airspace Opacity**	**Pleural Effusion**	**Support Devices**	**Average**
OLT	Dirichlet VAE	IoU	0.00	0.00	0.00	0.00
		Dice	0.24	0.09	0.22	0.18
		HoM	0.15	0.14	0.36	0.22
	Sparse VAE	IoU	0.01	0.01	0.01	0.01
		Dice	0.25	0.13	0.17	0.18
		HoM	0.17	0.17	0.35	0.23
	Gaussian VAE	IoU	0.00	0.00	0.01	0.00
		HoM	0.14	0.13	0.40	0.22
		Dice	0.14	0.06	0.22	0.14
	Sparse BIVA	IoU	0.01	0.00	0.01	0.01
		HoM	0.11	0.07	0.24	0.14
		Dice	0.19	0.06	0.13	0.13
	Gaussian BIVA	IoU	0.01	0.00	0.01	0.01
		HoM	0.28	0.25	0.42	0.32
		Dice	0.08	0.01	0.09	0.06
GGLT	Dirichlet VAE	IoU	0.01	0.00	0.00	0.00
		HoM	0.16	0.16	0.37	0.23
		Dice	0.26	0.10	0.22	0.19
	Sparse VAE	IoU	0.01	0.01	0.01	0.01
		HoM	0.15	0.12	0.34	0.20
		Dice	0.25	0.09	0.19	0.18
	Gaussian VAE	IoU	0.00	0.00	0.00	0.00
		HoM	0.14	0.15	0.36	0.22
		Dice	0.13	0.06	0.19	0.13
	Sparse BIVA	IoU	0.01	0.01	0.00	0.01
		HoM	0.10	0.05	0.23	0.13
		Dice	0.21	0.04	0.12	0.12
	Gaussian BIVA	IoU	0.00	0.00	0.01	0.00
		HoM	0.23	0.28	0.42	0.31
		Dice	0.10	0.02	0.08	0.07

TABLE 7.7: **VAE localisation of CheXlocalise disease features by OLTs and GGLTs methods across different labels.** *Abbrvs: Optimised Latent Traversal (OLT); Gradient Guided Latent Traversal (GGLT); Hit-or-Miss (HoM).*

	Metric	**Aortic Enlargement**	**Pleural Thickening**	**Cardiomegaly**	**Average**
Sparse-prior BIVA	*HoM*	0.11	0.39	0.17	0.22
	IoU	0.01	0.00	0.01	0.01
Gaussian-prior BIVA	*HoM*	0.07	0.00	0.09	0.05
	IoU	0.01	0.00	0.00	0.00
Multi-label deep CNN	*HoM*	0.98	0.88	1.00	0.95
	IoU	0.10	0.21	0.21	0.17

TABLE 7.8: **Localisation performance of Sparse-prior BIVA and Gaussian-prior BIVA GGLTs on VinDr-CXR data compared to localisation of multi-label deep CNN GradCAM++.** Classes with low prevalence in VinDr-CXR test data are omitted. *Abbrvs: Gradient-guided latent traversals (GGLTs)*

VAE explanations Figure 7.12 presents multi-label CheXlocalise examples of disease feature localisation using the GGLT method. While the CXR reconstructions generated by VAE methods are generally of low quality and difficult to interpret directly, with bounding boxes generated by post-processing feature changes over the traversal can be observed. The Dirichlet-prior VAE GGLTs produce feature changes relating to the diseases of interest. However, these changes are not fully isolated, examples show additional changes outside disease-relevant areas. For instance, Figure 7.13C shows a distinct pattern of bounding box activations that highlights changes related to the *Support Device* (label of interest) in the left lobe (right of image), with false positive bounding boxes predicted over the pleural edge. Figure 7.13A gives an *Airspace Opacity* example, this shows disease-specific changes, particularly around the heart margin, with additional false positive bounding boxes predicted over the diaphragm and pleural edge. There is a similar pattern of disease feature localisation with false positives in Sparse VAE GGLTs (Fig. 7.14). There is a general trend of limited disease feature localisation in Gaussian VAE GGLTs. Figure 7.12A gives a visual explanation of an *Airspace Opacity* prediction where bounding boxes are largely localised outside the lung field and no significant morphological changes can be viewed in the disease area of interest in the traversal reconstructions. The same observation can be made for Figure 7.12D, a Support Device example in which the features of the support device i.e., the intubation tube, are ignored.

Despite qualitative evaluation showing that the Sparse VAE and Dirichlet VAE GGLTs give meaningful clinical explanations with some degree of disease feature localisation, metrics suggest overall performance of VAE GGLTs is poor (Table 7.7). GGLT explanations give low IoU scores, with all model scores at zero or near-zero for all classes, although HoM and Dice scores are comparatively much better, with scores averaged over all classes ranging from 0.20-0.23 and 0.13-0.0.19, respectively.

Comparison of Dice scores suggests that Dirichlet VAE and Sparse VAE GGLT explanations localise disease features better than Gaussian VAE GGLTs; this is most apparent in the *Airspace Opacity* class in which Dirichlet VAE and Sparse VAE explanations identify disease features with a Dice score of 0.26 and 0.25, while Gaussian VAE explanations achieve only 0.13. However, HoM scores suggest a smaller gap in performance, with this metric VAEs give similar scores for all classes. For *Support Devices* localisation, there is a substantial increase in HoM scores compared to other classes, with scores upwards of 0.34.

VAE OLT explanations perform similarly to GGLT according to disease feature localisation performance metrics. IoU scores are universally low, and Dice and HoM scores exist in a similar range (Table 7.7). I provide a more in-depth comparison of GGLT and OLT methods in the evaluation of BIVA explanations.

Bounded BIVA GGLT explanations Figure 7.4 shows the importance of using the bounded BIVA for generating clear changes in traversal-based visual explanations.

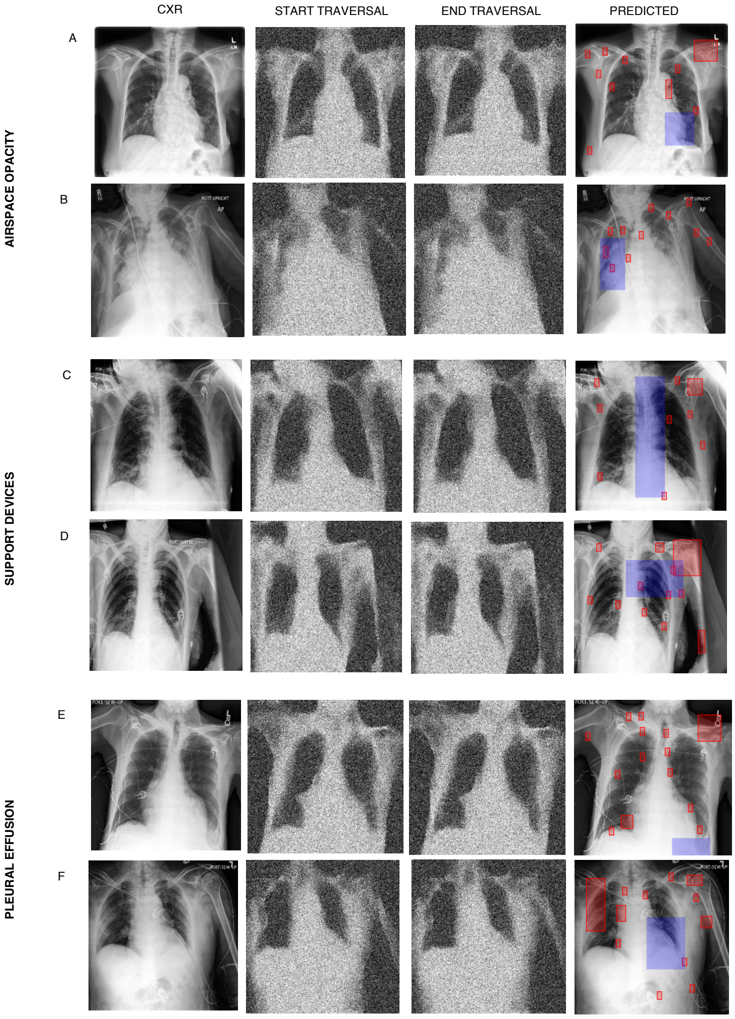


FIGURE 7.12: **Gaussian-VAE GGLTs for each evaluated label of the CheXlocalise dataset.** *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. Red boxes are predicted disease regions (derived from differences between start traversal and end traversal) and blue boxes are radiologist annotated ground truths. *Abbrvs: Gradient-guided latent traversals (GGLTs).*

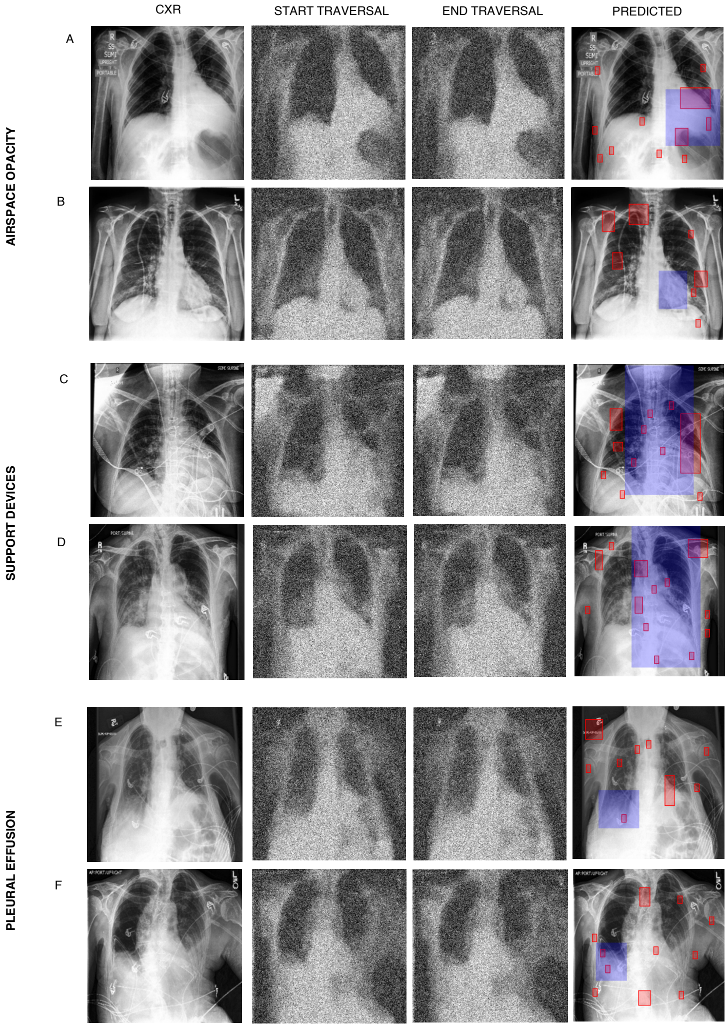


FIGURE 7.13: **Dirichlet-VAE GGLTs for each evaluated label of the CheXlocalise dataset.** *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. Red boxes are predicted disease regions (derived from differences between start traversal and end traversal) and blue boxes are radiologist annotated ground truths. *Abbrvs: Gradient-guided latent traversals (GGLTs).*

