SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY
INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Towards Fair and Accurate Medical Image Embeddings

Florian Hunecke

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Towards Fair and Accurate Medical Image Embeddings

Ein Beitrag zur Verbesserung von Fairness und Genauigkeit Medizinischer Image Embeddings

Author:	Florian Hunecke
Examiner:	Prof. Dr. Daniel Rückert
Supervisor:	Robert Graf
Submission Date:	April 15, 2024

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Munich, April 15, 2024 Florian Hunecke

Abstract

As machine learning models get more complex and profound in their structure, they rely even more on large and representative datasets. Recent years have brought up new medical studies driving the development of image classifiers within the medical domain to astonishing accuracy. However, these datasets and deep learning models still function like some black box and need more insights into their decision-making, especially regarding fair and unbiased treatment of individuals and groups. Most previous work evaluating bias in models or datasets has been done with human-reliant workflows speculating for hidden bias or manual surveys.

This thesis proposes to create new representations of image datasets using embeddings, obtained through unsupervised training of autoencoders, to extract semantic features and make the data more insightful. We train two recent autoencoder architectures based on Variational Bayes and Diffusion on the two medical image datasets, German National Cohort and CheXpert. We compare their accuracy with existing classifiers and use the obtained embeddings for further qualitative and quantitative analysis.

Visualizing the embeddings using t-SNE plots and the provided labels, our analysis shows that the autoencoders can confidently extract and encode the differences in protected variables like the sex or age of the patient, even without supervision in the training process.

Utilizing the newly obtained representations, we reveal interesting, systematic, and unwanted clustering of data samples based on variations in the imaging processes in both datasets. Furthermore, we quantitatively check for any unfair behavior in the predictions of chosen diseases for any patient's sex subgroup and apply automated bias mitigation by directly removing it from the embeddings.

Contents

1 Introduction

Deep learning started to impact our daily lives years ago. It began with simple handwriting detection with the first convolutional networks for automatic zip code recognition [41] and developed into highly capable image, music, or text generation and translation tools [53, 16, 52], often already publicly accessible and in daily use. This rapid development is not only based on newly discovered model architectures and methods like diffusion [51] or attention [64]. As these new deep learning models are getting more complex and deeper, with billions of parameters [3], the datasets they are trained on also need to get more and more extensive.

Motivated by this, efforts have been made to collect large datasets to fuel the development of deep learning methods even more. But as astonishing as the results and accuracy of modern models are, as inscrutable is their concrete decision-making. Deep models occur to us as big black boxes proven to work [27] but are not quite understood yet. The same goes for large datasets, which are growing in size to help build better models and get increasingly less insightful to the plain eye. Whereas model explainability has experienced a rise in interest in academic efforts in recent years, careful examination of the underlying data is as important and often neglected. This can very quickly lead to incorrect and unfair behavior of such models.

Fair treatment and unbiased decision-making are legal requirements of many deep learning applications [2]. Especially when artificial intelligence is used for more than zip code recognition or image creation, e.g., influencing decisions in employment processes, permission to education, or in health care. Thus, identifying and explaining unfair behavior is of utmost importance. Some work has already been done on discovering unfairness in models or datasets. However, most rely on detection processes supported by human workers. This is costly and does not scale to large datasets and models. Therefore, other methods and metrics are needed to improve the process.

This thesis takes an approach to automatically discovering biases in datasets by making use of Autoencoders (AEs) to obtain alternative representations of the data with reduced dimensionality and complexity. Therefore, we will thoroughly introduce the principles of AEs and how they are utilized to create useful representations of images for further analysis. In particular, we describe and evaluate the performance of two of the most used AEs in recent years: Variational Autoencoders (VAEs) and Diffusion Autoencoders (DAEs). These models will be trained to encode medical images into vector representations by assessing their performance in reconstructing the original images from their representations, which are called embeddings.

To compare the use of such embeddings to existing standard classifiers for medical datasets, they will then be evaluated for accuracy by training additional predictors to extract important features out of the latent space of these encodings. Additionally, AE embeddings can more easily be used to evaluate any known or unknown bias found visually or with common metrics for bias quantification. Also, an attempt is made to mitigate bias by automatically calculating and removing the biasing impact found in the latent representation of specified protected attributes on the prediction accuracy and fairness metrics.

This analysis will be performed on two recent and large medical image datasets. Namely the German National Cohort (GNC) [15], a collection of T2-weighted Magnetic Resonance Imaging (MRI) of over 30,000 subjects examined across 18 German study centers, as well as CheXpert [30], a collection of X-ray images featuring 224,000 chest radiographs from over 65,000 patients with uncertainty labels for over 14 different findings [30].

Chapter 2 will provide background for the used models and will give an overview of the datasets examined. We cover the foundations of bias and fairness in machine learning and reiterate a selection of widely used fairness metrics. In Chapter 3, we subsequently describe the training process and data preparation, followed by the visual and quantitative analysis of the obtained reconstructions and dataset representations for accuracy and fairness. Finally, we conclude this thesis with a summary of our results and an outlook on future work in Chapter 4.

2 Background

The following chapter gives an overview of the deep learning architectures and datasets used in this thesis, followed by an introduction to fairness in machine learning and its assessment.

2.1 Autoencoders

AEs are a specific type of deep learning model architecture used to learn efficient encodings, i.e., they represent an unsupervised approach for learning lower dimensional feature representations of unlabeled training data. A first proposition of an AE was published as a generalization of principal component analysis by Kramer [38]. Although traditionally used for dimensionality reduction and feature extraction, the most popular use case of AEs nowadays are generative models, especially when conditioned on some input to, for instance, create a completely new image from a line of text.

Intuition

AEs usually consist of two main parts: an encoder that maps some input, like an image, to a code, also called latent space, and a decoder that tries to reconstruct the original input from this code. The structure can be seen in Figure 2.1. The goal is to optimize this reconstruction task but with the constraint that the dimension of the latent space is much smaller than the input dimension. Thus, the middle part of such an encoder is often referred to as the bottleneck layer of the model. The encoder and decoder can each consist of multiple layers. Common are fully connected layers of neurons if the data is in numeric form, or Convolutional Neural Networks (CNNs) and deconvolutions in case the input consists of images. [56]

Training the encoder-decoder pair works by forwarding a sample through the entire model and then comparing the original sample with its reconstruction. According to the differences, small iterative updates are made to the parameters of the encoder and decoder to ensure continuous improvement of the reconstructions. This can be accomplished utilizing a process called backpropagation, which is mainly done using gradient descent and is repeated for all training samples.

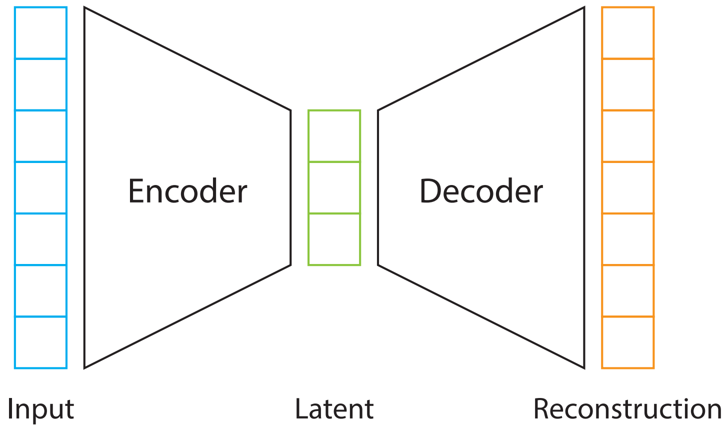


Figure 2.1: Structure of an AE

As mentioned, AEs are widely used as generative models. However, in this thesis, the interest lies in its feature extraction capabilities, i.e., in the latent representation of the input data before it gets decoded again.

Mathematical definition

If we describe our data as X, the space of original and decoded data, and denote Z, the space of the encoded data, AEs can mathematically be seen as two parameterized functions: One for the encoder $E_\phi : X \to Z$ parameterized by ϕ and one for the decoder $D_\theta : Z \to X$ parameterized by θ. A sample $x \in X$ would then be encoded into $z = E_\phi(x)$, the latent variable or latent vector. The reconstruction of x is denoted chiefly as $x' = D_\theta(z)$. [56]

2.1.1 Reconstruction

To assess the quality of the reconstructions and to mathematically define a task to be optimized while training, a so-called loss function is introduced. This function measures how well the reconstruction represents the input sample. With a function for the reconstruction quality, i.e., distance from the original, $d : X \times X \to [0, \infty]$, we can measure how much x' differs from x as $d(x, x')$. With that, we can define a general loss function of the AE as the expected distance:

$$L(\theta, \phi) := \mathbb{E}\left[d(x, D_\theta(E_\phi(x)))\right] \tag{2.1}$$

An optimal AE is achieved by minimizing this loss as $\arg\min_{\theta,\phi} L(\theta, \phi)$. [56]

Since no data labels are necessary, this represents an unsupervised feature extraction mechanism. This yields another use case for whenever data is not labeled or only labeled for a small portion of the dataset. Then, an AE can be used to learn the critical features of the distribution of the data to enable easier, further processing or training on this new representation.

2.1.2 Embeddings

Embeddings are typically known by their mathematical definition as the instance of a mathematical structure contained within another instance, such as a group. When embedding some object X into another object Z, the embedding is usually given by some injective map function $f : X \to Z$. This concept can be transferred to the AE domain, where some input data X gets encoded - or embedded - into some latent representation Z. So, the latent space is just an embedding of the data obtained by forwarding it through the encoder part of the model.

Unfortunately, to reconstruct an image properly, much information is still required to be stored in this latent representation. Even though this representation is used for dimension reduction, it can consist of a numeric vector with a dimension of hundreds to thousands. Preechakul et al. [51] gives an overview of the quality of different models compared to their latent space dimension when used for the FFHQ, a high-quality image dataset of human faces. The NVAE, a VAE producing high-fidelity reconstructions, needs a latent dimension of up to 6 million. Although the trend, especially using the DAEs proposed in the same paper that separate the details in the latent space from the semantic information, is going downwards, it is not straightforward to interpret a space with 512 dimensions. Further dimensionality reduction is required.

t-distributed Stochastic Neighbor Embedding

Even though the AE can reduce the data dimensionality, the latent space is still too highly dimensional for common visualization or plotting techniques like scatter plots. Fortunately, algorithmic dimensionality reduction is no new field for statistics. The well-known Principal Component Analysis (PCA) was already invented in 1901 by KPFRS [37]. With PCA, the data is linearly transformed into a new coordinate system such that the new axes capture the most significant variation of the data, and the main features can thus be easily identified. If transformed to the first two principal components containing the highest variation, it is possible to construct a two-dimensional plot that captures the most essential variation in the data.

But since PCA relies on a linear model, for a dataset with hidden, nonlinear patterns, it can no longer reliably create correct transformations. This problem was solved with further techniques like Stochastic Neighbor Embedding (SNE) [25], which proposes a new probabilistic approach to rearranging objects into a low-dimensional space by comparing their pairwise dissimilarities in their high-dimensional description while preserving their neighbor identities.

An improved version, called t-Distributed Stochastic Neighbor Embedding (t-SNE) [63], uses a symmetrized version of the SNE cost function with simpler gradients introduced by Cook et al. [11] and a Student's t-distribution rather than a Gaussian to compute the similarity between two points in the low-dimensional space. This alleviates both the crowding and optimization problems of SNE.

2.1.3 Variational Autoencoders

A significant limitation of AEs is that it is challenging to ensure that the encoder will smartly organize the latent space, which hinders the gathering of new insights into the data or its usage as a generative model. Since the AE is solely trained to encode and decode with minimal loss, the latent space is not regularized or forced to exhibit any particular order. Thus, the encoded data points and their positions have no relation or context to each other. As the task of the standard AE is not to enforce such organization, the need for regularized AEs arises.

The expected regularity of the latent space should fulfill continuity and completeness. This implies the following: the closer two points in the latent space are, the more similar their encoded inputs should be. Additionally, any point sampled from a chosen distribution of the latent space should give a meaningful output when decoded [55]. One type of regularized AE that tries to address this is introduced by Kingma and Welling [36] with a proposal of the VAE, an auto encoding variational Bayesian algorithm.

Within a VAE, the input gets encoded as a distribution over the latent space instead of as a single data point. For the reconstruction, a random point is sampled from that distribution and decoded to compute the loss and allow backpropagation through the network. In practice, Gaussian distributions are used so that the encoder trains to return the mean and covariance matrix describing it.

Formal scenario

According to Kingma and Welling [36], we describe the problem scenario as follows: we assume some dataset X consisting of N i.i.d. samples of some continuous or discrete variable x generated by some random process involving an unobserved continuous random variable z. Thus, the encoding process generates a value z from some prior distribution $p_{\theta^*}(z)$. In contrast, the decoding part then generates a value x from some conditional distribution $p_{\theta^*}(x|z)$. However, the true parameters θ^* and the latent variables z are unknown.

Two common problems arise with this concept: the integral of the marginal likelihood to compute $p_\theta(x)$ is most likely intractable if the likelihood function $p_\theta(x|z)$ is of moderate complexity [36]. This is the case for all neural networks with at least one nonlinear hidden layer. In addition, minibatch optimization should be possible for large datasets since parameter updates using the entire dataset are too costly. [36]

To account for these challenges, Kingma and Welling [36] describe the recognition model $q_\phi(z|x)$, an approximation to the intractable true posterior $p_\theta(z|x)$. The parameters ϕ will not be computed in a closed form but instead are learned jointly with the generative model parameters θ. Referring to the intuition above, the recognition model represents the encoder, encoding a data sample x into a latent space variable z, and $p_\theta(x|z)$ refers to the decoder producing a distribution over possible corresponding values of x.

The training task is as follows: the existing reconstruction loss remains, enforcing good reconstructions on the final layer. The reconstruction loss is usually calculated with the log-likelihood $\log p_\phi(x|z)$. However, an additional term for regularizing the latent space has been added. This term is expressed using the Kulback-Leibler-divergence [39], a measure for the distance between the approximate and the true posterior. This regularization often comes with an increased reconstruction error and needs to be adjusted and balanced properly. According to Kingma and Welling [36], using the closed form of the KL-Divergence, rearranging with the help of Jensen's inequality and the definition of evidence lower bound, this leads to the following model estimator:

$$\mathcal{L}(\theta, \phi) = -D_{KL}(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \tag{2.2}$$

One problem still arises: backpropagation is impossible through random sampling from the encoded distribution. Therefore, Kingma and Welling [36] proposes a novel solution: the reparameterization trick. Intuitively, the random sampling gets outsourced to another auxiliary vector ϵ, which is then used to compute the continuous random variable z. This allows backpropagation through the mean and covariance. For instance, when dealing with univariate Gaussian noise and if $z \sim p(z|x) = \mathcal{N}(\mu, \sigma^2)$, a valid representation is given by $z = \mu + \sigma\epsilon$, with the auxiliary noise variable $\epsilon \sim \mathcal{N}(0, 1)$. The resulting structure is shown in Figure 2.2.

Another variation of VAEs is the β-Variational Autoencoder (β-VAE), introduced by Higgins et al. [24], in which an additional hyperparameter β is introduced to balance the two components of the loss function. Usually, it is used to push the latent space toward a more disentangled representation of the input distribution. Disentanglement will further be described in Section 2.1.5.

