FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Disentanglement Representation Learning for Generalisability in Medical Multi-Centre Data

Daniel José Barros Silva

Mestrado Integrado em Bioengenharia

Supervisor: Luís Teixeira

Second Supervisor: Wilson Silva

October 16, 2023

Disentanglement Representation Learning for Generalisability in Medical Multi-Centre Data

Daniel José Barros Silva

Mestrado Integrado em Bioengenharia

Approved in public examination by the Jury:

President: Miguel Velhote Correia

Referee: Mara Graziani

Referee: Luís Teixeira

October 16, 2023

Resumo

O poder inerente ao uso de métodos de *Deep Learning* está a alterar o futuro da sociedade. Processos que envolvem redes neuronais, tais como a classificação de imagem, são usados em várias frentes e atingem resultados comparáveis aos dos seres humanos. A tarefa de classificação de imagens pode ser o catalisador para o avanço do funcionamento dos hospitais e clínicas de saúde, sendo uma ferramenta fulcral para os profissionais de saúde, principalmente no diagnóstico médico. Estes métodos reduzem o tempo necessário para o diagnóstico e ajudam os médicos no planeamento de tratamentos, levando a um aumento da qualidade de vida dos doentes e reduzindo os custos de operação das infraestruturas hospitalares. O uso de sistemas de Inteligência Artificial na saúde é visto com incerteza e cautela devido ao alto risco e responsabilidade associados a posições médicas. Esta incerteza advém de inconsistências presentes nas previsões destes modelos quando os dados de inferência não seguem a mesma distribuição de treino. Consequentemente, estes modelos não têm a capacidade de generalizar tão bem como os técnicos de saúde na avaliação de exames que seguiram diferentes processos de aquisição. Isto leva a graves consequências, particularmente no contexto médico. Assim sendo, soluções para melhorar a capacidade de generalização de modelos de *Deep Learning* têm sido alvo de investigação, de forma a mitigar quebras de rendimento na análise de dados fora da distribuição e promover confiança e aceitação destes modelos em aplicações médicas. Os esforços na investigação podem ser separados em soluções *data* e *model-centric*. O foco nos dados usa particularidades nas imagens para induzir mais diversidade e melhor extração de características nos modelos. Ao mesmo tempo, práticas focadas no modelo fazem mudanças nas arquiteturas e nos objetivos de treino de forma a prevenir que o modelo alcance um estado de demasiada confiança, também designado por *overfitting*.

Nesta dissertação, o nosso objetivo primário é desenvolver modelos mais robustos num ambiente médico multicentro, onde a distribuição dos dados muda nos passos de inferência. Este fenómeno é alcançável usando técnicas focadas nos dados, como a variabilidade existente nas características dos exames explorados, permitindo a criação de tarefas paralelas. Adicionalmente, usamos métodos focados no modelo, como a regularização usando *disentanglement* e mapas de atenção, promovendo invariância no modelo no que toca à previsão de doenças. As nossas experiências levaram à seleção de fontes de variabilidade nos dados, que, integrado com alterações nas arquiteturas, resultaram em ganhos de rendimento em cenários fora de distribuição, relativamente a um modelo estabelecido como base.

Em suma, este estudo foca-se em modificações de modelos de *Deep Learning*, cruciais para promover a generalização, de forma a aumentar a taxa de aceitação e confiança em soluções de Inteligência artificial na prática clínica.

Abstract

The power of Deep Learning applications is currently shaping the future of society. Deep neural approaches, such as image classification, are used on several fronts and achieve results comparable to human reasoning. Remarkably, the image classification scenario can significantly advance how hospitals and clinics operate, aiding overloaded medical professionals primarily for diagnostic purposes. These methods save doctors a substantial amount of time and help them with treatment planning, leading to a general increase in patient's quality of life and decreasing the facilities' operating costs.

The use of AI systems in healthcare is met with uncertainty and caution due to the high risk and responsibility inherent in medical roles. This is mainly due to inconsistencies in predictions when inference data is not precisely curated to training distributions. As a result, these models cannot generalise as effectively as human practitioners when evaluating exams that follow different acquisition procedures. This can lead to severe consequences, particularly in the medical field. Consequently, thoughtful solutions are currently being researched to improve the generalisability of deep neural algorithms, mitigating any performance drops when dealing with out-of-distribution data and promoting trust and acceptance of these models in medical applications. Research efforts for improved generalisation can be separated into data-centric and model-centric branches. Data-centrism takes advantage of data particularities to induce more diversity and better feature extraction. At the same time, model-specific practices make architectural and training objective changes to prevent the model from reaching an overconfidence state, better known as overfitting.

In this thesis, our primary goal is to develop more robust models in a medical multi-centre environment where the distribution of the data changes in inference steps. We can achieve this by using data-centric approaches, such as the explored feature variability in the training data, allowing the creation of valuable side tasks. Additionally, one can use model-focused techniques, namely disentanglement regularisation and attention maps, to promote model invariance in disease prediction. Our experiments led to the selection of favourable sources of variability in data, which, integrated with architecture modifications, led to performance gains in out-of-distribution data compared to the established baseline.

In conclusion, this study delves into modifications to Deep Learning models that promote generalisation, further increasing the acceptance and trust in AI solutions for daily use in the medical field.

Acknowledgments

My first word of thanks goes to my supervisors, Professor Luís Teixeira and Researcher Wilson Silva, who accompanied me through the dissertation, listening to all my crazy ideas and providing moments of insightful discussions that allowed me to explore the most exciting topics without any limitations. Additionally, thank you for your patience and motivation during this extended deadline period.

I want to thank all my colleagues at the NKI for helping me unblock new ideas and providing one of the best working environments I ever experienced. I also will remember all the great moments you offered me, especially during the borrels.

I want to extend my gratitude to the VCMI group at INESC TEC, which motivated me to think outside the box when I felt I could not do anything more to improve my results.

I am also thankful for the Causality-driven Generative Models for Privacy-preserving Case-based Explanations (CAGING) project at INESC TEC, under which I received a research grant. Being under a project increased my ambitions to the roof. I tried to give my everything to share something with the scientific community.

Last but not least, I want to thank my family and friends.

To my mother, father, brother and sister: I know these 5 years were rough, and I might not have always been the best person to be surrounded, especially when exams (and dissertations) are around. Thank you for your teachings that shaped me into who I am today. All was worth it, and I will give you more than the double you gave me. I will always remember. Also, Ollie, you are the best dog in the world, my mate, for all the daily walks and runs. You deserve to be eternal.

To my friends: I couldn't stand the pressure life throws at me without your support. Through thick and thin, you were always there, whether by having some drinks, going to parties, or having the most comforting conversations ever. Sometimes, you know me better than even myself. All the funny and crazy moments are stitched to my memory, and I know many adventures are still waiting for us. You guys are the best.

Daniel José Barros Silva

”Deep in the human unconsciousness is a pervasive need for a logical universe that makes sense.
But the real universe is always one step beyond logic.”

from "the sayings of Muab’Dib", by the Princess Irulan

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Motivation . . . 3
- 1.3 Objectives . . . 4
- 1.4 Main Contributions . . . 4
- 1.5 Dissertation Structure . . . 5

2 Background **7**
- 2.1 Introduction . . . 7
- 2.2 X-rays in the modern World . . . 7
  - 2.2.1 X-ray Characteristics and Generation . . . 7
  - 2.2.2 X-rays in Healthcare: Post-Processing . . . 8
- 2.3 Foundational Deep Learning . . . 10
- 2.4 Generative Models . . . 11
  - 2.4.1 Variational AutoEncoders . . . 11
  - 2.4.2 Generative Adversarial Networks . . . 12
- 2.5 The Attention Mechanism . . . 13
  - 2.5.1 Self-Attention . . . 14
  - 2.5.2 Multi-Head Attention . . . 15
  - 2.5.3 Transformer Architecture . . . 15
  - 2.5.4 Vision Transformers . . . 16
- 2.6 Disentanglement Representation Learning . . . 17
- 2.7 Conclusion . . . 18

3 Literature Review **19**
- 3.1 Generalisability in Deep Learning Applications . . . 19
  - 3.1.1 Data-Centric Approaches . . . 19
  - 3.1.2 Model-Centric Approaches . . . 20
- 3.2 Generative Models . . . 21
  - 3.2.1 Variational AutoEncoders . . . 21
  - 3.2.2 Generative Adversarial Networks . . . 22
- 3.3 Attention-Based Mechanisms . . . 23
- 3.4 Disentanglement Representation Learning . . . 25
- 3.5 Conclusion . . . 27

4 Under The Hood: Data Choices and Preliminary Implementations **29**
- 4.1 Introduction . . . 29
- 4.2 Datasets . . . 29

4.2.1 BRAX . . . 30
4.2.2 CheXpert . . . 30
4.2.3 MIMIC-CXR . . . 30
4.2.4 VinDr-CXR . . . 31
4.3 Implementation Details . . . 32
4.3.1 Disease Task . . . 32
4.3.2 Data Sampling . . . 32
4.3.3 Metrics . . . 32
4.3.4 Training and Optimization . . . 33
4.4 Baseline Selection . . . 33
4.4.1 CNNs vs. Transformers . . . 34
4.4.2 CNNs vs. custom Encoder . . . 35
4.4.3 Final Remarks . . . 35
4.5 Scanner Features Evaluation . . . 35
4.5.1 Manufacturer ID classification . . . 36
4.5.2 Windowing Settings cluster classification . . . 37
4.5.3 Scanner Features Task Selection . . . 38
4.6 Conclusion . . . 39

5 Multi-Task Encoding and Equal Probability Loss 41
5.1 Introduction . . . 41
5.2 Equal Probability Regularisation . . . 41
5.3 Materials and Methods . . . 42
5.4 Results and Discussion . . . 43
5.5 Conclusion . . . 45

6 Attention-based Regularisation 47
6.1 Introduction . . . 47
6.2 Contrastive Attention for Early Feature Separation . . . 47
6.3 Methods and Implementation . . . 48
6.3.1 Proposed Approach . . . 48
6.3.2 Regularisation Process . . . 49
6.4 Results and Discussion . . . 50
6.4.1 Ablation Study . . . 51
6.5 Conclusion . . . 52

7 Learning Neural Discrete Representations with Attention: A Unified Approach 53
7.1 Introduction . . . 53
7.2 VQ-VAE for Image Generation . . . 54
7.2.1 Implementation Details . . . 54
7.2.2 Results and Discussion . . . 54
7.3 VQ-VAE for Disentangled Disease Classification) . . . 55
7.3.1 Disease-Only Classification . . . 55
7.3.2 Multi-Task Scenario without Regularisation . . . 55
7.3.3 Multi-Task Scenario with Attention-Based Embeddings . . . 56
7.3.4 Attention-Based Embeddings with Contrastive Learning . . . 56
7.3.5 Attention-Based Embeddings with Adversarial Learning . . . 57
7.4 Results and Discussion . . . 57
7.5 Conclusion . . . 60

8 Conclusions and Future Work **63**
8.1 Conclusion 63
8.2 Future Work 64
8.2.1 Generating all windowing settings samples for each radiograph 64
8.2.2 Experimenting with other training datasets and scanner features 65
8.2.3 Performing a thorough study of the training dynamics from epoch to epoch 65
8.2.4 Using other generative models as sources for disentanglement 65

A Extensive 5-fold Cross Validation Results for VQ-VAE Experiments **67**

References **71**

List of Figures

2.1 Schematics of an X-ray tube . . . 8
2.2 Different windowing settings applied to the same CT scan to highlight different structures. Source: [1]. . . . 9
2.3 High-level representation of encoder and decoder blocks . . . 11
2.4 Architecure of the Variational AutoEncoder. . . . 12
2.5 General Architecture of a GAN. . . . 12
2.6 Architecture of the transformer model, proposed by [2] . . . 15
2.7 Overview of the Vision Transformer architecture, with its patch embedding [3] . 16
2.8 Example of the possibilities of disentanglement in changing a car's characteristics. 17

3.1 Early stopping based on the validation set [4] . . . 20
3.2 Overview of the VQ-VAE architecture. Source [5] . . . 22
3.3 Representation of the hierarchical VQ-VAE for two levels of discretisation [6] . . 22
3.4 Summary of the VQGAN, encompassing the VQ-VAE quantisation controlled by Transformers, reconstructing an image that will be discriminated in an adversarial setting [7]. . . . 23
3.5 The proposed convolutional self-attention module for the Self-Attention GAN [8]. 24
3.6 Contrastive setting applied to pairs of attention maps based on transformations done to the original input image [9]. . . . 25
3.7 Unified Adversarial Invariance model architecture. . . . 26

4.1 Differences between original DICOM (top) and sampled PNG versions (bottom) of the same radiographs. . . . 30
4.2 Random samples from each dataset . . . 31
4.3 Number of positive (green) and negative (red) cases for Atelectasis in each dataset. 32
4.4 GradCAM (top) and Guided BackPropagation (bottom) interpretability heat-maps for the manufacturer ID classifier. Each column represents one image example. Images were randomly selected and will be the same for the following comparations. 36
4.5 GradCAM (top) and Guided BackPropagation (bottom) interpretability heat-maps for the manufacturer ID classifier using a random crop of 140x140. Each column represents one image example. Images were randomly selected and will be the same for the following comparations. . . . 37
4.6 Windowing Settings Clusters . . . 37
4.7 GradCAM (top) and Guided BackPropagation (bottom) interpretability heat-maps for windowing settings classifier. Each column represents one image example. . . . 38

5.1 Architectural procedure of the Equal Probability Loss. The encoding is shared until the latent space. Then, each task is separated into an independent network in a MT setting. . . . 42

6.1 Diagram detailing the blocks involving the Proposed Architecture. 49
6.2 Proposed Training Procedure. The original/sampled image pairs go through the model in each training cycle, generating the environment for contrastive learning and equal probability regularisation. 50

7.1 Original (top) and reconstructed images (bottom) from BRAX testing set. 55
7.2 Architecure for the multi-task scenario using an attention-based embedding space. 56
7.3 Original (top) and Reconstructed (bottom) images from the out-of-distribution datasets. 59

List of Tables

4.1 Atelectasis Prediction AUC Scores in percentage (%) for the four baseline candidates. . . . 34

5.1 5-fold AUC results in percentage(%) - Testing inference for the simple multi-task setting model with no explicit regularisation. . . . 43
5.2 5-fold AUC results in percentage(%) - Testing inference for the simple multi-task setting model with equal probability regularisation. . . . 44
5.3 Average AUC results in percentage (%). Comparison between the baseline model and the two multi-tasking approaches. . . . 44
5.4 Average AUC results in percentage (%). Addition of the average AUC results for the new training dynamics. . . . 44

6.1 AUC scores for Atelectasis in percentage (%) between the baseline encoder and the attention-based model - In-distribution testing. . . . 51
6.2 AUC scores for Atelectasis in percentage (%) between the baseline encoder and the attention-based model - Out-of-distribution testing. . . . 51
6.3 AUC scores for Atelectasis in percentage (%) - Ablation Study. . . . 52

7.1 Average AUC scores in percentage (%) - Comparison between the baseline and the proposed methodologies. . . . 57
7.2 AUC scores for the two best models, in percentage (%) . . . 60

A.1 5-fold AUC scores in percentage (%) for VQ-VAE with Disease-Only Classification 67
A.2 5-fold AUC scores in percentage (%) for VQ-VAE Multi-Task with Shared Embedding Space . . . 67
A.3 5-fold AUC scores in percentage (%) for VQ-VAE Multi-Task with Independent Embedding Spaces . . . 68
A.4 5-fold AUC scores in percentage (%) for VQ-VAE with Attention-based Embedding Space . . . 68
A.5 5-fold AUC scores in percentage (%) for VQ-VAE with Attention-based Embedding Space - Contrastive Regularisation . . . 68
A.6 5-fold AUC scores in percentage (%) for VQ-VAE with Attention-based Embedding Space - Adversarial Regularisation . . . 69

Abbreviations

ALARA	As Low As Reasonably Achievable
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Curve
BRAX	Brazilian labelled chest X-ray dataset
CAGING	Causality-driven Generative Models for Privacy-preserving Case-based Explanations
CNN	Convolutional Neural Network
CT	Computed Tomography
CXR	Chest X-ray
DL	Deep Learning
DICOM	Digital Imaging and Communications in Medicine
ECG	Electrocardiograph
EEG	Electroencephalograph
GAN	Generative Adversarial Network
HIAE	Hospital Israelita Albert Einstein
HU	Hounsfield unit
KL	Kullback Leibler Divergence
kVp	kiloVoltage peak
MIMIC	Medical Information Mart for Intensive Care
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NLP	Natural Language Processing
PACS	Picture Archiving and Communication System
PET	Positron Emission Tomography
ReLU	Rectified Linear Unit
VAE	Variational AutoEncoder
VQ	Vector Quantisation
VQ-VAE	Vector Quantised - Variational AutoEncoder

Chapter 1

Introduction

1.1 Context

Our society is constantly evolving. Throughout the many years of Human existence, we continuously develop tools to help our daily lives. Certain innovations may have a minor impact, while others can potentially transform the lives of billions. In today's modern world, the rapid advance of technology has opened up unprecedented opportunities, especially in healthcare. With the development of diverse machines and systems, we now possess the potential to achieve remarkable levels of wellness and well-being.

Deep Learning (DL) is a relatively new field that emerged from tremendous advances in computational systems. At its core, DL attempts to reverse-engineer aspects of the human brain, tailored for specific use cases and driven by the quest to uncover meaningful correlations in vast pools of data [10]. Fortunately, since we are in a data-driven era, DL is becoming increasingly prevalent in most applications. One example is healthcare, where the need to store patient data [11], such as text reports, imaging studies and diagnostic information, created an extensive database for training models. Once a model can perform a task with reliable performance, it can be a valuable tool for doctors and patients.

Integrating DL algorithms into healthcare fosters a promising market, evaluated at 9.64 billion USD in 2022 and expected to reach 272.91 billion USD by 2030 [12]. One of the main drivers for this trend is the rise in the automation of medical services, strongly promoted by the unprecedented times caused by the COVID-19 pandemic [13]. Often given at hospitals and clinics, medical care delivery drifted to a decentralised system based on telemedicine and remote patient monitoring. The integration of AI systems allowed a decreased workload for medical professionals. The other driver revolves around the strategies around product development, where companies and startups take advantage of the current media attention to promote their new algorithms.

Deep Learning approaches can coordinate with medical care on several fronts [13, 14], such as genomics, analysis of sensorial data and clinical text data. The medical genomics discipline uses AI [15] to predict the 3D structuration of the genome [16], identify the transcription start sites [17,

18], detect DNA methylation [19] and predict genetic expression from genotype data [20]. Sensor-based algorithms use signals acquired from diverse sensors, like the Electroencephalograph (EEG) [21] or the Electrocardiograph (ECG) [22], to extract patterns and make predictions about human motor activity, seizures and emotions, or to diagnose particular diseases, such as Myocardial Infarction or Atrial Fibrillation, respectively. Regarding clinical text data, medical reports can be written by deep neural models using speech recognition-[23], or the available clinical reports can be used for feature extraction and improved disease classification [24].

Clinical imaging is an additional department where most DL applications are integrated with the medical field [14]. It is possible to analyse and extract information from distinct imaging modalities, such as Computer Tomography (CT) [25], X-ray [26], Magnetic Resonance Imaging (MRI) [27], Ultrasound [28] or Positron Emission Tomography (PET) [29]. This analysis is helpful for multiple tasks:

- **Image Segmentation** is a technique that partitions an image into distinct and meaningful regions or objects [30]. This process can identify and isolate different structures within medical images, allowing doctors to accurately delineate anatomical structures, tumours or calcifications for measuring or diagnostic purposes [31, 32].

- **Image Classification** categorises groups of pixels, vectors or even the entire image into one or multiple classes. It can be used for disease classification [33, 24, 27], like in this thesis, or for patient risk assessment.

- **Object Detection** is beneficial for mammography and histopathology, for example, since it finds bounding boxes around objects and classifies them [34, 33]. It can detect and identify cancerous masses or specific cell types.

- **Image Registration** ensures spatial similarity between two or more images. It can spatially combine medical images from different modalities or acquisition times in healthcare [35], useful for precise navigation in image-guided surgery, for example.

- **Image Reconstruction** involves generating new images based on raw data or other images. Medical image reconstruction can improve image quality [36], interchange modalities and enhance diagnostic value.

These aspects collectively encapsulate the majority of medical image analysis solutions using DL. Their profound impact on the healthcare sector empowers medical professionals with essential tools that enhance patient care, streamline diagnoses, and enable personalised treatment plans. By automating routine tasks, DL solutions alleviate pressure on healthcare practitioners, allowing them to focus on more demanding aspects of their work. Furthermore, this automation paves the way for early disease detection, leading to timely treatments with higher success rates and reduced patient impact, ultimately improving the quality of life and reducing costs [14, 13].

That said, it comes as no surprise that DL solutions are in the spotlight, gaining much attention, especially from investors. Recent reports show that global funding for AI startups has surged

by 108% from 2020 to 2021, reaching an impressive 66.8 billion USD. Health-focused companies account for 18% of these fundings [37]. Several news outlets have reported on the collaborations between hospitals and AI companies. For example, Baptist Health System has joined forces with Nuance to incorporate its speech recognition software [38]. Diagnostikum is also utilising the *AI-Rad Companion* software developed by Siemens Healthineers to automate chest X-ray analysis [39].

1.2 Motivation

Despite the immense prospect of AI systems for medical diagnosis and its current popularity, there must be some cautious procedures for integrating these systems into healthcare, especially when dealing with potentially life-threatening situations.

Medical decisions significantly impact patients' quality of life, whether it is in diagnosing their condition or planning their treatment. As a result, medical professionals carry a heavy responsibility as they are solely responsible for any mistakes, even if they were made in a group decision during multidisciplinary meetings [40]. Understandably, any changes to their workflow can cause scepticism among the workforce.

In contrast, AI systems lack sentience and are withheld from any liability for their predictions, whether the model performs mundane tasks, such as distinguishing cats from dogs or sensitive and high-risk predictions of diseases. In this perspective, integrating DL algorithms in healthcare implies sole accountability for the human agent [41]. Thus, for models to be accepted in medical centres, doctors must develop a deep trust in these technologies to vouch for their performance and reliability.

Despite all the development for building trustworthy deep neural solutions with incredible performances, the complexity, data sensitivity, cognitive biases and black-box reasoning behind these models hinder their confidence and adoption in healthcare. The general public also demonstrates signs of distrust, with a survey revealing that about 60% of US adults are uncomfortable with healthcare providers relying on AI [12].

When focusing on reliability, these architectures suffer significant prediction inconsistencies when evaluated in scenarios where the data input slightly changes from the regular distribution [42]. In other words, they do not hold similar generalisation capabilities to humans, and a simple alteration of acquisition parameters of a particular exam could make their predictions meaningless.

The acquisition procedures in medical imaging are not standardised, so different hospitals or medical centres use distinct protocols, resulting in exams with diverging data distribution [43]. This phenomenon also poses a problem to the widespread implementation of AI models in different medical institutions.