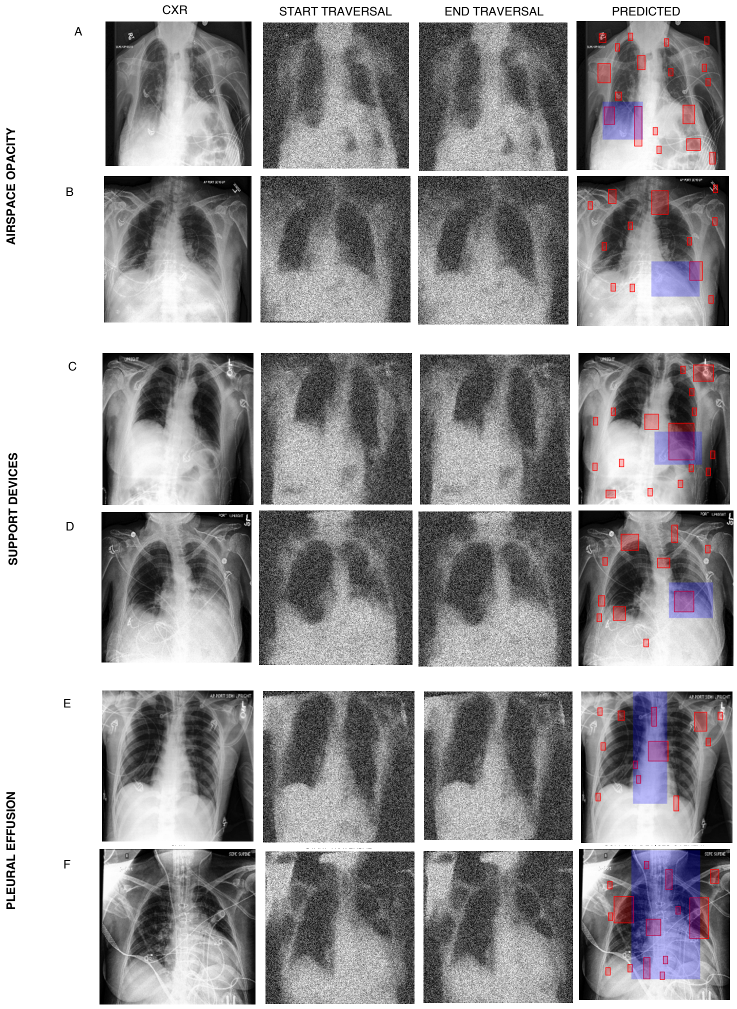


FIGURE 7.14: **Sparse-VAE GGLTs for each evaluated label of the CheXlocalise dataset.** *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. Red boxes are predicted disease regions (derived from differences between start traversal and end traversal) and blue boxes are radiologist annotated ground truths. *Abbrvs: Gradient-guided latent traversals (GGLTs).*

The GGLTs produced by the unbounded Sparse-BIVA fail to show any significant changes in images generated across the traversal [3]. With the bounded Sparse-BIVA ($\mathcal{L} > 2$) I observe clear, isolated changes in the generated images. Red arrows and residual images highlight where these changes appear disease-specific. Figure 7.16 shows examples of the GGLTs produced by bounded ($\mathcal{L} > 2$) Gaussian BIVA and Sparse BIVA models. With higher quality image generation, side-by-side comparison of Sparse-prior BIVA and Gaussian-prior BIVA shows the advantages of learning sparse posteriors for explainable prediction. I compare examples of *Aortic Enlargement* GGLTs generated by bounded BIVA models. For Gaussian-BIVA ($\mathcal{L} > 2$) non-specific changes that appear to correlate to increasing noise in the generated images. Contrastingly, with Sparse-BIVA ($\mathcal{L} > 2$) localised, interpretable changes in the CXR can be observed. Changes are isolated to the heart and mediastinum region, suggesting these features are correlated with *Aortic Enlargement* and are disease specific. Further qualitative evaluation suggests that regions highlighted in the residual image are clinically-relevant to the class of interest. Subsequently all BIVA GGLT explanations are obtained from the bounded BIVA $(L > 2)$ BIVA model.

The GGLT method is non-directional, this means changes to the latent representation may increase the probability of observing the evaluated label. With BIVA models providing higher quality image reconstructions, it is possible to observe changes to the generated image that reflect this. Figure 7.17 clearly shows that Sparse-BIVA GGLTs gives two types of feature change, the intensification and de-intensification of changing features. I consider *Aortic Enlargement* examples for their salient disease features and ease of interpretation (Figure 7.17). Explanations of *Aortic Enlargement* in Example C shows morphological changes that reduces the appearance of aortic enlargement in the CXR i.e., the pronounced structure around the aortic knuckle is reduced. While another GGLT explanation of an *Aortic Enlargement* example shown in Figure 7.17D give changes that intensify the appearance of the aortic knuckle through increased opacification of the pronounced structure.

Figure 7.17 gives additional examples of Sparse-BIVA GGLT, with these examples disease isolation within the same multi-label CXR can be observed. In addition to traversal images I generate the *expected residual* image, which captures variance in image generation over repeated sampling so as to demonstrate disease features are not simply highlighted because they are observed less frequently in the training data and are therefore more challenging to reconstruct. Figure 7.17C and Figure 7.17D show GGLT explanations of co-occurring *Cardiomegaly* prediction and *Aortic Enlargement* prediction, which co-occur frequently in the VinDr-CXR dataset. While there is some overlap in highlighted image features, it is possible to observe independent, disease-specific changes in each traversal. Comparison of the final image reconstruction (end traversal) demonstrates that performing GGLT in a disease-specific manner generates two distinctly different destination images.

---

[3] The same was observed in preliminary experiments with step size > 1.

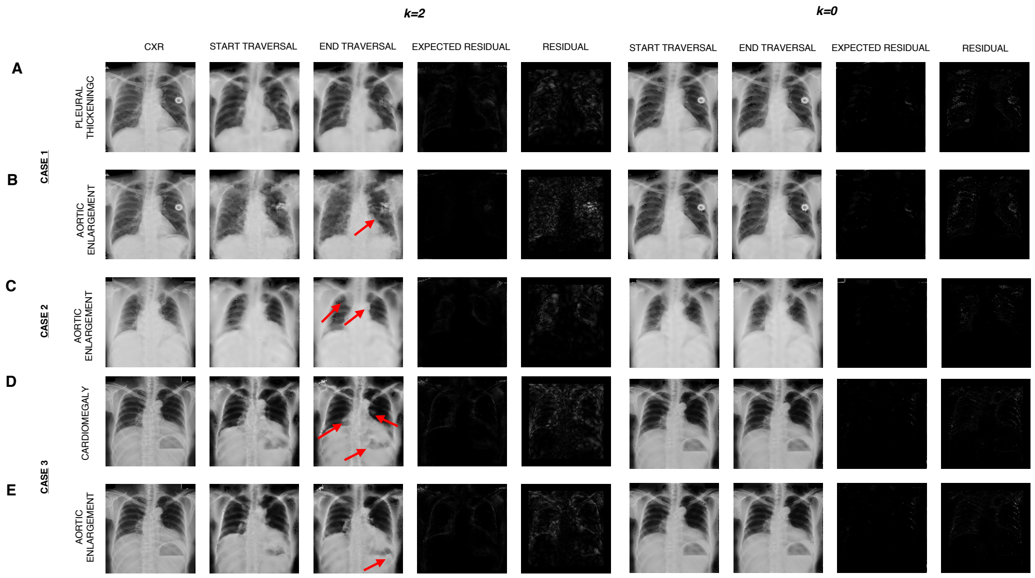


FIGURE 7.15: **Comparison of GGLT generated by unbounded Sparse-BIVA (k=0) and bounded $\mathcal{L} > 2$ Sparse-BIVA (k=2).** Each case corresponds to a different multi-label CXR. *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. *Expected residual* is a map of pixel variance over repeated reconstructions without traversal. *Residual* is the absolute difference in pixel values between the start traversal and end traversal images. Red arrows point to observable changes in the images generated over the traversal. *Abbrvs: Gradient-guided latent traversal (GGLT)*

Similarly Figure 7.17G and 7.17H shows a CXR with co-occurring *Pleural Thickening* and *Aortic Enlargement*. The *Pleural Thickening* GGLT gives a destination image with changes in 'border' areas around the edge of the lung field and decreased contrast over the clavicle, where *Pleural Thickening* is observed in this case. While *Aortic Enlargement* GGLT generates a destination image with significant changes to the heart margins and aortic knuckle.

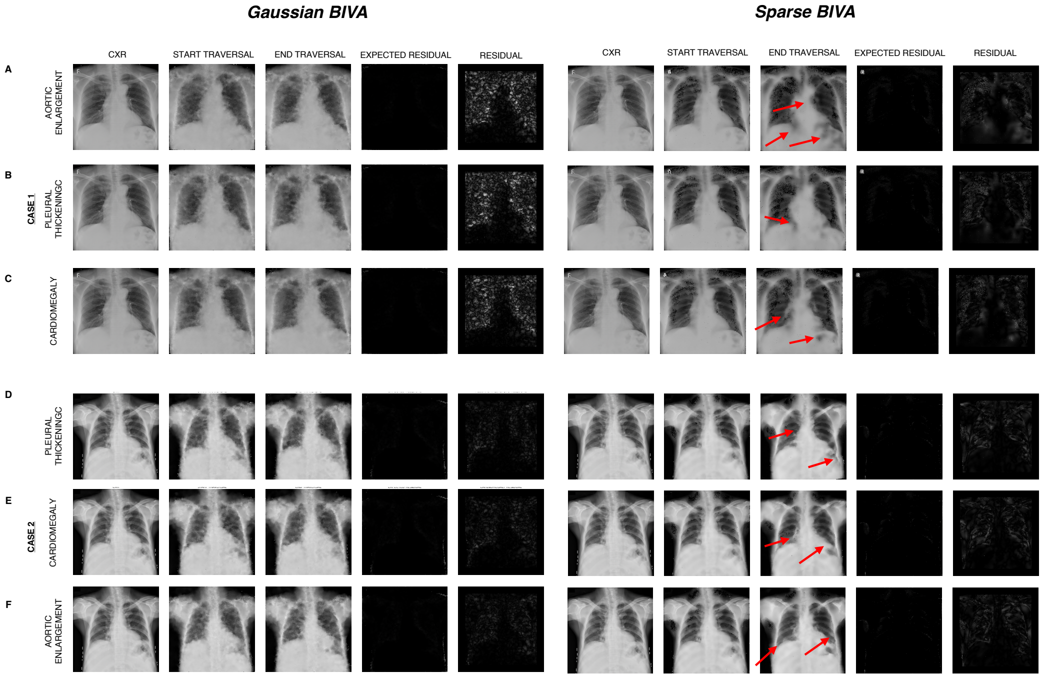


FIGURE 7.16: **Side-by-side comparison of Gaussian BIVA ($\mathcal{L} > 2$) and Sparse BIVA ($\mathcal{L} > 2$) GGLTs applied to multi-label VinDr-CXR examples.** Each case corresponds to a different multi-label CXR. *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. *Expected residual* is a map of pixel variance over repeated reconstructions without traversal. *Residual* is the absolute difference in pixel values between the start traversal and end traversal images. Red arrows point to areas of significant change. *Abbrvs: Gradient-guided latent traversal (GGLT); Chest X-ray (CXR).*

Bounded OLT explanations Similar results are observed for bounded BIVA OLTs. Figure 7.18 shows examples of Sparse-BIVA ($\mathcal{L} > 2$) OLTs applied to multi-label examples of VinDr-CXR. Qualitatively I observe that changes to the generated images are directional i.e., updates are made to the latent representation to minimise disease probability, image changes correspond to this and show the partial removal of disease features.