This work also showed one first use case beyond feature extraction of image datasets. It enables sampling and manipulation of latent features like the degree of a person's smile or the rotation of an object in an image if learned by the model. Impressive results have been shown on the celebA dataset. [24]

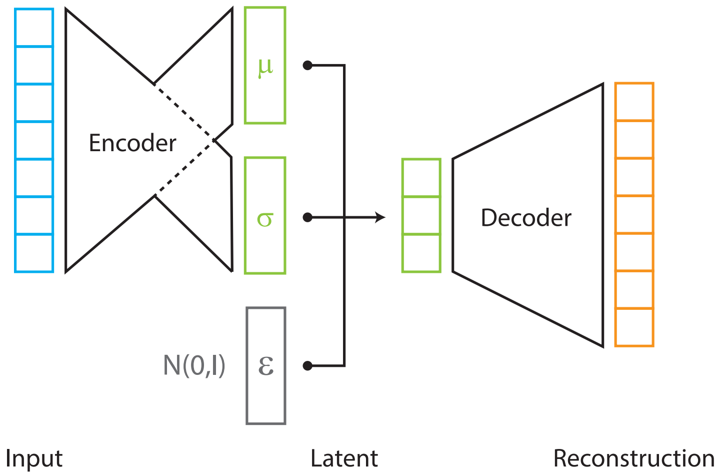


Figure 2.2: Structure of a VAE

2.1.4 Diffusion Autoencoders

Even though the latent representation of input data is now regularized and can be used for feature analysis and sample generation, the reconstructions of input data are not detailed enough. Images, especially, will not be reconstructed to be visually appealing. As seen in Section 3.1.1, the reconstructions tend to be blurry. The reason is that only global features are extracted, but the details of the input sample are not encoded into the reduced latent space. To improve the reconstruction quality, techniques using a hierarchy of latent codes have been proposed [54, 62]. However, the latent codes only encode spatial or local features for lossless reconstruction and can no longer capture global features or semantics.

A solution brings the work of Preechakul et al. [51], which introduces DAE, a new architecture of AE based on diffusion. This is a concept that was first proposed by Sohl-Dickstein et al. [57], initially inspired by processes from thermodynamics, and used by Ho, Jain, and Abbeel [26] for the first type of Diffusion Probabilistic Models (DPMs).

Intuitively, a DPM takes an input and adds, step by step, more noise to it until pure noise is obtained as a latent variable. This stochastic representation is then decoded by a trained decoder, who tries to reverse this process by progressively denoising the latent code to obtain the original image.

Formally, DPMs contain a Gaussian diffusion process with time steps $t \in T$ that increasingly add noise to an input image x_0 as

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}) \tag{2.3}$$

with hyperparameters β_t representing the noise levels [51]. The version of an image x_0 at time step t is then another Gaussian

$$q(x_t|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I}) \tag{2.4}$$

with $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$. To learn the autoencoding process, we are interested in the reverse process of this encoding: the distribution $p(x_{t-1}|x_t)$. If the gap between $t-1$ and t is not infinitesimally small, i.e., the amount of steps T is infinitely large, the complexity of this function increases. The encoding, however, has no parameters that need training, which makes optimization of this model type easier. [51]

To leverage DPMs for representation learning and to learn semantics in the latent space, Preechakul et al. add an encoder for high-level semantics to a DPM. This way, a two-part latent code is created. One part is semantical, meaningful, and linear and the other captures the stochastic details for near-exact reconstruction. Specifically, a conditional variant of the Denoising Diffusion Implicit Model (DDIM) [58] is used. Iterating over all denoising steps for image decoding or sampling with DPM takes significantly longer than with existing adversarial networks. This problem was tackled by Song, Meng, and Ermon [58] by modifying the forward process to be non-Markovian implicit probabilistic models but allowing deterministic encoding with the same training objectives as DPMs.

To achieve meaningful latent code, Preechakul et al. propose the following architecture: in addition to the DDIM stochastic encoder (inferred by reversing the generative process of DDIM), another semantic encoder, $z_{sem} = \text{Enc}_\phi(x_0)$, is added, mapping an input image x_0 to a semantically meaningful representation z_{sem}. Then, the DDIM image decoder $p(x_{t-1}|x_t, z_{sem})$ that is now conditioned on an additional latent variable z_{sem}, takes as input a latent variable $z = (z_{sem}, x_T)$, representing the high-level semantic subcode z_{sem} and the low-level stochastic subcode x_T. The authors provide a visual representation of a DAE in Figure 2.3.

Preechakul et al. [51] also conducted experiments with keeping the semantic code of an image but sampling different noises as stochastic latent code. This yields different

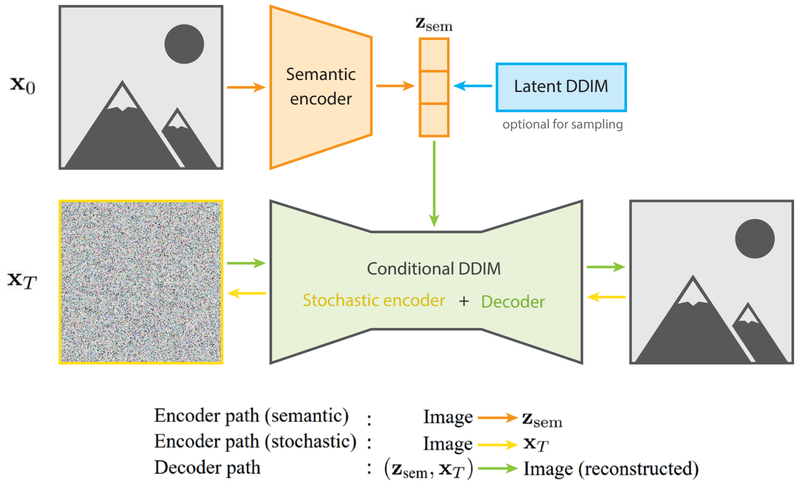


Figure 2.3: Structure of an DAE

images that all look the same but are different pictures in detail. Another experiment to show the capabilities of their DAE included predicting semantic attributes that the images were labeled with, only using the semantic code. This was accomplished with surprising accuracy, even outperforming some of the latest GAN-based models like StyleGan2 or VQ-GAN. [51]

With this, the model allows for even more demanding use cases. Conditioned image generation or image inpainting are probably the most popular ones [53]. Leveraging the new semantic encoding in combination with the high fidelity reconstruction that the VAEs did not offer, attribute manipulation and image interpolation is as easy as mutating or interpolating the semantic code and generating a slightly different image.

The focus of this thesis lies in the ability of both, VAE and DAE, to create high-quality feature extraction to be used for embedding creation of large datasets.

2.1.5 Disentanglement Strategies

Since high-quality embeddings aim to identify and visualize underlying factors hidden in the data, Disentangled Representation Learning (DRL) plays a major role in training the AEs for this data representation. DRL is the process of further separating these underlying factors of variation into variables with semantic meaning [65]. In other words, it limits the mutual information between different code dimensions. The importance of disentangled representations lies in their interpretability and generalization ability.

Disentangling yields more understandable and more easily readable results.

Common challenges and methods have been proposed by works of Bengio, Courville, and Vincent [4] or Locatello et al. [44] with detailed insights and comparisons of existing strategies. Locatello et al. [43] could even show that increased disentanglement in representation learning correlates with increased fairness in prediction models. There exist specific disentanglement strategies for domains like computer vision [34] or language models [9]. However, most proposed disentanglement strategies are model-specific and need to be adapted or will not even work with a different model architecture.

Hessian Penalty

This problem is addressed by Peebles et al. [49], introducing the Hessian Penalty, a simple - model and dataset unspecific - regularization term, showed to cause axis-aligned disentanglement and shrinkage of latent space.

The idea is to minimize off-diagonal entries in the Hessian matrix of the model's forward function. Considering, for simplicity, the scalar-valued model function $G : \mathbb{R}^{|z|} \to \mathbb{R}$ with the input vector z and its dimensionality $|z|$. Disentanglement is enforced when each component of z controls only a single aspect of variation in G. So varying any component z_i should produce a change in the output of G, mostly independently of other components $z_{j \neq i}$ [49]. Since the Hessian matrix represents all combinations of second derivatives of a function, setting an off-diagonal entry to 0 results in the following

$$H_{ij} = \frac{\partial^2 G}{\partial z_i \partial z_j} = \frac{\partial}{\partial z_j}\left(\frac{\partial G}{\partial z_i}\right) = 0 \tag{2.5}$$

This means if the outer derivative with respect to z_j of the inner derivative is zero, then $\frac{\partial G}{\partial z_i}$ is not a function dependent on z_j, i.e., as we change z_i, z_j does not affect the change of G's output [49].

The proposed regularization is then to minimize the squared off-diagonal entries of the Hessian matrix resulting in the Hessian Penalty

$$\mathcal{L}_H(G) = \sum_{i=1}^{|z|} \sum_{j \neq i}^{|z|} H_{ij}^2 \tag{2.6}$$

Since most deep networks are not vector-valued functions, dealing with images, video, or text, this formulation needs to be extended. A simple approach is to penalize the Hessian matrix of each scalar component in the output $x = G(z)$ individually. Denoting the collection of $|z| \times |z|$ Hessian matrices as $\mathbf{H}$, where $\mathbf{H}_i$ is the Hessian matrix of x_i, we get

$$\mathcal{L}_H(G) = \max_i \mathcal{L}_{H_i}(G) \tag{2.7}$$

In practice, computing the Hessian matrices during training is slow when $|z|$ is large. Thus an unbiased stochastic approximator is used to calculate the loss as

$$\mathcal{L}_H(G) = \mathrm{Var}_v(v^T H v) \tag{2.8}$$

where v are Rademacher vectors where each entry has an equal probability of being -1 or $+1$ and $v^T Hv$ is the second directional derivative of G in the direction of v times $|v|$ [49]. To further quickly compute the second directional derivative, another approximation is made using the second-order central finite difference as follows:

$$v^T Hv \approx \frac{1}{\epsilon^2}[G(z+\epsilon v) - 2G(z) + G(z-\epsilon v)] \tag{2.9}$$

With the hyperparameter $\epsilon > 0$ controlling the granularity of the second directional derivative estimate. This results in a regularizing penalty term that can easily be added to the existing loss of the model, regardless of the dataset or model architecture.

2.2 Medical Datasets

With the rise of deep learning for all varieties of tasks and the magnificent improvement in computer vision throughout the last decade, many have tried to apply these newly found methods to the medical domain. However, since modern computer vision models' sizes mostly range to several billion parameters, a vast amount of labeled data is necessary for successful training. Especially for medical appliances and the evaluation of medical images, there is also a high need for reference standards and metrics by expert humans for comparison. The prolonged absence of standardized and large-scale studies has hindered the progress of medical deep learning from the beginning. But in recent years, more and more studies and datasets have been published that try to meet these requirements.

This thesis deals with two recent datasets collected over the last few years. The GNC (2014 to 2022) and CheXpert (2019). These datasets provide high-resolution MR images and chest X-rays. MRI and chest radiography are the most common imaging examinations globally and are critical for screening, diagnosing, and managing life-threatening diseases [30]. Assessing, evaluating, and automatically interpreting these medical images would provide substantial benefits in many medical settings.

2.2.1 German National Cohort

The GNC is an interdisciplinary study from multiple research institutes across Germany to investigate causes for the development of major chronic diseases, like cardiovascular diseases, cancer, diabetes, neurodegenerative/-psychiatric diseases, musculoskeletal diseases, respiratory and infectious diseases, and their pre-clinical stages or functional health impairments. From 2014 on, distributed over 18 regional study centers, a total of 100,000 women and 100,000 men aged 20 to 69 years have been examined. [15]

The data collection of GNC comprises two levels of intensity. All participants were assessed through a recruiting protocol, a computer-assisted personal face-to-face interview questionnaire, and basic physical and medical examinations (Level 1). A random sub-sample of 40,000 participants participated in an extended protocol that included more in-depth physical and medical examinations (Level 2). Additionally, at five MRI centers in Augsburg, Berlin, Essen, Mannheim, and Neubrandenburg, a total of 30,000 participants were asked to undergo a high-resolution 3T MRI protocol for the acquisition of whole body, cardiac, and brain images (MRI program). After four to five years of baseline assessment (2014 to 2018), all participants were invited for a re-assessment (2018 to 2022). Although most of the assessment was done by 2022, the entire time frame for the GNC covers a period of 25 to 30 years for further follow-up studies on participants. [15, 50]

This thesis focuses on the intensified subcohorts of Level 2 and the MRI program for computer vision purposes. We use sagittal T1-weighted MR images of three regions of interest to train our AEs. The neck, chest, and lumbar regions. Those were obtained from a total of 11,186 participants. The MR images have an original resolution of 448 by 448 pixels and overlap between the regions. Also, the MR intensity decreases towards the edges. Therefore, cropping transformations were applied to a final resolution of 256 by 256 pixels.

For each imaging process, information that was directly collected and is available without missing values includes the region of the scan, the survey center or institution where it was acquired, and patient-specific information about sex, age, height, and weight. With the help of the digital survey, additional but less reliable information about the patient's well-being and habits has been collected.

2.2.2 CheXpert

CheXpert (Chest eXpert) is a medical image dataset that contains over 224,000 chest radiographs from over 65,000 patients. It was assembled by personnel of the Stanford University Hospital between 2004 and 2017 and published by researchers of Stanford University in 2019. In addition to making the dataset publicly available, a labeler was designed to detect the presence of 14 common chest radiographic observations from free-text radiology reports. The presence is classified into three different types of uncertainties: either negative, positive, or uncertain. The labeler was then evaluated using a set of 1,000 randomly sampled radiology reports annotated by two board-certified radiologists without access to additional patient information. [30]

This multiclass labeling approach leaves room for different approaches to handling uncertain cases. Irvin et al. [30] propose five different methods. One is to ignore all samples with uncertain labels. Trying to keep as many samples for training, a second way is to perform a binary mapping of all uncertain labels to either 0 or 1. This can be extended by later work to use an approach called Label-Smoothing Regularisation [60]. Here, the labels are randomly sampled using a uniform distribution to prevent the model from using wrongly labeled examples with excessive confidence during training.

Together with these automatically labeled samples, a smaller set of 200 images, manually annotated by expert board-certified radiologists, was published for validation purposes. Additional information is available about the patient's age and sexuality and the projection type of the X-ray process. Chest X-ray images can be obtained by frontal or lateral imaging. Frontal imaging is commonly divided into Posterior Anterior (PA) and Anterior Posterior (AP) projection. For PA imaging, the patient needs to be standing against a vertical device, and the rays run through the back first and front second. This type of projection is always preferred because it yields better results. For instance, patients usually find it easier to breathe in so that the diaphragm gets pushed down and more parts of the lung are visible. In cases where the patient is not able to stand or reach the X-ray facilities, portable X-ray machines are used, which can create images of patients lying in their beds or on a special table. Here, the rays run through the patient from front to back. This type of projection used with portable devices is mostly of poorer quality.

The researchers also developed models as a baseline to predict the 14 observations from a single-view chest radiograph. They were compared using the Area Under the Curve (AUC) metric (see Section 2.3.4) and focused on evaluating five observations, called the competition tasks, selected based on clinical importance and prevalence: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. [30]

As shown by Irvin et al. [30], some significant differences in the performance (e.g., in predicting Cardiomegaly) were found for a few approaches to dealing with uncertainties. But no best model was significantly better than the worst, especially on Consolidation, Edema, or Pleural Effusion. The authors also hold back a test set composed of 500 images annotated by a consensus of five board-certified radiologists to evaluate and compare new works submitted. The CheXpert dataset now serves as a standard benchmark to evaluate the performance of all different types of radiograph interpretation models. [30]

A lot of additional work was done using this dataset. Structured overviews of the data distribution have been created [19], and an improved labeling version, CheXpert++, was developed [45]. But the most significant work was done by Yuan et al. [67] for publishing a model in 2020 that is still ranked number 1 on the CheXpert leaderboard for the highest AUC score [31].

The work by Crespi, Loiacono, and Chiti [13] also already used VAEs to reconstruct these images and analyze the model's feature space. They used popular and pre-trained CNN architectures as backbones and replaced the last connected layer with layers providing the mean and standard deviation of the learned distribution. However, they kept the weights up until the last layer as they were pre-trained on CheXpert, which was previously done in a supervised manner. Instead, this work focuses on creating the embeddings without available labels.