We aim to understand the variability in exams of the same modality and create novel techniques to adapt models to generalisation scenarios, improving their reliability and trustworthiness in the

high-stakes landscape of medical diagnosis. This study is inserted in a project entitled Causality-driven Generative Models for Privacy-preserving Case-based Explanations (CAGING) [44]. This is an exploratory research project funded by the FCT - *Fundação para a Ciência e a Tecnologia, I.P.* (Portuguese Foundation for Science and Technology). This project focuses on explainable artificial intelligence, privacy-preserving machine learning and causality.

1.3 Objectives

This dissertation aims to explore Deep Learning techniques to improve Generalisation in distinct distributions of medical data revolving around Disentanglement Representation Learning. For this purpose, we formulate two questions that our work should resolve.

The first question is: "**Which factors of variability influence the disease prediction performance of Deep Learning solutions in out-of-distribution medical data?**". We outline particular tasks that we consider essential for promoting invariability in medical imaging classification and prepare a shared research ground for the upcoming question.

The second question is: "**Which deep neural mechanisms and training procedures promote feature independence for improved Generalisation?**". We investigate current techniques and their intuitions for enforcing the transformation of an image into independent feature sets. These techniques will ultimately be integrated in intricate ways to solve the selected tasks, creating different architectures whose primary goal is maintaining performance across Multi-Centre Medical datasets.

Finally, we endorse a discussion encircling the techniques experimented throughout this Dissertation, aiming to provoke interest in the scientific community for the next steps towards better models in healthcare.

1.4 Main Contributions

The Thesis' main contributions can be summarised as:

- We explore data heterogeneity in chest X-rays sourced from different datasets, selecting features that may be the underlying root for variability. Additionally, we define a set of rules and baseline models to make this research comparable throughout the entire study.
- We investigate the impact of modifying the baseline to be able to perform separate and independent tasks with the hopes of implicitly enforcing feature separation. Subsequently, we promote explicit regularisation for feature disentanglement.
- We propose a novel approach that utilises the previous findings and attention-based contrastive regularisation of pairs of images sampled differently from the same radiograph to promote disease prediction invariance. We submitted a research paper [45] with this work to the Deep Learning Special Session, held as part of the 22nd International Conference on Machine Learning and Applications (ICMLA 2023).

- We evaluate the ability of generative model methodologies to extract meaningful patterns by assessing image reconstruction quality. Then, we use these methodologies as a backbone to infer their strengths and weaknesses when incorporated with the contributions above.

1.5 Dissertation Structure

Besides the Introduction, this dissertation contains seven more chapters:

- Chapter 2 provides background to the concepts that are going to be explored;
- Chapter 3 reveals some literature studies that tackle the same objectives as our work;
- Chapter 4 defines a set of ground rules and assumptions to serve as a control group and Baseline for all the experiments;
- Chapter 5 presents the first experiments on architectures and regularisations to reach disentangled representations;
- Chapter 6 integrates attention mechanisms into the framework coupled with contrastive learning to improve the out-of-distribution performance of the model.
- Chapter 7 explores experiments done to a standard Variational AutoEncoder with and without the previous methodologies and assesses the generative models' disentanglement capability.
- Chapter 8 presents the main conclusions obtained in the work developed under the scope of this Dissertation and motivates the scientific community for further work.

Chapter 2

Background

2.1 Introduction

The previous chapter hints at the main techniques for achieving this dissertation's goals. Deep Learning processes prorate groundbreaking tools across many fields in the modern world, and their robustness in different, out-of-distribution scenarios generally links to the training procedures chosen, predominantly data and model architecture. Thus, selecting proper methods to ensure good generalisation capabilities is pivotal, resulting in more reliable and meaningful Deep Learning models that become more competitive in the promising market of Artificial intelligence applications.

This chapter provides the theoretical foundation for subsequent discussion, diving into all prominent topics surrounding X-rays, Deep Learning modules and Disentanglement Representation Learning.

2.2 X-rays in the modern World

2.2.1 X-ray Characteristics and Generation

X-rays are similar to visible light since they are a form of electromagnetic radiation but differ in two key aspects: energy and wavelength [46]. The "Röentgen light" has higher energy values ($100 - 1,0x10^5 eV$ compared to $1 - 3eV$), inversely corresponding to lower wavelengths ($0.01 - 10nm$ versus $380 - 700nm$)[46, 47]. These lower wavelengths allow the X-ray beam to pass through most solid objects, including the human body, losing a certain amount of energy that depends on the characteristics of the material [48]. This reduction of energy is the foundation of X-ray Imaging: if a body is between an X-ray generator and a detector, the detector will quantify the amount of radiation that hits it, corresponding to the radiation that was not absorbed by the body, resulting in different contrasts along the obtained film. Objects composed of atoms with a high atomic number, such as bone (calcium atoms), will absorb most of the radiation, resulting in high-contrast images, compared to fat, tissues, and air-filled cavities, that generate sparse shades of grey in the detector [49].

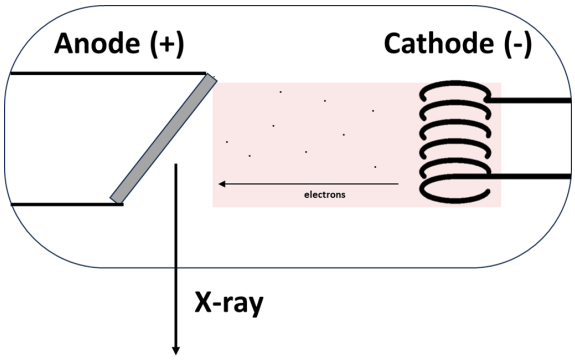


Figure 2.1: Schematics of an X-ray tube

As a general rule, the generation of X-rays occurs in an X-ray vacuum tube, comprised of two principal components: a cathode, usually a filament, that with electric current heats up, releases electrons via thermionic emission, which are accelerated by high voltage and strike the anode. Most of these electrons interact with the anode, slow down and stop, heating the latter and generating X-ray radiation [46, 50]. Figure 2.1 displays a simple example of an x-ray tube.

This radiation does not have constant energy values; it is a distribution of the number of photons created as a function of their energy, designated as the X-ray spectrum [51]. The X-ray spectrum depends on several factors, such as the electric current applied to the cathode, the potential voltage of the X-ray tube, the materials used in the cathode and anode, the detector type, etc. [52]. These factors subsequently influence the final image's contrast, resolution and noise levels. For example, a high potential voltage of 120 kVp (kiloVoltage peak) generally increases exposure, resulting in a brighter final image.

2.2.2 X-rays in Healthcare: Post-Processing

For diagnostic purposes, X-ray radiography uses these electromagnetic waves to check for anatomical abnormalities in our body, namely bone fractures, calcifications, infections, and tumours.

As mentioned above, the characteristics of the X-ray machine influence the final image and need to be carefully defined before each exam, taking into account the "As Low As Reasonably Achievable" principle (ALARA) [53]. This principle arises because X-ray is ionizing radiation that, when absorbed by organisms, can damage molecular structures and cells, promoting a range of effects, such as loss of skin, hair, vomiting or cancer. Therefore, radiographs focus on the best image quality attainable by the lowest possible exposure to X-ray radiation [54]. In other words, this double-edged sword should be appropriately handled by experienced technicians concomitantly with radiologists who specify what they want to obtain in this exam.

Additionally, there are some parameters that the radiologist can change after the radiograph [55]. Using the correct software, the practitioner can adjust the image's brightness and contrast to highlight the wanted areas and provide a wider intensity gamut. One typical setting generally adjusted is the Windowing Level, which separates in a pair of values, Window Center and Window Width, represented in Hounsfield Units (HU) [56], a tissue density unit. The air has a score of -1000 HU;

the lung, being composed mostly of air, scores around -500 HU; fat, water, and soft tissues portray -50 HU, 0 HU and 50 HU, respectively; bone is much more dense, having values of more than 1000 HU [57].

Window Center, or Window Level, corresponds to the midpoint of the range of the Housefield Units displayed. If increased, the final image turns brighter, and vice-versa. Window Width is the gamut itself. A wide Window Width will hold structures with various densities. Thus, the transition from black to white will occur across several compositions, meaning that subtle changes in density will not display observable pixel intensity alterations [58].

Radiologists change the windowing parameters to highlight or differentiate specific regions of interest, such as tissues, bones, or air cavities, each with its variation of ideal intensities. Image 2.2 compares different windowing width ranges applied to the same Computed Tomography (CT) scan.

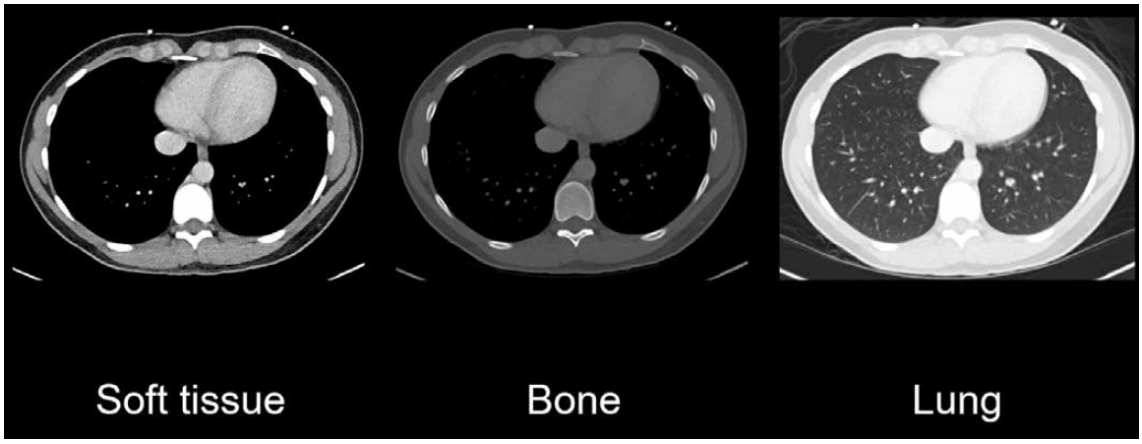


Figure 2.2: Different windowing settings applied to the same CT scan to highlight different structures. Source: [1].

Practitioners store the radiographs after no longer being needed for medical purposes. For insurance and law fulfilment, exams must be archived for at least 5 to 10 years [11]. Violation of this rule can result in hefty fines for the radiologist and the hospital or clinic in question. Therefore, efforts arise to create an efficient device for storing, retrieving and accessing radiographs [55], all medical images and procedures done to a particular patient. Picture Archiving and Communication System (PACS) is the most common device used in healthcare for this purpose [59].

Several formats and standards are available for storing X-ray imaging compatible with the PACS system, but the most widely accepted is the Digital Imaging and Communications in Medicine (DICOM) [60]. DICOM images contain the exam itself and metadata detailing all information about the patient, the image acquisition, the hospital and the vendor.

The climbing prominence of Deep Learning techniques is nurtured by the rising availability of big data. This premise stands on top of the numerous datasets curated for particular tasks. Regarding healthcare, creating datasets using medical information is generally centre-wise and heavily prioritizes patient anonymity to prevent the leak of any sensitive details [61]. The data goes through a particular pipeline of retrieval, storing and clustering that usually ensures that each

entry in the dataset fulfils specific requirements, leading to a high-quality dataset. Nevertheless, these requirements are typically only centre-wise, and while some post-processing techniques are similar, others make datasets from different sources diverge in their data distribution.

One significant difference is the medical image storing format [62]. As discussed previously, the DICOM standard ensures that the radiograph is in its raw form and carries other relevant information to detail the exam made that can explain some of the fluctuations between two different radiographs, for example, the exposure, differential voltage or X-ray tube current. However, saving files in DICOM format is quite demanding since the size is notably more extensive than a simple PNG or JPEG image. Thus, some organisations prefer to discard all details stored in the DICOM's metadata and apply some post-processing techniques to the radiograph, further polarizing the distributions of different datasets.

One of these techniques involves changing the radiograph's grayscale intensities to fit the windowing level chosen by the practitioner. Despite being beneficial for the radiologist at the moment of evaluation, it can sometimes introduce some variability to the data, especially considering that there are several X-ray visualisation software, customarily coupled with proprietary X-ray machines, which use different ranges for the same windowing.

2.3 Foundational Deep Learning

Deep Learning, as a subfield of Machine Learning, inherently drives the trajectory toward Artificial Intelligence. Within its sphere, Deep Learning encapsulates synthetic intelligent algorithms that extract intricate patterns and make decisions using complex networks driven by large volumes of data. These algorithms are heterogeneous and apply to almost all situations and modalities. The subsequent sections delve into some prevalent practices in Deep Learning, portraying some of its capabilities.

The typical Deep Learning architecture comprises several layers, sometimes grouped into blocks or modules. Convolutional, fully connected, batch normalisation and pooling layers constitute the fundamental operations within a deep neural network, often coupled with activation functions, such as Rectified Linear Units (ReLUs) and Sigmoid.

The process of extracting features from an input can be called encoding. Thus, the input gets encoded into meaningful, generally more concise representations called feature maps. For example, throughout the encoding pipeline in the imaging modality, the input gets spacially smaller but gains depth - the layers in an architecture compress the local information into a single pixel. This whole encoding process translates to the encoder block in architectures.

In some cases, the high-dimension feature space obtained by the encoder is called the latent space. The latent space corresponds to an abstract space representing the input's information in its most compact state, usually as a vector. The compact information is the stem for most Deep Learning applications, branching into multiple undertakings, such as classification, regression, or reconstruction tasks.

The antagonist of the encoder block is the decoder block. This new block tries to reconstruct the input using only the latent space. Therefore, the opposite flow happens - the feature maps expand gradually, dispersing concise information locally. For Convolutional Neural Networks (CNNs), the transposed convolution operation substitutes the convolutional layer.

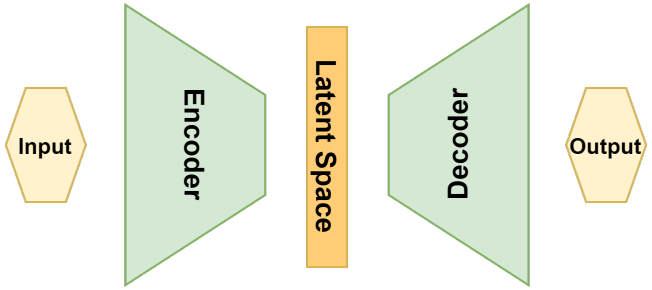


Figure 2.3: High-level representation of encoder and decoder blocks

The preliminaries explained in this section, depicted in Figure 2.3, are the foundations for the following sections that will dive into intricate ways to combine these blocks to generate insightful models.

2.4 Generative Models

Unlike deterministic models, generative models learn to extract meaningful features from the input images in an unsupervised setting. Throughout the training process, these models encode the image into the latent space, which depicts the characteristics of the image in its latent variables. Using a training dataset, the main objective of this model is to find a probabilistic distribution capable of describing the main factors of variation in the images [63]. After a successful training procedure, the model can generate new images following the training distribution.

These models are crucial for several fields, such as natural language processing, computer vision, and speech recognition, holding the potential to automatically augment datasets by providing appropriate new data points. The unsupervised feature extraction settings also hold promising scenarios for promoting feature separability since the latent variables produced can be related to factors of variation of the training distribution. Variational AutoEncoders (VAEs) and Generative Adversarial Networks (GANs) are the most popular generative models and will be further discussed.

2.4.1 Variational AutoEncoders

Variational AutoEncoders are architectures [64, 65] composed mainly of an encoder and a decoder, as portrayed in Figure 2.4. The encoder network maps the input data x into a latent space following a prior distribution. The decoder will take this latent distribution and return a reconstruction into the original data space, $\hat{x}$. One can argue that the latent space could be regularised using single data

points; however, this constricts the model and eliminates the possibility for variance. Therefore, probabilistic distributions are used, such as the Gaussian.

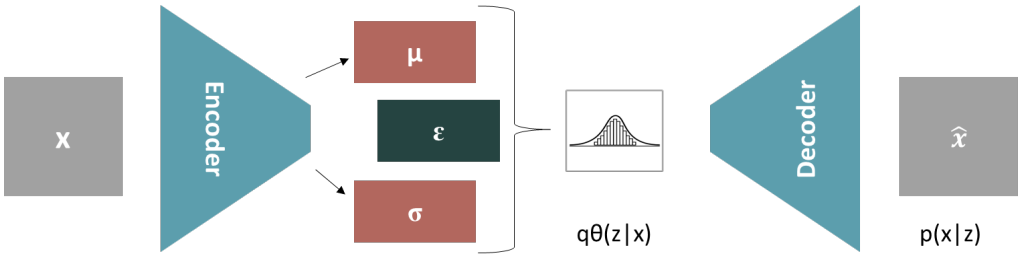


Figure 2.4: Architecure of the Variational AutoEncoder.

The training objective of VAEs uses a reconstruction term, which compares the reconstructed image with the original using mean squared error loss or cross-entropy functions. Additionally, the Kullback-Leibler Divergence (KL) approximates the encoder's distribution ($q\theta(z|x)$) with the original image distribution ($p(z|x)$). Equation 2.1 represents the loss function.

$$L = -\mathscr{E}_{z\sim q_\theta(z|x)}[\log p(x|z)] + KL(q_\theta(z|x)||p(z|x)) \tag{2.1}$$

2.4.2 Generative Adversarial Networks

Generative Adversarial Networks [66] are characterised by a pair of networks that will compete with each other. Unlike VAEs, GANs do not try to estimate the probabilistic data distribution. Instead, it uses an adversarial approach, training two architectures with opposite objectives. The main goal is to achieve an equilibrium between the two networks. One of the networks is the generator, G, which will construct images based on a simple distribution, z. The other network is the discriminator, D, which will infer if the input image corresponds to an original ($D(x)$) or generated ($D(G(z))$) instance. The main objective of the generator is to try fooling the discriminator that simultaneously learns to discriminate better. Figure 2.5 summarises the GAN architecture.

This adversarial training is made possible by a minimax loss, where the generator tries to minimise it while the discriminator maximises it. Equation 2.2 depicts the training objective of a GAN.

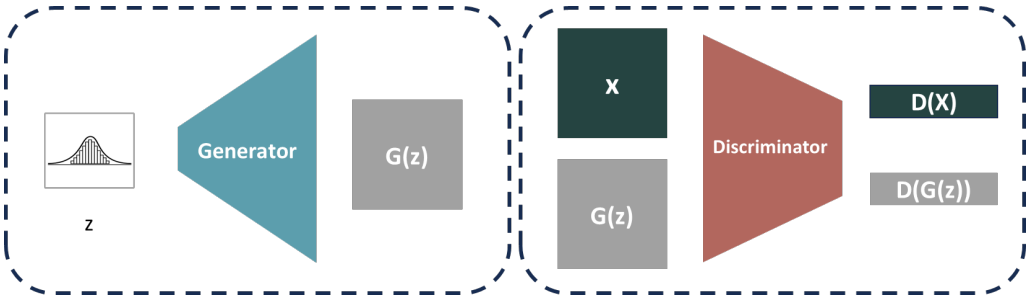


Figure 2.5: General Architecture of a GAN.

$$\min_G \max_D = \mathbb{E}_{x\sim p(x)}[log(D(x)] + \mathbb{E}_{z\sim p(z)}[log(1-D(G(z)))] \tag{2.2}$$

Compared to VAEs, GANs can produce reconstructions with finer details due to the differences in the training objective [67]. The adversarial training is an intuitive procedure to emphasise a high-quality image reconstruction. However, the strict control of the latent space given by VAEs gives an advantage to these networks regarding feature separability.

2.5 The Attention Mechanism

Despite their tremendous success, CNNs keep getting more extensive and comprehensive, contributing to soaring computational costs that scale with the number of image pixels. Mnih *et al.* introduced the attention mechanism [68] to force the model to only focus on certain image regions instead of the entire picture.

The attention mechanism gets inspiration from the human brain. The overwhelming amount of information coming through the retina to the brain would overcome the limited amount of energy available in the brain to process all the neuronal activity involved in processing the visual stimuli [68]. Therefore, selective attention arises where the previously conceived notions about the environment allow the brain to focus on particular locations of the visual apparatus, processing vital information while discarding meaningless stimuli [69]. By combining the perception from different fixations over time, our brain can efficiently build a representation of the surroundings, saving energy resources.

In Deep Learning, the idea behind attention mechanisms is to prioritize the information in particular locations over others to give the model a grasp on what matters in a specific context.

The general attention module [70] takes as input the feature vectors obtained by an encoder, F, and the query, q. The matrix F is where the attention mechanism will extract the most relevant information, guided by the query, that tells the mechanism where to focus.

As the name implies, the query corresponds to an inquiry or a question. Depending on the task, there are several ways to define this matrix - some can use hidden states obtained throughout the encoding process, the model's previous predictions, tabular characteristics, or even a combination of the feature vectors with randomly instantiated and learnable weights [70].

The feature vectors are the starting point for extracting the keys, K, and values, V, matrices. These names correlate with the notion of a dictionary of key and value pairs. As equation 2.3 shows, K and V result from a linear combination of the feature vectors, F, and learnable weights, W_K and W_V.

$$K = W_K \times F; V = W_V \times F \tag{2.3}$$

The end goal of the attention mechanism is to obtain a weighted average of the values vectors, Attention Pooling, constrained by the relevancy of the keys according to a particular query, Attention Scoring [70].

Attention scoring is a technique that outputs the attention scores vector, a, representing the degree of interest of each key regarding the query in question, as defined in 2.4.

$$a_l = score(q, k_l), \tag{2.4}$$

where a_l is the attention score translating the importance of k_l to the query. Several attention score functions are available [71, 70], and some are listed below:

- **Additive** — also known as Bahdanau Attention [72], combines the query and the keys vectors using an addition operation parametrized as a feedforward neural network. The two variables are combined using the following expression 2.5:

$$a(q, k_l) = v^T \times tanh(W_1 \times k_l + W_2 \times q), \tag{2.5}$$

  where v, W_1 and W_2 are learnable weights.

- **Multiplicative** — Being the most common function implemented, it joins the query and keys vectors by multiplying themselves with the help of a weight matrix, W [73].

$$a(q, k_l) = k_l^T \times W \times q \tag{2.6}$$

  If the variables have the same length, one can simplify the expression and use the dot-product operation instead. The multiplicative attention can scale by the factor $\frac{1}{\sqrt{d_k}}$, where d_k corresponds to the length of vector k_l, for vectors with a significant length.

- **Similarity** — Other typical similarity measurements can take place, such as the cosine similarity (equation 2.7) and the euclidean distance (equation 2.8).

$$a(q, k_l) = \frac{q \cdot k_l}{\|q\| \times \|k_l\|} \tag{2.7}$$

$$a(q, k_l) = \sqrt{\sum_{i=1}^{d_k} (q_i - k_{l_i})^2} \tag{2.8}$$

After the softmax operation [74], the obtained vector of attention scores is then used concomitantly to the values vector for the weighted average calculation in the attention pooling step. This context vector is the final step in the general attention mechanism.