Figures 7.19 and 7.20 show the probability changes of an OLT and GGLT when

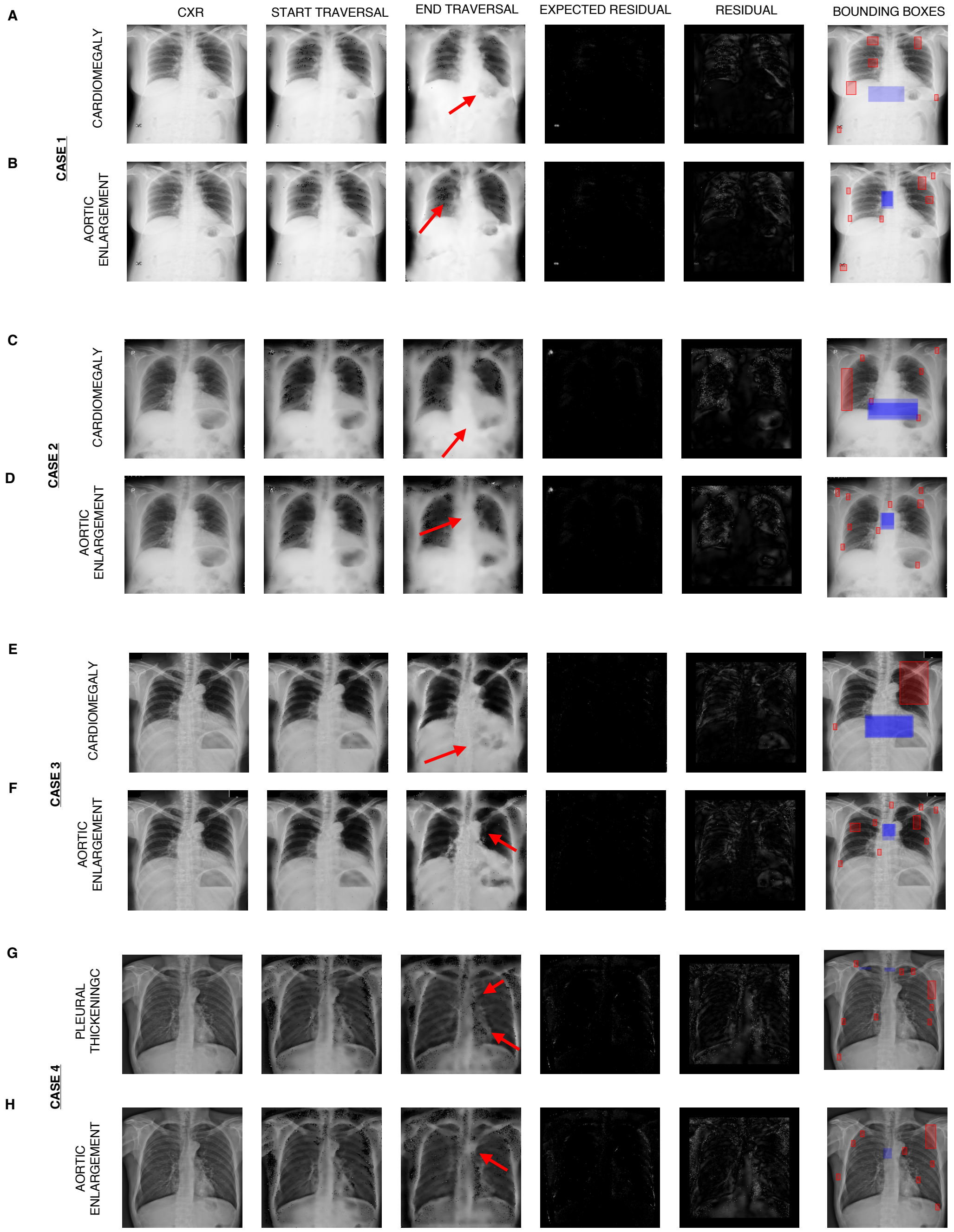


FIGURE 7.17: **Sparse-BIVA ($\mathcal{L} > 2$) GGLTs applied to multi-label VinDr-CXR examples.** Each case corresponds to a different multi-label CXR. *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. *Expected residual* is a map of pixel variance over repeated reconstructions without traversal. *Residual* is the absolute difference in pixel values between the start traversal and end traversal images. Red arrows point to structural changes that relate to the evaluated disease in generated images. Red boxes are predicted disease regions (derived from the residual) and blue boxes are radiologist annotated ground truths.

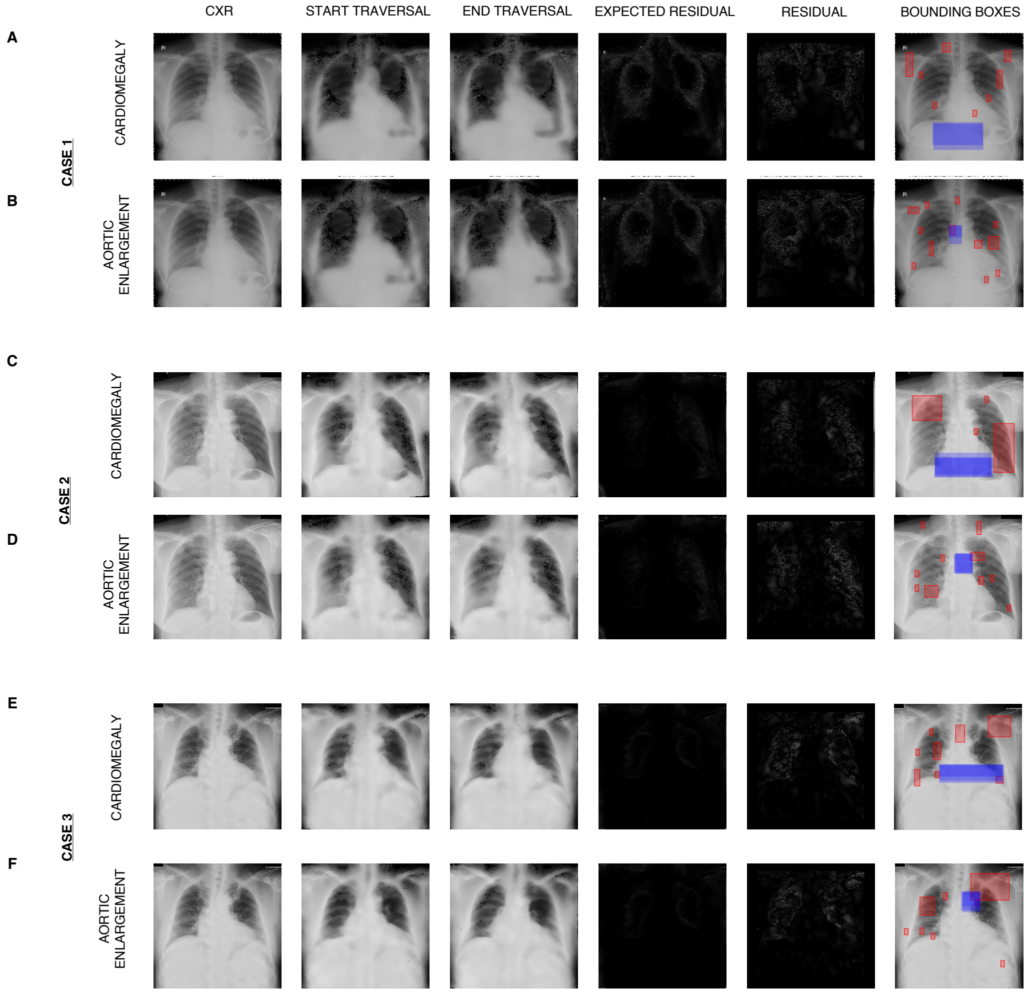


FIGURE 7.18: **Sparse-BIVA ($\mathcal{L} > 2$) OLTs applied to multi-label VinDr-CXR examples.** *Start traversal* is the original reconstruction prior to GGLT. *End traversal* is the final reconstruction post-GGLT. *Expected residual* is a map of pixel variance over repeated reconstructions without traversal. *Residual* is the absolute difference in pixel values between the start traversal and end traversal images. Red boxes are predicted disease regions (derived from the residual) and blue boxes are radiologist annotated ground truths. *Abbrvs: Optimised Latent Traversals (OLTs).*

applied to a multi-label CXR for explanation of *Aortic Enlargement* and *Pleural Thickening* prediction. Comparison of 7.19A with 7.20A clearly shows OLT leads to reduced probability of *Pleural Thickening*, while GGLT leads to increases in probability. For example, in Figure 7.19A it is possible to observe decreasing probabilities in disease classes alongside Pleural Effusion, e.g., *Pleural Effusion* (0.60 to 0.49) and *Cardiomegaly* (0.78 to 0.60). Some disease probability changes are in fact more significant than changes to the disease of interest, which is shown clearly in the *Aortic Enlargement* example, where the probability of *Aortic Enlargement* decreases from 0.95 to 0.88 while the probability of *Pleural Thickening* decreases from 0.73 to 0.65.

Localisation performance I compare VAE and BIVA explanations against CheXlocalise ground truth annotations. Both traversal-based methods show limited performance in disease localisation metrics (Table 7.5). While Gaussian-prior BIVA achieved the highest HoM (0.42), followed by Dirichlet-prior VAE (0.37), Gaussian-prior VAE (0.36), and Sparse-prior VAE (0.34), all models show poor Dice scores due to limited overlap with ground truth annotations and frequent false positive predictions. The OLT method shows similar performance patterns across models and disease classes. The Sparse-prior BIVA gives reduced localisation performance compared to Sparse-prior VAE, with IoU scores near zero for OLT in both models. Dice scores decrease from 0.18 to 0.13 and HoM drops substantially from 0.23 to 0.14, with similar declines in GGLT. Conversely, Gaussian BIVA improves upon Gaussian VAE HoM scores, with 0.31 compared to 0.22, despite a slight decrease in Dice scores (0.13 to 0.07).