2.3 Fairness in Machine Learning

Since related work has yet not come up with a common or united definition of fairness, this thesis tries to outline basic assumptions and categorize different causes and consequences of unfair behavior of statistical or self-learning models.

Artificial intelligence can learn and predict unfairly for a variety of reasons. The phenomenon of a systematic deviation from the norm is as old as human civilization itself and is often referred to as bias. It can be seen as an inclination toward something or a prejudice or preference. It can express itself on a daily basis as cognitive or cultural bias, where an individual's perception of reality is never objective but dependent on their perception of circumstances and how they learned to judge them.

This concept also applies to bias in machine learning, where the training data through the selection and preparation process, the training algorithm and its design choices, and the user who interprets and uses the results all influence the system's decision-making. Since machine learning models learn specific tasks by extracting relevant statistics from data, it is common that they unintendedly learn and use spurious correlations between protected attributes like age, gender, race, or other important features or even the target feature. Having the best, unbiased system in a technical sense is very likely to produce a biased result in a legal sense if trained on the wrong data. This lets us categorize bias in different categories that can also amplify each other's contribution to biasing the results.

2.3.1 Types of Bias

A categorization of bias has already been outlined by Suresh and Guttag [59] and Olteanu et al. [48]. As suggested in Mehrabi et al. [46], this section provides explanations and explores a non-exhaustive selection of biases from a technical perspective by classifying their causes into three types of bias:

Data to Algorithm Bias

As mentioned above, biased data can, quite intuitively, lead to a biased model and biased algorithmic outcomes. This includes the following typical occurrences of dataset bias:

- Measurement Bias. This type of bias describes differences through how particular features are chosen, utilized, and measured. [59]
- Representation Bias. This bias can arise by sampling from a population during data collection. [59]
- Sampling Bias. In contrast to representation bias, it arises from non-random sampling of subgroups. Estimations or representations learned for one group might not generalize to a new group of samples. [46]
- Aggregation Bias. False conclusions about individuals that are derived from observing an entire population are known as aggregation bias. [59]

Algorithm to User Bias

This category deals with biases resulting from an algorithmic point of view. Since algorithms also modulate user behavior and thinking, this type should not be neglected.

- Algorithmic Bias. When the bias is not present in the data but added by the algorithm, one speaks of algorithmic bias. This could result from design choices like optimization functions, regularizations, or loss metrics. [1]
- User Interaction Bias. This bias can be influenced by imposing self-selected biased behavior and interaction, or presentation and ranking of information the user consumes. [1]
- Popularity Bias. The popularity of items or results has a natural impact on an individual's perception. For example, fake news spread over search engines or social media can influence this. [10]
- Evaluation Bias. Depending on the selection of benchmarks or scores, the evaluation can also bias the outcomes of a model's decisions. [59]

User to Data Bias

Since many data sources are user-generated, any bias in user behavior can inherently be present in the generated data. In the case of medical datasets and patients, a natural bias occurs because certain observations or samples are only gathered because a patient was already found ill.

- Historical Bias. Even if the data is flawlessly obtained, this bias can arise if the circumstances as they are lead to unwanted outcomes. For instance, having the perfect dataset about alcohol consumption in the population could still reflect historical factors that the older and male population tends to have increased consumption rates. Even if this accurately reflects the world, it can still harm certain subgroups or individuals in the population. [59]
- Population Bias. This type of bias occurs when statistics or characteristics differ in the user population from the original target population. [48]
- Self-Selection Bias. A subtype of selection bias, where subjects of research select themselves. This is common for voluntary surveys or polls, where a certain interest or enthusiasm for the research is conducive to participation probability. [46]
- Social, Behavioral, Temporal, and Content Bias. These are types of bias resulting from social interaction, mutual influence on judgment, blending different domains or contexts, or changes and differences over time. [1, 48]

2.3.2 Fairness

Whereas bias describes the more technical approach of reasoning what causes can lead to a model's incorrect behavior, fairness is often seen as a term for the social impacts on affected users or subgroups. Therefore, it is common to categorize unfairness into the types of harms that result from it.

Types of Harms

An excerpt of existing harms is presented by [12] and summarized in [32]. The following list reiterates the most important types.

- Allocation Harms. This harm occurs when a model's behavior leads to opportunities or resources being withheld or extended to certain subgroups. The most common domain is decision-making regarding hiring or lending.

- Quality of Service Harms. If a model performs significantly better or worse for one subgroup or individual, we speak of harm in the quality of service.
- Stereotyping Harms. Often related to recommendation or auto-completion algorithms, stereotyping harm arises when a model makes suggestions based on stereotypes and does not treat people as individuals.
- Erasure Harms. This phenomenon results from a model that does not consider certain groups, individuals, or their work for their decisions based on their unpopularity.

Intuitively, unfair behavior can result from all possible combinations of bias, leading to harms that are not exclusive. Multiple different biases can lead to one specific harm, but one specific bias can result in various social harms.

For instance, Buolamwini and Gebru [6] showed imbalances in popular datasets like IJB-A, a facial image dataset widely used for early computer vision model training. Within IJB-A, about 80% of the samples are light-skinned subjects. This leads to a representation bias for underrepresented dark-skinned groups.

Also, the way of analyzing data can lead to bias when different subgroups are not considered individually. By intuition, dividing the data into male-female groups should give the correct insights into whether the data is biased towards one group. However, as Buolamwini and Gebru [6] mentions, only when dividing the data further by race into light-skinned females, light-skinned males, dark skinned-females, and dark-skinned males, a clear bias towards dark-skinned females can be observed. This bias was previously hidden and comprised of light-skinned males. Usage of this biased dataset can lead to multiple social harms such as quality of service harm because the model cannot be adequately trained on the underrepresented subgroups and performs worse, or stereotyping harms where the model makes suggestions generally favored towards light-skinned male individuals.

Another example of common bias occurs in the case of medical applications. Already in data collection, one has to consider the healthy user bias. This is a selection bias with potentially unrepresentative participants. People volunteering for clinical studies are likely more concerned about their health and predisposed to follow medical advice. This problem can also occur the other way around. When specific medical investigations are only carried out if a patient's health status demands it, a particular bias towards observations of ill patients could arise. For instance, X-rays are unlikely to be taken of patients without any symptoms or indications of a potential disease. This would lead to quality of service harm since the average individual is not represented correctly and will not be assessed right by a model trained on this data.

This concludes the brief overview of the causes and impacts of unfair models. Studies in this area are as critical as the models themselves. Only with a good understanding of how data, algorithms, and users interact and influence each other's tendencies, one can discover, understand, and counteract unintentional model behavior.

2.3.3 Assessment

Assessing fairness is not an easy task. Fairness is a mostly unobservable theoretical construct that cannot be directly measured but must instead be inferred through observable measurements [32]. Much work has already been done to extract these observations. But most work is model- or dataset-related. Le Quy et al. [40] is an example of an extensive study on dataset fairness. Amongst others, optimization of Bayesian Networks has been used to understand relationships between dataset features. However, observing fairness in image datasets is more complicated due to the images' sheer size and complexity.

Common pipelines exist for discovering bias in big image datasets, which are mostly human-reliant [29]. The proposed crowdsourcing process consists of three steps. First, crowd workers are presented with random subsamples of images from the dataset and are asked to find similarities, which will serve as candidate-biased attributes for the next step. Second, workers would examine different subsamples and check whether the extracted information from the previous step is present in most images of the subsample. In the third step, the workers are asked whether they believe that the statement of the prior step accurately reflects the real world based on their knowledge and subjective beliefs. Sorting these statements in decreasing order of the fraction of workers who indicate it as not accurately reflecting the real world gives a ranked list of potentially biased attributes of the image dataset.

However, as promising as the results are, this method is not scalable to big datasets and has to be done manually for each new dataset, which needs bias evaluation. So the urge for automated bias detection rises. One of the first approaches of analytically determining bias and making it easier to scale to more extensive datasets was created by Li and Xu [42]. In their work, the authors reduce the problem of finding a biased attribute to optimizing a hyperplane in the generative model's latent space. This can thus be automated and no longer relies on a crowd of workers analyzing images.

In this work, we use a similar approach. Using AEs - also generative models - the image datasets get encoded into embeddings to make them more insightful. Since these embeddings are just an extraction of the features found in the image, they can be used to discover unknown biases by visualizing them or performing common metrics and comparing quantitatively.

Many measurements need to be combined to quantify a model's fairness. The most widely used metrics are reviewed in the following section.

2.3.4 Fairness Metrics

To quantify fairness, we define a collection of measures below. Within the definitions, we use X for any input features that are used for making predictions from a certain data sample, Y for the corresponding true features, and $\hat{Y}$ for the predicted output features. V is a set of sensitive or protected features that should be investigated for fair behavior within the dataset or model. It is important to mention that, in general, X and V may or may not share features. Hence, V might be used by the predictor. This gets increasingly complex for medical image datasets since most images already implicitly contain protected features specific to the patient. However, like Dunkelau and Leuschel [17] define, a protected attribute is a property of an individual that must not influence the decision process of the machine learning algorithm.

Bias Amplification

The bias amplification of a predictor is described by Zhao et al. [68], comparing some bias prevalent in the dataset with bias prevalent in the predictions made. Accordingly, the bias score of a given output, $y \in Y$, with respect to a protected variable, $v \in V$, is given as:

$$b(y,v) = \frac{N_y^v}{\sum_{v' \in V} N_y^{v'}} \tag{2.10}$$

where N_y^v is the count of samples of class y with protected attribute v. The bias amplification score is helpful to evaluate the degree of bias amplification by the used predictor. Therefore, we additionally define the dataset bias, $b^*(y,v)$, for the respective counts in the dataset, i.e., the true labels, and the prediction bias, $\hat{b}(y,v)$, for the data annotated by the predictor.

Then, if y is positively correlated with v (i.e. $b^*(y,v) > \frac{1}{|V|}$ and $\hat{b}(y,v)$ is larger than $b^*(y,v)$, we can say that bias has been amplified.

With the indicator function $\mathbb{I}$, the score used for comparing the predictions of the different embeddings is the mean bias amplification for all pairs of y and v:

$$\frac{1}{|Y|} \sum_{v \in V} \sum_{y \in Y} \mathbb{I}_{b^*(y,v) > \frac{1}{|V|}} \left(\hat{b}(y,v) - b^*(y,v) \right) \tag{2.11}$$

Intuitively, the higher the score, the more the bias was amplified. A score less than zero means a bias reduction.

Demographic Parity

As introduced by Dwork et al. [18], demographic parity is a score to assess the independence of a prediction $\hat{y} \in Y$ from a protected binary attribute $v \in V = \{0, 1\}$ such that the following equation holds:

$$p(\hat{y} = y' | v = 0) = p(\hat{y} = y' | v = 1) \tag{2.12}$$

In other words, the prediction of the examined predictor should not depend on the value of the protected attribute. According to Beutel et al. [5], this can be transformed into a parity difference score:

$$\frac{1}{|Y|} \sum_{y \in Y} \left| \frac{TP_y^1 - FP_y^1}{N^1} - \frac{TP_y^0 + FP_y^0}{N^0} \right| \tag{2.13}$$

Here, TP_y^v and FP_y^v are the counts of true positives and false positives of class y with protected attribute v, and N^v is the number of images with protected attribute v. A score of 0 would imply demographic parity and fair behavior accordingly.

Equalized Odds

Equalized Odds is a slight modification of demographic parity. As introduced by Hardt, Price, and Srebro [23], it also considers the true y label. Therefore, it assesses the independence such that

$$p(\hat{y} = y' | y = y, v = 0) = p(\hat{y} = y' | y = y, v = 1) \tag{2.14}$$

holds. Here, the prediction $\hat{y}$, conditional on the ground truth y, should be equal for all protected attribute V values. Using this, we can measure the prevalent bias as:

$$\frac{1}{2} \left(\left| FPR_y^1 - FPR_y^0 \right| + \left| TPR_y^1 - TPR_y^0 \right| \right) \tag{2.15}$$

This yields values between 0, where Equation 2.14 is fulfilled and no bias is detected, and 1, which implies maximum bias.

ABROCA

To further analyze the fairness of a predictor over the underlying probabilities leading to its decision-making, one can compare the Receiver Operating Characteristic (ROC) curve across subgroups of a protected attribute. The ROC curve is a plot of the false positive rate against the true positive rate of a classifier's prediction across all possible thresholds $t \in [0, 1]$, where t determines at which probability or certainty the model predicts that $\hat{y} = 1$. From this curve, one can derive the AUC as a score for performance. Its values vary between 0 and 1. A higher score implies a better model performance. A score of 0.5 is equivalent to random guessing. This measure is one of the most commonly used model performance indicators and is known to be robust to imbalanced data [33].

A naive approach to comparing the scores for two subgroups would be to calculate the direct difference:

$$AUC_b - AUC_a \tag{2.16}$$

However, Gardner, Brooks, and Baker [20] point out the problem of cancellations of areas wherever the ROC curves cross and proposes a measurement called Absolute Between-ROC Area (ABROCA). It is the total difference across all possible thresholds between two points on each of the ROC curves from the two compared subgroups of the protected attribute. It is formally defined as:

$$\int_0^1 |ROC_b(t) - ROC_a(t)| \, dt \tag{2.17}$$

Intuitively, ABROCA is the area between two ROC curves and can vary between 0 and 1, where lower values imply fewer differences between the subgroups.

2.3.5 Mitigation

Although this thesis mainly focuses on finding and explaining bias in the examined datasets and the succeeding prediction models, mitigation of the prevalent bias is as important and necessary to create fair model predictions.

Literature has proposed numerous strategies for bias mitigation. Like the causes for bias that can appear in the data, the algorithm, or the user space, mitigation strategies can be distinguished similarly, based on where in the process from the data to the predictions they apply. The following three categories emerge: Pre-processing, In-processing, and Post-processing techniques. They deal with manipulating the input data, adapting the algorithms used to train and build the model, and adjusting the outputs or predictions that the models make, respectively. Dunkelau and Leuschel [17] and Hort et al. [28] have produced comprehensive studies that further categorize mitigation strategies depending on the type of changes made to the process:

Pre-processing Bias Mitigation

Pre-processing strategies usually affect the data used for training. This can be done by altering values of the ground truth labels (relabelling) or other features (perturbation) as used by Calders, Kamiran, and Pechenizkiy [7]. Sampling instead tries to counteract data imbalances by changing the distribution of samples by over- or undersampling (e.g., Kamiran and Calders [35]) as well as reweighing samples (e.g., Calders, Kamiran, and Pechenizkiy [7]).

In-processing Bias Mitigation

Applying changes to the model or its process of prediction is called an In-processing strategy. Here, for instance, bias mitigation can be achieved by regularization or constraints in the model's architecture. In addition to an existing loss function of the model, a regularisation term or constraint is added to penalize any unfair behavior or determine thresholds that must not be exceeded respectively. The idea of adversarial learning is also helpful for automatically mitigating bias. Here, an existing classifier tries to predict the ground truth, while an adversary model is trained to exploit fairness issues arising from the prediction [14]. Training a composition of multiple classifiers can also help mitigate bias. One classifier gets trained per group of individuals and is then used for predictions for that specific group. As a last effort to mitigate bias, one can adjust the learning methods and create novel algorithms like Dunkelau and Leuschel [17].

Post-processing Bias Mitigation

Post-processing mitigation can deal with so called input corrections, which are similar strategies to relabelling or perturbation but done to the testing data. Output corrections are usually done by modifying the predicted labels. Also, once the classifier is trained, for instance, Calders and Verwer [8] adapt the model after training to perform classifications independent with respect to a given sensitive attribute.

As this brief overview suggests, numerous mitigation techniques have been proposed. Hort et al. [28] gathered a total of 341 publications relevant to the context of bias mitigation in machine learning models. The effort of establishing a systematic comparison and visual benchmarking of bias mitigation strategies was made by Z. Wang et al. [66].