2.5.1 Self-Attention

Self-attention is a variant of the attention mechanism that extrapolates the relationship between each vector with the other feature vectors. Every feature vector will go through an attention mechanism, where the attention scoring process gets calculated using that vector's query representation

and the key vectors of itself and the other feature matrices. As a result, each context vector encapsulates the relevance of all feature vectors, offering a comprehensive perspective on the input sequence's internal dependencies and relationships [2].

2.5.2 Multi-Head Attention

The context vector corresponds to one representation subspace, extracted from the query, key and value vectors. To unlock the ability to have different representation subspaces from different positions, Vaswanie *et al.* proposed creating h learned linear projections of the vectors mentioned above, each one of them having its attention mechanism [2]. The h context vectors are calculated parallelly, concatenated and linearly transformed into the expected output dimension. This module is called multi-head attention, and its intuition is to enrich the model's capability to focus on different input positions, akin to an ensemble of attention.

2.5.3 Transformer Architecture

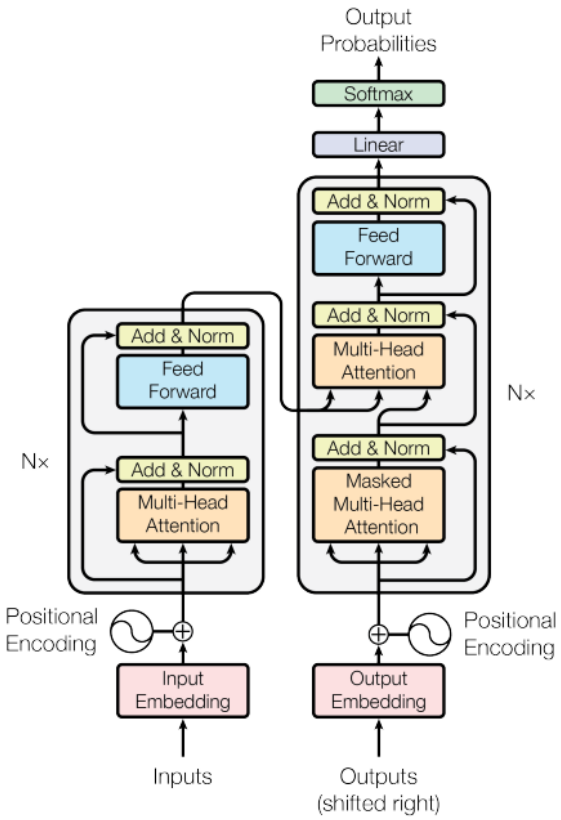


Figure 2.6: Architecture of the transformer model, proposed by [2]

In the landscape of modern Deep Learning, transformer architectures are garnering much acclamation and focus. With its foundation in the self-attention mechanism, capturing intricate patterns and contextual relationships, these architectures allowed a new era of Natural Language

Processing (NLP) tasks and further expanded to computer vision problems.

Unlike CNNs, transformer architectures are solely based on attention modules and consist of an encoder and a decoder. The encoder consists of multiple identical stacked blocks, each composed of two layers: a multi-head self-attention mechanism and a position-wise fully connected feedforward network. Around each layer, there is a residual connection, followed by layer normalisation. The decoder shares the same architecture as the encoder, adding a third layer known as encoder-decoder attention that receives the queries from the decoder's multi-head self-attention mechanism and the keys and values from the encoder outputs. The decoder multi-head self-attention uses outputs of the previous decoding step, shifted to the right alongside a mask that hides the future positions, allowing this new decoding step to only attend to earlier positions.

The position-wise fully connected feedforward network utilizes positional encoding. Each input vector is coupled to a positional encoding vector to give the model information about the vector's order related to the input feature vectors. These positional vectors can be learned or fixed *a priori* based on sine or cosine functions, enriching the model's comprehension of sequence context.

2.5.4 Vision Transformers

The performance scenario regarding Transformers in the field of NLP motivated the adaption of this attention-based architecture to computer vision. Thus, the Vision Transformers (ViTs) emerged, using a transformer-like encoding of patches obtained from the image input. The patch embedding process involves splitting the input image into several patches with a pre-determined size. These patches are then flattened and linearly projected to vectors with a particular dimension, using a convolutional operation with kernel size and stride set to the patch size. The vector dimension is maintained throughout the whole encoding process. Image 2.7 illustrates the ViT architecture. Comparisons between ViTs and state-of-the-art CNNs reveal no clear superiority in

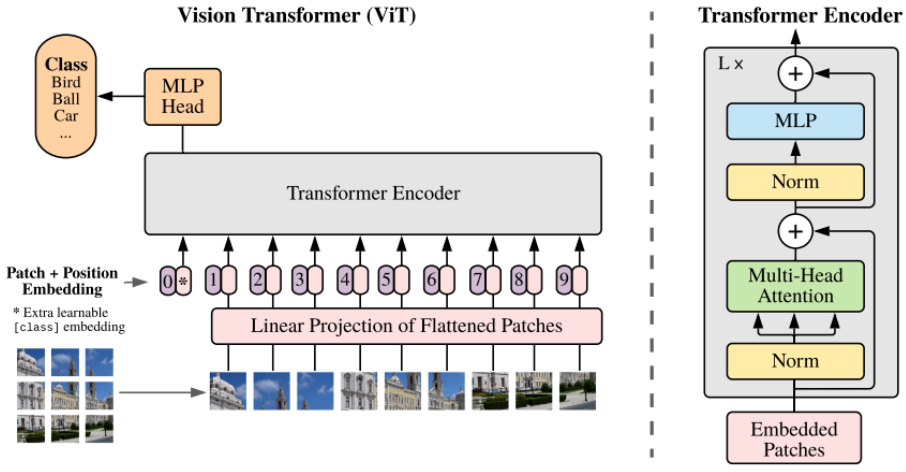


Figure 2.7: Overview of the Vision Transformer architecture, with its patch embedding [3]

one of these networks since both have unique strengths and weaknesses. Observations [75] state that on small-scale datasets, ViTs will have less generalisation ability and worse performance than CNNs. The transformer encoding lacks convolutional principles, like translation invariance and locality. Especially with locality, CNNs take advantage of a high inductive bias, which stands for the set of assumptions made by the model related to the neighbouring pixels of an image. However, in large-scale datasets, ViTs surpass the performance of CNNs since the attention operations are appropriately fitted and can capture global dependencies and contextual understanding.

2.6 Disentanglement Representation Learning

Disentanglement in the context of deep learning is a crucial pursuit to make sense of complex data representations. An image contains several characteristics. When encoding an image to a latent representation, these characteristics intertwine inside this space, which means they are not independent, as they should be (at least in most cases). For example, in a picture representing a car, the car's colour should not be influenced by the chassis shape. These conclusions are clear for human reasoning since we deeply understand colour, shape, and their unarguable separation. Our experiences in this world gave us enough context to completely disentangle the image to its fundamental characteristics. However, DL models do not have this context and have difficulties separating the main factors of variation inside a representation, mainly due to data scarcity. Figure 2.8 illustrates the car example.

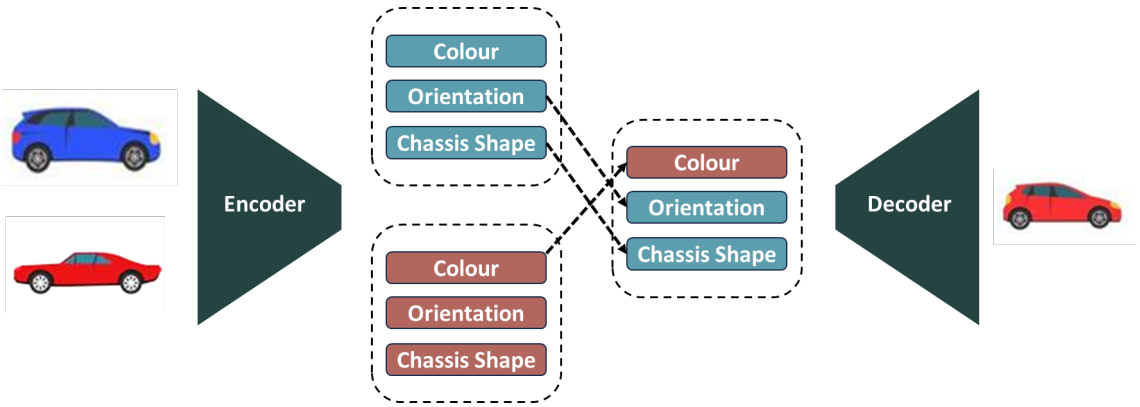


Figure 2.8: Example of the possibilities of disentanglement in changing a car's characteristics.

This is where disentanglement and feature separability come into play: the essence of disentanglement lies in encoding each dimension of the latent space to represent a single, independent feature. This ensures that any modification to a specific feature will not inadvertently affect others.

The standard procedure for implementing disentangled learning is to encode the image into the wanted independent features. These dimensions of the latent representation are subject to regularisation by giving them specific tasks that link them to the wanted features. In the car example, to ensure that a specific dimension is responsible for the chassis shape, it would be advisable to use that dimension's parameters for parallelly classifying the correct shape of the

chassis. Then, using a generative modelling approach, a decoder would reconstruct an image from this latent space. Some dimensions can be modified or omitted in the reconstruction according to the task objective. The final image may differ from the original but still shares most components.

By having independent feature spaces, changes made to one vector will not propagate to the others. That said, in a classification task, promoting feature disentanglement can translate into a better classifier that does not fluctuate its predictive power when presented with variability in the data distribution.

2.7 Conclusion

This chapter has provided an overview of the foundational concepts and techniques that form the basis of our research. We began by exploring the fundamental principles of X-rays and their use in the medical field. Then, we encompassed the main knowledge about Deep Learning, exploring models and techniques that have potential use for this thesis, namely the attention mechanism, generative models and disentanglement representation learning. These concepts motivate the next chapter, which reflects the main literature contributions that built our work.

Chapter 3

Literature Review

The previous chapter introduced the base ground for the concepts that will serve as foundations for this dissertation. In this chapter, we explore essential strategies and innovations that researchers have developed that link the background concepts with the prospective work aimed to be done in this study.

3.1 Generalisability in Deep Learning Applications

Research in improving generalisation can be divided into two main fields: data-centric approaches focus on optimising the training dataset, while model-centric methods refine the architecture and training techniques of neural networks. This section delves into these approaches, highlighting the strategies and innovations contributing to improved generalisation in deep learning.

3.1.1 Data-Centric Approaches

Data-centric approaches for enhancing generalisation in machine learning have gained substantial attention in recent years. These strategies focus on improving a model's performance and adaptability by carefully managing and augmenting the training data. Several key methodologies and studies have contributed to this field:

- **Data Augmentation:** These methods generate new training samples by applying various transformations to existing data. Classic augmentation methods [76, 77, 78] include rotation, translation, scaling, and image flipping. By artificially increasing the diversity of the training dataset, it helps models learn more robust and invariant features.

- **Feature Engineering:** Selection, transformation, or creation of new features from the raw data to improve machine learning models' performance [79]. These techniques play a crucial role in enhancing the generalisation of machine learning models by focusing on the relevance of input features [80]. While traditional techniques involve manual feature crafting and selection, recent advancements [81] in deep learning have shown the potential to

automatically extract relevant features from raw data, further improving model adaptability and performance.

- **Data Splits:** Performing dataset splits comprises dividing datasets into training, validation, and testing sets [82]. The model will only be able to learn its weights by using the training set, selecting the best hyperparameters based on the validation set and assessing performance with the testing set. These splits are essential in enabling accurate performance estimation and model selection [83].
- **Cross-Validation:** By repeatedly partitioning the dataset into training and validation subsets, cross-validation mitigates overfitting and minimises bias in model assessment [84].

3.1.2 Model-Centric Approaches

Pursuing model-centric approaches to improve generalisation in machine learning has resulted in innovative strategies that focus on enhancing the models' ability to adapt to diverse datasets. These techniques, separate from data-centric strategies, involve exploring the model's architecture, regularisation, and training methodologies. Some traditional approaches for improving generalisability are as follows:

- **Early-Stopping:** This is a widely utilised model-centric approach in machine learning, primarily employed to prevent overfitting. It involves monitoring a model's performance on a validation dataset during training and halting the process once the performance starts deteriorating [4]. Figure 3.1 illustrates the intuition behind early-stopping.

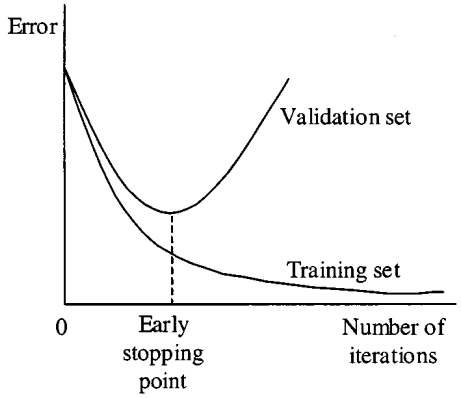


Figure 3.1: Early stopping based on the validation set [4]

- **Dropout:** Introduced by Hinton *et al.*, dropout [85] works by randomly deactivating a subset of neurons during each training iteration, not participating in the prediction. This random element helps the network become more robust and less reliant on any specific set of neurons.
- **Regularisation (L1 and L2):** Some techniques work by enforcing the model to decrease its complexity and confidence in its weights. L1 regularisation, also known as Lasso [86]

regularisation, encourages sparsity in neural network weights by adding the absolute values of weights as a penalty term to the loss function. In contrast, L2 or Ridge [87] regularisation mitigates overfitting by adding the squared weights as a penalty term.

3.2 Generative Models

Generative models have witnessed remarkable advancements, enabling diverse applications such as image synthesis and disentanglement. This section will present promising works that follow the scope of this dissertation, presenting interesting architectural modifications for VAEs and GANs.

3.2.1 Variational AutoEncoders

- **Beta-VAE [88]:** This architecture introduces a disentanglement factor, β, to the KL term in the VAE loss function (equation 2.1, as seen in equation 3.1:

$$L = -\mathscr{E}_{z\sim q_\theta(z|x)}[\log p(x|z)] + \beta KL(q_\theta(z|x)||p(z|x)) \tag{3.1}$$

  This factor helps control the trade-off between reconstruction accuracy and the degree of feature disentanglement. The higher the value for beta, the higher the degree of disentanglement. The beta-VAE model proves effective in learning semantically meaningful factors of variation in data. It is a valuable tool for applications requiring interpretable and controllable latent representations.

- **Vector Quantised-VAE (VQ-VAE):** Oord *et al.* introduce the Vector Quantized Variational Autoencoder architecture [5], a novel approach for learning discrete data representations. In traditional Variational Autoencoders (VAEs), latent representations are continuous and difficult to interpret. The key idea in VQ-VAE is to map continuous data into discrete codes, enabling more interpretable and efficient representations. It achieves this by using an embedding space of discrete latent vectors and training an encoder to map input data to the nearest vector in the embedding space. This approach allows for better disentanglement of features, making it easier to control and manipulate specific attributes in the data. Figure 3.2 illustrates the architecture of the VQ-VAE.

  The loss function in VQ-VAE (equation 3.2) consists of three components. The first term is reconstruction loss, measuring the difference between the input data, x, and the reconstruction, $\hat{x}$, typically using MSE or a similar measure. The second term moves the embedding vectors, e, towards the encoder outputs, $z_e(x)$ using $l2$ loss. The last term is the commitment loss, which makes sure the encoder outputs commit to the embeddings. The stopgradient operator is represented by sg.

$$L = MSE(x,\hat{x}) + \|(sg[z_e(x)] - e)\|_2^2 + \beta \cdot \|z_e(x) - sg[e])\|_2^2 \tag{3.2}$$

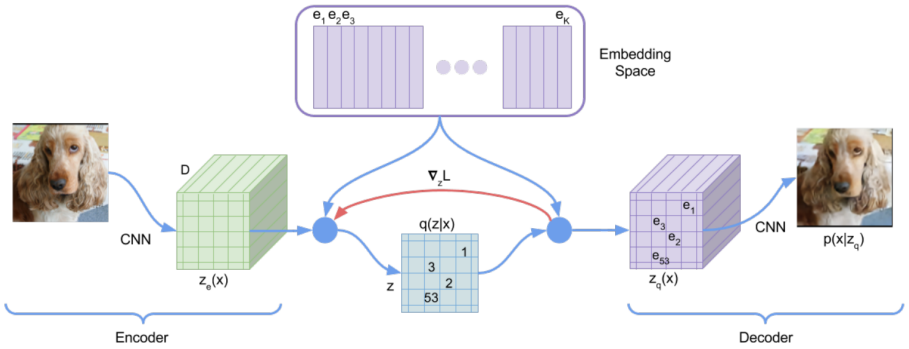


Figure 3.2: Overview of the VQ-VAE architecture. Source [5]

- **VQ-VAE 2:** This is a more complex and evolved version of the VQ-VAE, with multiple quantisation levels structured hierarchically. The image input is quantised into several levels of embedding spaces, gradually into higher-level discretised representations [6]. This hierarchy enables capturing complex features at different levels. Figure 3.3 demonstrates the hierarchical levels present in the VQ-VAE 2. It tends to outperform VQ-VAE regarding reconstruction quality, disentanglement of features, and generation of diverse and high-fidelity images.

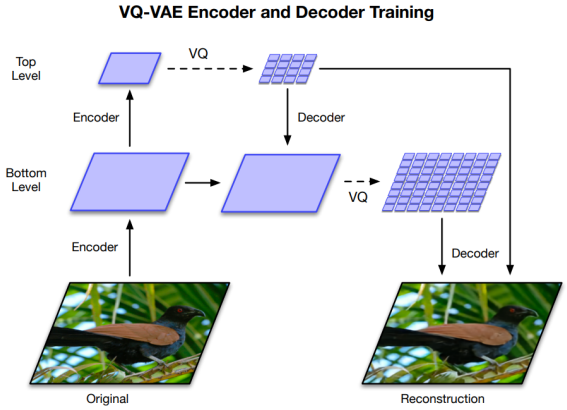


Figure 3.3: Representation of the hierarchical VQ-VAE for two levels of discretisation [6]

3.2.2 Generative Adversarial Networks

- **InfoGAN [89]:** This model extends the traditional GAN architecture by introducing a mutual information regularisation term, encouraging the generator to learn generative and informative representations about specific data attributes. The generator receives as input random Gaussian noise, z (as seen in the typical GAN), and latent code c, which is initialised as a random distribution. However, the latent code will learn to have meaningful representations throughout training. These meaningful representations are created by adding a fully connected layer to the discriminator, denoted as the auxiliary classifier, Q. This fully

connected layer predicts the distribution $Q(G(z))$, given the generated data. Thus, the mutual information term corresponds to the KL divergence between $Q(G(z))$ and c. λ controls the trade-off between the GAN loss and the mutual information regularisation. The overall training objective is represented in equation 3.3.

$$\begin{aligned}\min_{G,Q}\max_{D} = \mathbb{E}_{x\sim p(x)}[\log(D(x))] + \mathbb{E}_{z\sim p(z)}[\log(1-D(G(z)))] \\ -\lambda \cdot KL(c, Q(G(z)))\end{aligned} \tag{3.3}$$

InfoGAN enables the unsupervised discovery of meaningful and independent features within the data, making the learned representations more interpretable.

- **VQGAN:** This work is inspired by VQ-VAE, GAN and Transformers. GANs, with their adversarial training paradigm, excel at generating realistic images, while VQ-VAEs facilitate disentangled and structured representations.

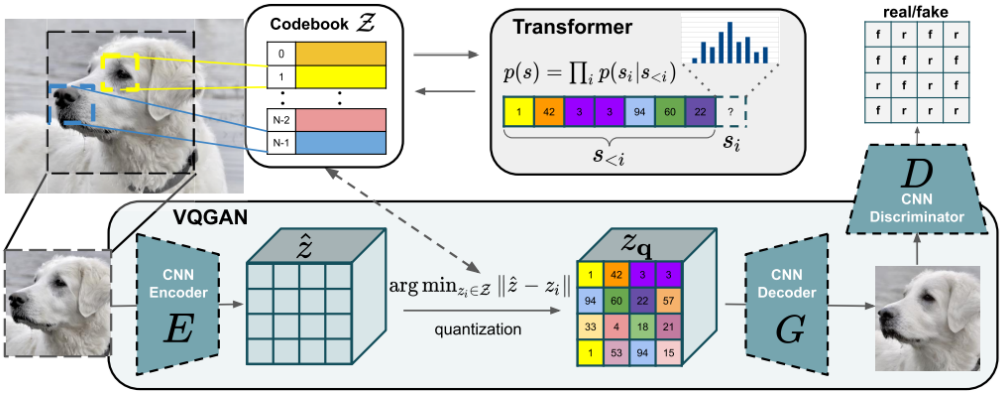


Figure 3.4: Summary of the VQGAN, encompassing the VQ-VAE quantisation controlled by Transformers, reconstructing an image that will be discriminated in an adversarial setting [7].

This article incorporates both training procedures [7], using a modified VQ-VAE objective that evaluates the reconstruction using a discriminator network in an adversarial setting. The transformers regulate the embedding space by modelling a sequence of these discrete latent variables. The quantised encoding of an image is represented by a sequence of embeddings in the embedding space. Thus, by applying the self-attention mechanism in the Transformers, one can enforce the learning of complex dependencies and patterns between the embeddings. This way, the model learns high-level representations of the input data, allowing it to generate higher-quality and fidelity images. Figure 3.4 summarises the VQGAN architecture.

3.3 Attention-Based Mechanisms

This section explores two intricate ways of using the concept of attention modules to induce the desirable effects on the training dynamics:

- **Self-Attention GAN:** Zhang *et al.* introduce the convolutional attention module [8], which employs a mechanism similar to the self-attention mechanism but adapted for convolutional neural networks. This work is applied to GANs for improved image generation. Convolutional operators of kernel size and stride of 1 map the input into query, key and values feature maps. Then, the same principle in attention is applied, and the output of the convolutional attention module represents the input focused on the obtained query. Figure 3.5 illustrates the scheme of this implementation.

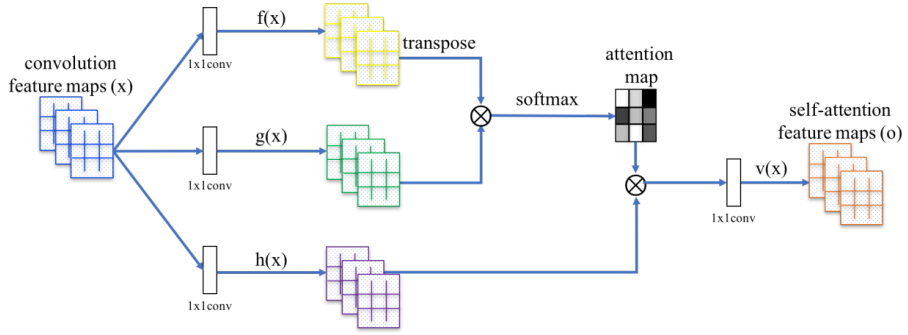


Figure 3.5: The proposed convolutional self-attention module for the Self-Attention GAN [8].