On VinDr-CXR, both Sparse-prior and Gaussian-prior BIVA models achieve higher HoM scores (0.22) compared to their VAE counterparts (0.05). However, the multi-label deep CNN significantly outperforms all BIVA models with near-perfect HoM (0.95), particularly in *Cardiomegaly* (1.00) and *Aortic Enlargement* (0.98). The CNN also achieves substantially higher IoU scores (0.17) compared to BIVA models ($\leq$0.01).

While GradCAM++ achieves perfect HoM for the *Cardiomegaly* class of VinDr-CXR, its low IoU scores reflect limitations in localisation precision. Qualitative evaluation confirms GradCAM++ generates overly large, non-specific bounding boxes compared to both traversal-based methods and ground truth annotations. Figure 7.21 shows a comparison in GGLT and GradCAM++ visual explanations for a mix of VinDr-CXR examples. For the same examples, the GGLT explanations are much more localised. Overall traversal-based methods give an advantage in identifying irregularly-shaped or very small structures.

Qualitative evaluations show that post-processing of pixel changes via traversal to binarised activations and ultimately bounding boxes is ineffective, and results in a loss of information. The predicted bounding boxes are small and spread diffusely, which can be difficult to interpret and leads to a loss of information compared to

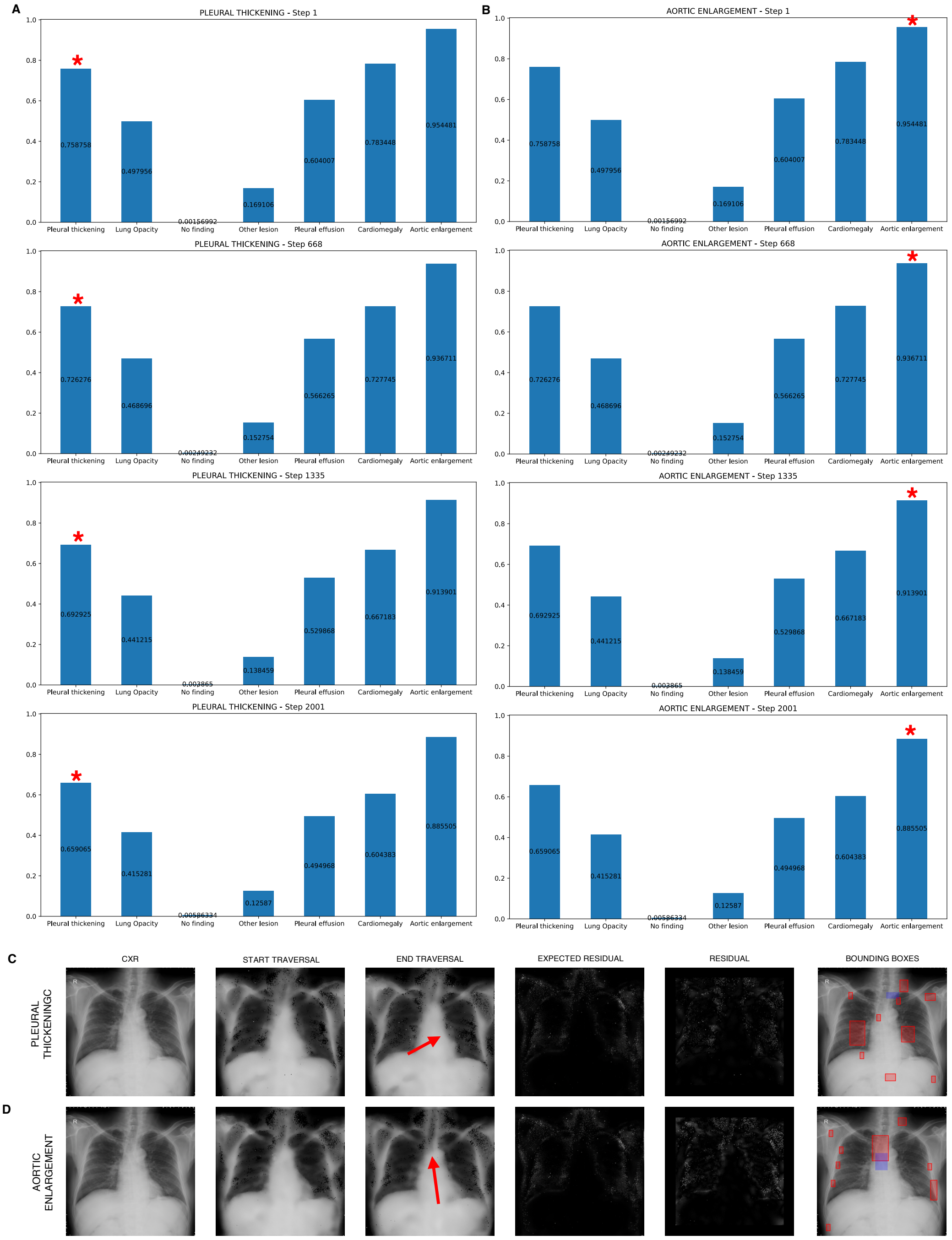


FIGURE 7.19: **Probability changes over label-specific OLT for multi-label example of VinDr-CXR.** (A) Changes in probabilities over OLT for Pleural Thickening. (B) Changes in probabilities over OLT for Aortic Enlargement. (C) Changes in image generation over OLT for Pleural Thickening. (D) Changes in image generation over OLT for Aortic Enlargement. Red asterisks * point to the evaluated class. *Abbrvs: Optimised Latent Traversals (OLTs)*

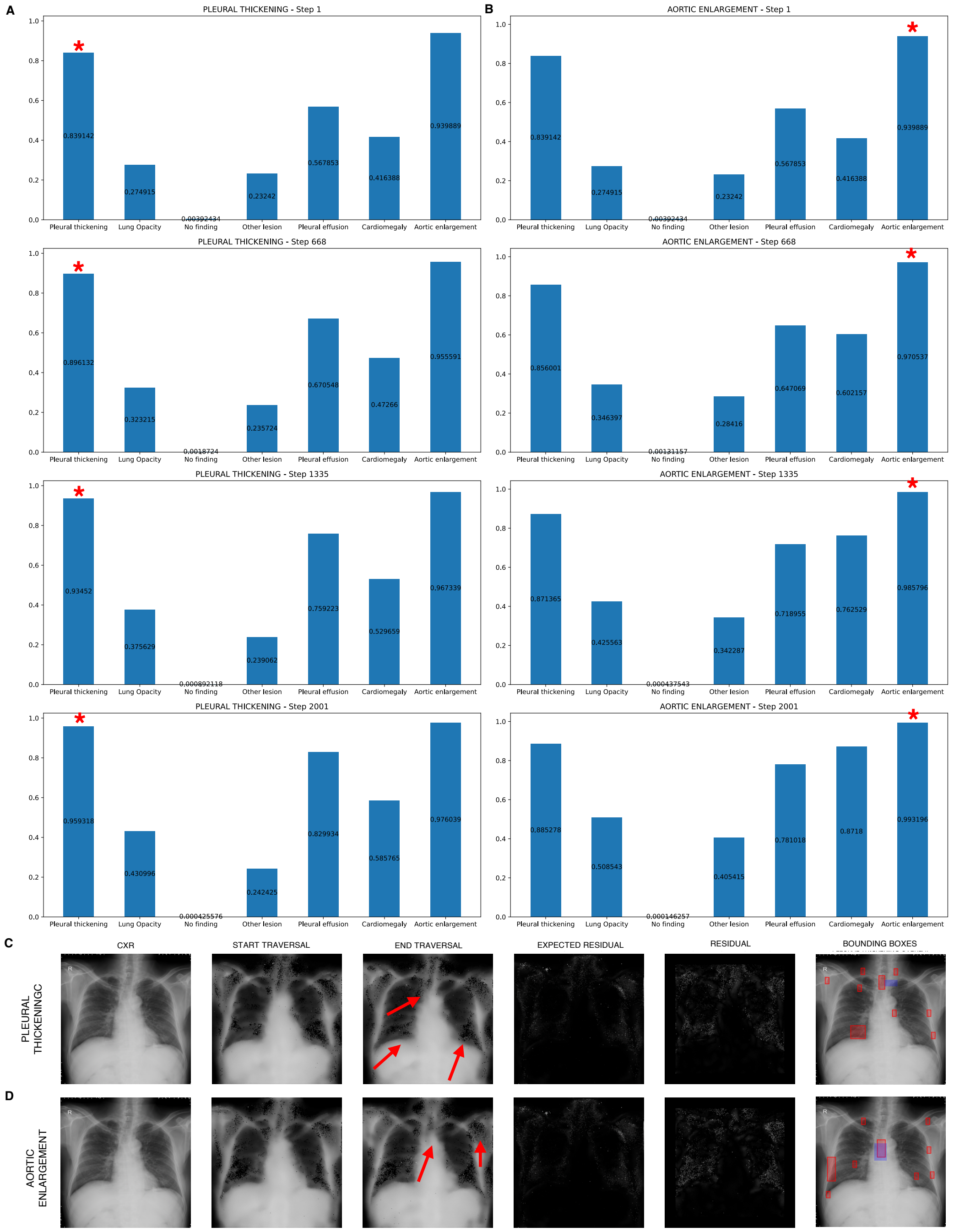


FIGURE 7.20: **Probability changes over a GGLT for a multi-label example of VinDr-CXR.** (A) Changes in probabilities for GGLT for Pleural Thickening. (B) Changes in probabilities for GGLT for Aortic Enlargement. (C) Changes in image generation over GGLT for Pleural Thickening. (D) Changes in image generation over GGLT for Aortic Enlargement. Red arrows point to areas of significant change in the traversal. The residual image is the absolute pixel-wise difference between the start traversal and end traversal images. Bounding boxes are identified from the binarised residual image. Red asterisks * point to the evaluated class. *Abbrvs: Gradient-guided latent traversals (GGLTs).*

direct visualisation of morphological changes to image (shown in the residual image of a traversal). Moreover, morphological changes by their nature happen at the boundaries of structures, not the centre, and therefore pixel changes are typically around the edges of disease feature regions, which is translated into off-centred bounding boxes. Evaluations show examples of noisy reconstructions produced by bounded BIVA models, in which changes to 'grainy' areas overshadow disease-related feature changes. This is carried over in translation of pixel changes into bounding boxes, and is perhaps another possible cause of poor localisation metrics despite observation of consistent disease-related morphological changes (Fig. 7.17).