Image Classifier Bias Mitigation

The problem with most strategies for bias mitigation is that they are not quite scalable, especially for image datasets. For instance, Nguyen, Bouzerdoum, and Phung [47] use a weighted clustering technique to resample the training dataset into a smaller set of representative training examples. However, such methods are not feasible for high dimensional data like images, where, without significant pre-processing, there is no notion of data clusters. This further incentivizes the use of image embeddings as a general pre-processing technique.

Thong and Snoek [61] continue on this thought and propose a pre-processing bias mitigation method specific to image classifiers. They use the feature representation of images within any standard CNN classification architecture to apply modifications before the final classification step. Since AE supported classification is focused on finding the most suitable feature representation first, this method can easily be leveraged without any further changes to the process.

The proposed method includes the following steps: Suppose an encoder provides the feature representation of any input sample x: $h = \mathrm{Enc}(x)$. First, we calculate the average representation of each predicted class value $y \in Y$ and protected class value $v \in V$ in the feature space.

$$\mu_y^v = \frac{1}{N_y^v} \sum_i \mathbb{I}[y_i = y, v_i = v]\mathrm{Enc}(x_i) \tag{2.18}$$

Second, the bias direction is computed, which is the difference between the averages for each predicted class. For a binary-protected attribute, we get:

$$\boldsymbol{\Delta} = \{\mu_y^1 - \mu_y^0 | y \in Y\} \tag{2.19}$$

For the mitigation, we use the first principal component of Δ, which we call b, and perform a simple mathematical projection of the features h on this bias direction b and obtain h_b. This will then be subtracted from the original feature space h:

$$\tilde{h} = h - h_b = h - \frac{h \cdot b}{\|b\|} \frac{b}{\|b\|} \tag{2.20}$$

3 Methods and Results

In this chapter, we describe the training process of the AEs and how each dataset was prepared. Subsequently, we summarize the results of the reconstructions and the learned embeddings, followed by a visual analysis of the embeddings to discover unknown biases in the data. Finally, a quantitative analysis is performed to compare the obtained embeddings with existing results regarding both accuracy and fairness.

3.1 Model Training

The VAE and DAE were each trained on the GNC and CheXpert datasets. Both datasets were transformed into one-channel grayscale images. The GNC images were randomly cropped to the desired resolution and random horizontal flip was applied for data augmentation. For CheXpert, we used resizing and center cropping to obtain the correct resolution for the used model architectures. The Adam optimizer was used for all models.

The Hessian penalty was used and, as described in Section 2.1.5, added to the loss term of both model types to further enforce disentanglement within the latent space. The weight was first chosen to scale the penalty according to the other loss scores and was fine-tuned later.

First, the hyperparameters were chosen strategically to enforce the convergence of each model, and then manual and automated grid search was used to fine-tune the most important parameters like learning rate, latent dimension, and shape of the input images or other model-specific parameters. The β-parameter of the VAE was found to work best at a value of 4.

An overview of the most important parameters we used can be taken from Table 3.1. The entire practical part of this work was performed on a single GPU and thus not focused on finding the optimal parameters to surpass current deep learning models. The main interest lies in showing the concept of differences between the different model types and their performances on the selected recent datasets, especially with respect to possible fairness issues.

The quality of the models was assessed not only by quantitive comparison of their respective losses but also by visually inspecting their reconstructions. A selection of

Table 3.1: Overview of parameters used for all examined model architectures

	NAKO		Chexpert	
	VAE	DAE	VAE	DAE
Learning Rate	5×10^{-3}	1×10^{-3}	25×10^{-4}	5×10^{-4}
Hess. Pen. Weight	1×10^{-6}	1	5×10^{-7}	1
Input Shape (px)	256	256	256	256
Batch Size	64	5	128	64
Latent Dim.	128	512	128	512

reconstructions for each model can be seen in Figure 3.1 for GNC and Figure 3.2 for CheXpert.

3.1.1 Variational Autoencoder

For the VAEs, additionally, the β-value and hidden dimensions were optimized in the described way. Structurewise, we used a symmetric encoder-decoder structure with five CNN layers each. While encoding, the filter dimension gets doubled for each layer while the resolution gets halved on each side. After each two-dimensional Convolution, a Batchnorm and leaky ReLU are added. After the last convolutional layer, two fully connected layers are trained to extract the mean and variance of the encoded distribution. After the reparameterization, a single fully connected layer is used before the decoder part, which is completely symmetric to the encoder.

GNC Reconstruction

Figure 3.1 shows the originals and reconstructions of six randomly selected MRI of, in this order, the cervical, thoracic, and lumbar regions. We can see that the VAE is far from reconstructing the image with all its details. The main outlines of the spine or the spinal canal containing the spinal cord are reconstructed and recognizable. In reconstruction example a), even a little ridge on the patient's neck was correctly reconstructed. However, further details about vertebrae or intervertebral discs are lost in the reconstruction process. Since the bottleneck of the VAE is designed to only hold most semantic information about the encoded image, the VAE was expected to show less potential in reconstructing the samples in detail than rather extracting a usable feature representation as will be discussed in the following sections.

CheXpert Reconstruction

Originals and reconstructed chest X-rays of the CheXpert dataset can be seen in Figure 3.2. The selected views are, in this order, AP, PA, and lateral. The reconstructions show again that the VAE struggles with keeping all details of an image throughout the latent dimension. The X-ray dataset CheXpert is generally more challenging for AE because even more details, like electrodes, cables, or heart devices, must be captured for a correct reconstruction. Nevertheless, the key elements of the input images are obtained. The outlines and edges of the spine, the lung, and the diaphragm are visible in the reconstructions. Even bigger-sized hearts in examples b) or d) are captured. Also, just by examining the reconstructions, the VAE achieves to clearly separate and correctly reconstruct frontal and lateral X-ray images. Further details are again lost in the process of reconstruction.

Another problem occurred with this richer and higher-resolution dataset. The decoder of the VAE struggled to upsample coherent images from the latent space that would not look like different chunks patched together. We experimented with swapping the transpose convolution layers with upsampling followed by a convolutional layer and observed a slight mitigation of this behavior.

3.1.2 Diffusion Autoencoder

Regarding the DAEs, the hyperparameter optimization was mainly done for the GNC dataset. Important parameters, like input and output dimensions and the learning rate, were adjusted to CheXpert. The model architecture is based on the implementation published by Preechakul et al. [51].

GNC Reconstruction

As explained in Section 2.1.4, the DAEs uses an additional latent space independent from the semantic information, storing the specific details of the input image. As expected, the reconstructions (Figure 3.1) show that the DAEs can obtain most of this information from the input to create near-perfect reconstructions. For instance, the reconstruction of b) shows how each vertebra and intervertebral disc is well reconstructed. But still, differences can be seen in the lower end of the spinal cord where it thins out. Possibly a learned concept from other images in the lumbar region. Also, the positioning and edges of the ligament and spinous processes are sometimes blended and not well differentiated. The reconstruction of image c) even created an additional vertebra in an unusually wide intervertebral space. However, reconstructions can still be used to assess patients' health status, and they are thus a significant and necessary step towards perfect reconstruction and automated assessment.

CheXpert Reconstruction

In the case of CheXpert (Figure 3.2), most of the bones, prominent tissue, and the shape of the heart or diaphragm are well-reconstructed. The only details that this model cannot fully reconstruct are some devices or, most of the time, the wires of the electrodes used for electrocardiography. Those can mainly be found on the AP view of a patient's X-ray, which is generally only taken if the patient cannot stand up for a proper PA scan. This leads to almost always less perfect reconstructions with this type of view. We can also see that an additional difficulty is presented with frontal X-ray scans of female patients, where the breast, as an additional organ, is in front of important other chest regions that could have to be examined. The reconstruction of b) especially shows that the outlines of the breast are not perfectly reconstructed and affect the reconstruction of the underlying diaphragm. Subject to skepticism is the model's behavior in example f), where it reconstructs the breast of a female patient larger than it was. Considering that there was no patient information available for the AE apart from the original image, it must have learned that this type of scan is of a female patient and was biased from other images it had learned from so far to reconstruct the breast to a more average size. This could be attributed to the fact that more frontal images are available, which will be further illuminated in our bias analysis.

3.2 Embeddings

The goal of training the AEs in this thesis was to use the new representation of the images - their embedding - for further analysis in the following section.

3.2.1 Creation

The embeddings are obtained by forwarding the images only through the encoder part of the AE models. They are then embedded into the latent code and represent the original images as a numeric vector. We then reduce the dimensionality of the vectors using t-SNE, iteratively moving similar data points together. We are then left with two highly meaningful dimensions that can be plotted in an ordinary coordinate system, where each point represents the encoded version of one image.

When inspecting these plots, we notice that they are often arranged into groups which we call clusters. To visualize how these encodings represent the original information only given by the image itself, we colorize the points according to the most important attributes known about the patients. So, for each inspected attribute, we obtain an additional plot of the same embedding, showing how well the arrangement of the points correlates with the different values for this attribute. Note again that the model

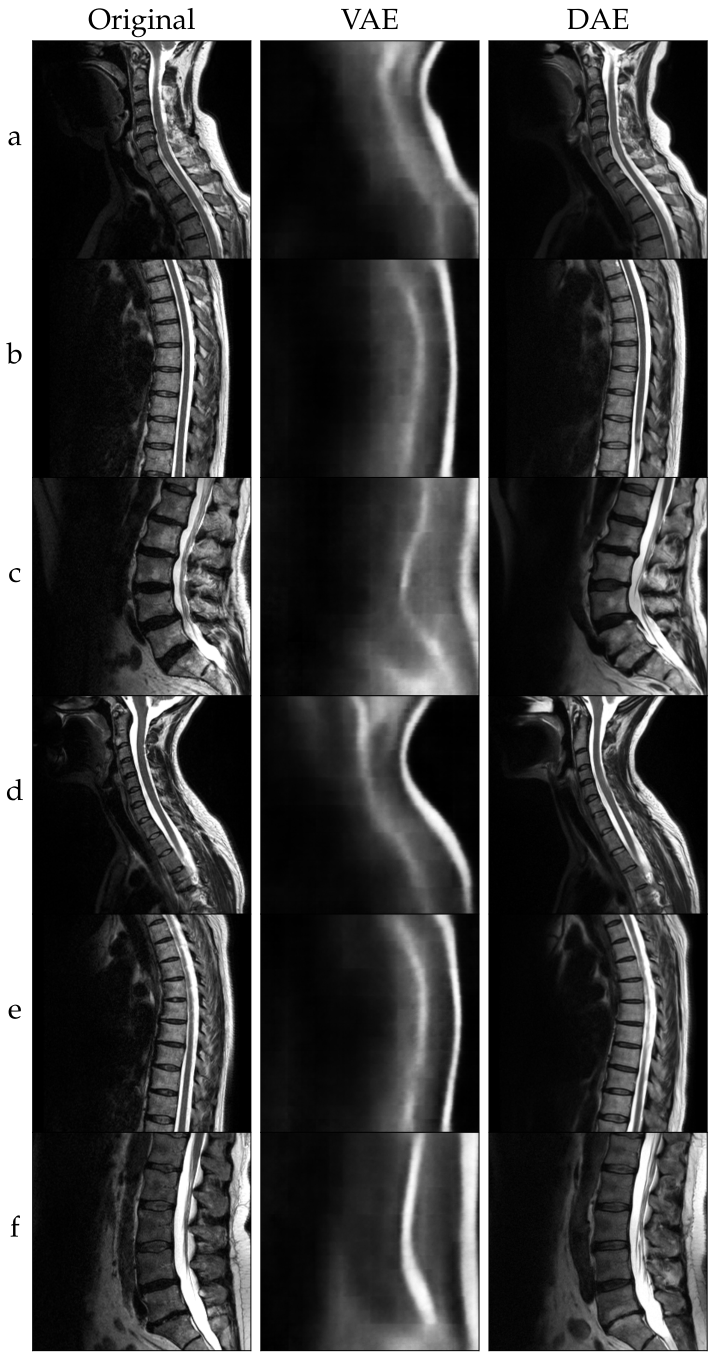


Figure 3.1: GNC reconstructions

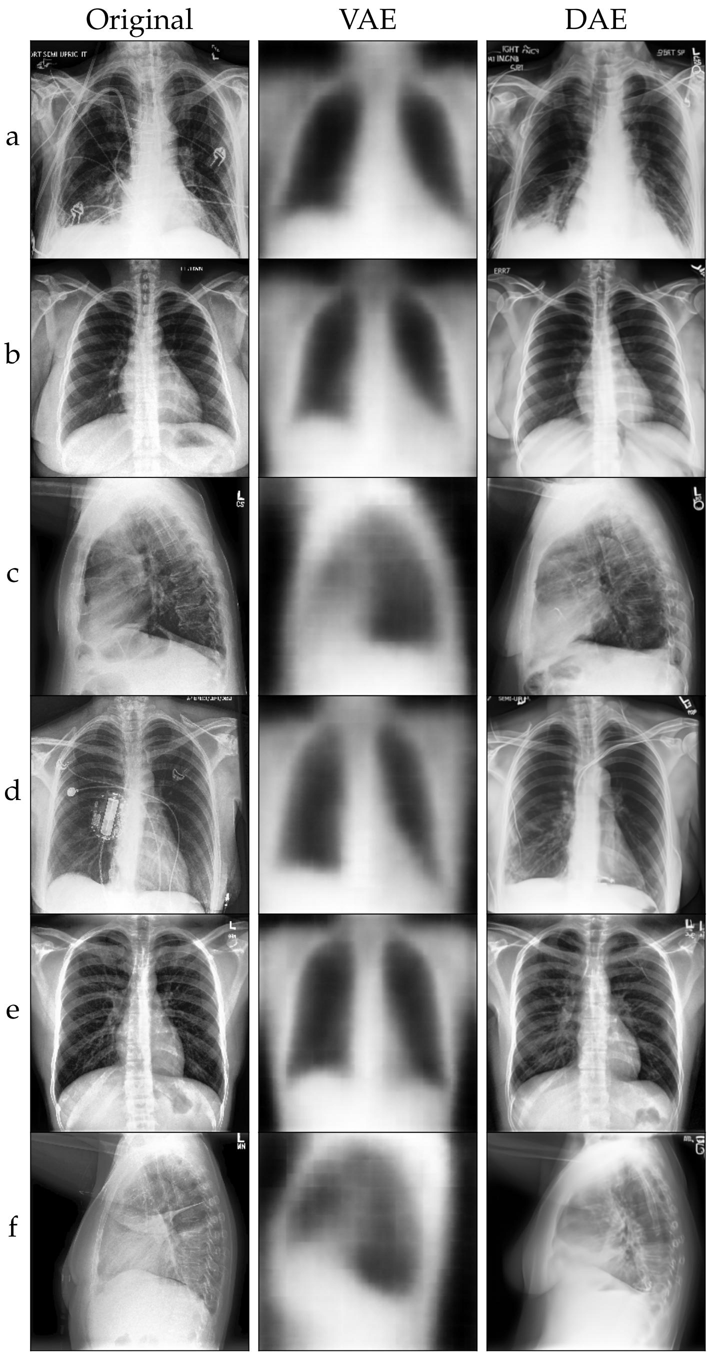


Figure 3.2: CheXpert reconstructions

derived all information encoded into the latent code from the images themselves since no supervision with any additional information was performed. The resulting plots for each dataset are discussed in the following sections.

GNC Embeddings

For the GNC dataset, we decided to inspect the main patient information given together with the images in the dataset, depicted in Figure 3.3 and Figure 3.5 for the VAE and DAE respectively. This information includes the following:

a) the region of the imaging, separated into cervical, thoracic, and lumbar views,

b) the biological sex of the patient,

c) the institution where the MRI was taken,

d) the patient's age in years,

e) the patient's weight in kg, and

f) the patient's height or size in m.