This module allows each pixel or feature map location in the generated image to attend to distant regions within the same image efficiently, capturing long-range dependencies and improving the coherence and quality of generated images. Using convolutional attention modules significantly facilitates the incorporation of the attention mechanism in CNNs.

- **Constrastive Attention Maps:** The authors propose a method based on creating a self-supervised setting using contrastive attention maps [9] to help identify and localise specific objects or features within an image without relying on explicit annotations. This method encodes multiple views of the input image, such as rotated transformations, and performs attention pooling on the obtained feature maps.

  A transformation T is applied to the original image to obtain the transformed image. Then, the model generates the attention maps for the original (A_{ori}) and the transformed input images (A_{trans}). By applying the transformation T to A_{ori}, we obtain $A_{ori2trans}$. The inverse procedure happens to A_{trans}, generating $A_{trans2ori}$. A_{ori} and $A_{trans2ori}$ are positive pairs, such as A_{trans} and $A_{ori2trans}$. The negative pairs are the backgrounds from the A_{ori} and A_{trans}. Figure 3.6 elucidates these transformations.

$$
\begin{aligned}
L = \mathbb{E}_x \Big(& \max(\|A_{trans2ori} - A_{ori}\|^2 - \|A_{trans2ori} - A_{ori2bg}\|^2 + m, 0) \\
& + \max(\|A_{ori2trans} - A_{trans}\|^2 - \|A_{ori2trans} - A_{trans2bg}\|^2 + m, 0) \Big),
\end{aligned}
\tag{3.4}
$$

where m indicates the margin. Following Equation 3.4, these attention maps are regularised in a contrastive setting, maximising the similarity between positive pairs and minimising

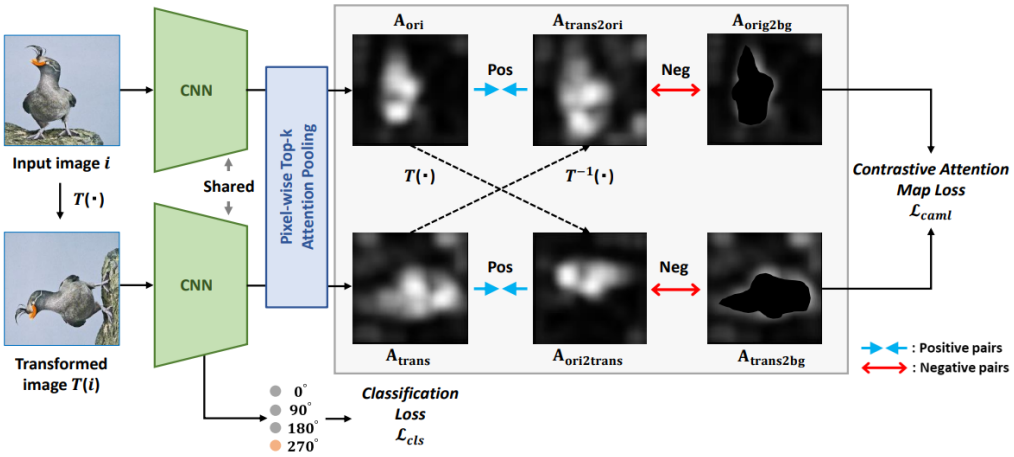


Figure 3.6: Contrastive setting applied to pairs of attention maps based on transformations done to the original input image [9].

that between negative pairs. This objective encourages consistency between the attention maps before and after the transformation of the input image. Also, it penalises the attention maps of two anchors being activated in backgrounds. In essence, this approach enables the network to focus on informative image regions while suppressing less relevant areas, significantly boosting the accuracy of self-supervised co-localisation tasks. This contrastive setting can be easily adapted to other scenarios, proving its promising uses.

3.4 Disentanglement Representation Learning

Disentanglement refers to the process of disentangling and separating latent factors of data representation, fostering the creation of interpretable and factorised models. The following techniques enforce disentanglement representation learning using an adversarial approach:

- **Adversarial Disentanglement:** This framework promotes independence from variability factors within data, ultimately facilitating fair predictions [90]. The main objective involves competitive training between a prediction, *Pred*, and a reconstruction, *Dec*, task.

  As seen by Figure 3.7, this involves encoding x ($Enc(x) = e$) and learning a split representation of data as $e = [e_1, e_2] = [Enc(x)_1, Enc(x)_2]$, such that information for the prediction task is pulled to e_1. In contrast, information for reconstruction goes to e_2. Then, two adversarial *disentanglers*, Dis_1 and Dis_2 are incorporated into the network. While Dis_1 aims to predict e_2 from e_1, Dis_2 does the inverse. If e_1 and e_2 are genuinely independent, it would be impossible for these *disentanglers* to achieve their goal. Equation 3.5 resumes the training objective.

$$\begin{aligned}\min_{Enc,Pred,Dec}\max_{Dis_1,Dis_2} &= \alpha L_{pred}(y, Pred(e_1)) + \beta L_{dec}(x, Dec(e_1, e_2)) \\ &+ \gamma\{L_{dis_1}(e_2, Dis_1(e_1)) + L_{dis_2}(e_1, Dis_2(e_2))\},\end{aligned} \tag{3.5}$$

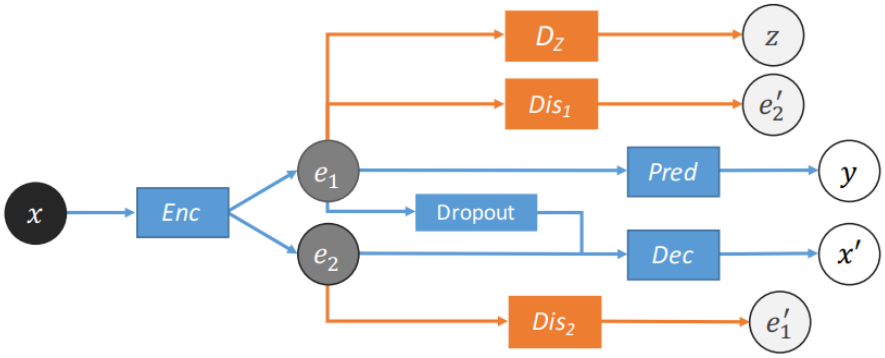


Figure 3.7: Unified Adversarial Invariance model architecture.

where L_{pred} and L_{dec} are the losses for the prediction and image reconstruction tasks, respectively, and L_{dis_1} and L_{dis_2} are MSE losses. α, β and γ are coefficients to control the trade-offs between the regularisations applied. This approach successfully proposes disentanglement to any supervised learning setting, leading to more robust and generalisable predictions.

- **Disentangled Representation Learning for Privacy** Montenegro *et al.* introduce a model [91] capable of disentangling identity and medical features from images, allowing the generation of privatised explanations to justify the model's decision. This framework comprises a generative, an identity and an explanatory module. The generative module is a GAN responsible for generating the anonymised image. The identity module is a pre-trained identity features extractor, acting in the generated image, promoting anonymisation. The explanatory module is a pre-trained medical features extractor and acts in the generated image, preserving the medical features relevant for serving explanatory evidence for the model's decision. The feature extractors are trained to promote invariance to the opposing task. For example, the medical feature extractor, F_{ext}, is identity-invariant since it was trained in an adversarial setting so that a disease classifier, C_{dis}, can recognise the disease of the image. However, an identity classifier, C_{id}, cannot identify the patient. The training objective to promote the invariance is defined in Equation 3.6.

$$L_{F_{ext}} = \mathbb{E}(-y_{dis} \cdot log(C_{dis}(F_{ext}(I))) + y_{id} \cdot log(C_{id}(F_{ext}(I)))), \tag{3.6}$$

where I is the input image and y_{dis} and y_{id} are the ground-truth labels of the images in regard to disease and identity recognition, respectively. The results obtained translate the model's ability to generate privatised images, meaning that the adversarial training successfully separated the disease from the identity features. This framework is promising for promoting disentanglement in order to achieve invariability in model predictions.

3.5 Conclusion

Our comprehensive literature review has delved into research development in generalisation, generative modelling, attention mechanisms, and disentanglement techniques. These components are essential for this dissertation and will be explored in our methodology. The following chapter begins this dissertation's experiments by delving into the assumptions and choices made in the preliminary work.

Chapter 4

Under The Hood: Data Choices and Preliminary Implementations

4.1 Introduction

The paths to achieve the ultimate goal of this dissertation are multifold and need to be better defined. Disentangled representation learning is generally hard to demonstrate quantitatively, and coupling it with generalisation may yield unexpected results. While the theory and all work done on synthetic datasets show promising results, the implementation in real-world scenarios is somewhat troublesome since there are too many variables to control; hence, disentangled representations may not project better generalisation.

The paths chosen for all the work are not unique and were picked based on continuous forethoughts, so some trajectories are arguable. Throughout the following chapters, we elucidate the reasoning behind the elected methods. Therefore, this chapter serves as an introductory work for this dissertation, with the sole objective of defining and explaining choices around datasets and ground rules for the upcoming implementations.

4.2 Datasets

Evaluating generalisability is a tricky task, and there are several ways to do it. However, since the central intuition behind a generalisable model is shared performance traits between the training and out-of-distribution data, we define that we aim to improve the out-of-distribution scores of a model compared to its baseline.

Chest X-ray is the chosen imaging modality since there is a plethora of publicly available data, and the main task is binary disease classification. Due to time and computational limitations, using all the available datasets was impossible, so we determined four datasets, one for training and the others for out-of-distribution testing. The principle behind this selection was to have reputable datasets and some outliers from different regions worldwide. The selected datasets, detailed in the coming subsections, are BRAX, CheXpert, MIMIC-CXR and VinDr-CXR.

4.2.1 BRAX

The Brazilian labelled chest X-ray dataset (BRAX) [92] is an automatically labelled dataset containing 40,967 images from 24,959 radiography studies of 19,351 unique patients. The radiographs were extracted from the PACS of Hospital Israelita Albert Einstein (HIAE) in Brazil, and experienced radiologists reviewed the fourteen generated labels. One critical particularity of this dataset is that it comprises a manufacturer ID for the X-ray machine.

BRAX contains the radiographs in both DICOM and PNG format. While processing the images in PNG format, the authors interpolated the pixel values according to the windowing level used by the radiologist. In other words, the images in DICOM format may differ from those in PNG format since they have distinctive brightness and contrast. Figure 4.1 compares some samples between the original DICOM images and PNG interpolated ones.

Figure 4.1: Differences between original DICOM (top) and sampled PNG versions (bottom) of the same radiographs.

This dataset will be used for training and validating all the models since it has useful information for a classification task responsible for regularising the scanner features. In Section 4.5, we validate this task by selecting a valid task for the scanner features.

4.2.2 CheXpert

CheXpert [93] is a public chest radiograph dataset, with 224,316 radiographs of 65,240 patients from Stanford Hospital in the United States of America. It was one of the first large chest X-ray (CXR) datasets publicly available. CheXpert radiographs are represented in PNG format, and the authors did not provide information regarding the post-processing methods.

4.2.3 MIMIC-CXR

Medical Information Mart for Intensive Care (MIMIC)-CXR comprises 377,110 images of 227,835 studies from 65,379 Beth Israel Deaconess Medical Center Emergency Department patients between 2011 and 2016, together with free-text clinical reports [94]. This dataset from the United

States of America is one of the most extensive and popular CXR datasets publicly available and provides radiographs in DICOM and JPG formats. The JPG images are interpolated from the DICOM originals without using the windowing levels.

The DICOM metadata gives insightful information about the X-ray machine. However, this dataset was not used for training purposes due to its size: the DICOM-format version occupies around 5.5 Terabytes, which would take a substantial amount of time to transfer and process, and its storage would be costly.

4.2.4 VinDr-CXR

The last dataset consists of 18,000 images manually annotated by 17 radiologists with 22 labels, collected from Hospital 108 and the Hanoi Medical University Hospital in Vietnam [95]. It encompasses scans from a different region, which may introduce heterogeneity and difficult generalisation.

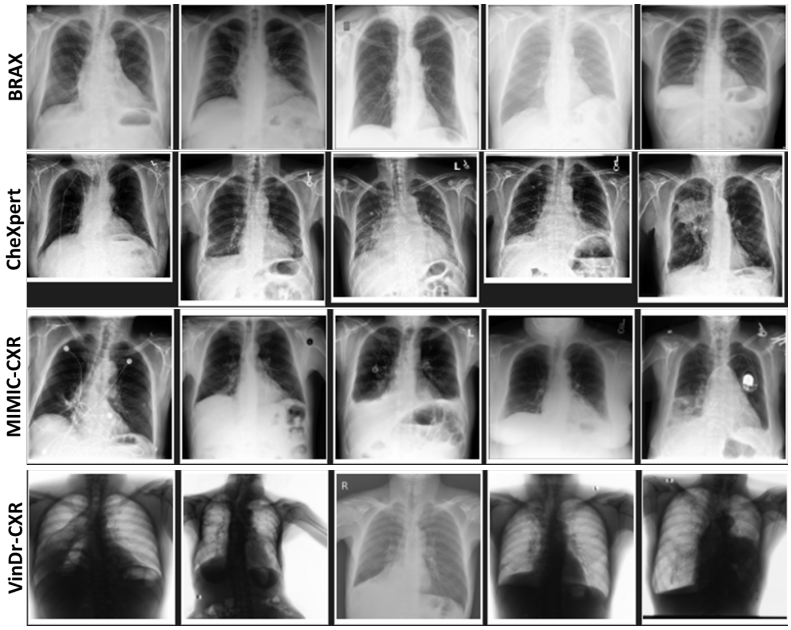


Figure 4.2: Random samples from each dataset

As shown in Figure 4.2, the radiographs from these datasets differ substantially. These differences cause models to underperform in alternative datasets of the same modality. After specifying the implementation details in section 4.3, we will demonstrate the first results for baseline models to demonstrate the performance drops.

4.3 Implementation Details

4.3.1 Disease Task

Atelectasis is described as the collapse of the lung tissue caused by a blockage of the air pathways or pressure on the lungs [96]. CXR is a helpful tool to diagnose this disease, and Atelectasis is shared among the four datasets used for this work. There are other concomitant diseases in the datasets, but Atelectasis has more positive cases in the training dataset. Figure 4.3 describes the number of samples for positive and negative cases of Atelectasis for each dataset.

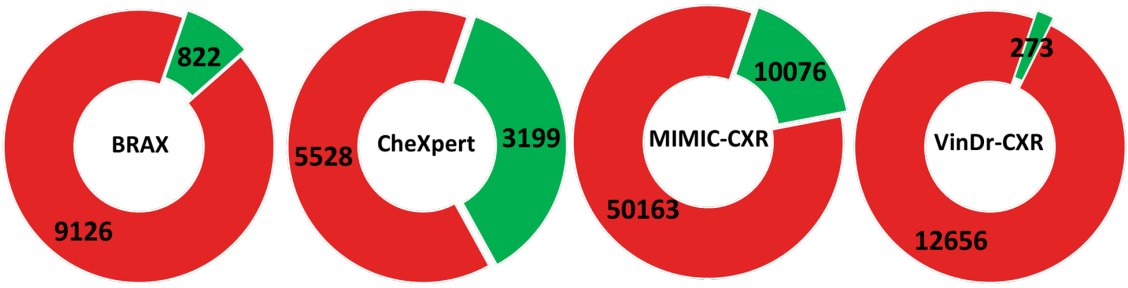


Figure 4.3: Number of positive (green) and negative (red) cases for Atelectasis in each dataset.

4.3.2 Data Sampling

A quick analysis shows that there is a significant data imbalance. Positive cases for Atelectasis correspond only to around 10% of the total training images. Using a pre-trained architecture resulted in a biased model that would only predict negative instances. To tackle this problem, we implemented a sampling procedure given by Pytorch's library, *WeightedRandomSampler*, that randomly samples the images according to their weighted distribution. In other words, this method selects an equal distribution of positive and negative cases at each minibatch, oversampling the minority class while undersampling the majority class. Throughout the epochs, all the negative examples should be selected, and to prevent overfitting due to repeated positive images, Data Augmentation took place, namely random crops, translations and small rotations. The validation and testing set distribution sampling was the original, and no data augmentation occurred. Images were resized to 256x256 to increase computational efficiency.

4.3.3 Metrics

The metrics are an indirect way to measure the model's ability to perform a particular task. Its choice heavily impacts the overall statements one can make about a model's performance, especially when comparing it against others. This subsection explains the insights behind the metrics chosen for evaluating and comparing all models.

Typical evaluations use Accuracy, F1-Score and Area Under the Receiver Operation Characteristic (AUROC or AUC) for binary classification tasks [97]. Accuracy projects the percentage of

correct predictions without distinguishing each class's prevalence. It is one of the most used empirical measures, but it can be meaningless in medical applications or scenarios with data imbalance. Thus, this work will discard accuracy scores. F1-score is a metric that focuses on the positive class and evenly balances the estimation of correctly classified examples and the misclassified ones (recall) and the rate of true positives against samples misclassified as positives (precision). AUC is the area under the ROC curve that plots the true positive rate against the false positive rate at various threshold values. These thresholds define the probability limit between the negative and positive prediction. In the end, this metric measures the ability of a model to distinguish classes and is widely used in learning with imbalanced datasets.

The initial thought in this work was to use the F1-Score. However, this metric's feasibility for this task was discussed after poor initial performance. Many pieces in the literature share a common approach, using the AUC score as the evaluation metric for objectives similar to ours, with no mention of the F1-Score. Furthermore, specific articles do disclose the F1-Score; however, upon examination, it becomes evident that the F1 scores in these cases are notably low and comparable to the values obtained in our work. After some deliberation, we will use the AUC as the central evaluation metric for all further discussion and comparison between models. For medical diagnostic purposes, AUC represents a model's ability to understand divergences between the positive and the negative classes and the degree of overlapping information. In contrast, the focus on the false negatives and false positives by the F1-Score can induce the model to a high state of criticism in more complicated cases where the certainty of a prediction is arguable. For example, some radiograph studies are labelled as positive by the radiologist after additional information or exams since the presence of the disease was dubious.

4.3.4 Training and Optimization

The hyperparameters across all implementations vary, depending on the task and complexity of the idea explored. This subsection discloses the base hyperparameters, and any variations to these values are referenced in each corresponding chapter. Models were trained with a batch size of 16, trying to reduce the binary cross entropy loss value. Adam optimiser [98] is used for updating the weights of the model's parameters, with an initial learning rate of 1×10^{-4} that decays by a factor of 0.1 after the plateauing of the AUC metric, using Pytorch's *ReduceLROnPlateau* object. After 50 epochs of training, the checkpoint of interest is the one that portrays the highest AUC score. We use 5-fold cross-validation for the implementations discussed in the coming chapters; in this chapter, only one fold was used since it was for baseline comparisons and definitions.

4.4 Baseline Selection

Research begins with the definition of a baseline, serving as the comparator against all proposed changes in this dissertation. This baseline should be tailored to the main objectives of the work, being fair and straightforward so that no unwanted disturbances affect the results. Thus, we trained four disease classifiers for Atelectasis using the BRAX dataset and conducted out-of-distribution

testing. Three were well-established pre-trained frameworks, while the last was a custom encoder. Each network is detailed in the following paragraphs, culminating in presenting performance results and nominating the Baseline architecture.

The first two models are ResNet-18 [99] and DenseNet-121 [100], the shallowest variations of some of the most used CNNs for Deep Learning, hence being strong candidates for the baseline network. Recalling the discussion in section 3.1, we use Pytorch's implementation of the pre-trained superficial variations with fewer parameters to reduce overfitting and improve generalisability.

The third model is based on the ViT architecture, discussed in subsection 2.5.4. The attention-focused candidate is a solid choice to compare performance between Transformers and CNNs, the main foundations of deep networks. Furthermore, this network is one of the top performers on the popular image classification benchmark, ImageNet [101]. We used the *ViTModel* framework made available in the HuggingFace's library.

Finally, the last model is a simple custom encoder, with one initial convolutional layer followed by four ResNet's basic encoding blocks. The primary motivation behind this encoder was to create a more controllable scenario for better discussion throughout the experiments taking place in the following chapters. Furthermore, the simplicity of this encoder allows straightforward implementations for further tests. We used the minimum number of basic encoding blocks that achieve similar in-distribution results compared with the other networks.

Table 4.1 displays the in and out-of-distribution AUC scores of these networks.

Table 4.1: Atelectasis Prediction AUC Scores in percentage (%) for the four baseline candidates.

	In-Distribution	Out-of-Distribution		
Model	BRAX	CheXpert	VinDr-CXR	MIMIC-CXR
DenseNet-121	85.36	83.01	**69.81**	**82.64**
ResNet-18	**87.22**	**83.32**	65.09	81.53
Custom Encoder	85.78	75.50	59.04	72.00
ViT	75.01	66.29	63.22	62.63

This early analysis reveals similar results in the in-distribution test set across three models, with the ViT underperforming; hence, fitting an architecture to extract meaningful patterns for predicting the disease in the BRAX dataset is possible, and most architectures can do it. However, the out-of-distribution results show a substantial difference between DenseNet-121, ResNet-18 and the others. The pre-trained CNNs achieve much better results than the ViT and the custom encoder.

4.4.1 CNNs vs. Transformers

Comparing the three architectures proposed in the literature, the CNN models outperform ViTs considerably. Looking at the number of parameters of each network, a particular justification arises: the number of parameters of the ViT is significantly higher than the CNNs. While ViTs

achieved better scores for the ImageNet [101] benchmark, it has been shown that these networks fail to propagate the local relations to the bottom layers for specific tasks with low data availability [75].

The convolutional operations in CNNs project filters across local regions of an image, creating the assumption that neighbouring pixels relate to each other [102]. This assumption translates into higher spatial inductive bias than with ViT, which focuses more on global dependencies and relationships between different patches of the image (neighbours and distant). Ultimately, the high inductive bias of the CNNs leads to better generalisation performance compared to the ViT architecture when there is data limitation [103]. However, this high inductive bias is only sometimes advisable since it can lead to overfitting [104].

4.4.2 CNNs vs. custom Encoder

The custom encoder also underperforms in out-of-distribution testing compared to the literature CNNs. The complex relations between each layer and blocks of the very deep CNNs allow these architectures to learn higher-order dependencies and biases that can help with their generalisation ability.

4.4.3 Final Remarks

ResNet-18 and DenseNet-121 can effortlessly solve the disease classification task and generalise well, to a certain degree. However, for the sake of this work, we chose to select the custom encoder as the baseline. While this decision may contradict the results obtained, some valid arguments elucidate our reasoning, such as:

- The custom encoder is a flexible network, easily adjustable for all approaches;
- There are no *a priori* network assumptions to overshadow the impact of the modifications;
- The custom encoder has similar performance for in-distribution inference;
- It reflects the real-world problems in AI companies that usually use shallower architectures for cost savings.