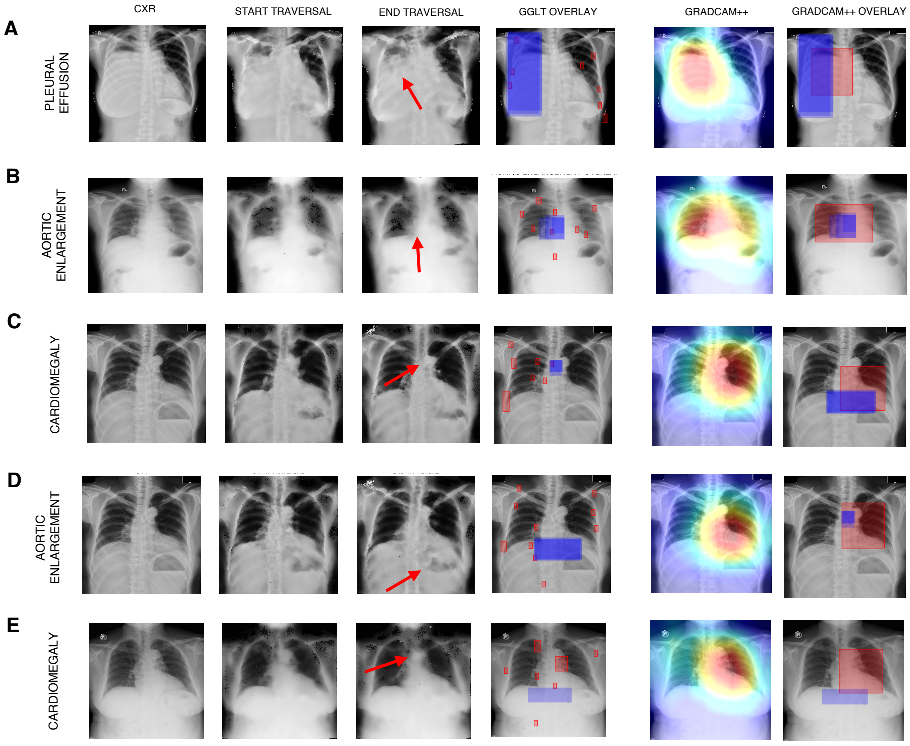


FIGURE 7.21: **Comparison of BIVA GGLT visual explanations and multi-label deep CNN GradCAM++ with post-processed localisation.** *Start traversal* is the original image reconstruction prior to GGLT. *End traversal* is the image generated post-GGLT. GGLT overlay displays the GGLT-derived bounding boxes (red) and ground truth radiologist annotations (blue), likewise for GradCAM++ overlay. GradCAM++ was selected as the baseline saliency method due to its improved localisation performance over standard Grad-CAM and its ability to pass sanity checks that verify dependence on learned model parameters (Chattopadhay et al., 2018; Adebayo et al., 2020). Red arrows point to structural changes that relate to the evaluated disease in generated images. *Abbrvs: Gradient-guided latent traversals (GGLTs).*

Shortcut features I use qualitative evaluation of visual explanations to gain insight to model reliance on shortcut features. I observe model reliance on shortcut features with traversal-based methods and GradCAM++. *Aortic Enlargement* and *Cardiomegaly* co-occur frequently within the VinDr-CXR dataset and I observe related shortcut learning when evaluating explanations for each target class. While latent dimensions isolate predictive features i.e., image features learned by the model to be associated with the target class, results suggest variational methods are unable to separate these features from shortcut features, particularly where these co-occurrences are dominant in the dataset, i.e., they exist together more frequently than independently or with any other label. Figure 7.17 presents examples of GGLTs applied to a multi-label CXR, where explanations are generated for each class. Explanations for *Aortic Enlargement* examples highlight features around both the aortic knuckle and heart margins, disease features in these areas are typically related to aortic enlargement and cardiomegaly, respectively. This pattern of correlated feature change is also observed in explanations for *Cardiomegaly* (Fig. 7.17). Moreover, the *Pleural Effusion* sample in Figure 7.21 shows an interesting example of GGLT feature changes. The opacity in the upper regions of the CXR and edge of the effusion begins to disappear, at the same time a dense, circular object forms, which could relate to the *Support Device* class.

For traversal-based approaches I also include visualisation of changing probabilities for each class-specific traversal (Fig. 7.20). I observe significant changes in probability in non-target classes. There is a large increase in *Cardiomegaly* probability, even more so than the increase in *Aortic Enlargement* probabilities (the target class). I see similar results for *Pleural Thickening*, with large increases in *Pleural Effusion* and *Cardiomegaly* probabilities. This is suggestive of shortcut learning of co-occurring pathologies. I attempt to visually evaluate shortcut learning with GradCAM++ explanations of *Aortic Enlargement* but the identified region of interest is large and non-specific, overlapping with *Pleural Thickening* features (Fig. 7.20).

7.4.2 Discussion

While VAE and BIVA models offer reasonable multi-label prediction performance, these models significantly underperform compared to the deep CNN. This drop in performance compared to CNN models is a key limitation of my proposed BIVA approach. In clinical settings, CXRs often present multiple co-occurring pathologies, thus accurately identifying all relevant conditions from a single image is crucial for effective patient management. In practice, a multi-label prediction model that can correctly predict the presence of multiple conditions simultaneously provides a more accurate and actionable representation of the patient's health status, and is a more useful tool for clinicians. This performance gap could stem from the inherent complexity of multi-task optimisation applied to variational models. While CNNs focus solely on classification, VAEs and BIVAs must simultaneously optimise for input reconstruction, latent space structure (through KL divergence), and multi-label classification. Future work should explore methods to better balance these competing objectives, such as dynamic loss weighting schemes or enhanced training protocols, to improve classification performance while maintaining the models' generative capabilities.

Despite using regularisation through multi-label prediction and KL divergence, the visual explanations revealed that both variational models and CNNs still relied on shortcut features. To address this, future research should explore incorporating additional objectives like adversarial training and synthetic counterfactuals into the multi-task framework. The persistence of shortcut learning is particularly concerning in clinical settings, where it can both degrade model performance and create dangerous divergences from medical expertise. These findings emphasise the importance of transparent visual explanations to identify such model behaviours.

Variational methods, particularly the BIVA approaches, demonstrated their own merits as explainable models. Through exploration of well-structured latent representations, the sparse-prior BIVA provides good visual explanations that offer insights into the underlying relationships between image features in complex data. Comparison with the popular post-hoc visual explainer GradCAM++ highlights the improved precision of BIVA explanations. These advantages are particularly valuable in cases of co-occurring pathologies, the diagnostic features which may exist close together in the CXR. However, quantitative localisation performance places GradCAM++ above traversal-based methods. This goes against what is observed with direct interpretation of residual/saliency maps. I propose that quantification that relies on post-processing steps disadvantages the pixel-wise explanations of variational methods, for which many small bounding boxes are predicted, against class-activation map explanations, which most frequently produce a single, large bounding box. Cohen et al. (2021), who follow a similar traversal-based strategy for generating explanations with VAEs, observe similar results in their study, reporting low overlap with ground truth pathology masks despite their reader (panel of expert radiologists) study indicating that the models are generally looking at the correct features.

While explanation precision is greatly improved, I identify a number of limitations in Sparse BIVA explanations. Traversal-based explanations appear generally limited to the boundaries of the pathology, which make interpretation more difficult and would further disadvantage these methods in this quantification workflow. In Cohen et al. (2021), radiologists also comment, that like my method, their method looks at the boundaries of the abnormality. Additionally, bounded BIVA traversals can suffer noisy reconstructions, which can lead to false positive activations when summarising feature changes.

7.5 Conclusion

This study examined the strengths and weaknesses of variational deep learning models, compared against deep CNN methods, for multi-label prediction of CXRs. While VAEs and BIVAs have promise as explainable models, they struggle with multi-label prediction due to the difficulty in balancing classification with reconstruction and regularisation objectives during training. This trade-off limits their ability to effectively predict co-occurring pathologies, which is critical in clinical settings. Furthermore, both VAE/BIVA and CNN models were prone to shortcut learning, despite efforts to mitigate this with regularisation. Despite these challenges, the sparse-prior BIVA models were shown to provide valuable interpretability through structured latent representations. While qualitative evaluation showed that these models improve on GradCAM++ in explanation precision, they lagged behind when quantitatively evaluated for disease feature localisation. This suggests that localisation metrics may not fully capture the nuances of multi-label prediction explanations through generative methods. Overall, while VAEs and BIVAs offer advantages in explainability, their performance in multi-label prediction tasks remains limited. Further research is needed to improve the multi-label prediction performance of variational methods, fully mitigate shortcut learning, and reduce false positive activations in traversal-derived explanations. Pursuing these research directions would ultimately serve to improve the clinical utility of multi-label prediction models in medical imaging.

Chapter 8

Conclusion

8.1 Key Findings

This thesis explored the challenges and advancements in deep learning models for the detection of complex lung pathologies from chest X-rays (CXRs). My primary findings can be summarised as follows:

Chapters 4 and 5 underscore the critical importance of proper evaluation and clinical guidance in developing reliable predictive models for disease diagnosis and medical image interpretation, particularly in the context of COVID-19 detection. I found that models trained on the open source COVID-19 datasets (e.g., COVIDX) generalised poorly to real-world hospital populations. I showed that the early studies that used open data and reported optimistic model performance results were overly reliant on non-clinical features, such as image resolution and annotations. My evaluation of the COVIDX dataset as a "Frankenstein" dataset, which combined multiple data sources, highlighted the significant risks of bias and confounding factors in the absence of proper metadata.

My evaluation of hospital data trained COVID-19 detection models showed performance comparable to radiologists, but inferior to the gold standard RT-PCR test. One of the major challenges identified in my evaluation is the difficulty in detecting COVID-19 in complex clinical cases, particularly among patients with comorbidities and co-occurring pathologies. These models are prone to learning shortcut features, which compromises their ability to generalise effectively and limits predictive performance. Moreover, I identified the need for precise visual explanations for complex cases with multiple pathologies to build trust in model prediction and identify model bias, which often arises due to shortcut learning. To address this, I proposed the integration of multi-label training objectives. Multi-label classification forces models to learn a more comprehensive understanding of the data, perhaps preventing shortcut learning, thereby improving their robustness and generalisability across varied clinical populations.