Furthermore, a selection of additional patient information was obtained from a digital survey. We focused on meaningful attributes with respect to analyzing the model's ability to predict certain diseases and correlations that might be important for later inspection and finding representative bias in the data. The accordingly colored embeddings are depicted in Figure 3.4 and Figure 3.6 and include:

a) the presence of increased blood lipids like cholesterol or triglycerides,

b) the presence of a thyroid disease,

c) the presence of arthrosis,

d) whether the patient has suffered back pain in the last three months, separated into no up to low pain and moderate up to strong pain,

e) whether the patient is currently employed full-time,

f) the average daily alcohol consumption in g/d.

Note that in the depictions, missing values or the refusal to provide this information is omitted. This leads to sparse plots for very subjective information like the feeling of back pain and illustrates the difficulties with such subjective assessments.

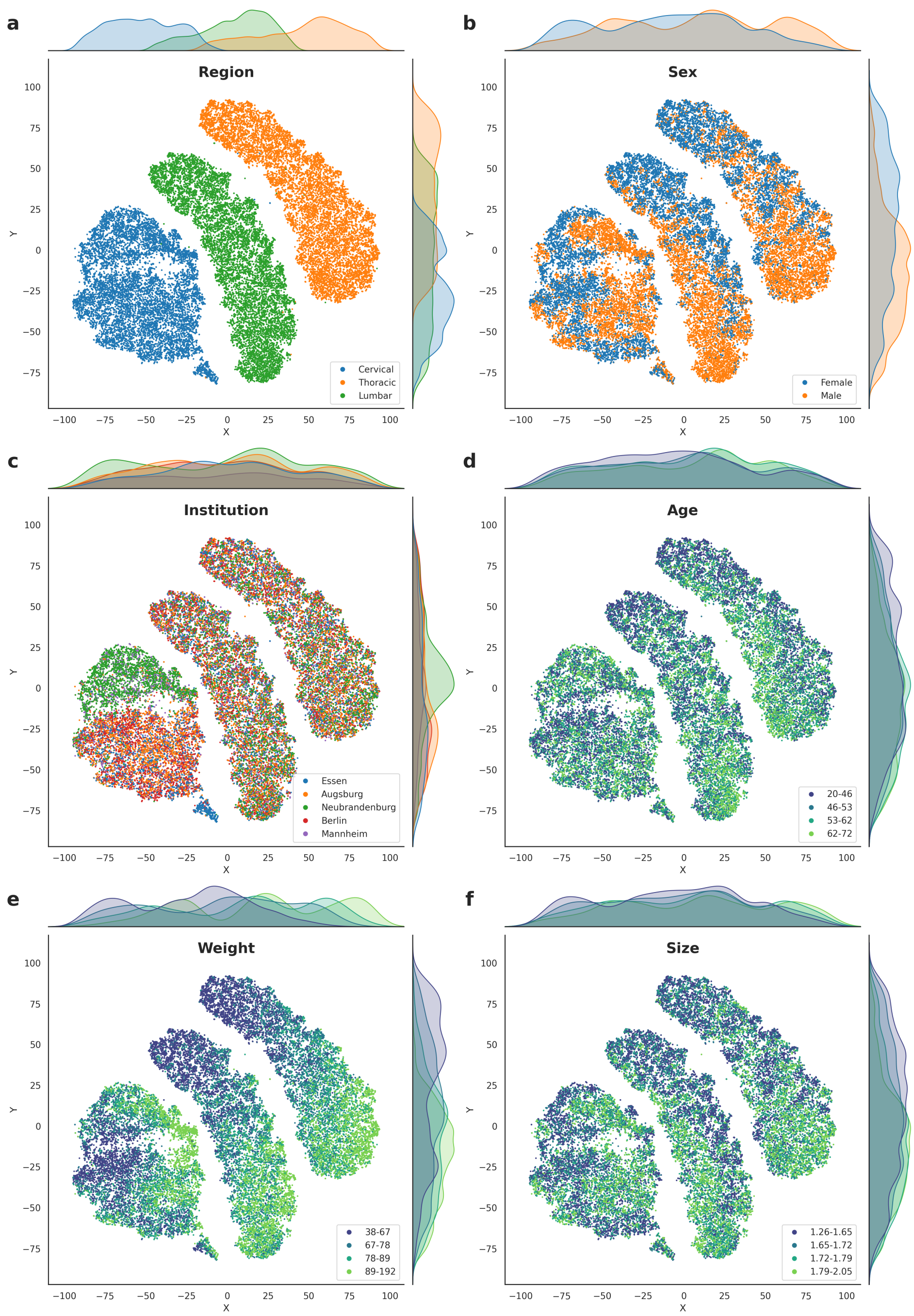


Figure 3.3: VAE embeddings t-SNE of GNC 1

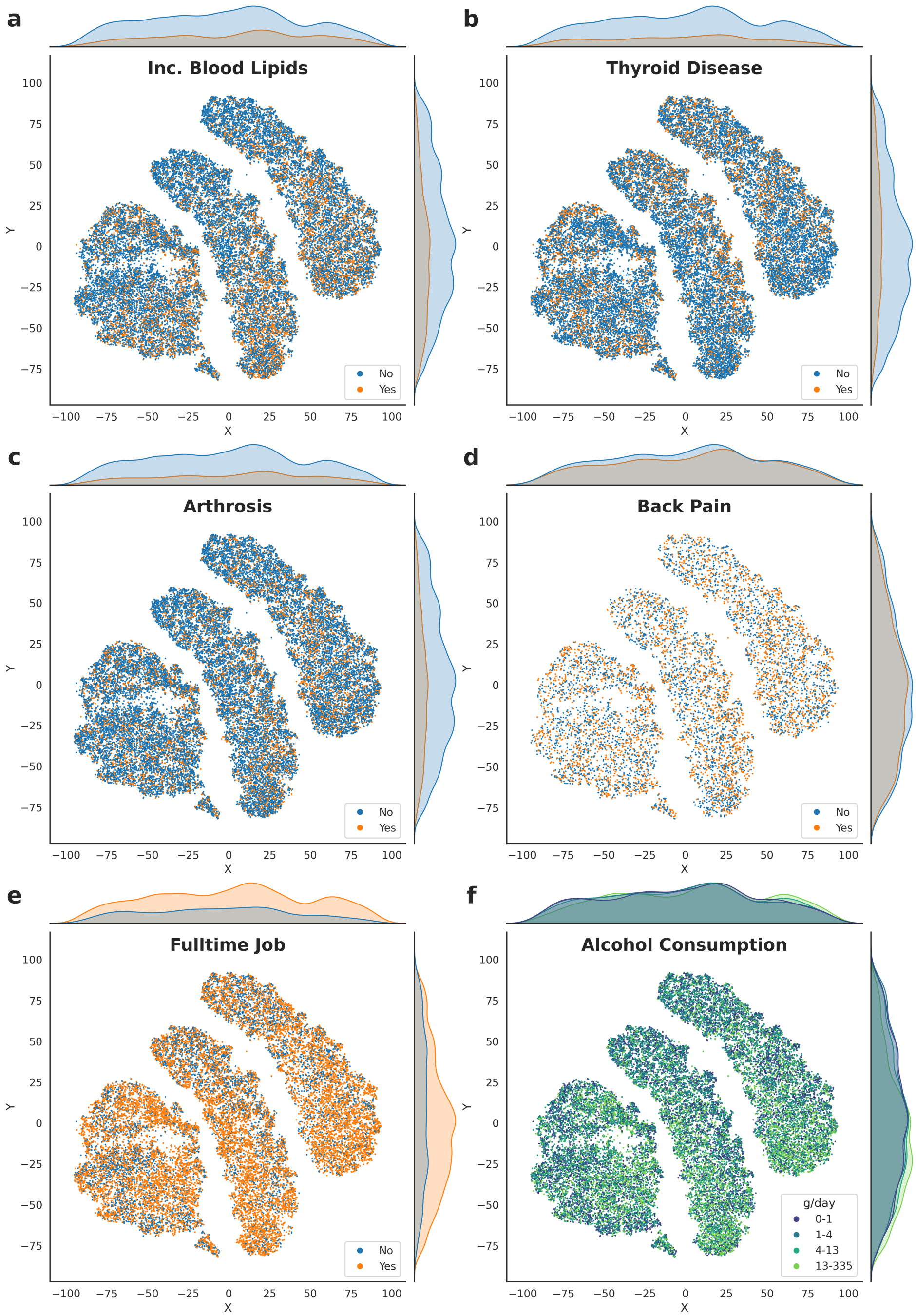


Figure 3.4: VAE embeddings t-SNE of GNC 2

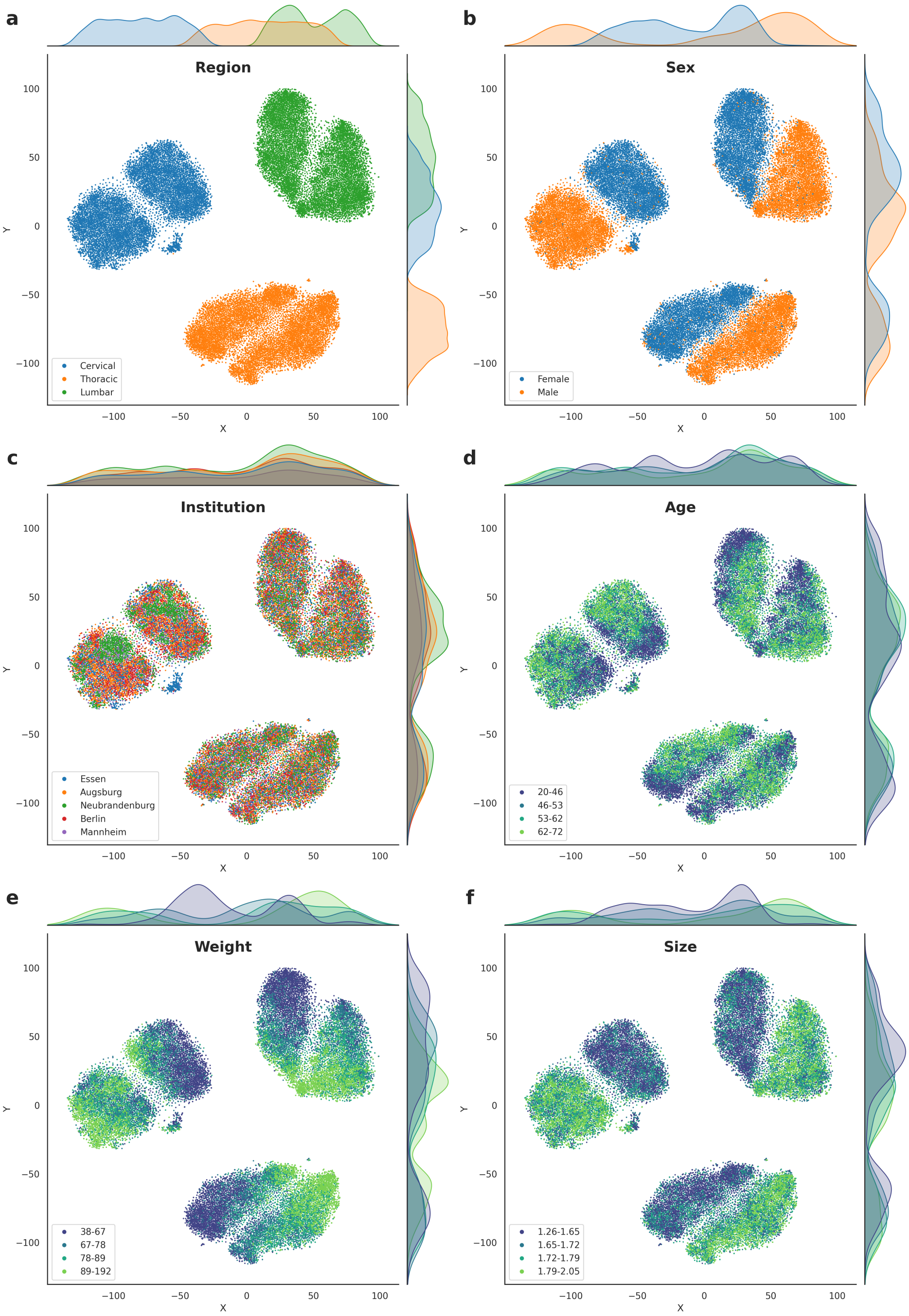


Figure 3.5: DAE embeddings t-SNE of GNC 1

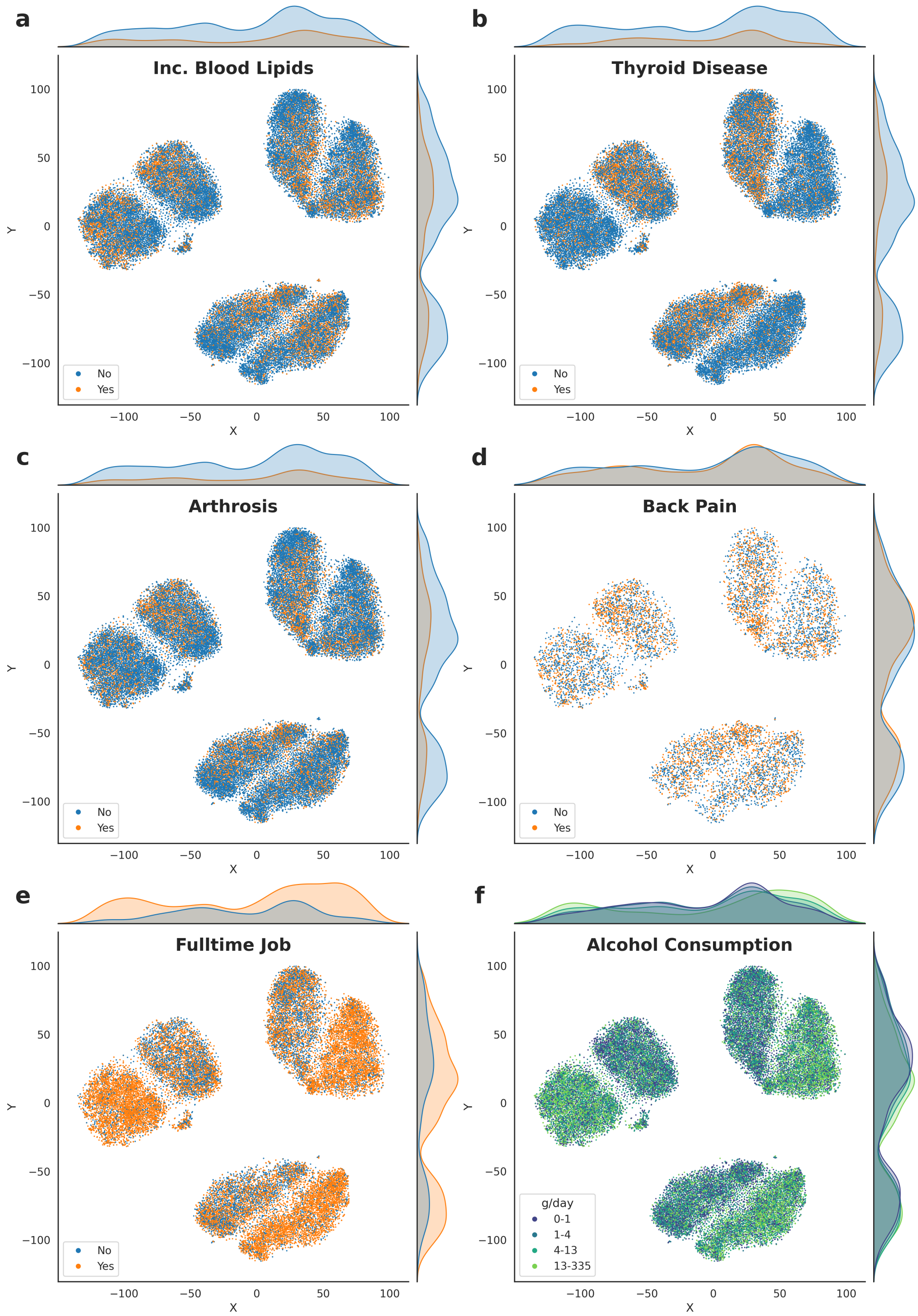


Figure 3.6: DAE embeddings t-SNE of GNC 2

CheXpert Embeddings

The CheXpert dataset contains similar information about the patient's sex (a) and age (b), as well as the view of the X-ray (c). The latter is divided into the frontal views AP and PA and the lateral view. As described in Section 2.2.2, the dataset gives additional information about 14 different findings. Since most of these are highly imbalanced towards *uncertain* or *no statement*, we decided to combine them into a three-class attribute stating whether there was a certain finding of Pleural Effusion, any other disease, or no finding at all (d). The embeddings of the VAE and DAE are depicted in Figure 3.7 and Figure 3.8.