4.5 Scanner Features Evaluation

The process of disentangling characteristics begins with defining the tasks that interpolate to these features. The choice of task is crucial as it must relate accurately to the traits we wish to distinguish and be entirely separate from the other task. This subsection analyses possible side tasks and selects the elected one to regularise disentanglement for better disease classification generalisation.

The first task corresponds to the already-defined disease classification, the model's primary goal. In order to make this task's performance invariant to different datasets, one should pick patterns that distinguish CXR images from the data pool. Figure 4.2 clearly depicts the discrepancy

between the radiographs from other distributions, raising the assumption that this variability is related to the X-ray scanner. Therefore, the second task will correlate to a certain extent to the scanner variability. The following segments discuss two possible classification tasks explored to promote scanner feature regularisation.

4.5.1 Manufacturer ID classification

As mentioned in subsection 4.2.1, the BRAX dataset identifies the manufacturer from the X-ray machine. Different manufacturers tune their X-ray scanners in distinctive ways, leading to variability in the final image. Thus, the manufacturer ID classification task seems a good contender for the scanner features regularisation.

There are five different manufacturers, each assigned to an integer to promote anonymity. We deployed a classifier with a DenseNet-121 backbone to evaluate the pattern extraction capabilities this task encouraged. This classifier performed well, with a testing accuracy of 99.75%, so meaningful image patterns resonated with the scanner manufacturer. We generated GradCAM and Guided BackPropagation interpretability maps using MONAI's library [105] to understand the regions the model focused on, displayed in Figure 4.4.

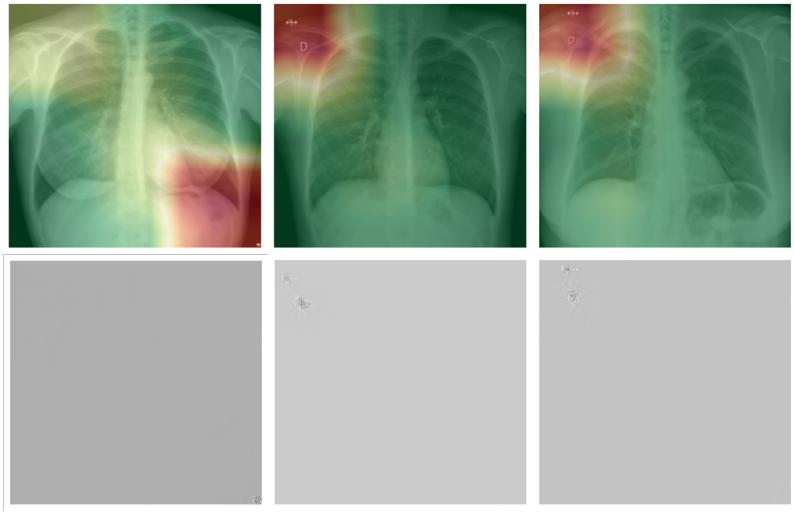


Figure 4.4: GradCAM (top) and Guided BackPropagation (bottom) interpretability heat-maps for the manufacturer ID classifier. Each column represents one image example. Images were randomly selected and will be the same for the following comparations.

Evaluating the GradCAM interpretability maps, the model focused predominantly on some details and watermarks outside the region of interest of the radiograph. The Guided BackPropagation maps have minor activations only on the watermarks. We can say that this classification was out of scope regarding the disease classification task. To prevent this behaviour, we added $140x140$ cropping in random locations of the image to the training data augmentation. The performance dropped to 97.04% accuracy, but the GradCAM interpretability maps shown in Figure 4.5 indicate a better pattern extraction by the new classifier, meaning that performing random crop to the images for ID manufacturer classification induced the model to focus on regions-of-interest

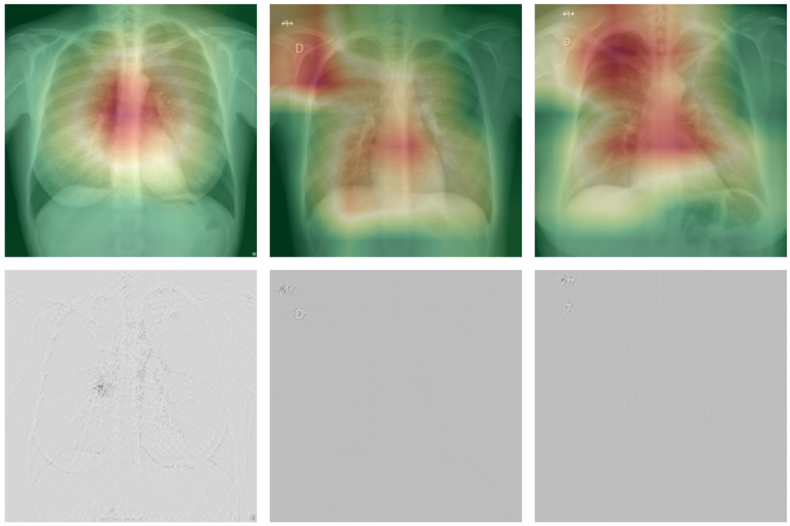
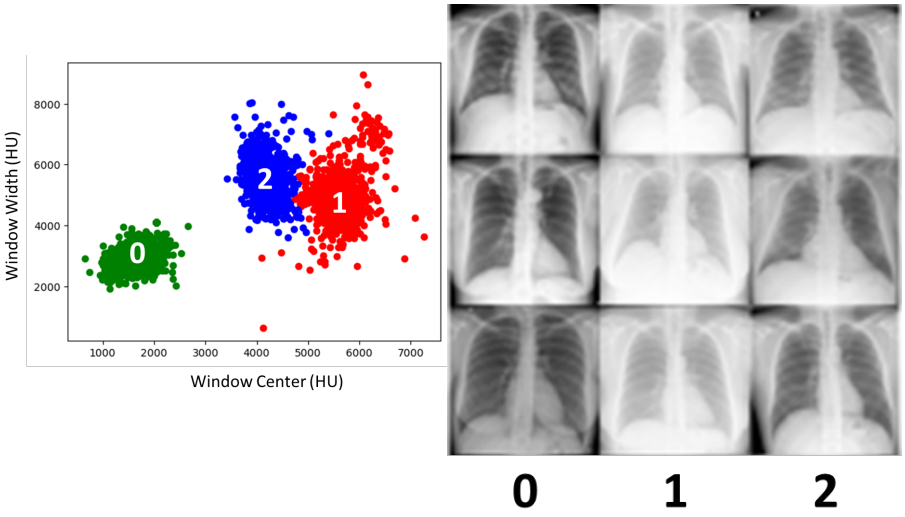


Figure 4.5: GradCAM (top) and Guided BackPropagation (bottom) interpretability heat-maps for the manufacturer ID classifier using a random crop of 140x140. Each column represents one image example. Images were randomly selected and will be the same for the following comparations.

more compatible with the disease task. However, the Guided BackPropagation maps are still not ideal.

4.5.2 Windowing Settings cluster classification



(a) Window Width versus Window Level plot.

(b) Image Examples from each cluster. Each number represents a cluster created.

Figure 4.6: Windowing Settings Clusters

In Section 2.2.2, we delved into radiologists' techniques to refine specific tissues of interest in radiography analysis using windowing settings. The PNG images from the BRAX dataset

were sampled considering these settings, meaning the radiographs vary in contrast and brightness levels. Given that the Window Center and Window Width values are available in the metadata, the idea was to use this information as the scanner features indicator. However, the values for the windowing settings have a broad range, and a linear regression model may not be stable for coupled training with a classifier. To address this, we plotted the Window Width against the Window Center in Figure 4.6a and determined that a cluster classification approach would be more suitable.

We generated three clusters for windowing settings, and the differences between the radiographs of each group are illustrated in Figure 4.6b.

We obtained similar predictive performance after implementing a classifier with the same backbone as the previous subsection. However, the interpretability maps shown in Figure 4.7 translate a broader perception of the radiograph, compared to the last task, highlighting the changes in contrast and brightness throughout each image.

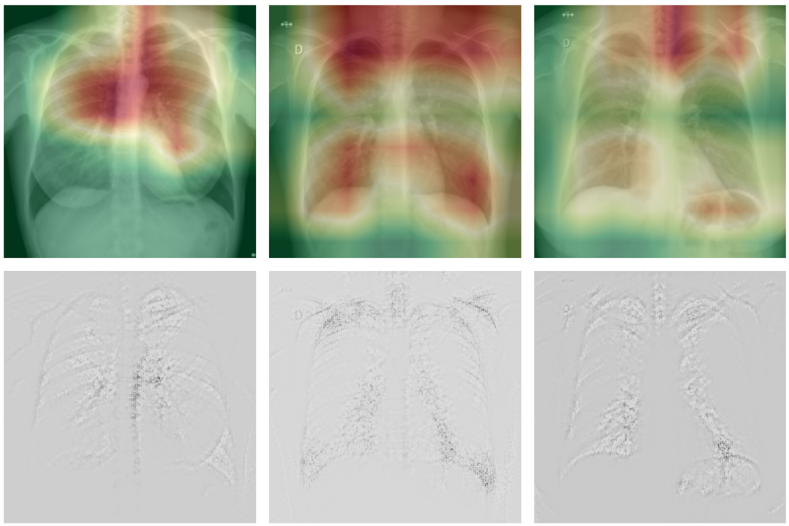


Figure 4.7: GradCAM (top) and Guided BackPropagation (bottom) interpretability heat-maps for windowing settings classifier. Each column represents one image example.

4.5.3 Scanner Features Task Selection

Both tasks evaluated promote feature extraction since the model reached convergence for both scenarios. Nevertheless, the produced interpretability maps from the windowing settings cluster classification delineate similar intentions to our assumptions of how the scanner features variability should behave. The broad focus on the chest and regions of high-frequency values for intensity indicates the source of variability we intend to study. Additionally, although the different manufacturers can introduce variability, the same manufacturer's scanner can apply dissimilar acquisition parameters and windowing levels. This may lead to conflicts, for example, in cases of similar images with distinct manufacturers and vice-versa. Therefore, the windowing settings cluster classification is the chosen task to act as the scanner features evaluation.

4.6 Conclusion

Generalisation in Deep Learning has a variety of solutions. Multiple approaches can effectively improve models' performance in out-of-distribution scenarios. Therefore, it was crucial to establish some ground rules to obtain a controllable environment for all experiments in this exploratory project.

Most established settings are based on assumptions and discussions at an early stage of the project, so their rationale can suffer changes throughout the following chapters. However, changing the pathway delineated in this chapter in the middle of the dissertation due to newly formed impressions would lead to higher variability and entropy in all implementations. These implementations would be hard to compare, and the thesis would lose its structural integrity and flow. Thus, the tasks selected for disease and scanner features extraction, alongside the metrics used and the baseline model, will remain constant throughout this work.

The subsequent chapters demonstrate all the solutions explored for generalisation based on this chapter's foundations. We begin by using a simple multi-task training setting that can undergo some regularisation. Then, we take advantage of the attention module to promote further feature separation. Finally, in a unified manner, we incorporate the main findings into a VQ-VAE backbone.

Chapter 5

Multi-Task Encoding and Equal Probability Loss

5.1 Introduction

The initial experiments to improve generalisation involve making subtle modifications to the baseline encoder. As we work with two distinct classification tasks, the encoder must integrate two classifying branches into its architecture. Therefore, this chapter examines the performance differences between the disease-only baseline and a model trained in a multi-task (MT) setting.

This multi-task approach is presented in two iterations. The first iteration corresponds to a simple MT scenario without explicit regularisation. The second one integrates one extra term in the training objective, explained in the following section.

5.2 Equal Probability Regularisation

This attempt to improve the generalisability of the baseline encoder is based on an essential rule of disentanglement: the factors of variation of a particular feature should not hold any information about other features. The abstract values displayed at the bottom of an encoder network make checking for feature overlap between vectors troublesome.

Theoretically, a latent space containing information about a particular feature is valuable for a correct prediction; hence, one can train a classifying head to learn how to make this inference. The opposite affirmation also applies, so if a latent space does not contain information about a feature, the classifying head will not reach any predictive value. Thus, this method tries to replicate this idea. The model has a pair of classifiers, each with its particular latent space and classifying head. The classifying head uses its correspondent latent space to learn how to extract meaningful information for the desired task.

Moreover, a classifying head in the other latent space should output no informative prediction. In other words, this cross-classification (or fake classification) output should be similar to a random prediction.

For example, taking the disease-classifying head and using it in the windowing cluster latent space would result in the output of two logits. This fake prediction, after activation, should portray two neurons with a probability of 50%. The windowing clusters classifying head should output three neurons with around 33% probability each.

5.3 Materials and Methods

An architecture was created with the custom encoder backbone that results in a fully connected output. This fully connected layer is proceeded by two independent fully connected layers, each with a classifying head responsible for its task.

The first iteration uses a simple Multi-Tasking training objective, demonstrated by equation 5.1.

$$\begin{aligned} L &= BCE(\hat{y}_{dis}, y_{dis}) + CE(\hat{y}_{clus}, y_{clus}) \\ &= -\frac{1}{N}\sum_{i=1}^{N}[y_{i_{dis}} \cdot \log(\hat{y}_{i_{dis}}) + (1 - y_{i_{dis}}) \cdot \log(1 - \hat{y}_{i_{dis}})] \\ &\quad -\frac{1}{N}\sum_{i=1}^{N} y_{i_{clus}} \cdot \log(\hat{y}_{i_{clus}}), \end{aligned} \tag{5.1}$$

where $\hat{y}_{dis}$ and y_{dis} are the estimated and ground truth labels for disease, and $\hat{y}_{clus}$ and y_{clus} are the predicted and ground truth labels for the windowing cluster. Regarding the second training procedure, we introduce equal probability loss, which enforces equal probabilities when performing the fake prediction. After obtaining the fake logits for each prediction, an activation function transforms the logits into probabilities for each class that undergoes a Mean Squared Error estimation compared with the respective value for equal likelihood across the different classes. Figure 5.1 illustrates the procedure of this regularisation, and equation 5.2 describes the updated training objective with the added term.

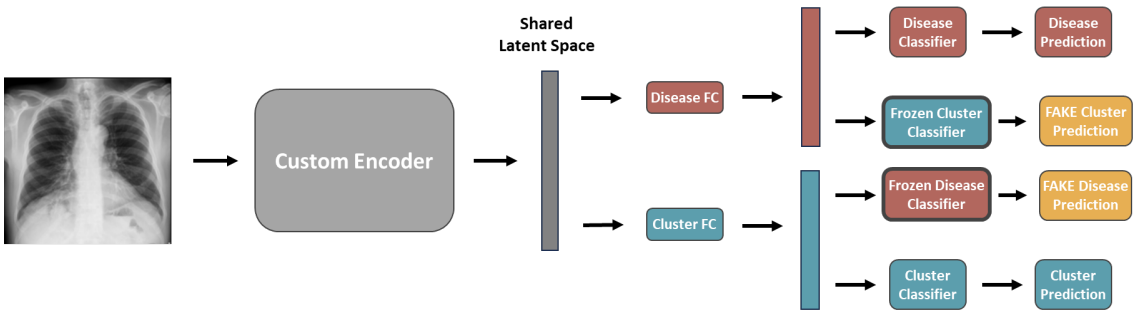


Figure 5.1: Architectural procedure of the Equal Probability Loss. The encoding is shared until the latent space. Then, each task is separated into an independent network in a MT setting.

$$
\begin{aligned}
L &= BCE(\hat{y}_{dis}, y_{dis}) + CE(\hat{y}_{clus}, y_{clus}) + Eq\_Prob(\hat{y}_{fake}, eq\_probabilities) \\
&= -\frac{1}{N}\sum_{i=1}^{N}[y_{i_{dis}} \cdot \log(\hat{y}_{i_{dis}}) + (1 - y_{i_{dis}}) \cdot \log(1 - \hat{y}_{i_{dis}})] \\
&\quad -\frac{1}{N}\sum_{i=1}^{N} y_{i_{clus}} \cdot \log(\hat{y}_{i_{clus}}) + \frac{MSE(\hat{y}_{fake_{dis}}, eq\_prob_{dis}) + MSE(\hat{y}_{fake_{clus}}, eq\_prob_{clus})}{2},
\end{aligned} \tag{5.2}
$$

where $\hat{y}_{fake_{dis}}$ and $\hat{y}_{fake_{clus}}$ are the fake output probabilities for disease and windowing cluster, and eq_prob_{dis} and eq_prob_{clus} are vectors containing the equal probabilities for each task: 50% for disease and around 33% for the windowing cluster.

The two classifying heads are responsible for extracting the relevant information for predictive purposes and outputting the fake logits. These assignments can disaccord at the early stages of training, and the objective of equal probability regularisation is to promote independence in the latent space. Therefore, each training step involves two forward propagations. The first performs the classification tasks and updates the weights for all model's parameters. The second forward utilises frozen classifying heads to obtain the values for the equal probability term. This training step configuration ensures that the changes do not hinder classification performance in the classifier heads, motivating weight updates in the rest of the network.

5.4 Results and Discussion

Table 5.1: 5-fold AUC results in percentage(%) - Testing inference for the simple multi-task setting model with no explicit regularisation.

	In-Distribution	Out-of-Distribution		
Fold	BRAX	MIMIC-CXR-JPG	CheXpert	VinDr-CXR
0	85.06	73.57	75.65	63.46
1	84.56	79.06	78.17	69.65
2	85.22	77.48	77.91	66.45
3	86.25	74.55	77.58	62.09
4	87.11	70.71	72.47	63.24
Average	**85.64**	**75.07**	**76.36**	**64.98**
STD	1.03	3.29	2.39	3.07

Tables 5.1 and 5.2 show the 5-fold cross-validation AUC scores obtained for the MT model without and with equal probability regularisation, respectively. Table 5.3 compares the average AUC results with the performance of the Baseline model.

In-distribution testing results show a slight decrease in performance for both models. This indicates that the added task affects the model's convergence ability, but not significantly. Regarding out-of-distribution, the MT settings fail to overcome the Baseline in the CheXpert and MIMIC-CXR testing while obtaining performance gains of 1.5% in VinDr-CXR inference. Focusing on

Table 5.2: 5-fold AUC results in percentage(%) - Testing inference for the simple multi-task setting model with equal probability regularisation.

	In-Distribution	Out-of-Distribution		
Fold	BRAX	MIMIC-CXR-JPG	CheXpert	VinDr-CXR
0	85.17	75.05	76.80	56.59
1	86.75	72.36	74.48	67.42
2	86.11	75.87	79.06	62.32
3	85.20	74.26	77.21	65.91
4	85.27	74.20	74.00	72.61
Average	**85.70**	**74.35**	**76.31**	**64.97**
STD	0.70	1.30	2.08	5.97

Table 5.3: Average AUC results in percentage (%). Comparison between the baseline model and the two multi-tasking approaches.

	In-Distribution	Out-of-Distribution		
	BRAX	MIMIC-CXR-JPG	CheXpert	VinDr-CXR
Baseline	$\mathbf{86.07 \pm 0.41}$	$\mathbf{76.05 \pm 1.98}$	$\mathbf{76.95 \pm 1.91}$	63.35 ± 3.30
MT - no regularisation	85.64 ± 1.03	75.07 ± 3.29	76.36 ± 2.39	$\mathbf{64.98 \pm 3.07}$
MT - Eq. Prob.	85.70 ± 0.70	74.35 ± 1.30	76.31 ± 2.08	64.97 ± 5.97

the Equal probability regularisation, one can see that this model underperformed the simple MT architecture against expectations.

One possible explanation for the underperforming model may be contradicting training dynamics due to the joint training procedure. Having a shared encoder for disease and windowing cluster classification that suffers updates for both tasks in the same batch cycle may induce overshadowed backpropagation. Therefore, we developed a third model with modified training dynamics. This model underwent two separate training cycles for each epoch. The first cycle focused on adapting the model to the windowing cluster classification, and the second cycle centred on fitting the model for disease classification. The equal probability loss term is split at each cycle. The head for disease classification freezes when adapting the model to windowing cluster classification and vice versa. This new training procedure allows the model to gain more confidence in its updates, promoting shared parameters with more meaningful information.

Table 5.4: Average AUC results in percentage (%). Addition of the average AUC results for the new training dynamics.

	In-Distribution	Out-of-Distribution		
	BRAX	MIMIC-CXR-JPG	CheXpert	VinDr-CXR
Baseline	$\mathbf{86.07 \pm 0.41}$	76.05 ± 1.98	$\mathbf{76.95 \pm 1.91}$	63.35 ± 3.30
MT - no regularisation	85.64 ± 1.03	75.07 ± 3.29	76.36 ± 2.39	64.98 ± 3.07
MT - Eq. Prob. regularisation	85.70 ± 0.70	74.35 ± 1.30	76.31 ± 2.08	64.97 ± 5.97
MT - Eq. Prob. new dynamics	84.86 ± 0.24	$\mathbf{76.93 \pm 1.21}$	76.69 ± 1.95	$\mathbf{66.17 \pm 3.19}$

As seen in Table 5.4, the AUC scores for in-distribution testing dropped. However, out-of-distribution results demonstrate a significant performance increase across all datasets compared with the previous MT approaches. This approach is also more competitive against the Baseline, surpassing it in the MIMIC-CXR inference while closing the gap for CheXpert.

The positive results indicate that with this alternated procedure, the model is not controlled by possible contradicting gradients and can explore more weight configurations to solve the main training objective. This leads to a stronger encoder fitted adequately for extracting meaningful information for both classifiers. Using the equal probability term decreases variability caused by different windowing settings, improving the disease classifier performance during out-of-distribution testing.

5.5 Conclusion

Incorporating a multi-tasking setting into the baseline encoder proved beneficial to generalisation. While the first two approaches did not meet the expected results, changing the training dynamics significantly improved out-of-distribution performance.

Promoting disentanglement between the two tasks using equal probability regularisation effectively removed some factors for variability in the disease classification task. Thus, we confirm assumptions made regarding the prevalence of windowing clusters' classification and the alignment with our overall objectives for the dissertation.

However, regularising the bottom layers may prove to be insufficient to ensure feature independence. The following chapters provide more profound studies revolving around the same foreground, making changes in higher layers of the encoder while linking attention mechanisms and generative models.

Chapter 6

Attention-based Regularisation

6.1 Introduction

The variability in the windowing settings started to be explored in Section 2.2 and displayed promising results for generalisation improvements in medical multi-centre data. However, the equal probability regularisation, presented in Chapter 5, can only promote disentanglement at the bottom of the encoder since most parameters are shared between the disease and windowing cluster classification tasks.

This chapter introduces a novel approach to try enforcing the feature separation early in the encoding process, using data and model-centric techniques. The combination of particularities in the windowing settings sampling, previously discussed, with attention modules in a contrastive scenario can mitigate the effect of variability in windowing levels for disease classification. Ultimately, the model should be more robust in different distributions, hence promoting generalisability.