In Chapter 6 I introduced a novel approach to explainable multi-label classification using the Dirichlet-prior VAE model, where the Dirichlet distribution is parametrised for extreme sparsity. By leveraging gradient guided latent traversals, I was able to provide precise visual explanations of model predictions, making the decision-making process more transparent. I demonstrated that the Dirichlet-prior VAE was able to isolate disease-specific features much better than the standard Gaussian-prior VAE model. This approach has the potential to improve clinical decision-making by ensuring that AI-driven predictions are interpretable and reliable. However, I observed a limited capacity for model fit to the complex medical imaging data, leading to poor quality reconstructions, and reduced prediction capacity and explanation quality. To address this I examined the use of HVAEs, such as BIVAs, in multi-label prediction tasks (Chapter 7). I demonstrated model capacity for precise explanations through latent traversal methods by quantifying disease localisation against radiologist annotations, and compared these against GradCAM++ explanations (a popular post-hoc approach to visual explanations) from a multi-label deep CNN. I again showed improved disease feature isolation with the use of sparse prior distributions compared to the regular dense Gaussian prior.

Overall, with these improvements I make strides toward enhancing the interpretability of deep learning models in medical imaging, offering a framework that balances predictive accuracy with explainability. My findings highlight the potential of structured latent variable models like HVAEs in addressing the limitations of current deep learning approaches, particularly in multi-label classification tasks where feature isolation is crucial.

8.2 Limitations & Future Works

While this thesis presents significant advances in understanding and addressing the limitations of deep learning models in pulmonary disease detection, several limitations remain, and future work is required to build upon these findings.

Chapter 5 demonstrated that deep learning models remain vulnerable to learning shortcut features that may not represent clinically relevant information. Further research is needed to improve model robustness, particularly in the presence of confounding factors, like co-occurring pathologies. Multi-label classification, as I suggest in Chapter 6, may help mitigate this issue, but further exploration is needed to refine multi-label training objectives to prevent shortcut learning, and further investigation is needed to verify reliance on shortcut learning.

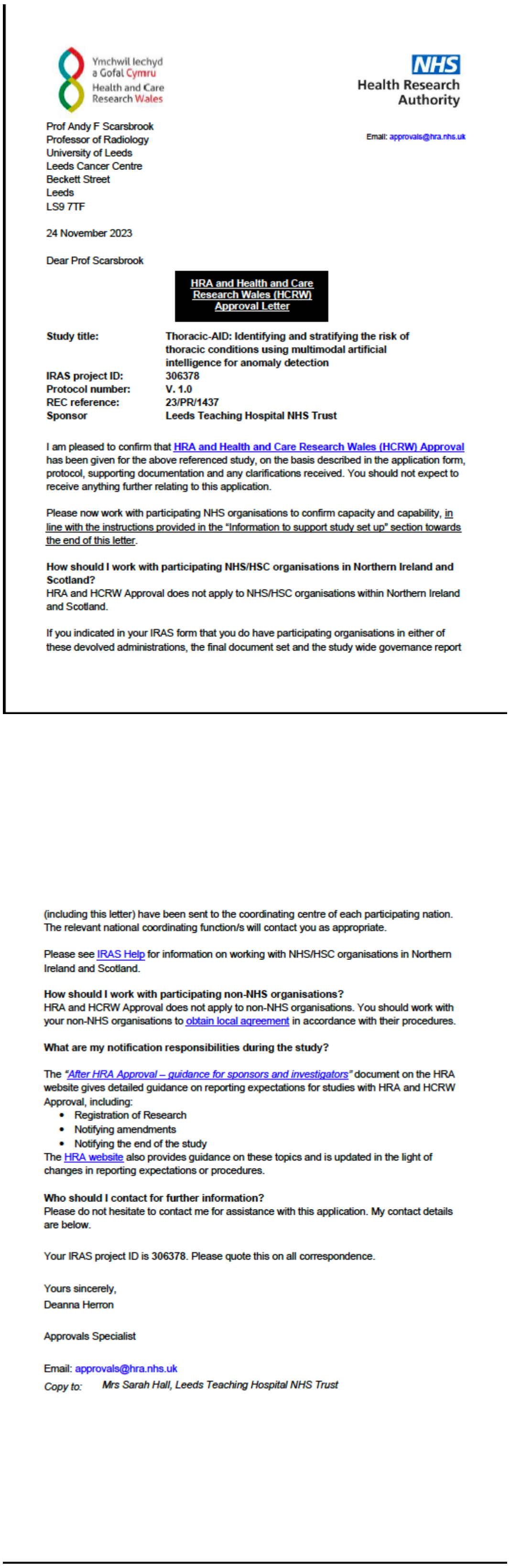
The explainable AI methods proposed in Chapter 6, while promising, require further refinement. Future work should focus on developing metrics to quantitatively assess explainability, as well as improving image reconstruction and generation quality in order to make latent traversals more interpretable. These improvements will enhance the practical utility of explainable models in clinical decision-making. I propose that an approach that combines the directional OLT method with the selective GGLT method may improve visual explanations further.

While variational methods like DirVAE and Sparse BIVA show promise, their performance in multi-label prediction tasks is still limited. The trade-off between classification accuracy, reconstruction quality, and regularisation needs to be addressed through more advanced training techniques and model architectures. Comparison of multi-label prediction performance with a deep CNN highlighted this as a key issue. Additionally, efforts should be made to reduce false positive activations and improve the localisation of disease features to enhance the clinical utility of these models.

In conclusion, while deep learning models for complex pulmonary disease detection hold great promise, there remain significant challenges to overcome, particularly in terms of model generalisation, bias mitigation, and explainability. Addressing these challenges will require continued collaboration between AI researchers and clinicians to ensure that AI tools are not only accurate but also transparent and trustworthy in real-world clinical environments.

Appendix A

Ethics Approval



Ymchwil Iechyd
a Gofal Cymru
Health and Care
Research Wales

NHS
Health Research
Authority

Email: approvals@hra.nhs.uk

Prof Andy F Scarsbrook
Professor of Radiology
University of Leeds
Leeds Cancer Centre
Beckett Street
Leeds
LS9 7TF

24 November 2023

Dear Prof Scarsbrook

HRA and Health and Care Research Wales (HCRW) Approval Letter

Study title: **Thoracic-AID: Identifying and stratifying the risk of thoracic conditions using multimodal artificial intelligence for anomaly detection**
IRAS project ID: **306378**
Protocol number: **V. 1.0**
REC reference: **23/PR/1437**
Sponsor **Leeds Teaching Hospital NHS Trust**

I am pleased to confirm that HRA and Health and Care Research Wales (HCRW) Approval has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications received. You should not expect to receive anything further relating to this application.

Please now work with participating NHS organisations to confirm capacity and capability, in line with the instructions provided in the "Information to support study set up" section towards the end of this letter.

How should I work with participating NHS/HSC organisations in Northern Ireland and Scotland?
HRA and HCRW Approval does not apply to NHS/HSC organisations within Northern Ireland and Scotland.

If you indicated in your IRAS form that you do have participating organisations in either of these devolved administrations, the final document set and the study wide governance report (including this letter) have been sent to the coordinating centre of each participating nation. The relevant national coordinating function/s will contact you as appropriate.

Please see IRAS Help for information on working with NHS/HSC organisations in Northern Ireland and Scotland.

How should I work with participating non-NHS organisations?
HRA and HCRW Approval does not apply to non-NHS organisations. You should work with your non-NHS organisations to obtain local agreement in accordance with their procedures.

What are my notification responsibilities during the study?

The *"After HRA Approval – guidance for sponsors and investigators"* document on the HRA website gives detailed guidance on reporting expectations for studies with HRA and HCRW Approval, including:

- Registration of Research
- Notifying amendments
- Notifying the end of the study

The HRA website also provides guidance on these topics and is updated in the light of changes in reporting expectations or procedures.

Who should I contact for further information?
Please do not hesitate to contact me for assistance with this application. My contact details are below.

Your IRAS project ID is 306378. Please quote this on all correspondence.

Yours sincerely,
Deanna Herron

Approvals Specialist

Email: approvals@hra.nhs.uk
Copy to: *Mrs Sarah Hall, Leeds Teaching Hospital NHS Trust*

FIGURE A.1: Required evidence of ethics approval granted by the Health Research Authority for use of NHS data.

Bibliography

Abdar, Moloud et al. (2023). “UncertaintyFuseNet: Robust uncertainty-aware hierarchical feature fusion model with Ensemble Monte Carlo Dropout for COVID-19 detection”. In: *Information Fusion* 90, pp. 364–381. ISSN: 1566-2535. DOI: `https://doi.org/10.1016/j.inffus.2022.09.023`. URL: `https://www.sciencedirect.com/science/article/pii/S1566253522001609`.

Adebayo, Julius et al. (2020). *Sanity Checks for Saliency Maps*. arXiv: `1810.03292 [cs.CV]`. URL: `https://arxiv.org/abs/1810.03292`.

Afshar, Parnian et al. (2020). “Covid-caps: A capsule network-based framework for identification of COVID-19 cases from X-ray images”. In: *Pattern Recognition Letters* 138, 638–643. DOI: `10.1016/j.patrec.2020.09.010`.

Albiol, Alberto et al. (2022). “A comparison of covid-19 early detection between Convolutional Neural Networks and radiologists”. In: *Insights into Imaging* 13.1. DOI: `10.1186/s13244-022-01250-3`.

Arevalo, John et al. (2015). “Convolutional neural networks for mammography mass lesion classification.” eng. In: *Annu Int Conf IEEE Eng Med Biol Soc* 2015, pp. 797–800. ISSN: 2694-0604 (Electronic); 2375-7477 (Linking). DOI: `10.1109/EMBC.2015.7318482`.

Arun, Nishanth et al. (2021). *Assessing the (Un)Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging*. arXiv: `2008.02766 [cs.CV]`. URL: `https://arxiv.org/abs/2008.02766`.

Banerjee, I. et al. (2023). “"Shortcuts" Causing Bias in Radiology Artificial Intelligence: Causes, Evaluation, and Mitigation”. In: *Journal of the American College of Radiology* 20.9, pp. 842–851. DOI: `10.1016/j.jacr.2023.06.025`. URL: `https://doi.org/10.1016/j.jacr.2023.06.025`.

Biffi, Carlo et al. (2020). “Explainable Anatomical Shape Analysis Through Deep Hierarchical Generative Models”. In: *IEEE Transactions on Medical Imaging* 39.6, pp. 2088–2099. DOI: `10.1109/TMI.2020.2964499`.

Burda, Yuri, Roger Grosse, and Ruslan Salakhutdinov (2015). “Importance weighted autoencoders”. In: *arXiv preprint arXiv:1509.00519*.

Burgess, Christopher P et al. (2018). “Understanding disentangling in β-VAE”. In: *arXiv preprint arXiv:1804.03599*.

Carbonneau, Marc-André et al. (2022). *Measuring Disentanglement: A Review of Metrics*. arXiv: `2012.09276 [cs.LG]`. URL: `https://arxiv.org/abs/2012.09276`.

Chattopadhay, Aditya et al. (2018). “Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks”. In: *2018 IEEE Winter*

Conference on Applications of Computer Vision (WACV), pp. 839–847. DOI: 10.1109/WACV.2018.00097.