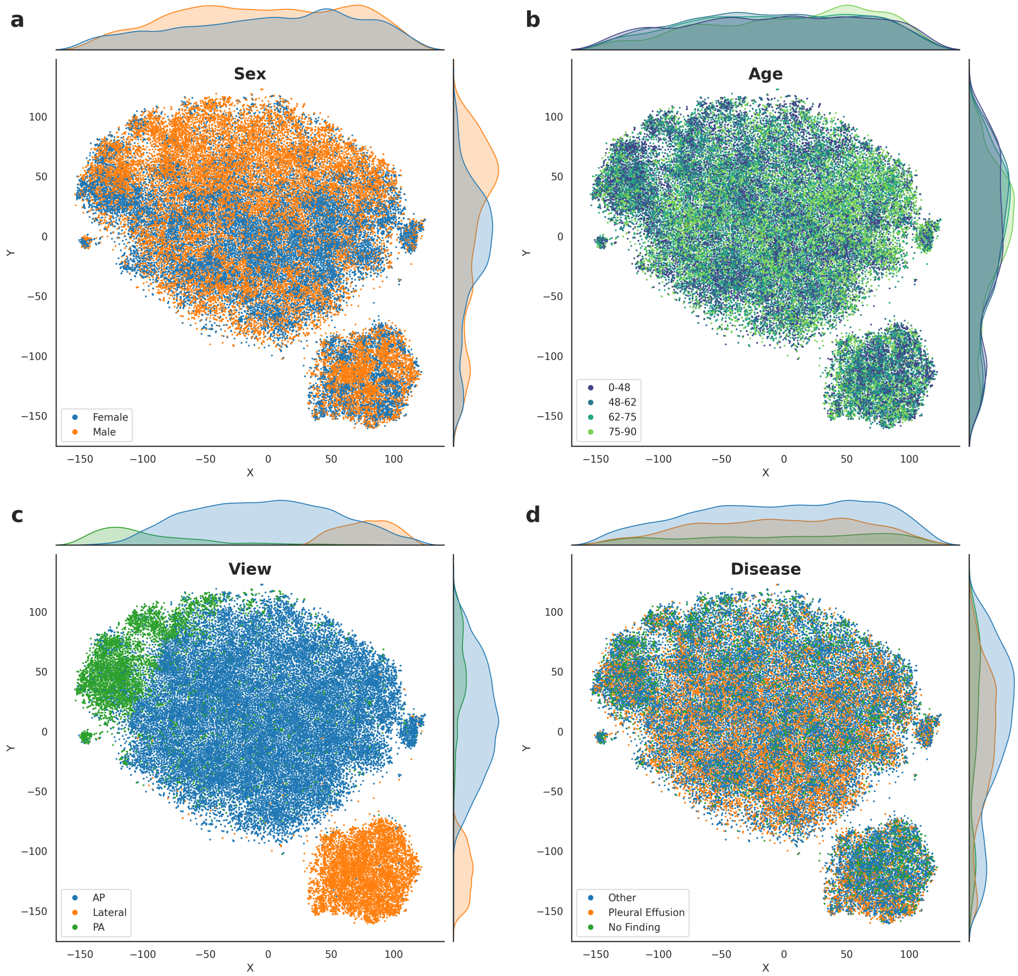


Figure 3.7: VAE embeddings t-SNE of CheXpert

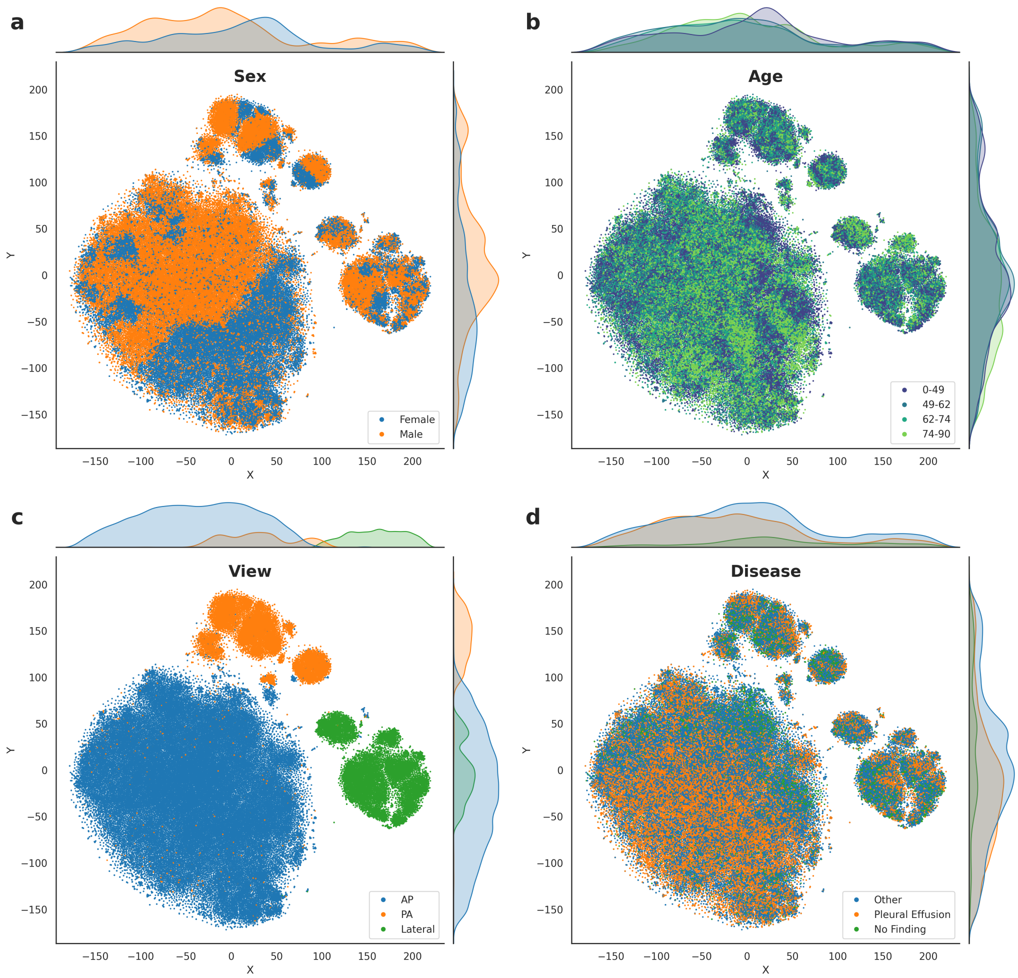


Figure 3.8: DAE embeddings t-SNE of CheXpert

3.2.2 Visual Interpretation

With these visualizations, we can already interpret how well the models performed to represent the semantic information hidden in the images as well as discover previously unknown anomalies or systematic differences between groups of samples.

GNC

For the VAE (Figure 3.3), subfigure a) shows a clear separation of the three regions captured in an image. Differentiating the male and female patients was already more difficult, but a clear tendency for separation is noticeable in plot b), with the male patients arranged near the bottom of the region clusters. For the Cervical region, however, the intermingling is significantly increased. This can be attributed to the arrangement of the institution in which the cervical images were acquired. There should be no discrepancy in an ideal world depending on the institution providing the images. Subplot c) shows that this is only the case for the Thoracic and Lumbar region images, where the institution does not noticeably affect the arrangement of the data points. The cervical images, in turn, exhibit a clear separation for the institutions Essen and Neubrandenburg. Conversely, within these unnatural clusters, we can again observe the expected separation between male and female patients. In the attributes of age, weight, and height, a distinct distribution can be observed according to their values, while the weight distribution stands out visually in particular.

Astonishing performance in the separation of the attributes can be observed in the embeddings by the DAE (Figure 3.5). The patient's sex is even separated into different subclusters of the region clusters. We observe the same distribution within the subclusters according to age, weight, and size. In particular, especially with the DAE plots, we notice that the attributes age and weight are distributed nearly orthogonal within the clusters. This can be explained since all combinations, from light and small patients to heavy and large patients, need to be represented by the embedding. Distributing them along the same axis would lead to significant informational loss. Therefore, they get encoded as the two principal components in this two-dimensional representation for maximum visual variance.

The arrangement of the different institutions is similar, with an extra cluster for Essen subdivided into male and female patients. Still, within the two main clusters of the cervical region, there are two clear aggregations of images from Neubrandenburg. This would be a typical measurement bias, where the images have been created with deviations from the standard protocol.

To understand the origin of these clusters, following Graf et al. [22], we take the affected samples and compute an average image for each cluster. For this comparison, we chose the DAE embedding and used the male clusters of Essen and Neubrandenburg compared to all other male samples. Figure 3.9 shows that a systematic shift in the head position of the patient causes the differences between these clusters. Compared to all other images, examination in Essen would result in a higher head position, whereas examination in Neubrandenburg caused less diffraction in the patient's neck.

Another interesting model behavior is that the VAE seems to attribute more significance to the variance in the patient's weight than the patient's sex. The patient's weight is perfectly distributed over the entire length of the two undisturbed clusters of thoracic and lumbar images. Furthermore, the clusters exhibit a significantly greater length along the direction of weight distribution, especially in the clusters of lumbar and

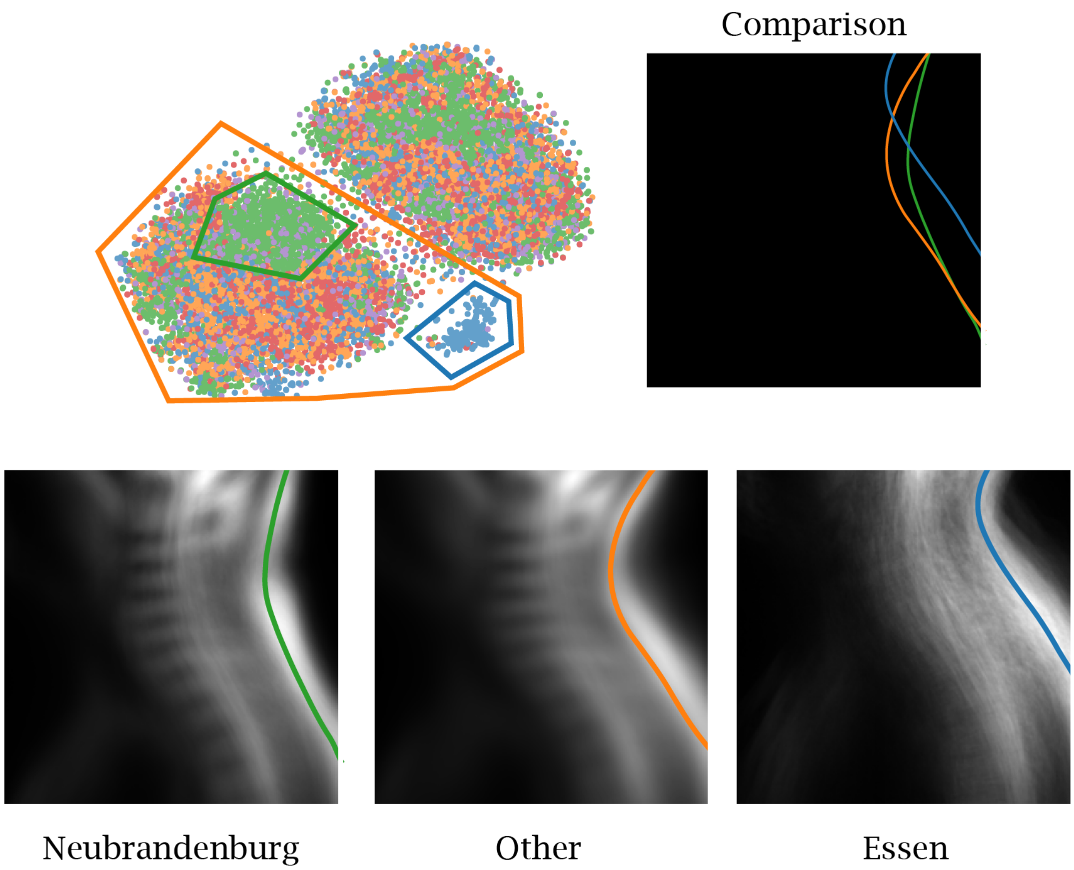


Figure 3.9: Comparison of local bias in GNC cervical images

thoracic images, where the patient's weight seems more expressed. In contrast, the sex is less clearly separated. This could result from the enforced disentangling, weighing the variance in patients' weight more.

Inspecting the presence of increased blood lipids, we can notice a clear skew towards the edges of clusters. The direction of the shift visually correlates with the weight of the patients. Even if this is a true medical correlation, it would be a strong bias if the model solely learned to conclude by examining the patient's weight and not by inferring its presence from the image itself.

This can also be observed for thyroid diseases. Because, commonly, thyroid diseases are more likely to affect the female population, it is not surprising to see that the positive labels are mainly distributed within the female clusters but now have the tendency to correlate with the patient's age. Again, this might be a correct correlation. Still, fair decision-making for future patients who do not fulfill this stereotype is not secured without ensuring the model learns to predict those diseases independently. The presence of arthrosis is also mainly correlated with the age of the patient, with a majority of female patients affected by this disease.

The plot of the embeddings colored according to the patient's declarations regarding back pain proves to be challenging for further analysis since most data is either invalid or the patient declined to provide this information. This shows the difficulty with subjective measurements like the degree of pain.

Another typical type of bias, the historical bias, can be seen within the plots of working hours and alcohol consumption. Both are clearly correlated to the patient being a male person who has a historical tendency to be the family member providing the main income and consuming more alcohol. This, of course, affects the creation of those large datasets and is a prime example of a user-to-data bias.

CheXpert

The images of the CheXpert dataset are generally much more detailed. Alignment and consistency are very hard to obtain throughout image capturing. This unwanted variance, of course, makes it harder for the model to extract information. Even the most simple information, like the view of the image, which a radiologist normally can identify at first sight, is quite challenging for the VAE.

As Figure 3.7 shows, while separating the lateral from the frontal images, it can barely identify the differences between AP and PA projections in Subfigure c). Within the frontal images, in Subfigure a), female and male patients are accumulating in the top and bottom areas, respectively. A clear separation, however, was not possible. The finding of a pleural effusion also seems skewed to the cluster's lower end. A distribution of the patient's age is not visible at all.

On the other hand, the DAE can easily distinguish between the different views of the X-ray images, shown in Figure 3.8 c).

Even though no independent clusters for the sexuality of the patient were created in Subfigure a), a very clear separation within the clusters is visible. Like the VAE, the pleural effusion was distributed to the bottom of the cluster, as Subfigure d) shows. A specific distribution of the patient's age is not apparent.

