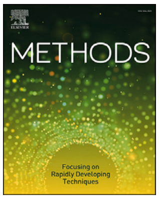
Methods 218 (2023) 110–117

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth



Interpretable thoracic pathologic prediction via learning group-disentangled representation

Hao Li [a,b], Yirui Wu [a,b,c,*], Hexuan Hu [a,b,c], Hu Lu [d], Qian Huang [a,b], Shaohua Wan [e]

[a] Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing 210093, China
[b] College of Computer and Information, Hohai University, Nanjing 210093, China
[c] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130015, China
[d] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China
[e] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, China

ARTICLE INFO

Keywords:
Disentangled representation learning
Group-disentangled feature representation
Thoracic pathologic prediction

ABSTRACT

Deep learning has brought a significant progress in medical image analysis. However, their lack of interpretability might bring high risk for wrong diagnosis with limited clinical knowledge embedding. In other words, we believe it's crucial for humans to interpret how deep learning work for medical analysis, thus appropriately adding knowledge constraints to correct the bias of wrong results. With such purpose, we propose Representation Group-Disentangling Network (RGD-Net) to explain the process of feature extraction and decision making inside deep learning framework, where we completely disentangle feature space of input X-ray images into independent feature groups, and each group would contribute to diagnose of a specific disease. Specifically, we first state problem definition for interpretable prediction with auto-encoder structure. Then, group-disentangled representations are extracted from input X-ray images with the proposed Group-Disentangle Module, which constructs semantic latent space by enforcing semantic consistency of attributes. Afterwards, adversarial constricts on mapping from features to diseases are proposed to prevent model collapse during training. Finally, a novel design of local tuning medical application is proposed based on RGB-Net, which is capable to aid clinicians for reasonable diagnosis. By conducting quantity of experiments on public datasets, RGD-Net have been superior to comparative studies by leveraging potential factors contributing to different diseases. We believe our work could bring interpretability in digging inherent patterns of deep learning on medical image analysis.

1. Introduction

To simplify the problem of performing medical image analysis is to build a desired intermediate representation like feature for hidden information in the input data. In fact, desired representation refers to that each feature should vary with others, where the most prefer form could be features as orthogonal basis. Under such feature basis, diseases can be easily classified without overlap or mutual information, resulting in wrong diagnose problem. Essentially, such problem is well studied in computer vision domain, named as disentangling factors of variation, where they try to learn a representation of the data which decomposes an observation into factors of variation which we can independently control.

For example, Bouchacourt et al. [1] propose a deep probabilistic model to learn a disentangled representation from a set of grouped samples, where they successfully separate the latent representation into two swappable and semantical parts. Later, Group Supervised Learning (GSL) [2] is trained on groups of semantically related images and reconstruction objectives, allowing to decompose inputs into swappable components. Components from different images thus can be recombined to synthesize new samples. Regarding that existing Self-Supervised Learning (SSL) only disentangles simple augmentation features such as rotation and colorization, Wang et al. [3] formulate an iterative SSL algorithm: Iterative Partition-based Invariant Risk Minimization (IP-IRM), which successfully grounds the abstract semantics and the group acting on them into concrete contrastive learning.

[*] Corresponding author at: College of Computer and Information, Hohai University, Nanjing 210093, China.
E-mail addresses: 211307030003@hhu.edu.cn (H. Li), wuyirui@hhu.edu.cn (Y. Wu), hexuan_hu@hhu.edu.cn (H. Hu), luhu@ujs.edu.cn (H. Lu), huangqian@hhu.edu.cn (Q. Huang), shaohua.wan@uestc.edu.cn (S. Wan).

https://doi.org/10.1016/j.ymeth.2023.08.001
Received 9 March 2023; Received in revised form 21 May 2023; Accepted 2 August 2023
Available online 4 August 2023

Inspired by these works, we want to use minimal supervision to learn a latent representation that reflects the disease semantics behind a specific grouping of X-ray images, where the samples within a group share a common factor of variation. In other words, we want to group semantics hidden in images into a relevant and disentangled representation that clinicians can easily understand and exploit, which is the original idea of applying group-disentangled representation for medical image analysis. Inspired by remarkable progress gained by deep learning methods [4,5], we further explore how to extracts group-disentangled disease representations with one typical deep learning structure.

In fact, existing deep probabilistic models often assume that the observations are independent and identically distributed, work as mappings from input factors to output classification results without explicit explanations. Therefore, they fail in promoting clinicians' and patients' confidence in trusting automatical diagnosis, thus preventing the usage of deep learning in medical domain. Most attempts [6,7] to explain deep learning focus on 'post-hoc' analysis by proving the importance of low-level visual features in producing accurate predictions. However, they couldn't directly link low-level visual features with high-level semantical diseases, and visually explain the decision making process. Essentially, both linking and explaining operations are valuable for clinicians to understand working patterns of deep learning for prediction.