6.2 Contrastive Attention for Early Feature Separation

As stated in section 2.3, the encoding process reduces the input's spatiality into richer high-dimension representations. This phenomenon occurs gradually throughout the network, meaning that feature maps at the top of the model still hold substantially sparse information about the input image. Furthermore, in a multi-task classification scenario like ours, an early bifurcation in the network gives each task a considerable amount of differentiated parameters, with weight updates reflecting only on that specific task's performance.

The sparse feature maps can undergo regularisation before the bifurcation. An attention module highlights the relevant information of an input based on a particular inquiry. Therefore, this attention module could use the early feature maps to select the regions of interest for further encoding. Additionally, by having two distinct attention modules, each responsible for a task, we can have two different perspectives on the same input.

If these maps receive coordination to be independent, we can promote disentanglement in the downstream processes. However, since they are located at the beginning of the network, a rigid control could destabilise the training procedure, as some regions of interest can still be shared between tasks.

On this premise, we set a contrastive learning setting that takes advantage of the availability of both original and sampled images in the BRAX dataset. These pairs of images present the same radiograph but with different contrast and brightness levels. At an abstract level, the disease is still present, so disease-related activations remain unchanged. However, the windowing cluster has changed, so windowing-related activations should be modified accordingly. Thus, our implementation rests on feeding the model pairs of images from the same radiograph, extracting activation maps from each task, and encouraging similarity between disease-related maps while discouraging similarity between windowing-related maps.

Section 6.3 provides further details about this implementation, while Sections 6.4 and 6.5 discuss the results obtained and the main takes from this approach, respectively.

6.3 Methods and Implementation

As introduced in the previous section, two images from the same case, representing the original and the radiologist's view of a radiograph, are inputted into a model. This model is regularisable in a contrastive setting by producing independent disease and windowing cluster attention maps for each interpretation. The contrastive loss term approximates the disease attention maps while differentiating the windowing settings' attention maps. Equal Probability Loss is also applied to ensure disentanglement at the bottom layers of the network. The following subsections detail our implementation and regularisation process.

6.3.1 Proposed Approach

To achieve the objectives set for this chapter, we made some changes to the baseline encoder architecture, defined in Section 4.4. After the first encoding block, the network undergoes a bifurcation, generating two branches, one for encoding the disease information and the other for windowing cluster classification. These branches are asymmetrical: the cluster predictor comprises one extra encoding block, while the disease classifier has three more encoding processes. We empirically observed that the windowing cluster classification is easier than disease prediction since it focuses on sparse information. Thus, one extra encoding block is enough for the cluster classification task.

The step that branches the network relies on the attention module. We included two independent convolutional attention maps inspired by Zhang *et al.*, each representing a task [8]. This way, we ensure that the model selects the relevant information for each task at an early stage, and further regularisation, explained in subsection 6.3.2, propagates invariable Atelectasis prediction. Figure 6.1 gives a technical perspective on the proposed architecture.

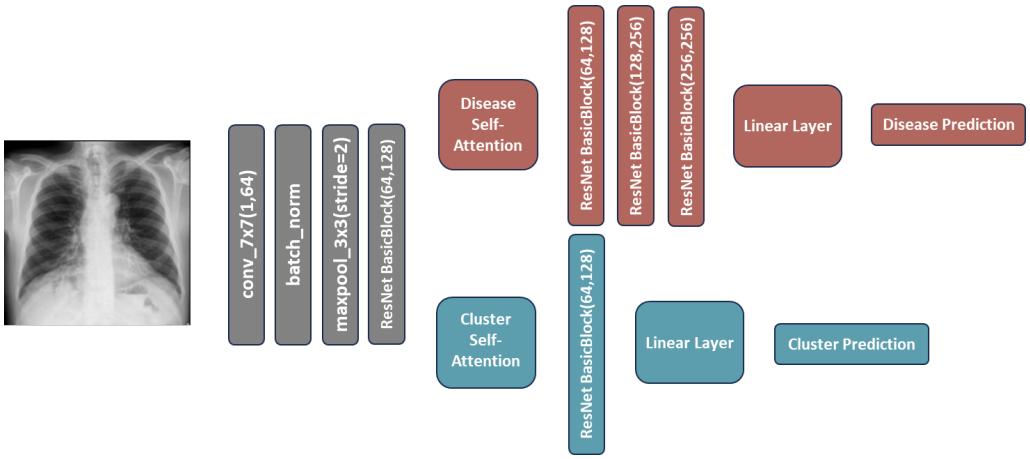


Figure 6.1: Diagram detailing the blocks involving the Proposed Architecture.

6.3.2 Regularisation Process

The prediction processes require using the Binary Cross-Entropy Loss for disease classification and the Cross-Entropy Loss for windowing cluster classification, depicted in equations 6.1 and 6.2, respectively.

$$L_{dis} = BCE(\hat{y}_{dis}, y_{dis}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_{i_{dis}} \cdot \log(\hat{y}_{i_{dis}}) + (1 - y_{i_{dis}}) \cdot \log(1 - \hat{y}_{i_{dis}})\right], \tag{6.1}$$

where $\hat{y}_{dis}$ and y_{dis} correspond to the estimated and ground truth labels for disease classification, respectively.

$$L_{clus} = CE(\hat{y}_{clus}, y_{clus}) = -\frac{1}{N}\sum_{i=1}^{N} y_{i_{clus}} \cdot \log(\hat{y}_{i_{clus}}), \tag{6.2}$$

where $\hat{y}_{clus}$ and y_{clus} are the predicted and ground truth labels for windowing cluster classification, respectively.

For the equal probability technique, represented in equation 6.3, as described in Chapter 5, we use two distinct Adam optimizers at each training step. One can access and update all model parameters, while the other discards any change on the classifying heads since they are frozen. This process implies that two propagation steps occur at each training step.

$$L_{eq\_prob} = Eq\_Prob(\hat{y}_{fake}, eq\_probabilities) \tag{6.3}$$

The contrastive learning term applies Mean Squared Error Loss between the pair of original, o_{dis}, and sampled, s_{dis}, flattened disease attention maps (equation 6.4), and Pytorch's Cosine Embedding Loss for the dissimilar vectors, o_{clus} and s_{clus}, to the flattened windowing attention maps, characterized by equation 6.5.

$$L_{attn_{dis}} = MSE\,(o_{dis}, s_{dis}) = \frac{1}{N}\sum_{n=0}^{N}(o_{dis_n} - s_{dis_n})^2 \tag{6.4}$$

$$L_{attn_{clus}} = CosEmbed\,(o_{clus}, s_{clus}) = max\left(0, \frac{o_{clus} \cdot s_{clus}}{|o_{clus}||s_{clus}|}\right) \tag{6.5}$$

Equation 6.6 summarises the training objective for this approach.

$$L_{total} = L_{dis} + L_{clus} + L_{attn_{dis}} + L_{attn_{clus}} + L_{eq\_prob} \tag{6.6}$$

For inference, only the disease output is considered. Figure 6.2 provides a visualization of the whole training procedure to facilitate comprehension.

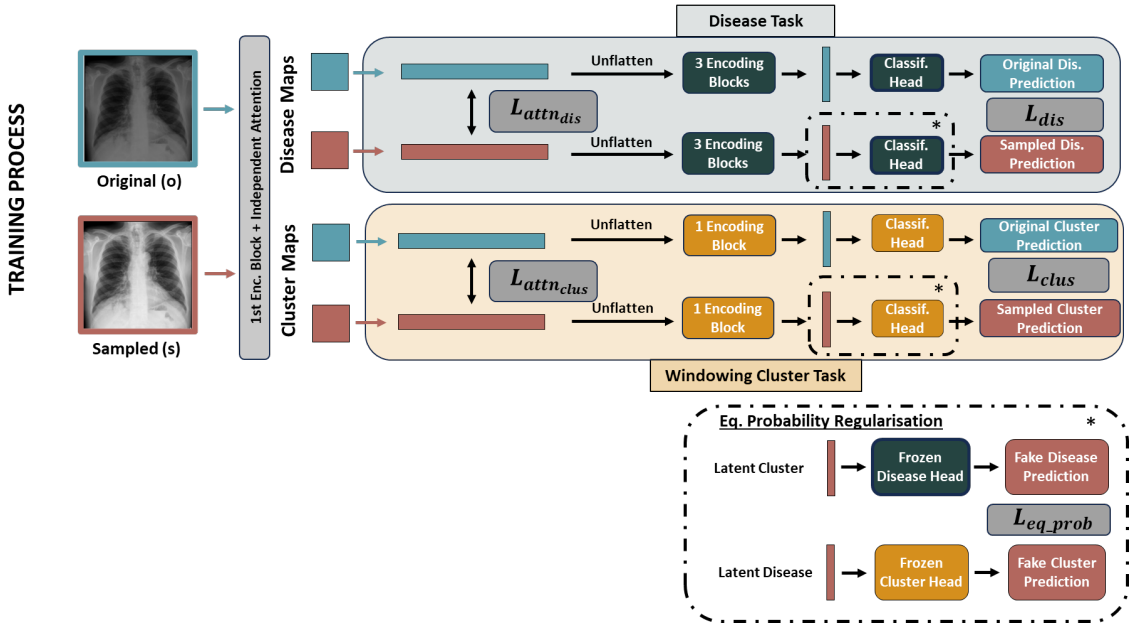


Figure 6.2: Proposed Training Procedure. The original/sampled image pairs go through the model in each training cycle, generating the environment for contrastive learning and equal probability regularisation.

6.4 Results and Discussion

The proposed technique is promising if it can score higher AUC scores in the out-of-distribution inference compared to the baseline network, with no compromises in the in-distribution testing. With that said, Table 6.1 displays the AUC scores in percentage for disease classification in the BRAX dataset, while Table 6.2 shows the results for out-of-distribution inference in the CheXpert, MIMIC-CXR and VinDr-CXR datasets.

The proposed framework slightly outperforms the baseline at in-distribution testing. Thus, the introduced regularisation does not negatively affect in-source performance, meaning that in a closed system with no data distribution fluctuations, the model is as functional as the baseline. This scenario is beneficial for single-centre clinical purposes.

Table 6.1: AUC scores for Atelectasis in percentage (%) between the baseline encoder and the attention-based model - In-distribution testing.

FOLD	In distribution testing	
	BRAX	
	Baseline	Proposal
1	**86.04**	84.88
2	85.41	**86.58**
3	86.09	**86.98**
4	86.34	**86.40**
5	86.46	**86.69**
Average	86.07 ± 0.41	$\mathbf{86.31 \pm 0.82}$

Table 6.2: AUC scores for Atelectasis in percentage (%) between the baseline encoder and the attention-based model - Out-of-distribution testing.

FOLD	Out-of-distribution testing					
	MIMIC-CXR-JPG		CheXpert		VinDr-CXR	
	Baseline	Proposal	Baseline	Proposal	Baseline	Proposal
1	72.77	**77.91**	**75.95**	75.17	61.07	**67.77**
2	75.83	**77.39**	**79.50**	76.21	65.23	**70.43**
3	76.47	**80.18**	75.41	**81.46**	61.46	**71.17**
4	77.58	**79.84**	78.46	**78.86**	60.69	**68.78**
5	**77.58**	75.87	**75.41**	74.08	**68.28**	67.90
Average	76.05 ± 1.98	$\mathbf{78.24} \pm 1.79$	76.95 ± 1.91	$\mathbf{77.16} \pm 2.99$	63.35 ± 3.30	$\mathbf{69.21} \pm 1.53$

Out-of-distribution testing evidences that the method improves AUC performance in the three datasets. The improvements are less noticeable in CheXpert and MIMIC-CXR compared to the VinDr-CXR; however, the quantity of samples in the first two is substantial, meaning that a slight change in scoring is noteworthy.

6.4.1 Ablation Study

Compared to the baseline model, this approach modified the architecture and the training procedure. Focusing on the training scheme, one can argue that the improvements verified in the previous subsection could arise from the availability of more data as a way of data augmentation. Consequently, we performed an ablation study to ensure that the performance gains are not only derived from doubling the amount of training samples. This study consisted of training the baseline encoder with both the original and the sampled versions of the radiographs and repeated the out-of-distribution inference. As Table 6.3 suggests, including more data improves the baseline encoder performance in the CheXpert dataset, surpassing our proposed architecture. However, it performs considerably worse for both MIMIC-CXR and VinDr-CXR. Therefore, the increased predictive power does not come from simple data augmentation but from the changes applied in our proposal.

Table 6.3: AUC scores for Atelectasis in percentage (%) - Ablation Study.

	Out-of-distribution testing					
	MIMIC-CXR-JPG		CheXpert		VinDr-CXR	
	W/ Augm.	Proposal	W/ Augm.	Proposal	W/ Augm.	Proposal
Avg.	75.75 ± 2.94	$\mathbf{78.24} \pm 1.79$	$\mathbf{79.12} \pm 2.10$	77.16 ± 2.99	63.40 ± 4.23	$\mathbf{69.21} \pm 1.53$

Using attention maps regularised by contrastive learning at an early stage of the network created a model more robust to out-of-distribution testing. This model's disease classification capability is not affected by the variability induced by the radiographs' different contrast and brightness levels.

6.5 Conclusion

Including attention modules in our baseline model for a contrastive learning setting resulted in an effective solution to overcome performance hits in medical multi-centre data. Our approach achieves higher AUC scores in distinct distributions by leveraging the variability in the training dataset induced by the radiologist's windowing settings. The windowing parameters, generally overlooked for Deep Learning solutions, hold great potential for improving generalisation.

Tinkering with attention modules showed a promising technique for improving generalisation. Their adaptability and functionality characteristics provide an opportunity to explore their potential further. Therefore, the next chapter will continue experiments using attention-related methods.

Chapter 7

Learning Neural Discrete Representations with Attention: A Unified Approach

7.1 Introduction

The previous chapter successfully combined different techniques, such as attention modules and contrastive learning, to take advantage of data variability, namely the windowing settings applied by radiologists, to improve disentanglement, producing better generalisation results for Atelectasis prediction. However, other models can benefit from these adaptations. Thus, this chapter focuses on a different class of model, namely one famous generative architecture, the VQ-VAE. This is a natural next step in this study, mainly because of all the work published referencing generative models for improved disentanglement practices, as presented in Sections 3.2.2 and 3.4.

In a nutshell, VAEs are generative networks composed of an encoder that learns how to parameterise the input data into a latent space, the latter being used by the decoder component to reconstruct the initial image. In VQ-VAE, the authors propose a new parameterisation method that uses discrete latent variables inspired by vector quantisation (VQ) instead of continuous, random latent variables.

Using VQ, the input of the decoder corresponds to samples drawn from an embedding table that are the closest to the representation provided by the encoder. In other words, slightly different inputs may be parameterised similarly because the most comparable feature vectors remain unchanged. Therefore, issues regarding variability can be filtered by this discretisation, allowing an increased robustness to the model. Additionally, our end goal involves disease classification, a discrete task with only two results; thus, applying discretisation at an early stage may improve performance.

In this chapter, our focus is on VQ-VAE. We begin implementing this network using the BRAX dataset solely for image generation purposes. As we progress, implementations gradually change the original architecture, fine-tuning it to improve generalisation qualities at Atelectasis prediction.

With that said, the changes applied represent similar techniques discussed in previous chapters to encapsulate all the work done in this dissertation without going out of scope. We also use an adversarial component in one of the methods, further explained in the following sections.

7.2 VQ-VAE for Image Generation

The first task involves checking if VQ-VAE achieves an acceptable image reconstruction quality using the BRAX dataset. Since we are using a VAE, widely used for disentanglement purposes, we need to guarantee successful image reconstruction to ensure that the model is correctly parameterising the input images. This section details the changes made to the training environment and displays some examples of the obtained reconstructions.

7.2.1 Implementation Details

We used the VQ-VAE Pytorch architecture implementation available on GitHub (https://github.com/zalandoresearch/pytorch-vq-vae). We tried to recreate the same initialisation parameters without compromising the depth of our baseline encoder. The main difference between the two architectures is the lack of Batch Normalisation layers in the VQ-VAE since they can negatively impact image reconstruction performance by removing intricate details of unique samples in a batch.

Some hyperparameters differ from the ones established in Chapter 4. The learning rate was set to 2×10^{-4} to follow the authors' implementation, and the number of training epochs increased to 100 due to a higher convergence window. The training objective is the one used in the original work, already portrayed in Equation 3.2.

7.2.2 Results and Discussion

Figure 7.1 illustrates the quality of the image reconstruction obtained. The model converged correctly, and compared with the original images, the reconstructions maintain the essential details. There is visible noise, and the images lose some sharpness. Nonetheless, the overall brightness and contrast are correctly transferred, and the anatomical structures are still visible, meaning that this VQ-VAE can extract meaningful distributions from the BRAX dataset.

While some may contend that the superior quality of the images is solely due to the short encoding pathway, it is essential to note that the objective of this endeavour is not to have state-of-the-art image generation capabilities. Instead, the focus is enforcing top layers to generate adequate Disentangled Representations. In conclusion, these preliminary findings effectively showcase the potential of VQ-VAEs.

Figure 7.1: Original (top) and reconstructed images (bottom) from BRAX testing set.

7.3 VQ-VAE for Disentangled Disease Classification)

The positive results in image generation led to a thorough analysis of VQ-VAE capabilities for Atelectasis prediction. Consequently, this architecture underwent several modifications, gradually increasing the implementation complexity. All the changes are based on the previous chapters and will be explained in the following subsections. Since there are many results to analyse, the discussion will be concentrated in a single subsection, making it easier to compare the different methodologies.

7.3.1 Disease-Only Classification

The first change was to take the quantised representation of the input image and put it through a disease classifier. Since there is only one encoding block at the VQ-VAE parameterisation, this classifier has three additional encoding blocks to mimic the structure of the baseline encoder.

This implementation does not have a multi-task scenario in order to evaluate the immediate impact of vector quantisation in disentanglement.

7.3.2 Multi-Task Scenario without Regularisation

Adding a windowing cluster classifier to the VQ-VAE brought two additional iterations. The first one uses a shared quantised parameterisation of the input for both disease and windowing cluster predictions. Since the vector quantisation is shared, there may be some overlap in the embeddings used for each task. The second iteration has independent vector quantisation procedures for each task to prevent information leakage and, subsequently, more variability in disease prediction. The windowing cluster classifier uses one encoding block.

These iterations can translate the impact of the shared discretisation pathway in the model's performance in out-of-distribution settings.

7.3.3 Multi-Task Scenario with Attention-Based Embeddings

Until now, the embedding space used for quantisation was randomly initialised and suffered updates so that the chosen vectors were closer to the input feature maps. This implementation utilises the attention mechanism inspired by Vision Transformers as the embedding space.

The main goal of this approach is to merge the convolutional feature extraction from the encoding block with the patch-oriented global attention pooling employed by the Vision Transformer. Convolutional feature extraction creates assumptions based on pixel interactions with neighbouring pixels. At the same time, the attention mechanism allows for a broader focus on the radiograph's areas of interest.

By incorporating attention mechanisms into the embedding process, the vector quantisation procedure can select the most relevant attention vectors based on their proximity to the convolutional feature vectors. This fusion of convolutional and attention-based approaches enriches image encoding, potentially enhancing disentanglement. We assign independent attention mechanisms

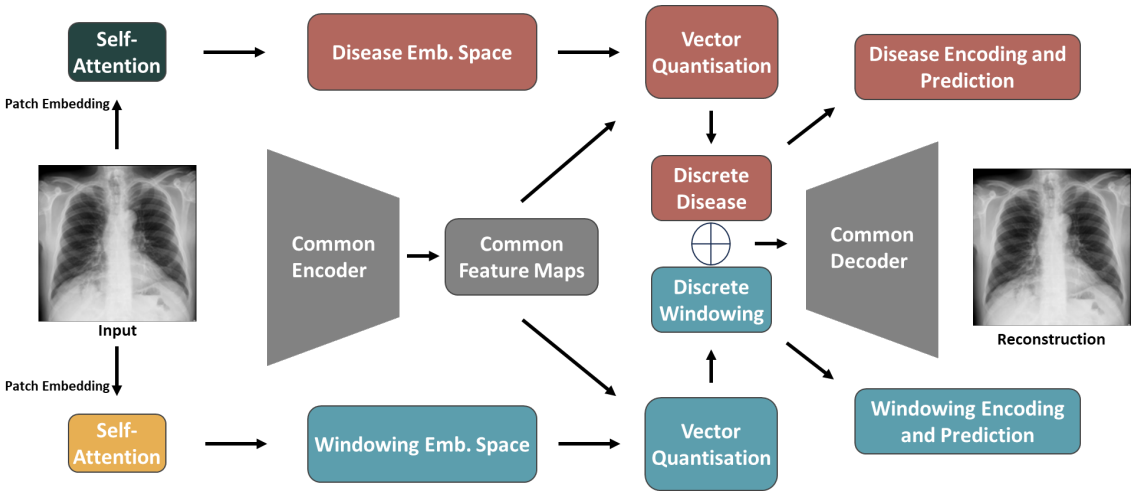


Figure 7.2: Architecure for the multi-task scenario using an attention-based embedding space.

to each classifier to ensure feature separation and independence between disease classification and windowing cluster prediction. This step allows the model to focus separately on disease-related and windowing-related aspects within the input data, facilitating disentanglement. For a visual representation of this architecture, refer to Figure 7.2. These independent attention maps can further undergo regularisation, which will be discussed in detail in the following subsections.

7.3.4 Attention-Based Embeddings with Contrastive Learning

This method uses the same foreground as in Chapter 6, so the model receives a pair of original/sampled images of the same radiograph. The attention maps undergo contrastive regulari-

sation, which ensures similarity between the disease attention maps while separating the ones responsible for the windowing clusters. The training loss terms do not change compared with Chapter 6.

7.3.5 Attention-Based Embeddings with Adversarial Learning

The final methodology introduces the concept of adversarial training introduced in Section 3.4. The adversarial environment involves two small networks, adv_{dis} and adv_{clus}, responsible for generating one task's attention map using the other. In other words, these networks aim to predict one set of attention maps using the features and information intended for a different task.

The objective of the adversarial networks is to minimise the MSE loss between the predicted attention map and the original one, contradicting the primary objective of having independent attention maps. Thus, by adding the adversarial term to the training loss, as seen in Equation 7.1, the model is enforced to widen the differences between the attention maps, further ensuring feature separation.

$$\min_{L_{dis},L_{clus},L_{vqvae}} \max_{adv_{dis},adv_{clus}} = L_{dis} + L_{clus} + L_{vqvae} + L_{adv_{dis}} + L_{adv_{clus}} \tag{7.1}$$

The adversarial network comprises four convolutional operations using a kernel size of 1 and gets updated after each training cycle using Adam's optimiser with a learning rate of 1×10^{-4}.

7.4 Results and Discussion

Table 7.1: Average AUC scores in percentage (%) - Comparison between the baseline and the proposed methodologies.