Interestingly, additional clusters emerge between and within the said and reasoned clusters. This suggests that the model captured systematic differences in the data not yet described by examined labels. To properly investigate the differences between certain clusters, Figure 3.10 shows several images averaged over a selection of the data. For easier comparison and less blur in the averaged images, only male patients were considered.

For instance, an additional small cluster emerged between the AP and PA clusters - indicated by image c) and the neighboring cluster - containing images of both projection types. Note that commonly, the AP images are mostly taken from ill patients who are not capable of standing up for a proper PA scan. These images are more likely to be misaligned or have additional noise caused by devices or cables. However, even though cables are supposed to be taken off before an PA scan, it can occur that this was missed by radiologists, leading to major confusion if the model bases its decisions on this vague correlation. Examining the two neighboring clusters, we can see that the AP images are just of uncommon quality with respect to contrast and alignment, leading the model to assess them as more similar to the PA images. The opposite goes for the other half of the cluster. This suggests that the model uses at least some of these image quality characteristics within the images to make assumptions about its capturing process, mitigating the previous assumption of some bias based on the presence of cables or devices.

Interestingly, in addition to the expected clusters, the DAE also captured other separations or at least significant differences within the major groups of samples. This is mainly to be seen within the PA view, where at least three clearly separated clusters with their own male-female separation are depicted, or, a similar separation, in the lateral images, with the main cluster separated into two large and an additional smaller cluster further apart.

Figure 3.10 shows averaged images over the most significant and most different clusters. Image a) and b) compare the male patients within the PA cluster. Here, we can observe significant differences in chest width and lung volume. Image b) seems to show thinner patients since black borders are visible to the left and right. Also, the lung seems significantly smaller. Having an overall smaller outline, it is reasonable that this cluster is located nearer the latent view cluster. A very similar observation can be made considering the lateral views. Here, the cluster located towards the frontal images indicates patients with, on average, larger chest circumferences. The difference can be seen in images d) and e).

Even though these differences seem to be explainable with patient characteristics, this would affect the distribution of data points in a continuous manner. Instead, the whole clusters' separation is noticeable. This suggests further systematic differences. Possible influences would be the setting of the X-ray facility, resulting in different quality or appearance or alignment issues of the sensitive area, including the surface the patient is lying on (specific for non-PA images) or the positioning of the panel detector. All these factors can and should be analytically analyzed but require specific knowledge and are left for further research.

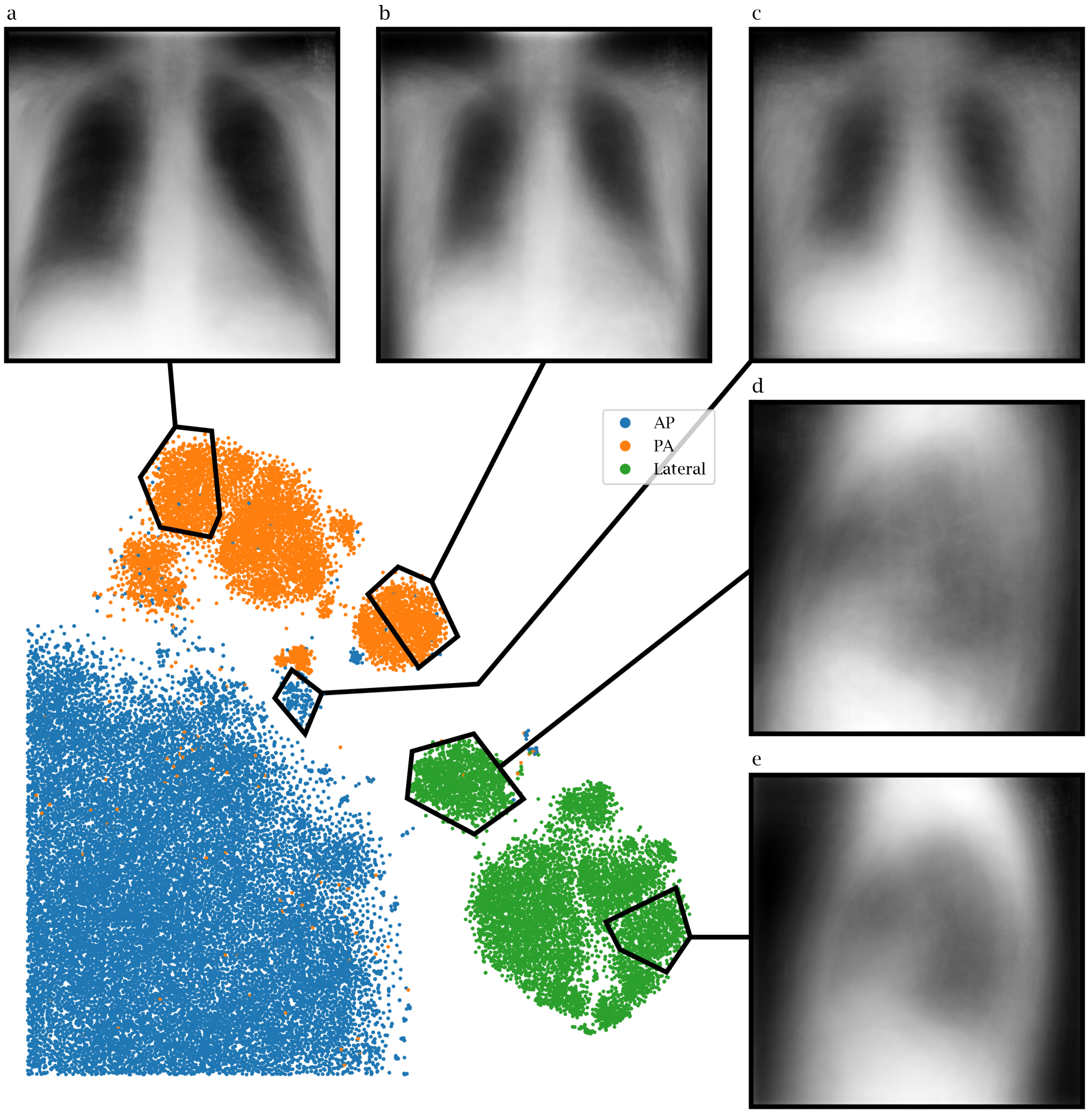


Figure 3.10: Comparison of clusters in CheXpert

3.3 Metrics for Quantitative Evaluation

After this intuitive and visual approach to analyzing the embeddings, we also quantify the achieved accuracy and fairness, described in the following section.

3.3.1 Accuracy

To quantitatively assess the performance of the AEs and the accuracy of the embeddings in representing the original data distribution, in a secondary step, we train common predictors on the encoded representations. This is done using the scikit-learn implementation of support vector machines for classification and regression.

GNC

For the GNC, we randomly sample 10,000 images from the original training set to fit the vector machines and test them on 2,500 randomly selected test samples. We predicted the same labels that were visually examined in the previous section to see if our assumptions on the accuracy would hold.

As in the previous plots, we omitted missing values. Since most classes are highly imbalanced, we used the automated balancing of the support vector classifiers of scikit-learn. For continuous variables, we calculated weights for each label l as $w_l = \frac{N}{N_l}$, where N_l represents the number of values equal to l and N the total number of samples trained on. The score is derived from a balanced accuracy score for the classification tasks and an L1-loss (mean absolute error) for the regression tasks. The results can be seen in Table 3.2.

Both models achieve astonishing accuracy in predicting the body region in an encoded image. The DAE even makes no mistakes at all. We also get near-perfect results for predicting the patient's sex from the representation. However, the VAE performs significantly worse. However this was already expected after the visual inspections of the embeddings, which mostly seemed to focus on the distribution of the patient's weight.

For the attributes weight, height, and age, the mean absolute error lies within a range of 2.8% to 13.7% (VAE) or 2.1% and 7.5% (DAE) of the mean value. Both models can determine the patient's height within an accuracy of 5cm, which is noteworthy since the prediction is only done on one of the three regions at a time. So, only a third of the spine can be examined by the model for a certain prediction. The DAE performs on average about 26% to 45% better.

The additional information provided with the patient survey is unfortunately not quite learnable without further adjustments. Simple reweighing was not enough to bring the balanced accuracy to a score consistently high enough to have any medical use. Highly imbalanced classes with highly noisy patient self-assessments are very challenging conditions.

Table 3.2: Predictions on GNC embedding

	Body Region accuracy ↑	Sex accuracy ↑	Weight ℓ_1 kg ↓	Height ℓ_1 meter ↓	Age ℓ_1 years ↓
VAE	0.999	0.921	6.38	0.049	7.20
DAE	1.000	0.986	4.28	0.036	3.94

	Blood Lipids accuracy ↑	Thyroid Disease accuracy ↑	Arthrosis accuracy ↑	Back Pain accuracy ↑
VAE	0.611	0.647	0.617	0.545
DAE	0.643	0.680	0.667	0.557

CheXpert

The accuracy assessment for CheXpert was done by fitting vector machines on 20,000 randomly selected training samples and testing them on the 234 samples published for validation, together with 2000 different samples from the training set, with no overlapping between training and testing data.

For simplicity, we used the U-Zeros approach, which means all missing values in the training data have been mapped to 0 since Irvin et al. [30] showed that only marginal differences occurred between the different mappings used. To make the results even more stable with less uncertain labels, we decided to engineer a new feature, disease, classifying each examination into no finding, pleural effusion, or any other disease. We also added a binary attribute differentiating only between no finding and any finding for a more straightforward fairness comparison in the next section,

The score for the patient information was again calculated with balanced accuracy and L1-loss. For a better comparison with the Stanford baseline and recent models predicting on CheXpert, we used the AUC score for classifying the presence of the five diseases used for the competition score. The results are summarized in Table 3.3.

Since the classes of findings were also highly imbalanced, the same reweighing was applied before classification as described for the GNC dataset.

For the protected attribute sex and the view, quite good accuracy is achieved with the DAE. The balanced accuracy of classifying the view is about 4.4% lower for the VAE, which complies with our assumptions made before since it already did not cluster the AP and PA views correctly, or as accurate as the DAE did. The even higher discrepancy shows the accuracy of the patient's sex classification. Here, the VAE performs over about 21.5% worse. Similar results for the age prediction, where the mean absolute error is about 40% reduced, using DAE.

However, the balanced accuracy is not as different between the models for predicting findings. Varying between 0.02 and 0, 1 in difference of the AUC score. Unfortunately, the scores are often close to the actual percentage of the imbalance of the classes, making the predictions not much more efficient than random guessing. Although the AUC scores of the DAE are on average 0.075 better than of the VAE, compared to the Standford baseline AUC of 0.9065 [30] and the current best model focussing specifically optimizing the AUC score with an AUC of 0.9305 [67] they do not quite match the accuracy. Only standard baselines, done with, for instance an DenseNet121, are surpassed in AUC score.

Table 3.3: Predictions on CheXpert embedding

	View accuracy ↑	Sex accuracy ↑	Disease accuracy ↑	Disease Binary accuracy ↑	Age ℓ_1 years ↓
VAE	0.942	0.736	0.544	0.719	13.06
DAE	0.986	0.951	0.642	0.778	7.90

	Cardiomegaly AUC ↑	Edema AUC ↑	Consolidation AUC ↑	Atelectasis AUC ↑	Pleural Eff. AUC ↑	Mean AUC ↑
VAE	0.698	0.726	0.655	0.634	0.736	0.690
DAE	0.791	0.829	0.726	0.654	0.824	0.765

The AEs represent a good approach for representation learning and understanding large datasets, and we notice that the created embeddings are highly valuable for inspecting prominent features like the patient's sex, weight, or age. But for much more complex correlations between details in the images and some very rarely occurring diseases, and to match the latest developments in image classification, further tuning and domain-specific expertise is necessary. However, since the training of AE is completely unsupervised and labels are only used within the second stage, this method is highly scalable, even with more, unlabeled data.

3.3.2 Bias

To assess the bias that is contained in any prediction based on the created embeddings, we use the in Section 2.3.4 defined bias metrics Bias Amplification, Demographic Parity, Equalized odds, and ABROCA. The fairness will always be assessed by predicting the same label against each value of a protected attribute. We decided to use the patient's sex as the protected attribute since both datasets provide it, and it is a binary attribute. For the predicted attribute, we use a binary label indicating the presence of a disease. For CheXpert, we use the above-described presence of any disease or no finding. With

GNC, we use Arthrosis for prediction, being the least imbalanced class out of the above observed.

As mentioned in Section 2.3, the protected attributes may or may not be part of the data that is used for training. Note that in this case, no additional patient information was used to train the AE, but, of course, the patient's information is provided implicitly by the image itself.

The resulting scores can be seen in Table 3.4 and Table 3.5 for the GNC and CheXpert dataset, respectively. The plots of the ABROCA plots can be seen in Figure 3.11.

Whereas the VAE performed at least slightly worse in any scenario than the DAE, it now prevails regarding almost all fairness metrics. Bias amplification does not exhibit major differences, but within the remaining metrics, an improvement of up to

This makes the VAE fairer in terms of proposed metrics, but as seen before, it comes at the cost of reduced accuracy in prediction tasks. For CheXpert, Figure 3.11 shows the ABROCA plot and reveals that the DAE behaves more differently for both types of sex. The female curve in Figure b) is almost always below the male curve, leading to quality of service bias for the female patient group and, therefore, an increased ABROCA score. Since we already noticed biases based on the patient's sex with respect to some of the diseases it is not surprising to also see them present in the quantitative analysis.

Table 3.4: Bias scores on GNC

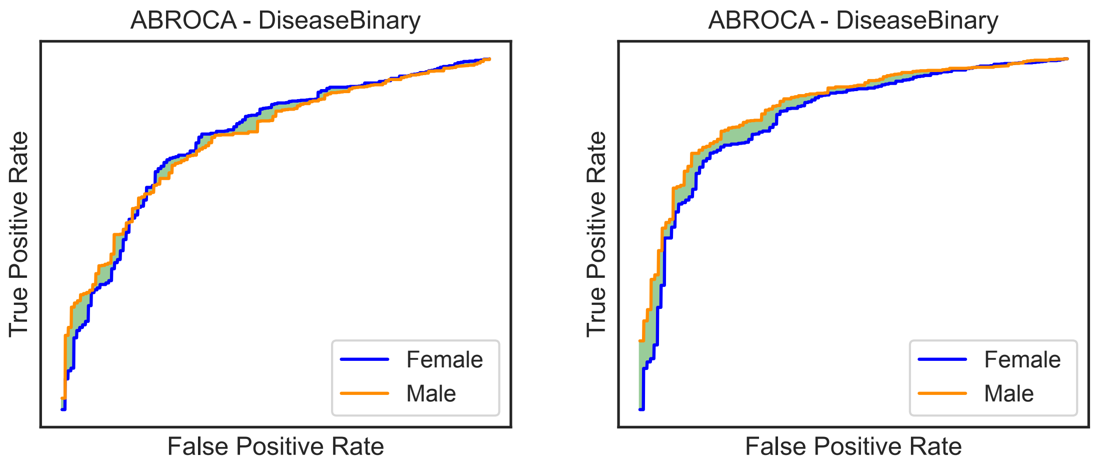
Model	Bias Ampl. ↓	Dem. Parity ↓	Eq. Odds ↓	ABROCA ↓
VAE	-0.194	0.091	0.078	0.021
DAE	-0.195	0.150	0.180	0.035

Table 3.5: Bias scores on CheXpert

Model	Bias Ampl. ↓	Dem. Parity ↓	Eq. Odds ↓	ABROCA ↓
VAE	-0.230	0.086	0.090	0.023
DAE	-0.186	0.030	0.107	0.031

Bias Mitigation

To guarantee fair decision-making, the prevalent bias needs to be removed. For this thesis, we decided on using the method proposed by Thong and Snoek [61], where we can utilize the already created embeddings to calculate and subtract the bias for a specific protected attribute from the embeddings. This was done in the following section for the CheXpert dataset providing significantly higher accuracy in the prediction tasks than GNC.