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

Chowdhury, Nihad Karim et al. (2021). "ECOVNet: A highly effective ensemble based deep learning model for detecting COVID-19". In: *PeerJ Computer Science* 7. DOI: 10.7717/peerj-cs.551.

Chung, Adrian (2020). *Actualmed COVID-19 Chest X-Ray Data Initiative*. https://github.com/agchung/Figure1-COVID-chestxray-dataset.

Cohen, Joseph Paul and Adrian Chung (n.d.). *COVID-19 Chest X-Ray Dataset Initiative*. https://github.com/agchung/Figure1-COVID-chestxray-dataset.

Cohen, Joseph Paul, Paul Morrison, and Lan Dao (2020). "COVID-19 image data collection". In: *arXiv preprint arXiv:2003.11597*.

Cohen, Joseph Paul et al. (2021). "Gifsplanation via Latent Shift: A Simple Autoencoder Approach to Counterfactual Generation for Chest X-rays". In: *International Conference on Medical Imaging with Deep Learning*. URL: https://api.semanticscholar.org/CorpusID:236168629.

Collaborators, GBD Chronic Respiratory Disease (2020). "Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017". In: *Lancet Respir Med* 8.6, pp. 585–596. DOI: 10.1016/S2213-2600(20)30105-3. URL: https://doi.org/10.1016/S2213-2600(20)30105-3.

Cozzi, Andrea et al. (2020). "Chest x-ray in the COVID-19 pandemic: Radiologists' real-world reader performance". In: *European Journal of Radiology* 132, p. 109272. ISSN: 0720-048X. DOI: https://doi.org/10.1016/j.ejrad.2020.109272. URL: https://www.sciencedirect.com/science/article/pii/S0720048X20304617.

Cushnan, Dominic et al. (2021). "An Overview of the National COVID-19 Chest Imaging Database: Data Quality and Cohort Analysis". In: *GigaScience* 10.11, giab076. DOI: 10.1093/gigascience/giab076. URL: https://doi.org/10.1093/gigascience/giab076.

DeGrave, Alex J., Joseph D. Janizek, and Su-In Lee (2021). "AI for radiographic COVID-19 detection selects shortcuts over signal". In: *Nature Machine Intelligence* 3.7, 610–619. DOI: 10.1038/s42256-021-00338-7.

Dorent, Reuben et al. (2023). "Unified brain MR-ultrasound synthesis using multimodal hierarchical representations". In: *International conference on medical image computing and computer-assisted intervention*. Springer, pp. 448–458.

Drukker, K. et al. (2023). "Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment". In: *J Med Imaging (Bellingham)* 10.6, p. 061104. DOI: 10.1117/1.JMI.10.6.061104. URL: https://doi.org/10.1117/1.JMI.10.6.061104.

Esteva, Andre et al. (Dec. 2021). "Deep learning-enabled medical computer vision". In: *npj Digital Medicine* 4, p. 5. DOI: 10.1038/s41746-020-00376-2.

Fabbri, L. M. et al. (2023). "COPD and multimorbidity: recognising and addressing a syndemic occurrence". In: *Lancet Respir Med* 11.11, pp. 1020–1034. DOI: 10.1016/S2213-2600(23)00261-8. URL: https://doi.org/10.1016/S2213-2600(23)00261-8.

Fallah, Kion and Christopher J Rozell (2022). "Variational Sparse Coding with Learned Thresholding". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 6034–6058. URL: https://proceedings.mlr.press/v162/fallah22a.html.

Fay, Louisa et al. (2023). "Avoiding Shortcut-Learning by Mutual Information Minimization in Deep Learning-Based Image Processing". In: *IEEE Access* 11, pp. 64070–64086. DOI: 10.1109/ACCESS.2023.3289397.

Forsyth, David A. and Jean Ponce (2002). *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference. ISBN: 0130851981.

Gal, Yarin and Zoubin Ghahramani (2016). "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. PMLR, pp. 1050–1059.

Gaube, S et al. (2023). "Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays". In: *Sci Rep* 13.1, p. 1383. DOI: 10.1038/s41598-023-28633-w.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. Adaptive computation and machine learning. MIT Press. ISBN: 9780262035613. URL: https://books.google.co.in/books?id=Np9SDQAAQBAJ.

Gyawali, Prashnna Kumar et al. (2019). *Improving Disentangled Representation Learning with the Beta Bernoulli Process*. arXiv: 1909.01839 [cs.LG]. URL: https://arxiv.org/abs/1909.01839.

Haddad, M. and S. Sharma (2023). *Physiology, Lung*. Updated 2023. Available from: https://www.ncbi.nlm.nih.gov/books/NBK545177/. Treasure Island (FL): StatPearls Publishing.

Havtorn, Jakob D et al. (2021). "Hierarchical vaes know what they don't know". In: *International Conference on Machine Learning*. PMLR, pp. 4117–4128.

He, Kaiming et al. (2015). *Deep Residual Learning for Image Recognition*. arXiv: 1512.03385 [cs.CV]. URL: https://arxiv.org/abs/1512.03385.

Higgins, Irina et al. (2017). "beta-vae: Learning basic visual concepts with a constrained variational framework." In: *ICLR (Poster)* 3.

Hosny, Ahmed et al. (2018). "Artificial intelligence in radiology." eng. In: *Nat Rev Cancer* 18.8, pp. 500–510. ISSN: 1474-1768 (Electronic); 1474-175X (Print); 1474-175X (Linking). DOI: 10.1038/s41568-018-0016-5.

Hunter, Tim B. et al. (2004). "Medical Devices of the Chest". In: *RadioGraphics* 24.6. PMID: 15537981, pp. 1725–1746. DOI: 10.1148/rg.246045031. eprint: https:

//doi.org/10.1148/rg.246045031. URL: https://doi.org/10.1148/rg.246045031.

Hussain, S. et al. (2022). “Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review”. In: *Biomed Res Int* 2022, p. 5164970. DOI: 10.1155/2022/5164970. URL: https://doi.org/10.1155/2022/5164970.

IEEE8023 (n.d.). *Covid Chest X-Ray Dataset*. https://github.com/ieee8023/covid-chestxray-dataset.

Irvin, Jeremy et al. (2019). “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison”. In: *CoRR* abs/1901.07031. arXiv: 1901.07031. URL: http://arxiv.org/abs/1901.07031.

Jabbour, Sarah et al. (2020). “Deep Learning Applied to Chest X-Rays: Exploiting and Preventing Shortcuts”. In: *Proceedings of the 5th Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez et al. Vol. 126. Proceedings of Machine Learning Research. PMLR, pp. 750–782. URL: https://proceedings.mlr.press/v126/jabbour20a.html.

Jankowiak, Martin and Fritz Obermeyer (2018). “Pathwise Derivatives Beyond the Reparameterization Trick”. In: *CoRR* abs/1806.01851. arXiv: 1806.01851. URL: http://arxiv.org/abs/1806.01851.

Joo, Weonyoung et al. (2020). “Dirichlet variational autoencoder”. In: *Pattern Recognition* 107, p. 107514.

Joseph, Nicholos P. et al. (2020). “Racial and Ethnic Disparities in Disease Severity on Admission Chest Radiographs among Patients Admitted with Confirmed Coronavirus Disease 2019: A Retrospective Cohort Study”. In: *Radiology* 297.3. PMID: 32673191, E303–E312. DOI: 10.1148/radiol.2020202602. eprint: https://doi.org/10.1148/radiol.2020202602. URL: https://doi.org/10.1148/radiol.2020202602.

Khan, Asif Iqbal, Junaid Latief Shah, and Mohammad Mudasir Bhat (2020). “Coronet: A deep neural network for detection and diagnosis of COVID-19 from chest X-ray images”. In: *Computer Methods and Programs in Biomedicine* 196, p. 105581. DOI: 10.1016/j.cmpb.2020.105581.

Khobahi, Shahin, Chirag Agarwal, and Mojtaba Soltanalian (2020). “CoroNet: A Deep Network Architecture for Semi-Supervised Task-Based Identification of COVID-19 from Chest X-ray Images”. In: *medRxiv*.

Kim, Hyunjik and Andriy Mnih (2018). “Disentangling by factorising”. In: *International conference on machine learning*. PMLR, pp. 2649–2658.

Kingma, Diederik P. and Max Welling (2013). “Auto-Encoding Variational Bayes”. In: *CoRR* abs/1312.6114.

Kingma, Diederik P, Max Welling, et al. (2019). “An introduction to variational autoencoders”. In: *Foundations and Trends® in Machine Learning* 12.4, pp. 307–392.

Kortela, Elisa et al. (May 2021). “Real-life clinical sensitivity of SARS-CoV-2 RT-PCR test in symptomatic patients”. In: *PLOS ONE* 16.5, pp. 1–19. DOI: 10.1371/

journal.pone.0251661. URL: https://doi.org/10.1371/journal.pone.0251661.

Krishna, Rachana et al. (2024). “Pleural Effusion”. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing. URL: https://www.ncbi.nlm.nih.gov/books/NBK448189/.

Lakhani, Paras and Baskaran Sundaram (2017). “Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks.” eng. In: *Radiology* 284.2, pp. 574–582. ISSN: 1527-1315 (Electronic); 0033-8419 (Linking). DOI: 10.1148/radiol.2017162326.

LeCun, Yann, Y. Bengio, and Geoffrey Hinton (May 2015). “Deep Learning”. In: *Nature* 521, pp. 436–44. DOI: 10.1038/nature14539.

Lee, Kyung Soo et al. (2013). “Consolidation”. In: *Radiology Illustrated: Chest Radiology*, pp. 221–233. DOI: 10.1007/978-3-642-37096-0_22.

Li, Aixuan et al. (2023). “Mutual information regularization for weakly-supervised RGB-D salient object detection”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 34.1, pp. 397–410.

Li, Jingxiong et al. (2021). “Multiscale attention guided network for covid-19 diagnosis using chest X-ray images”. In: *IEEE Journal of Biomedical and Health Informatics* 25.5, 1336–1346. DOI: 10.1109/jbhi.2021.3058293.

Liu, Xiao et al. (2022). “Learning disentangled representations in the imaging domain”. In: *Medical Image Analysis* 80, p. 102516.

Maaløe, Lars et al. (2019). “Biva: A very deep hierarchy of latent variables for generative modeling”. In: *Advances in neural information processing systems* 32.

Maaten, Laurens van der and Geoffrey Hinton (2008). “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Mathieu, Emile et al. (2019). “Disentangling disentanglement in variational autoencoders”. In: *International conference on machine learning*. PMLR, pp. 4402–4412.

Matthey, Loic et al. (2017). *dSprites: Disentanglement testing Sprites dataset*. https://github.com/deep
dataset/.

Middleton, P et al. (2021). “Characteristics and outcomes of clinically diagnosed RT-PCR swab negative COVID-19: a retrospective cohort study”. In: *Sci Rep* 11.1, p. 2455. DOI: 10.1038/s41598-021-81930-0.