	In-Distribution	Out-of-Distribution		
	BRAX	MIMIC-CXR	CheXpert	VinDr-CXR
Baseline	**86.07 ± 0.41**	76.05 ± 1.98	76.95 ± 1.91	**63.35 ± 3.30**
VQ-VAE Disease Only	85.27 ± 0.93	76.78 ± 2.40	77.82 ± 2.15	59.82 ± 7.59
MT - Shared Embeddings	84.52 ± 0.80	75.60 ± 1.89	79.48 ± 1.48	58.61 ± 4.59
MT - Separate Embeddings	84.77 ± 1.06	**77.23 ± 1.26**	**79.62 ± 2.11**	59.04 ± 4.06
Attention Embeddings	83.41 ± 1.20	75.46 ± 1.60	77.98 ± 1.68	59.07 ± 5.00
Attention Emb. - Contrastive	83.82 ± 0.92	75.16 ± 1.28	78.35 ± 2.09	58.59 ± 5.99
Attention Emb. - Adversarial	82.48 ± 1.21	76.31 ± 1.78	77.98 ± 1.94	58.59 ± 3.57

Table 7.1 offers a summarised perspective of the performance of the different techniques applied to VQ-VAE. Due to the extensive testing, only the average AUC scores are presented for each dataset. Appendix A entails the results for each fold and implementation.

In-distribution testing displays a significant drop in performance across all models compared to the Baseline. The decreased AUC scores in the BRAX dataset are consistent with the increased amount of regularisation performed. The VQ-VAE objective differs considerably from a common

encoder such as the baseline. The additional complexity inherent in the parameterisation techniques for the Encoder-Decoder architecture may undermine the extraction of meaningful patterns essential for disease classification. The data limitations of the BRAX dataset can also exacerbate this problem. In a real-world scenario, if the primary goal of a model is the straightforward predictive performance on a stationary dataset, this implementation is not advisable.

Regarding out-of-distribution inferences, the analysis compartmentalises into three discussions listed below.

- **Baseline and VQ-VAE Disease-Only Classification:** Despite lower in-distribution scores, the VQ-VAE Disease classifier outperforms the Baseline model in the MIMIC-CXR and CheXpert datasets. The VQ-VAE backbone efficiently removes variability for Atelectasis prediction, contributing to a better generalisation.

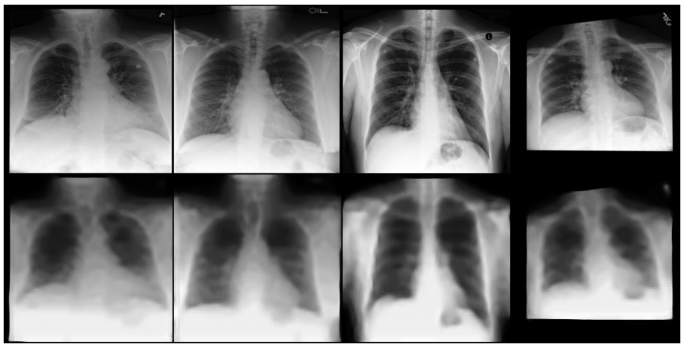
- **VQ-VAE Multi-Task Classification:** Comparing the two iterations proposed for the multi-task setting, it is clear that using a separate embedding space for each classifier increased performance in all datasets. This improvement resonates with the distinct embedding spaces preventing overlapping quantisation vectors for disease and windowing cluster classification. With the overlap, these embeddings suffer parameter updates according to both tasks, preventing task-exclusive weight modifications. Thus, independent discretisation modules ensure a more robust feature separation, promoting disentanglement.

- **VQ-VAE with Attention-based Embedding Space:** Substituting the embedding space for an attention module did not meet the expected results. The three models using attention-based embedding spaces achieved similar AUC metrics for out-of-distribution inference, indicating no significant difference between them. Unlike in Chapter 6, the contrastive learning apparatus did not effectively promote the invariability of the disease attention maps. We can draw the same conclusion from the adversarial regularisation. While it is impossible to state what happened clearly, one can argue that the increased complexity in the training objective overwhelms the number of images available for training and that there can be some contradiction between training terms. An instance of this is when the attention module undergoes regularisation to align with the convolutional feature maps. This causes the two different attention maps to try and merge into a similar representation of the convolutional feature maps, which goes against the primary goal of inter-independence.

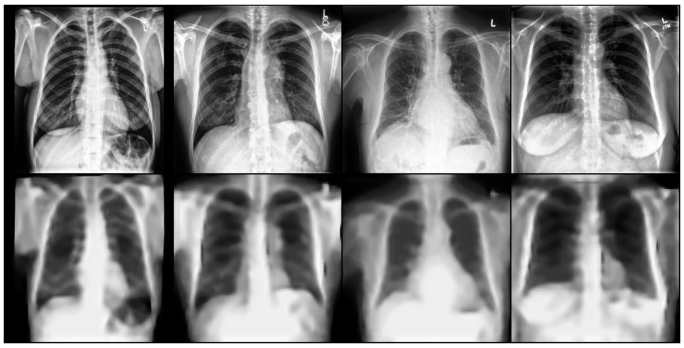
Overall, the results clearly show that the VQ-VAE model that performs multi-task classification using independent embedding spaces is the one that achieves higher out-of-distribution scores across all datasets, even surpassing the baseline.

The discussion did not mention the VinDr-CXR AUC scores due to the VQ-VAE's poor performance on this dataset. The high standard deviation values also translate the significant variability in AUC scores across folds. This effect is not prevalent in the other datasets, as well as the performance drop. This phenomenon may have to do with the VAE backbone of the VQ-VAE. One essential task the model performs is image encoding and decoding, which forces the model to

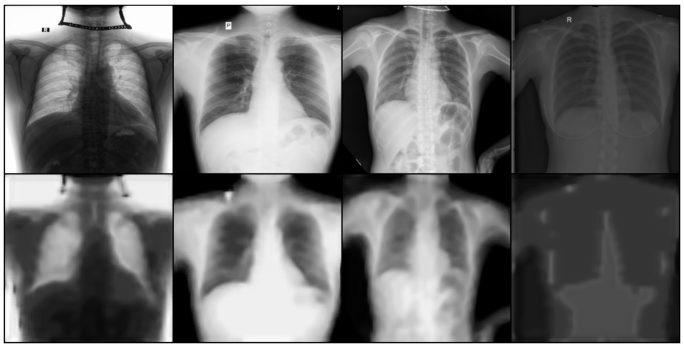
discard all the irrelevant variability in a dataset. However, the VinDr-CXR radiographs are significantly different from the ones used for the other datasets, as it was possible to see in Figure 4.2. This heterogeneity is not introduced to the VQ-VAE during training, so the model is not able to successfully encode enough information of the VinDr-CXR image to the latent space. We can observe this encoding difficulty by comparing the image reconstruction results from the testing sets of BRAX, MIMIC-CXR, CheXpert and VinDr-CXR, displayed in Figures 7.1, 7.3a, 7.3b and 7.3c, respectively.



(a) MIMIC-CXR testing set.



(b) CheXpert testing set.



(c) VinDR-CXR testing set.

Figure 7.3: Original (top) and Reconstructed (bottom) images from the out-of-distribution datasets.

The reconstructions hold less detail on VinDr-CXR compared with the other datasets. In some cases, even the lung cavity is not correctly delineated. In addition, this dataset contains samples with inverse pixel intensities not mentioned in the metadata. All these factors explain the

Table 7.2: AUC scores for the two best models, in percentage (%)

	In-Distribution	Out-of-Distribution		
	BRAX	MIMIC-CXR-JPG	CheXpert	VinDr-CXR
MT - Sep. Embed.	84.77 ± 1.06	77.23 ± 1.26	$\mathbf{79.62 \pm 2.11}$	59.04 ± 4.06
Attn. Contrast. Learning	$\mathbf{86.31 \pm 0.82}$	$\mathbf{78.24 \pm 1.79}$	77.16 ± 2.99	$\mathbf{69.21 \pm 1.53}$

low scores obtained in Chapters 5 and 6, further witnessed in VQ-VAE, a network with stricter encoding.

Table 7.2 displays the final comparison between this chapter's best model and the dissertation's current state-of-the-art. VQ-VAE with multi-task classification using distinct embedding spaces leads CheXpert AUC scores by around 2.5 percentual points while underperforming in MIMIC-CXR by 1%. The attention-based classifier is significantly better at in-distribution inference and VinDr-CXR testing.

That said, the Chapter 6 implementation remains the better performer globally. The strict parameterisation nature of the VQ-VAE severely impacted the performance in VinDr-CXR, being this an extreme example of data heterogeneity. Nevertheless, the VQ-VAE backbone shows potential, and its image generation capabilities could be advantageous for multimodal applications.

7.5 Conclusion

This chapter resulted in an extensive exploration of the architecture of VQ-VAE for Atelectasis prediction. Using the techniques studied in previous chapters, we wanted to conclude this dissertation by integrating everything and evaluating the performance of VAEs, known for their image generation qualities and disentanglement potential.

Regarding image reconstruction, VQ-VAE successfully captured meaningful information from the data through Vector Quantisation, and the generated images could effortlessly detail the majority of anatomical structures available in the radiograph. These preliminary results were promising for the Atelectasis prediction tasks.

Disease classification was inserted in VQ-VAE using several architectures and training procedures. After filtering all the results, the higher-complexity methods did not develop any improvements compared to the baseline. The superior out-of-distribution AUC scores came from the more straightforward techniques, namely the disease-only classification and the multi-task setting with separate embedding spaces. The attention-based approaches underperformed substantially, even when using contrastive or adversarial learning, pointing to a probable saturation of the training objective.

The comparison made between the two best techniques in this dissertation unveiled that the VQ-VAE is marginally inferior to the attention-based contrastive regularisation presented in Chapter 6. Nevertheless, VQ-VAE poses an interesting approach to tackle disentanglement, and further

fine-tuning could present exciting developments. Some of these techniques will be discussed in the subsequent chapter.

Chapter 8

Conclusions and Future Work

8.1 Conclusion

In this thesis, we aimed to explore possible pathways to improve deep neural models' ability to generalise using disentanglement representation learning. Promoting this feature separation can be done in several forms, so we searched for methodologies to help us achieve our goal. In the Introduction of this document, we formulated two investigation questions that would serve as motivation for the following developments. The development began by understanding the underlying topics for this dissertation, such as how X-rays are produced and which parameters affect the final image result or different methodologies that severely impacted the world of deep neural networks, like the attention module and generative models. Based on the gained knowledge, we aimed to prospect published work that shared some goals with our dissertation, such as achieving disentanglement with attention modules or VAEs or applying contrastive learning to data-centric scenarios. A thorough literature review referenced the findings that shaped our assumptions and line of work. Ultimately, the foundations set allowed us to conduct interesting experiments, which addressed the following questions:

- **Which factors of variability influence the disease prediction performance of Deep Learning solutions in out-of-distribution medical data?**

  In X-ray imaging, we understood that there were many possible occasions where variability could be introduced in the final radiograph, beginning at the generation of the X-ray radiation and ending at the type of processing and storing the X-ray exam undergoes. Each of these steps has many parameters that can be tuned and heavily influence the final result, such as the tube current and voltage, the exposure, the windowing settings applied by the radiologist, and the sampling procedure done in post-processing. The experiments we performed concluded that the windowing settings applied to a radiograph had predictive value to a deep neural network, thus leading us to assume the existence of meaningful characteristics that could impact the disease classification. This hypothesis was further confirmed when building models that focused on removing the impact of the windowing settings in

the disease classification, and the gains in performance reflect the influence of this parameter. Other factors were indirectly confirmed to influence the disease prediction, such as the demographics of the data and the patient positioning. Poor results for the VinDr-CXR dataset indicate that the disease prediction had some demographic bias. Additionally, the initial observations made in the VinDr-CXR radiographs pointed to heterogeneity in patient positioning compared to the other evaluated datasets. Limitations in the training dataset prevented further investigation into other possible factors of variation, such as exposure or tube voltage.

- **Which deep neural mechanisms and training procedures promote feature independence for improved generalisation?**

  After establishing a simple network without specific tuning for feature independence as the baseline, we made architectural modifications and tried different training procedures and regularisations to achieve disentangled features. In a multi-task scenario, adding a regularisation promoting no information leakage and alternated training improved some out-of-distribution AUC scores. However, since the bifurcation for the independent tasks was at the bottom of the encoder, there was room for improvement. Thus, we made the encoder separation at an early stage, controlled by attention maps regularised in a contrastive setting, promoting disease invariance while varying the windowing settings. This model substantially surpassed the baseline, meaning that these techniques promoted disentanglement, leading to a better generalisation. In the end, motivated by the disentangling capabilities of generative models, particularly VAEs, we implemented methods based on previous findings to enclose all the work made. The unified approach unexpectedly underperformed, leading to speculation regarding the use of a saturated training objective. However, the behaviour of the VQ-VAE in a multi-task setting with independent embedding spaces showed prominent results, indicating that further studies should occur regarding minor regularisations to the VQ-VAE for better disentanglement. Summarising all the findings, attention maps and the VQ-VAEs showed the most prospective results for promoting feature independence.

This thesis successfully explored and answered the initial research questions, opening new views on promoting more generalisable models in the medical field. In the subsequent section, we outline future research directions to inspire further advancements in the field of generalisation.

8.2 Future Work

8.2.1 Generating all windowing settings samples for each radiograph

Only one sampled image using the windowing settings was available for each original image. We performed preliminary tests to replicate the sampled version of the original radiograph using the metadata information. While slight differences existed between the images, applying the three representations of the windowing settings from the original image was possible. Therefore, if we

generated all the possible sampling levels for each original image, we could significantly increase the size of the training dataset. The higher number of samples would improve the generalisation capability through data augmentation. Additionally, the three distinct classes could motivate an ordinal regularisation, inspired by Albuquerque *et al.* [106], to the windowing settings cluster attention maps: the windowing levels can be ordered in terms of the degree of change of brightness and contrast. Thus, the amount of contrastive regularisation made to the attention maps of the windowing levels clusters could be smaller in samples closer to the original image, gradually increasing until the furthest sample. This new method could promote a more sensible approach to feature independence, further enhancing generalisability.

8.2.2 Experimenting with other training datasets and scanner features

The main reason behind not experimenting with other possible factors of variability, like exposure or tube voltage, was that the BRAX dataset did not have that information available for all images. Adding to its relatively small size, we had to compromise to use only the metadata information present in the majority of cases. With that said, MIMIC-CXR is an excellent dataset, with enough samples and interesting acquisition details. If one could ignore its heavy storage requirements, checking the effects of other variability candidates on the disease prediction could be interesting. Ultimately, several factors of variation could be encoded and combined into a vector, removing the maximum amount of confounders from the disease prediction.

8.2.3 Performing a thorough study of the training dynamics from epoch to epoch

One of the problems discussed in this thesis is the probable saturation of the training objective alongside contradicting terms for fitting the model. Thus, it would be interesting to compare the training dynamics of each model trained, particularly the evolution of model parameter weights from epoch to epoch. Some intriguing articles [107] create analytics systems that visually demonstrate the rich dynamics of training a model, and others even evaluate disentanglement using these same training dynamics [108]. These techniques facilitate the comprehension of our models' behaviour, elucidating some conflicting situations that may be hindering performance.

8.2.4 Using other generative models as sources for disentanglement

The study using VQ-VAE for promoting feature separation provided prospecting results about the capabilities of these models regarding disentanglement. Thus, other classes of generative models could be tested, such as VQ-VAE 2, GANs or the popular Diffusion Models. The higher complexity and computational costs limited their implementation in this dissertation. However, various peer-reviewed articles uncover the disentanglement abilities of GANs [109, 110] and Diffusion Models [111, 112]. Additionally, the quality of image reconstruction is superior to VAEs and could be helpful in the future prospects of the CAGING project regarding the generation of privatised radiographs.

Appendix A

Extensive 5-fold Cross Validation Results for VQ-VAE Experiments

Table A.1: 5-fold AUC scores in percentage (%) for VQ-VAE with Disease-Only Classification

	In-Distribution	Out-of-Distribution		
Fold	BRAX	MIMIC-CXR-JPG	CheXpert	VinDr-CXR
0	83.77	76.40	79.48	54.18
1	85.60	76.63	77.18	62.98
2	85.93	78.42	78.18	70.70
3	86.04	79.35	79.79	59.82
4	85.02	73.09	74.46	51.44
Average	**85.27**	**76.78**	**77.82**	**59.82**
STD	0.93	2.40	2.15	7.59

Table A.2: 5-fold AUC scores in percentage (%) for VQ-VAE Multi-Task with Shared Embedding Space

	In-Distribution	Out-of-Distribution		
Fold	BRAX	MIMIC-CXR-JPG	CheXpert	VinDr-CXR
0	83.83	76.85	80.38	55.76
1	85.83	75.66	78.30	66.29
2	84.71	75.56	80.75	56.08
3	84.10	72.53	77.50	55.47
4	84.14	77.40	80.47	59.47
Average	**84.52**	**75.60**	**79.48**	**58.61**
STD	0.80	1.89	1.48	4.59

Table A.3: 5-fold AUC scores in percentage (%) for VQ-VAE Multi-Task with Independent Embedding Spaces

	In-Distribution	Out-of-Distribution		
Fold	BRAX	MIMIC-CXR-JPG	CheXpert	VinDr-CXR
0	84.52	76.34	77.58	62.64
1	84.61	75.50	77.06	56.20
2	86.60	78.08	80.97	56.86
3	83.91	77.70	81.15	55.32
4	84.19	78.51	81.33	64.16
Average	**84.77**	**77.23**	**79.62**	**59.04**
STD	1.06	1.26	2.11	4.06

Table A.4: 5-fold AUC scores in percentage (%) for VQ-VAE with Attention-based Embedding Space

	In-Distribution	Out-of-Distribution		
Fold	BRAX	MIMIC-CXR-JPG	CheXpert	VinDr-CXR
0	84.96	74.59	77.20	60.43
1	81.89	74.17	77.20	51.45
2	84.08	78.17	80.98	62.89
3	82.64	74.79	77.32	56.96
4	83.49	75.56	77.21	63.64
Average	**83.41**	**75.46**	**77.98**	**59.07**
STD	1.20	1.60	1.68	5.00

Table A.5: 5-fold AUC scores in percentage (%) for VQ-VAE with Attention-based Embedding Space - Contrastive Regularisation

	In-Distribution	Out-of-Distribution		
Fold	BRAX	MIMIC-CXR-JPG	CheXpert	VinDr-CXR
0	84.36	75.98	78.63	50.49
1	84.54	75.50	80.50	57.50
2	83.50	74.32	76.21	60.02
3	82.33	76.59	80.22	57.73
4	84.35	73.43	76.19	67.20
Average	**83.82**	**75.16**	**78.35**	**58.59**
STD	0.92	1.28	2.09	5.99

Table A.6: 5-fold AUC scores in percentage (%) for VQ-VAE with Attention-based Embedding Space - Adversarial Regularisation

	In-Distribution	Out-of-Distribution		
Fold	BRAX	MIMIC-CXR-JPG	CheXpert	VinDr-CXR
0	82.90	78.33	78.75	60.43
1	80.74	74.67	78.70	61.83
2	82.22	76.89	79.79	56.45
3	84.10	74.23	74.71	53.37
4	82.42	77.41	77.93	60.86
Average	**82.48**	**76.31**	**77.98**	**58.59**
STD	1.21	1.78	1.94	3.57

References

[1] Abdominal CT: Windows basics • LITFL • Radiology library. URL: https://litfl.com/abdominal-ct-windows-basics/.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv:1706.03762 [cs] version: 7. URL: http://arxiv.org/abs/1706.03762.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs]. URL: http://arxiv.org/abs/2010.11929.

[4] Ramazan Gençay and Min Qi. Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping, and bagging. *Neural Networks, IEEE Transactions on*, 12:726–734, August 2001. doi:10.1109/72.935086.

[5] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning, May 2018. arXiv:1711.00937 [cs]. URL: http://arxiv.org/abs/1711.00937, doi:10.48550/arXiv.1711.00937.

[6] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating Diverse High-Fidelity Images with VQ-VAE-2, June 2019. arXiv:1906.00446 [cs, stat]. URL: http://arxiv.org/abs/1906.00446.

[7] Patrick Esser, Robin Rombach, and Björn Ommer. Taming Transformers for High-Resolution Image Synthesis, June 2021. arXiv:2012.09841 [cs]. URL: http://arxiv.org/abs/2012.09841, doi:10.48550/arXiv.2012.09841.

[8] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-Attention Generative Adversarial Networks, June 2019. arXiv:1805.08318 [cs, stat]. URL: http://arxiv.org/abs/1805.08318.

[9] Minsong Ki, Youngjung Uh, Junsuk Choe, and Hyeran Byun. Contrastive Attention Maps for Self-supervised Co-localization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2783–2792, Montreal, QC, Canada, October 2021. IEEE. URL: https://ieeexplore.ieee.org/document/9710782/, doi:10.1109/ICCV48922.2021.00280.

[10] A. S. Maida. Chapter 2 - Cognitive Computing and Neural Networks: Reverse Engineering the Brain. In Venkat N. Gudivada, Vijay V. Raghavan, Venu Govindaraju, and C. R. Rao, editors, *Handbook of Statistics*, volume 35 of *Cognitive Computing: Theory and Applications*,

pages 39–78. Elsevier, January 2016. URL: https://www.sciencedirect.com/science/article/pii/S0169716116300529, doi:10.1016/bs.host.2016.07.011.

[11] AHIMA. Retention and Destruction of Health Information. *Retention and Destruction of Health Information / AHIMA, American Health Information Management Association*, October 2013. Publisher: American Health Information Management Association. URL: http://library.ahima.org/PB/RetentionDestruction.

[12] Nick Rubright. AI in Healthcare – Statistics and Trends, March 2023. URL: https://resources.freeagentcrm.com/ai-in-healthcare-statistics/.

[13] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, 3:58–73, January 2022. URL: https://www.sciencedirect.com/science/article/pii/S2666603022000069, doi:10.1016/j.ijin.2022.05.002.

[14] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, November 2018. URL: https://doi.org/10.1093/bib/bbx044, doi:10.1093/bib/bbx044.

[15] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nature Genetics*, 51(1):12–18, January 2019. Number: 1 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/s41588-018-0295-5, doi:10.1038/s41588-018-0295-5.

[16] Jacob Schreiber, Maxwell Libbrecht, Jeffrey Bilmes, and William Stafford Noble. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture, January 2017. Pages: 103614 Section: New Results. URL: https://www.biorxiv.org/content/10.1101/103614v3, doi:10.1101/103614.

[17] Feng Liu, Hao Li, Chao Ren, Xiaochen Bo, and Wenjie Shu. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *Scientific Reports*, 6(1):28517, June 2016. Number: 1 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/srep28517, doi:10.1038/srep28517.

[18] Dimitrios Kleftogiannis, Panos Kalnis, and Vladimir B. Bajic. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Research*, 43(1):e6, January 2015. URL: https://doi.org/10.1093/nar/gku1058, doi:10.1093/nar/gku1058.

[19] Yiheng Wang, Tong Liu, Dong Xu, Huidong Shi, Chaoyang Zhang, Yin-Yuan Mo, and Zheng Wang. Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. *Scientific Reports*, 6(1):19598, January 2016. Number: 1 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/srep19598, doi:10.1038/srep19598.