(a) ABROCA plot - VAE

(b) ABROCA plot - DAE

Figure 3.11: ABROCA plots - CheXpert

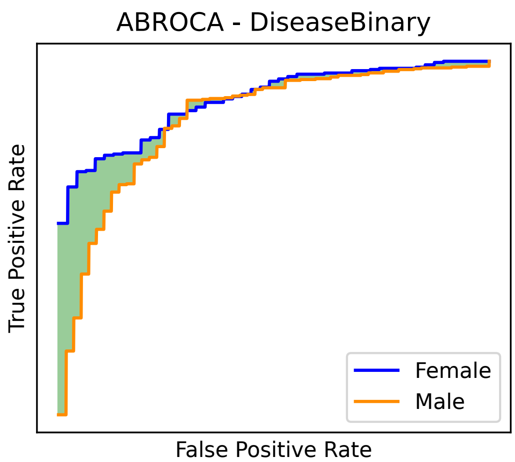
Table 3.6 shows the new scores obtained after the proposed bias mitigation. Removing the average bias for each value of patient sex, we expect the predictions to be more independent from this protected attribute, leading to improved fairness scores. The changes meet our expectations for the VAE. Demographic parity gets reduced by about 80% and equalized odds by about 60%. The ABROCA score, however, increases by more than double. Looking at the ROC plot in Figure 3.12, we see the origin of this change. The ROC of the female patient group experienced an unusual upward shift. This correlates with the increased balanced accuracy score of 0.769 for the VAE. But even though both parity and accuracy get improved, this unequal quality of predictions is no fairer behavior than before.

The differences for the DAE are significantly less salient. Apart from a slightly increased demographic parity difference, the metrics do not change noticeably. The same applies to the ABROCA plot, where no significant changes can be observed. The accuracy for the predicted disease does not change at all.

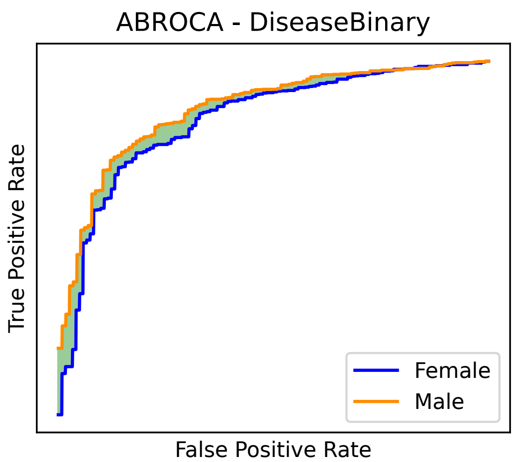
For proposed bias mitigation, this leads to the conclusion that VAEs react quite unstable to changes in the latent space but utilize the changes and yield better predictions. The DAEs seem to be much more robust. Further experiments utilizing more combinations of protected and predicted attributes and different mitigation strategies would be insightful.

Table 3.6: Bias mitigated scores on CheXpert

Model	Bias Ampl. ↓	Dem. Parity ↓	Eq. Odds ↓	ABROCA ↓
VAE	-0.178	0.018	0.030	0.048
DAE	-0.186	0.036	0.107	0.030



(a) ABROCA plot - bias mitigated - VAE



(b) ABROCA plot - bias mitigated - DAE

Figure 3.12: ABROCA plots - CheXpert - bias mitigated

4 Conclusion and Outlook

The findings in this thesis present the importance of proper dataset and model analysis with increasingly inscrutable models. We leveraged the nature of two common types of AEs to construct meaningful representations providing valuable semantic insight into two large medical image datasets. This enabled us to find strong invariances in the imaging processes, indicated by clearly distinctive clusters in the image embeddings of both examined datasets. Whereas multiple reasonings were explained, the visual analysis of the embeddings suggests that, especially for the CheXpert dataset, more systematic differences exist, requiring further analysis leveraging domain-specific knowledge. Furthermore, multiple types of bias have been detected with respect to the protected attribute of the patient's sex. Following a quantitative bias analysis using selected metrics, we applied an existing bias mitigation technique. The mitigation was observed to impact the types of AEs quite differently. Especially the DAE seems more robust to changes in its fairness when altering the latent space accordingly, whereas the VAE was also heavily impacted in its accuracy.

This thesis gives an overview of the most well-known metrics and compares only a selection of possible scenarios. Further comprehensive studies should be done regarding these datasets with more combinations of protected and predicted attributes. It could help to find more interrelations between clustering and the imaging process or patient information not noted within the dataset. From a technical perspective, with more resources or in a large-scale survey, other latent and hidden dimensions, input resolutions, and other hyperparameters should be compared and evaluated with respect to their impact on accuracy and fairness.

Future work could also focus on determining which information of an image a specific primary prediction task is using. Similar studies like Glocker et al. [21] could be done to check if any primary or secondary tasks correlate.

This thesis has also limited itself to binary classification and fairness evaluation. Multi-class scenarios are an essential generalization for both bias metrics and mitigation techniques.

Whereas the VAE is only capable of capturing the semantic information of the images, DAEs are, very capable of reconstructing the images in almost all their details. Whereas this thesis only used the semantic part of the latent code, one could leverage the detailed reconstruction for a generative approach. Encoding an image and altering only the semantic latent space could lead to surprising results in creating new images with missing protected attributes, like a gender-neutral X-ray. Interpolating between different semantic attributes of an image or recreating new, slightly different images with the same semantic information could be a valuable contribution to the medical domain, especially within medical education.

Abbreviations

AE Autoencoder

PCA Principal Component Analysis

SNE Stochastic Neighbor Embedding

t-SNE t-Distributed Stochastic Neighbor Embedding

VAE Variational Autoencoder

β-VAE β-Variational Autoencoder

DAE Diffusion Autoencoder

DPM Diffusion Probabilistic Model

DDIM Denoising Diffusion Implicit Model

DRL Disentangled Representation Learning

GNC German National Cohort

PA Posterior Anterior

AP Anterior Posterior

ROC Receiver Operating Characteristic

AUC Area Under the Curve

ABROCA Absolute Between-ROC Area

MRI Magnetic Resonance Imaging

CNN Convolutional Neural Network

List of Figures

List of Tables

Bibliography

[1] R. Baeza-Yates. "Bias on the web." In: *Communications of the ACM* 61.6 (2018), pp. 54–61.

[2] S. Barocas and A. D. Selbst. "Big data's disparate impact." In: *Calif. L. Rev.* 104 (2016), p. 671.

[3] M. Bastian. *GPT-4 has more than a trillion parameters - Report*. 2023. URL: `https://the-decoder.com/gpt-4-has-a-trillion-parameters/` (visited on 04/12/2024).

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

[5] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. "Data decisions and theoretical implications when adversarially learning fair representations." In: *arXiv preprint arXiv:1707.00075* (2017).

[6] J. Buolamwini and T. Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91.

[7] T. Calders, F. Kamiran, and M. Pechenizkiy. "Building classifiers with independency constraints." In: *2009 IEEE international conference on data mining workshops*. IEEE. 2009, pp. 13–18.

[8] T. Calders and S. Verwer. "Three naive bayes approaches for discrimination-free classification." In: *Data mining and knowledge discovery* 21 (2010), pp. 277–292.

[9] D. S. Carvalho, G. Mercatali, Y. Zhang, and A. Freitas. "Learning disentangled representations for natural language definitions." In: *arXiv preprint arXiv:2210.02898* (2022).

[10] G. L. Ciampaglia, A. Nematzadeh, F. Menczer, and A. Flammini. "How algorithmic popularity bias hinders or promotes quality." In: *Scientific reports* 8.1 (2018), p. 15951.

[11] J. Cook, I. Sutskever, A. Mnih, and G. Hinton. "Visualizing similarity data with a mixture of maps." In: *Artificial intelligence and statistics*. PMLR. 2007, pp. 67–74.

[12] K. Crawford. *The Trouble with Bias*. 2017. URL: https://www.youtube.com/watch?v=fMym_BKWQzk (visited on 03/22/2024).

[13] L. Crespi, D. Loiacono, and A. Chiti. "Chest X-Rays Image Classification from Variational Autoencoders Latent Features." In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2021, pp. 1–8.

[14] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma. "Adversarial classification." In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, pp. 99–108.

[15] G. N. C. (C. geschaeftsstelle@ nationale-kohorte. de. "The German National Cohort: aims, study design and organization." In: *European journal of epidemiology* 29.5 (2014), pp. 371–382.

[16] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang. "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[17] J. Dunkelau and M. Leuschel. "Fairness-aware machine learning." In: *An Extensive Overview* (2019), pp. 1–60.

[18] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. "Fairness through awareness." In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226.

[19] C. Garbin, P. Rajpurkar, J. Irvin, M. P. Lungren, and O. Marques. "Structured dataset documentation: a datasheet for CheXpert." In: *arXiv preprint arXiv:2105.03020* (2021).

[20] J. Gardner, C. Brooks, and R. Baker. "Evaluating the fairness of predictive student models through slicing analysis." In: *Proceedings of the 9th international conference on learning analytics & knowledge*. 2019, pp. 225–234.

[21] B. Glocker, C. Jones, M. Bernhardt, and S. Winzeck. "Algorithmic encoding of protected characteristics in chest X-ray disease detection models." In: *EBioMedicine* 89 (2023).

[22] R. Graf, F. Hunecke, S. Pohl, M. Atad, H. Möller, S. Starck, T. Kröncke, S. Bette, F. Bamberg, T. Pischon, H. Niendorf, C. Schmidt, J. Paetzold, D. Rueckert, and J. Kirschke. "Detecting Unforeseen Data Properties in Medical Imaging with Diffusion Autoencoder Embeddings using Spine MRI data from the German National Cohort." In: (2024), to appear.

[23] M. Hardt, E. Price, and N. Srebro. "Equality of opportunity in supervised learning." In: *Advances in neural information processing systems* 29 (2016).

[24] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. "beta-vae: Learning basic visual concepts with a constrained variational framework." In: *ICLR (Poster)* 3 (2017).

[25] G. Hinton. "Stochastic neighbor embedding." In: *Advances in neural information processing systems* 15 (2003), pp. 857–864.

[26] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models." In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[27] K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators." In: *Neural networks* 2.5 (1989), pp. 359–366.

[28] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro. "Bias mitigation for machine learning classifiers: A comprehensive survey." In: *ACM Journal on Responsible Computing* (2023).

[29] X. Hu, H. Wang, S. Dube, A. Vegesana, K. Yu, Y.-H. Lu, and M. Yin. "Discovering biases in image datasets with the crowd." In: *Proceedings of HCOMP* (2019).

[30] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597.

[31] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. *CheXpert: A Large Chest X-Ray Dataset And Competition*. 2019. URL: `https://stanfordmlgroup.github.io/competitions/chexpert/` (visited on 03/18/2024).

[32] A. Jalali, H. Weerts, M. Madaio, M. Dudik, R. Edgar, et al. *Fairness in Machine Learning*. 2023. URL: `https://fairlearn.org/v0.10/user_guide/fairness_in_machine_learning.html` (visited on 03/22/2024).

[33] L. A. Jeni, J. F. Cohn, and F. De La Torre. "Facing imbalanced data–recommendations for the use of performance metrics." In: *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE. 2013, pp. 245–251.

[34] X. Ji, J. F. Henriques, and A. Vedaldi. "Invariant information clustering for unsupervised image classification and segmentation." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 9865–9874.

[35] F. Kamiran and T. Calders. "Classification with no discrimination by preferential sampling." In: *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*. Vol. 1. 6. Citeseer. 2010.

[36] D. P. Kingma and M. Welling. "Auto-encoding variational bayes." In: *arXiv preprint arXiv:1312.6114* (2013).

[37] L. KPFRS. "On lines and planes of closest fit to systems of points in space." In: *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (SIGMOD)*. 1901, p. 19.

[38] M. A. Kramer. "Nonlinear principal component analysis using autoassociative neural networks." In: *AIChE journal* 37.2 (1991), pp. 233–243.

[39] S. Kullback and R. A. Leibler. "On information and sufficiency." In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[40] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi. "A survey on datasets for fairness-aware machine learning." In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12.3 (2022), e1452.

[41] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation applied to handwritten zip code recognition." In: *Neural computation* 1.4 (1989), pp. 541–551.

[42] Z. Li and C. Xu. "Discover the unknown biased attribute of an image classifier." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 14970–14979.

[43] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem. "On the fairness of disentangled representations." In: *Advances in neural information processing systems* 32 (2019).

[44] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. "Challenging common assumptions in the unsupervised learning of disentangled representations." In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.

[45] M. B. McDermott, T. M. H. Hsu, W.-H. Weng, M. Ghassemi, and P. Szolovits. "Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output." In: *Machine Learning for Healthcare Conference*. PMLR. 2020, pp. 913–927.

[46] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. "A survey on bias and fairness in machine learning." In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35.

[47] G. H. Nguyen, A. Bouzerdoum, and S. L. Phung. "A supervised learning approach for imbalanced data sets." In: *2008 19th international conference on pattern recognition*. IEEE. 2008, pp. 1–4.

[48] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman. *Social data: Biases, methodological pitfalls, and ethical boundaries.(2016)*. 2016.

[49] W. Peebles, J. Peebles, J.-Y. Zhu, A. Efros, and A. Torralba. "The hessian penalty: A weak prior for unsupervised disentanglement." In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer. 2020, pp. 581–597.

[50] A. Peters, G. N. C. (C. https://nako. de geschaeftsstelle@ nako. de, A. Peters, K. H. Greiser, S. Göttlicher, W. Ahrens, M. Albrecht, F. Bamberg, T. Bärnighausen, H. Becher, et al. "Framework and baseline examination of the German National Cohort (NAKO)." In: *European Journal of Epidemiology* 37.10 (2022), pp. 1107–1124.

[51] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn. "Diffusion autoencoders: Toward a meaningful and decodable representation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10619–10629.

[52] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. "Language models are unsupervised multitask learners." In: *OpenAI blog* 1.8 (2019), p. 9.

[53] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. "Hierarchical text-conditional image generation with clip latents." In: *arXiv preprint arXiv:2204.06125* 1.2 (2022), p. 3.

[54] A. Razavi, A. Van den Oord, and O. Vinyals. "Generating diverse high-fidelity images with vq-vae-2." In: *Advances in neural information processing systems* 32 (2019).

[55] J. Rocca. *Understanding Variational Autoencoders (VAEs)*. 2019. URL: `https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73` (visited on 03/27/2024).

[56] J. Schmidhuber. "Deep learning in neural networks: An overview." In: *Neural networks* 61 (2015), pp. 85–117.

[57] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. "Deep unsupervised learning using nonequilibrium thermodynamics." In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265.

[58] J. Song, C. Meng, and S. Ermon. "Denoising diffusion implicit models." In: *arXiv preprint arXiv:2010.02502* (2020).

[59] H. Suresh and J. V. Guttag. "A framework for understanding unintended consequences of machine learning." In: *arXiv preprint arXiv:1901.10002* 2.8 (2019), p. 73.

[60] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. "Rethinking the inception architecture for computer vision." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[61] W. Thong and C. G. Snoek. “Feature and label embedding spaces matter in addressing image classifier bias.” In: *arXiv preprint arXiv:2110.14336* (2021).

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

[63] L. Van der Maaten and G. Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is all you need.” In: *Advances in neural information processing systems* 30 (2017).

[65] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu. “Disentangled representation learning.” In: *arXiv preprint arXiv:2211.11695* (2022).

[66] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky. “Towards fairness in visual recognition: Effective strategies for bias mitigation.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8919–8928.

[67] Z. Yuan, Y. Yan, M. Sonka, and T. Yang. “Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3040–3049.

[68] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. “Men also like shopping: Reducing gender bias amplification using corpus-level constraints.” In: *arXiv preprint arXiv:1707.09457* (2017).