Mittermaier, M., M. M. Raza, and J. C. Kvedar (2023). “Bias in AI-based models for medical applications: challenges and mitigation strategies”. In: *NPJ Digit Med* 6.1, p. 113. DOI: 10.1038/s41746-023-00858-z. URL: https://doi.org/10.1038/s41746-023-00858-z.

Mondal, Arnab Kumar et al. (2022). “XViTCOS: Explainable vision transformer based COVID-19 screening using radiography”. In: *IEEE Journal of Translational Engineering in Health and Medicine* 10, 1–10. DOI: 10.1109/jtehm.2021.3134096.

Müller, Sarah et al. (Feb. 2024). "Disentangling representations of retinal images with generative models". In: *arXiv e-prints*, arXiv:2402.19186, arXiv:2402.19186. DOI: 10.48550/arXiv.2402.19186. arXiv: 2402.19186 [cs.CV].

Nafisah, Saad I. et al. (2023). "A Comparative Evaluation between Convolutional Neural Networks and Vision Transformers for COVID-19 Detection". In: *Mathematics* 11.6. ISSN: 2227-7390. DOI: 10.3390/math11061489. URL: https://www.mdpi.com/2227-7390/11/6/1489.

Nalisnick, Eric and Padhraic Smyth (2017). *Stick-Breaking Variational Autoencoders*. arXiv: 1605.06197 [stat.ML]. URL: https://arxiv.org/abs/1605.06197.

Nguyen, H. Q., K. Lam, L. T. Le, et al. (2022). "VinDr-CXR: An Open Dataset of Chest X-Rays with Radiologist's Annotations". In: *Scientific Data* 9, p. 429. DOI: 10.1038/s41597-022-01498-w. URL: https://doi.org/10.1038/s41597-022-01498-w.

Ong Ly, C., B. Unnikrishnan, T. Tadic, et al. (2024). "Shortcut Learning in Medical AI Hinders Generalization: Method for Estimating AI Model Generalization Without External Data". In: *npj Digital Medicine* 7, p. 124. DOI: 10.1038/s41746-024-01118-4. URL: https://doi.org/10.1038/s41746-024-01118-4.

Ozturk, Tulin et al. (2020). "Automated detection of COVID-19 cases using deep neural networks with X-ray images". In: *Computers in Biology and Medicine* 121, p. 103792. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2020.103792. URL: https://www.sciencedirect.com/science/article/pii/S0010482520301621.

Park, Junghoon, Il-Youp Kwak, and Changwon Lim (2021). "A deep learning model with self-supervised learning and attention mechanism for COVID-19 diagnosis using chest X-ray images". In: *Electronics* 10.16, p. 1996. DOI: 10.3390/electronics10161996.

Prior, F. et al. (Jan. 2020). "Open access image repositories: high-quality data to enable machine learning research". en. In: *Clinical Radiology* 75.1, pp. 7–12. ISSN: 0009-9260. DOI: 10.1016/j.crad.2019.04.002. URL: https://www.sciencedirect.com/science/article/pii/S0009926019301692 (visited on 05/30/2021).

Putcha, N. et al. (2015). "Comorbidities and Chronic Obstructive Pulmonary Disease: Prevalence, Influence on Outcomes, and Management". In: *Semin Respir Crit Care Med* 36.4, pp. 575–591. DOI: 10.1055/s-0035-1556063. URL: https://doi.org/10.1055/s-0035-1556063.

Rahman, Tawsifur (n.d.). *COVID-19 Radiography Database*. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

Roberts, Michael et al. (2021). “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans”. In: *Nature Machine Intelligence* 3.3, pp. 199–217. DOI: `10.1038/s42256-021-00307-0`. URL: `https://doi.org/10.1038%2Fs42256-021-00307-0`.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, pp. 234–241.

RSNA Pneumonia Detection Challenge (2018). `https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pneumonia-detection-challenge-2018`.

Rubin, Geoffrey D. et al. (2020). “The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society”. In: *Radiology* 296.1. PMID: 32255413, pp. 172–180. DOI: `10.1148/radiol.2020201365`. eprint: `https://doi.org/10.1148/radiol.2020201365`. URL: `https://doi.org/10.1148/radiol.2020201365`.

Rudin, Cynthia (2019). *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. arXiv: `1811.10154 [stat.ML]`. URL: `https://arxiv.org/abs/1811.10154`.

Saporta, A, X Gui, A Agrawal, et al. (2022). “Benchmarking saliency methods for chest X-ray interpretation”. In: *Nature Machine Intelligence* 4, pp. 867–878. DOI: `10.1038/s42256-022-00536-x`.

Selvaraju, Ramprasaath R. et al. (2017a). “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. DOI: `10.1109/ICCV.2017.74`.

— (2017b). “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. DOI: `10.1109/ICCV.2017.74`.

Sharma, Vishal and Curtis Dyreson (2020). “COVID-19 Screening Using Residual Attention Network an Artificial Intelligence Approach”. In: *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1354–1361. DOI: `10.1109/ICMLA51294.2020.00211`.

Shen, Dinggang, Guorong Wu, and Heung-Il Suk (2017). “Deep Learning in Medical Image Analysis.” eng. In: *Annu Rev Biomed Eng* 19, pp. 221–248. ISSN: 1545-4274 (Electronic); 1523-9829 (Print); 1523-9829 (Linking). DOI: `10.1146/annurev-bioeng-071516-044442`.

Sogani, Julie et al. (2020). “Artificial intelligence in radiology: the ecosystem essential to improving patient care.” eng. In: *Clin Imaging* 59.1, A3–A6. ISSN: 1873-4499 (Electronic); 0899-7071 (Linking). DOI: `10.1016/j.clinimag.2019.08.001`.

Sønderby, Casper Kaae et al. (2016a). “How to train deep variational autoencoders and probabilistic ladder networks”. In: *33rd International Conference on Machine Learning (ICML 2016)*.

— (2016b). “Ladder variational autoencoders”. In: *Advances in neural information processing systems* 29.

Springenberg, Jost Tobias et al. (2015). *Striving for Simplicity: The All Convolutional Net*. arXiv: `1412.6806 [cs.LG]`. URL: `https://arxiv.org/abs/1412.6806`.

Stone, Rebecca S et al. (2022). “Epistemic Uncertainty-Weighted Loss for Visual Bias Mitigation”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2897–2904. DOI: `10.1109/CVPRW56347.2022.00327`.

Strohm, Lea et al. (May 2020). “Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors”. In: *European Radiology* 30. DOI: `10.1007/s00330-020-06946-y`.

Sun, Ju et al. (June 2022). “Performance of a Chest Radiograph AI Diagnostic Tool for COVID-19: A Prospective Observational Study”. In: *Radiology: Artificial Intelligence* 4.

Sun, Susu et al. (2023). “Inherently interpretable multi-label classification using class-specific counterfactuals”. In: *arXiv preprint arXiv:2303.00500*.

Sverzellati, Nicola et al. (2020). “Integrated Radiologic Algorithm for COVID-19 Pandemic”. In: *Journal of Thoracic Imaging* 35.4, pp. 228–233. DOI: `10.1097/RTI.0000000000000516`.

Tabik, S. et al. (2020). “COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images”. In: *IEEE Journal of Biomedical and Health Informatics* 24.12, pp. 3595–3605. DOI: `10.1109/JBHI.2020.3037127`. URL: `https://doi.org/10.1109/JBHI.2020.3037127`.

Tahir, Anas M. et al. (2021). “COVID-19 infection localization and severity grading from chest X-ray images”. In: *Computers in Biology and Medicine* 139, p. 105002. ISSN: 0010-4825. DOI: `https://doi.org/10.1016/j.compbiomed.2021.105002`. URL: `https://www.sciencedirect.com/science/article/pii/S0010482521007964`.

Tonolini, Francesco, Bjørn Sand Jensen, and Roderick Murray-Smith (2020). “Variational sparse coding”. In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 690–700.

Tsai, Emily B. et al. (2021). “The RSNA International COVID-19 Open Radiology Database (RICORD)”. In: *Radiology* 299.1. PMID: 33399506, E204–E213. DOI: `10.1148/radiol.2021203957`. eprint: `https://doi.org/10.1148/radiol.2021203957`. URL: `https://doi.org/10.1148/radiol.2021203957`.

Türk, Fatih and Yasin Kökver (2023). “Detection of Lung Opacity and Treatment Planning with Three-Channel Fusion CNN Model”. In: *Arabian Journal for Science and Engineering*. Epub ahead of print, pp. 1–13. DOI: `10.1007/s13369-023-07843-4`.

Vafaii, Hadi, Jacob Yates, and Daniel Butts (2024). “Hierarchical VAEs provide a normative account of motion processing in the primate brain”. In: *Advances in Neural Information Processing Systems* 36.

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

Vercheval, Nicolas and Aleksandra Pižurica (2021). “Hierarchical Variational Autoencoders For Visual Counterfactuals”. In: *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2513–2517. DOI: `10.1109/ICIP42928.2021.9506780`.

Wang, Linda, Zhong Qiu Lin, and Alexander Wong (2020). “Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images”. In: *Scientific Reports* 10.1. DOI: `10.1038/s41598-020-76550-z`.

Wang, Xiaosong et al. (July 2017). “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv: 1705.02315, pp. 3462–3471. DOI: `10.1109/CVPR.2017.369`. URL: `http://arxiv.org/abs/1705.02315` (visited on 05/15/2021).

Warren, Melissa A et al. (2018). “Severity scoring of lung oedema on the chest radiograph is associated with clinical outcomes in ards”. In: *Thorax* 73.9, 840–846. DOI: `10.1136/thoraxjnl-2017-211280`.

Watson, Jessica, Penny F Whiting, and John E Brush (2020). “Interpreting a covid-19 test result”. In: *BMJ*, p. m1808. DOI: `10.1136/bmj.m1808`.

Weng, Nina et al. (2023). “Fast diffusion-based counterfactuals for shortcut removal and generation”. In: *arXiv preprint arXiv:2312.14223*.

Wold, Svante, Kim Esbensen, and Paul Geladi (1987). “Principal component analysis”. In: *Chemometrics and Intelligent Laboratory Systems* 2.1. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists, pp. 37–52. ISSN: 0169-7439. DOI: `https://doi.org/10.1016/0169-7439(87)80084-9`. URL: `https://www.sciencedirect.com/science/article/pii/0169743987800849`.

Xu, Kunxiong, Wentao Fan, and Xin Liu (2023). “Unsupervised Disentanglement Learning via Dirichlet Variational Autoencoder”. In: *IEA/AIE (1)*, pp. 341–352. URL: `https://doi.org/10.1007/978-3-031-36819-6_30`.

Zhou, Zongwei et al. (2018). “Unet++: A nested u-net architecture for medical image segmentation”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, pp. 3–11.