[20] Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, and Xiaohui Xie. Gene expression inference with deep learning. *Bioinformatics (Oxford, England)*, 32(12):1832–1839, June 2016. doi:10.1093/bioinformatics/btw074.

[21] Alexander Craik, Yongtian He, and Jose L. Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering*, 16(3):031001, April 2019. Publisher: IOP Publishing. URL: https://dx.doi.org/10.1088/1741-2552/ab0ab5, doi:10.1088/1741-2552/ab0ab5.

[22] B. Pyakillya, N. Kazachenko, and N. Mikhailovsky. Deep Learning for ECG Classification. *Journal of Physics: Conference Series*, 913(1):012004, October 2017. Publisher: IOP Publishing. URL: https://dx.doi.org/10.1088/1742-6596/913/1/012004, doi:10.1088/1742-6596/913/1/012004.

[23] Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art. *IEEE Reviews in Biomedical Engineering*, 14:342–356, 2021. Conference Name: IEEE Reviews in Biomedical Engineering. doi:10.1109/RBME.2020.3006860.

[24] Diogo Mata, Wilson Silva, and Jaime S. Cardoso. Increased Robustness in Chest X-Ray Classification Through Clinical Report-Driven Regularization. In Armando J. Pinho, Petia Georgieva, Luís F. Teixeira, and Joan Andreu Sánchez, editors, *Pattern Recognition and Image Analysis*, Lecture Notes in Computer Science, pages 119–128, Cham, 2022. Springer International Publishing. doi:10.1007/978-3-031-04881-4_10.

[25] Xiaohong W. Gao, Rui Hui, and Zengmin Tian. Classification of CT brain images based on deep learning networks. *Computer Methods and Programs in Biomedicine*, 138:49–56, January 2017. URL: https://www.sciencedirect.com/science/article/pii/S0169260716305296, doi:10.1016/j.cmpb.2016.10.007.

[26] C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi. Deep learning for automated skeletal bone age assessment in X-ray images. *Medical Image Analysis*, 36:41–51, February 2017. URL: https://www.sciencedirect.com/science/article/pii/S1361841516301840, doi:10.1016/j.media.2016.10.010.

[27] Nagaraj Yamanakkanavar, Jae Young Choi, and Bumshik Lee. MRI Segmentation and Classification of Human Brain Using Deep Learning for Diagnosis of Alzheimer's Disease: A Survey. *Sensors*, 20(11):3243, January 2020. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute. URL: https://www.mdpi.com/1424-8220/20/11/3243, doi:10.3390/s20113243.

[28] Seokmin Han, Ho-Kyung Kang, Ja-Yeon Jeong, Moon-Ho Park, Wonsik Kim, Won-Chul Bang, and Yeong-Kyeong Seong. A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Physics in Medicine & Biology*, 62(19):7714, September 2017. Publisher: IOP Publishing. URL: https://dx.doi.org/10.1088/1361-6560/aa82ec, doi:10.1088/1361-6560/aa82ec.

[29] Yiming Ding, Jae Ho Sohn, Michael G. Kawczynski, Hari Trivedi, Roy Harnish, Nathaniel W. Jenkins, Dmytro Lituiev, Timothy P. Copeland, Mariam S. Aboian, Carina Mari Aparici, Spencer C. Behr, Robert R. Flavell, Shih-Ying Huang, Kelly A. Zalocusky, Lorenzo Nardo, Youngho Seo, Randall A. Hawkins, Miguel Hernandez Pampaloni, Dexter Hadley, and Benjamin L. Franc. A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain. *Radiology*, 290(2):456–464, February 2019. Publisher: Radiological Society of North America. URL: https://pubs.rsna.org/doi/full/10.1148/radiol.2018180958, doi:10.1148/radiol.2018180958.

[30] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, July 2022. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. doi:10.1109/TPAMI.2021.3059968.

[31] K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, D. Rueckert, and B. Glocker. Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Bjoern Menze, and Mauricio Reyes, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Lecture Notes in Computer Science, pages 450–462, Cham, 2018. Springer International Publishing. doi:10.1007/978-3-319-75238-9_38.

[32] Zhaoye Zhou, Gengyan Zhao, Richard Kijowski, and Fang Liu. Deep convolutional neural network for segmentation of knee joint anatomy. *Magnetic Resonance in Medicine*, 80(6):2759–2770, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.27229. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.27229, doi:10.1002/mrm.27229.

[33] R. Krithiga and P. Geetha. Breast Cancer Detection, Segmentation and Classification on Histopathology Images Analysis: A Systematic Review. *Archives of Computational Methods in Engineering*, 28(4):2607–2619, June 2021. URL: https://doi.org/10.1007/s11831-020-09470-w, doi:10.1007/s11831-020-09470-w.

[34] Eduardo Castro, Jaime S. Cardoso, and Jose Costa Pereira. Elastic deformations for data augmentation in breast cancer mass detection. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 230–234, March 2018. doi:10.1109/BHI.2018.8333411.

[35] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52:128–143, February 2019. URL: https://www.sciencedirect.com/science/article/pii/S1361841518300495, doi:10.1016/j.media.2018.11.010.

[36] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, and David Firmin. DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1310–1321, June 2018. Conference Name: IEEE Transactions on Medical Imaging. doi:10.1109/TMI.2017.2785879.

[37] AI Funding Doubles in 2021, Especially for Healthcare. URL: https://www.brinknews.com/quick-take/ai-funding-doubles-in-2021-especially-for-healthcare/.

[38] Noah Schwartz. How Baptist is using Microsoft's AI tools to fight burnout, August 2023. URL: https://www.beckershospitalreview.com/innovation/how-baptist-is-using-microsofts-ai-tools-to-fight-burnout.html.

[39] How AI Is Transforming the Field of Radiology - SPONSOR CONTENT FROM SIEMENS HEALTHINEERS. *Harvard Business Review*, September 2023. Section: Technology and analytics. URL: https://hbr.org/sponsored/2023/09/how-ai-is-transforming-the-field-of-radiology.

[40] Ma Sidhom and Mg Poulsen. Group decisions in oncology: Doctors' perceptions of the legal responsibilities arising from multidisciplinary meetings. *Journal of Medical Imaging and Radiation Oncology*, 52(3):287–292, 2008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1440-1673.2007.01916.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1440-1673.2007.01916.x, doi:10.1111/j.1440-1673.2007.01916.x.

[41] Mario Verdicchio and Andrea Perin. When Doctors and AI Interact: on Human Responsibility for Artificial Risks. *Philosophy & Technology*, 35(1):11, 2022. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8857871/, doi:10.1007/s13347-022-00506-6.

[42] Onur Asan, Alparslan Emrah Bayrak, and Avishek Choudhury. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, 22(6):e15154, June 2020. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. URL: https://www.jmir.org/2020/6/e15154, doi:10.2196/15154.

[43] Viswanathan Venkataraman, Travis Browning, Ivan Pedrosa, Suhny Abbara, David Fetzer, Seth Toomay, and Ronald M. Peshock. Implementing Shared, Standardized Imaging Protocols to Improve Cross-Enterprise Workflow and Quality. *Journal of Digital Imaging*, 32(5):880–887, October 2019. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6737153/, doi:10.1007/s10278-019-00185-4.

[44] INESC TEC. Causality-driven Generative Models for Privacy-preserving Case-based Explanations, February 2023. URL: https://www.inesctec.pt/en/projects/caging.

[45] Silva, Daniel, Agrotis, Georgios, Silva, Wilson, F. Teixeira, Luís, and Beets-Tan, Regina. Attention-based Regularisation for Improved Generalisability in Medical Multi-Centre Data.

[46] Martin Berger, Qiao Yang, and Andreas Maier. X-ray Imaging. In Andreas Maier, Stefan Steidl, Vincent Christlein, and Joachim Hornegger, editors, *Medical Imaging Systems: An Introductory Guide*. Springer, Cham (CH), 2018. URL: http://www.ncbi.nlm.nih.gov/books/NBK546155/.

[47] Lee Prangnell. Visible Light-Based Human Visual System Conceptual Model.

[48] Horst Aichinger, Joachim Dierker, Sigrid Joite-Barfuß, and Manfred Säbel. Principles of X-Ray Imaging. In Horst Aichinger, Joachim Dierker, Sigrid Joite-Barfuß, and Manfred Säbel, editors, *Radiation Exposure and Image Quality in X-Ray Diagnostic Radiology: Physical Principles and Clinical Applications*, pages 3–7. Springer, Berlin, Heidelberg, 2012. URL: https://doi.org/10.1007/978-3-642-11241-6_1, doi:10.1007/978-3-642-11241-6_1.

[49] X-rays. URL: https://www.nibib.nih.gov/science-education/science-topics/x-rays.

[50] Horst Aichinger, Joachim Dierker, Sigrid Joite-Barfuß, and Manfred Säbel. Production and Measurement of X-Rays. In Horst Aichinger, Joachim Dierker, Sigrid Joite-Barfuß, and Manfred Säbel, editors, *Radiation Exposure and Image Quality in X-Ray Diagnostic Radiology: Physical Principles and Clinical Applications*, pages 13–20. Springer, Berlin, Heidelberg, 2012. URL: https://doi.org/10.1007/978-3-642-11241-6_3, doi:10.1007/978-3-642-11241-6_3.

[51] E L Nickoloff and H L Berman. Factors affecting x-ray spectra. *RadioGraphics*, 13(6):1337–1348, November 1993. Publisher: Radiological Society of North America. URL: https://pubs.rsna.org/doi/abs/10.1148/radiographics.13.6.8290728, doi:10.1148/radiographics.13.6.8290728.

[52] Horst Aichinger, Joachim Dierker, Sigrid Joite-Barfuß, and Manfred Säbel. *Radiation Exposure and Image Quality in X-Ray Diagnostic Radiology: Physical Principles and Clinical Applications*. Springer, Berlin, Heidelberg, 2012. URL: https://link.springer.com/10.1007/978-3-642-11241-6, doi:10.1007/978-3-642-11241-6.

[53] Oudiz, A., Croft, J., Fleishman, A., Lochard, J., Lombard, J., and Webb, G. What is ALARA? *centre d'étude sur l'évaluation de la protection dans le domaine nucléaire*, September 1986.

[54] Horst Aichinger, Joachim Dierker, Sigrid Joite-Barfuß, and Manfred Säbel. Image Quality and Dose. In Horst Aichinger, Joachim Dierker, Sigrid Joite-Barfuß, and Manfred Säbel, editors, *Radiation Exposure and Image Quality in X-Ray Diagnostic Radiology: Physical Principles and Clinical Applications*, pages 85–101. Springer, Berlin, Heidelberg, 2012. URL: https://doi.org/10.1007/978-3-642-11241-6_9, doi:10.1007/978-3-642-11241-6_9.

[55] D.R. Dance, S. Christofides, A.D.A. Maidment, I.D. McLean, and K.H. Ng. *Diagnostic Radiology Physics: A Handbook for Teachers and Students*. Digital Imaging. International Atomic Energy Agency, 2014.

[56] Tami D. DenOtter and Johanna Schubert. Hounsfield Unit. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023. URL: http://www.ncbi.nlm.nih.gov/books/NBK547721/.

[57] MyEndoConsult. Hounsfield Unit Chart - My Endo Consult, February 2022. Section: Endocrine disease - adrenal gland. URL: https://myendoconsult.com/learn/hounsfield-unit-chart/,https://myendoconsult.com/learn/hounsfield-unit-chart/.

[58] Andrew Murphy. Windowing (CT) | Radiology Reference Article | Radiopaedia.org, March 2017. URL: https://radiopaedia.org/articles/windowing-ct, doi:10.53347/rID-52108.

[59] Ricky K. Taira and H. K. Huang. A picture archiving and communication system module for radiology. *Computer Methods and Programs in Biomedicine*, 30(2):229–237, October 1989. URL: https://www.sciencedirect.com/science/article/pii/0169260789900758, doi:10.1016/0169-2607(89)90075-8.

[60] W. Dean Bidgood, Steven C. Horii, Fred W. Prior, and Donald E. Van Syckle. Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging. *Journal of the American Medical Informatics Association*, 4(3):199–212, 1997. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC61235/.

[61] Oliver Diaz, Kaisar Kushibar, Richard Osuala, Akis Linardos, Lidia Garrucho, Laura Igual, Petia Radeva, Fred Prior, Polyxeni Gkontra, and Karim Lekadir. Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. *Physica Medica*, 83:25–37, March 2021. URL: https://www.sciencedirect.com/science/article/pii/S1120179721000958, doi:10.1016/j.ejmp.2021.02.007.

[62] Michele Larobina and Loredana Murino. Medical Image File Formats. *Journal of Digital Imaging*, 27(2):200–206, April 2014. URL: https://doi.org/10.1007/s10278-013-9657-9, doi:10.1007/s10278-013-9657-9.

[63] Achraf Oussidi and Azeddine Elhassouny. Deep generative models: Survey. In *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–8, April 2018. doi:10.1109/ISACV.2018.8354080.

[64] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, December 2022. arXiv:1312.6114 [cs, stat] version: 10. URL: http://arxiv.org/abs/1312.6114, doi:10.48550/arXiv.1312.6114.

[65] Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. arXiv:1906.02691 [cs, stat]. URL: http://arxiv.org/abs/1906.02691, doi:10.1561/2200000056.

[66] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, June 2014. arXiv:1406.2661 [cs, stat]. URL: http://arxiv.org/abs/1406.2661, doi:10.48550/arXiv.1406.2661.

[67] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, November 2022. arXiv:2103.04922 [cs, stat]. URL: http://arxiv.org/abs/2103.04922, doi:10.1109/TPAMI.2021.3116668.

[68] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent Models of Visual Attention, June 2014. arXiv:1406.6247 [cs, stat]. URL: http://arxiv.org/abs/1406.6247.

[69] Marisa Carrasco. How visual spatial attention alters perception. *Cognitive processing*, 19(Suppl 1):77–88, September 2018. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6170011/, doi:10.1007/s10339-018-0883-4.

[70] Gianni Brauwers and Flavius Frasincar. A General Survey on Attention Mechanisms in Deep Learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3279–3298, April 2023. arXiv:2203.14263 [cs]. URL: http://arxiv.org/abs/2203.14263, doi:10.1109/TKDE.2021.3126456.

[71] Yuanyuan Shen, Edmund M.-K. Lai, and Mahsa Mohaghegh. Effects of Similarity Score Functions in Attention Mechanisms on the Performance of Neural Question Answering Systems. *Neural Processing Letters*, 54(3):2283–2302, June 2022. URL: https://doi.org/10.1007/s11063-021-10730-4, doi:10.1007/s11063-021-10730-4.

[72] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate, May 2016. arXiv:1409.0473 [cs, stat] version: 7. URL: http://arxiv.org/abs/1409.0473, doi:10.48550/arXiv.1409.0473.

[73] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation, September 2015. arXiv:1508.04025 [cs] version: 5. URL: http://arxiv.org/abs/1508.04025.

[74] John S. Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Proceedings of the 2nd International Conference on Neural Information Processing Systems*, NIPS'89, pages 211–217, Cambridge, MA, USA, January 1989. MIT Press.

[75] Haoran Zhu, Boyuan Chen, and Carter Yang. Understanding Why ViT Trains Badly on Small Datasets: An Intuitive Perspective, February 2023. arXiv:2302.03751 [cs]. URL: http://arxiv.org/abs/2302.03751, doi:10.48550/arXiv.2302.03751.

[76] P.Y. Simard, D. Steinkraus, and J.C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 958–963, August 2003. doi:10.1109/ICDAR.2003.1227801.

[77] Luis Perez and Jason Wang. The Effectiveness of Data Augmentation in Image Classification using Deep Learning, December 2017. arXiv:1712.04621 [cs]. URL: http://arxiv.org/abs/1712.04621, doi:10.48550/arXiv.1712.04621.

[78] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, July 2019. URL: https://doi.org/10.1186/s40537-019-0197-0, doi:10.1186/s40537-019-0197-0.

[79] Huan Liu and Hiroshi Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, USA, June 1998.

[80] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3(null):1157–1182, March 2003.

[81] Nicholas Pudjihartono, Tayaza Fadason, Andreas W. Kempa-Liehr, and Justin M. O'Sullivan. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*, 2, 2022. URL: https://www.frontiersin.org/articles/10.3389/fbinf.2022.927312.

[82] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, August 1995. Morgan Kaufmann Publishers Inc.

[83] T. G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7):1895–1923, September 1998. doi:10.1162/089976698300017197.

[84] Yoshua Bengio and Yves Grandvalet. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL: https://papers.nips.cc/paper_files/paper/2003/hash/e82c4b19b8151ddc25d4d93baf7b908f-Abstract.html.

[85] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, July 2012. arXiv:1207.0580 [cs]. URL: http://arxiv.org/abs/1207.0580, doi:10.48550/arXiv.1207.0580.

[86] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02080.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x, doi:10.1111/j.2517-6161.1996.tb02080.x.

[87] Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 42(1):80–86, 2000. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality]. URL: https://www.jstor.org/stable/1271436, doi:10.2307/1271436.

[88] I. Higgins, L. Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. November 2016. URL: https://www.semanticscholar.org/paper/beta-VAE%3A-Learning-Basic-Visual-Concepts-with-a-Higgins-Matthey/a90226c41b79f8b06007609f39f82757073641e2?p2df.

[89] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, June 2016. arXiv:1606.03657 [cs, stat] version: 1. URL: http://arxiv.org/abs/1606.03657.

[90] Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Unified Adversarial Invariance, September 2019. arXiv:1905.03629 [cs, stat]. URL: http://arxiv.org/abs/1905.03629.

[91] Helena Montenegro, Wilson Silva, and Jaime S. Cardoso. Disentangled Representation Learning for Privacy-Preserving Case-Based Explanations. In Jana Fragemann, Jianning Li, Xiao Liu, Sotirios A. Tsaftaris, Jan Egger, and Jens Kleesiek, editors, *Medical Applications with Disentanglements*, Lecture Notes in Computer Science, pages 33–45, Cham, 2023. Springer Nature Switzerland. doi:10.1007/978-3-031-25046-0_4.

[92] Eduardo P. Reis, Joselisa P. Q. de Paiva, Maria C. B. da Silva, Guilherme A. S. Ribeiro, Victor F. Paiva, Lucas Bulgarelli, Henrique M. H. Lee, Paulo V. Santos, Vanessa M. Brito, Lucas T. W. Amaral, Gabriel L. Beraldo, Jorge N. Haidar Filho, Gustavo B. S. Teles, Gilberto Szarf, Tom Pollard, Alistair E. W. Johnson, Leo A. Celi, and Edson Amaro. BRAX, Brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487, August 2022. Number: 1 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/s41597-022-01608-8, doi:10.1038/s41597-022-01608-8.

[93] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, January 2019. arXiv:1901.07031 [cs, eess]. URL: http://arxiv.org/abs/1901.07031, doi:10.48550/arXiv.1901.07031.

[94] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, November 2019. arXiv:1901.07042 [cs, eess]. URL: http://arxiv.org/abs/1901.07042, doi:10.48550/arXiv.1901.07042.

[95] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9(1):429, July 2022. Number: 1 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/s41597-022-01498-w, doi:10.1038/s41597-022-01498-w.

[96] Kelly Grott, Shaylika Chauhan, and Julie D. Dunlap. Atelectasis. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023. URL: http://www.ncbi.nlm.nih.gov/books/NBK545316/.

[97] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In Abdul Sattar and Byeong-ho Kang, editors, *AI 2006: Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 1015–1021, Berlin, Heidelberg, 2006. Springer. doi:10.1007/11941439_114.

[98] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs]. URL: http://arxiv.org/abs/1412.6980, doi:10.48550/arXiv.1412.6980.

[99] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. arXiv:1512.03385 [cs]. URL: http://arxiv.org/abs/1512.03385, doi:10.48550/arXiv.1512.03385.

[100] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks, January 2018. arXiv:1608.06993 [cs]. URL: http://arxiv.org/abs/1608.06993, doi:10.48550/arXiv.1608.06993.

[101] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919. doi:10.1109/CVPR.2009.5206848.

[102] José Maurício, Inês Domingues, and Jorge Bernardino. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences*, 13(9):5521, January 2023. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute. URL: https://www.mdpi.com/2076-3417/13/9/5521, doi:10.3390/app13095521.

[103] Hong Vin Koay, Joon Huang Chuah, and Chee-Onn Chow. Convolutional Neural Network or Vision Transformer? Benchmarking Various Machine Learning Models for Distracted Driver Detection. In *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, pages 417–422, December 2021. ISSN: 2159-3450. doi:10.1109/TENCON54134.2021.9707341.

[104] Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. A Study on Overfitting in Deep Reinforcement Learning, April 2018. arXiv:1804.06893 [cs, stat]. URL: http://arxiv.org/abs/1804.06893.

[105] Project MONAI — MONAI 1.2.0 Documentation. URL: https://docs.monai.io/en/stable/.

[106] Tomé Albuquerque, Ricardo Cruz, and Jaime S. Cardoso. Ordinal losses for classification of cervical cancer risk. *PeerJ Computer Science*, 7:e457, April 2021. Publisher: PeerJ Inc. URL: https://peerj.com/articles/cs-457, doi:10.7717/peerj-cs.457.

[107] Dongyu Liu, Weiwei Cui, Kai Jin, Yuxiao Guo, and Huamin Qu. DeepTracker: Visualizing the Training Process of Convolutional Neural Networks. *ACM Transactions on Intelligent Systems and Technology*, 10(1):6:1–6:25, November 2018. URL: https://dl.acm.org/doi/10.1145/3200489, doi:10.1145/3200489.

[108] Chester Holtz, Gal Mishne, and Alexander Cloninger. Evaluating Disentanglement in Generative Models Without Knowledge of Latent Factors, October 2022. arXiv:2210.01760 [cs, stat]. URL: http://arxiv.org/abs/2210.01760, doi:10.48550/arXiv.2210.01760.

[109] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2004–2018, April 2022. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. doi:10.1109/TPAMI.2020.3034267.

[110] Helena Montenegro, Wilson Silva, and Jaime S. Cardoso. Privacy-Preserving Generative Adversarial Network for Case-Based Explainability in Medical Image Analysis. *IEEE Access*, 9:148037–148047, 2021. Conference Name: IEEE Access. doi:10.1109/ACCESS.2021.3124844.

[111] Tao Yang, Yuwang Wang, Yan Lv, and Nanning Zheng. DisDiff: Unsupervised Disentanglement of Diffusion Probabilistic Models, January 2023. arXiv:2301.13721 [cs]. URL: http://arxiv.org/abs/2301.13721, doi:10.48550/arXiv.2301.13721.

[112] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the Disentanglement Capability in Text-to-Image Diffusion Models. pages 1900–1910, 2023. URL: https://openaccess.thecvf.com/content/CVPR2023/html/Wu_

Uncovering_the_Disentanglement_Capability_in_Text-to-Image_
Diffusion_Models_CVPR_2023_paper.html.