As an alternative way, interpretable deep learning [8,9] considers the inherent requirement of interpretation to embed clues based explanations in their neural network design. For example, Ouyang et al. [10] propose Longitudinal Neighborhood Embedding (LNE), which is defined as a refinement of group-learning representation by replacing the linear modeling of brain aging with one that is consistent in local neighborhoods in the latent space. With LNE, they successfully obtain an encoding so that neighborhoods are age-consistent and progression-consistent for further applications.

Most of them built their framework on variational auto-encoder (VAE), which achieve significant process towards explainable deep learning by performing linking and explaining steps with help of visual clues represented as feature groups. However, they generally ignore independence of learned clues, where they map visual samples onto a latent space that overlapped separates the information belonging to different attributes. Therefore, they only achieve partly disentangled effects with overlapping and coarse-grained low-level features as shown in Fig. 1(a), resulting in confused explanations and low accuracy classification results.

To achieve completely group-disentangled latent space as shown in Fig. 1(b), it's proved to enforce semantic consistency of attributes, thus facilitating to leverage semantic links between samples. In other words, a completely disentangled latent representation space should be consist of subspaces, each encoding one attribute and each pair sharing no feature components. Following designs of completely disentanglement, training such a model usually faces one fundamental challenge, i.e., shortcut problem, that models may learn degenerate encodings by focusing on local minimum instead of global minimum, especially equipped with relatively free-form encoding network, such as VAE. Last but not least, how to develop task-specified interpretable deep learning methods remains an open question, due to the lack of involvement of existing clinical knowledge for either decision making or post explanation.

In this paper, we propose RGD-Net for interpretable thoracic pathologic prediction. We firstly achieve completely group-disentangled representations of diseases through the proposed Group-Disentangle Module. Such module is designed with group-swap and linking operations to leverage semantic links between input X-ray images and diseases, enforcing semantic consistency of attributes. To mitigate shortcut problem, we propose adversarial constricts, which borrows the idea of GAN to retain informative features during iteratively updating via group-swap and linking operations. Such constricts guarantee the model to seek for global minimum by forcing nash equilibrium between free-form grouping and convinced diagnosis, thus preventing model collapse. Significantly, we have developed a local-tuning medical application to demonstrate the effectiveness of interpretable thoracic pathologic prediction using RGD-Net. This application is capable of making informed decisions by updating only a subset of the subspace, thus reducing the computational burden of training from scratch when encountering new samples or unsatisfactory results.

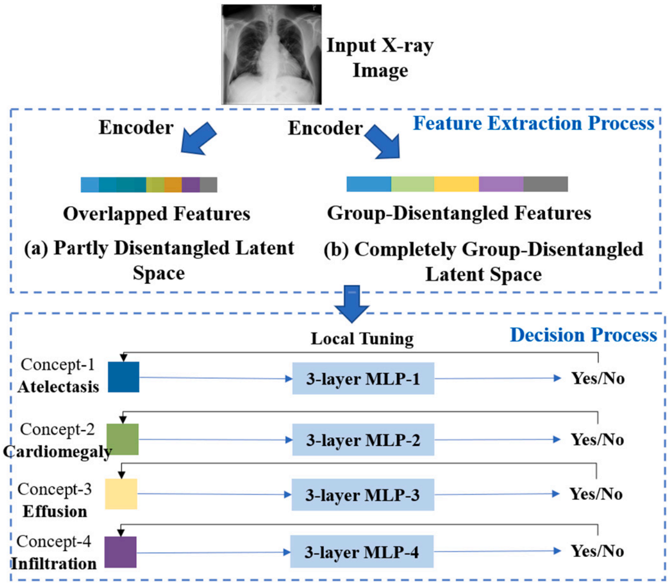


Fig. 1. (a) Partly disentangled latent space, where its feature groups are overlapped and coarse-grained, leading to confused explanations. (b) Completely disentangled latent space provided by RGD-Net, where feature groups are completely decomposed into independent subspaces, each of which corresponds to a specific disease.

To sum up, our contributions are as follows:

- We propose *Representation Group-Disentangling Network* (RGD-Net), which completely extracts group-disentangled disease representations with fine-grained and non-overlapping features, thus promoting both interpretability and prediction accuracy.
- To resist shortcut problem caused by trapping in local minimum, an adversarial constraint is proposed to retain informative features during iteratively updating, thus forcing global minimum and avoiding model collapse.
- We experimentally demonstrate that RGD-Net can significantly improve classification accuracy, and showcase the potential local tuning medical application of RGD-Net, which not only enhances interpretable capability, but also relieves the burden of re-training.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 presents an overview structure and details of the RGD-Net. Section 4 conducts quantity of experiments to show the effectiveness of RGD-Net. Finally, Section 5 concludes the paper and shows the prospect.

2. Related work

The existing methods related to RGD-Net can be categorized into the following two types: Disentangled Representation Learning, and Thoracic Pathologic Prediction.

2.1. Disentangled representation learning

Disentangled Representation Learning [11,9] aims to separate the latent space of data into several parts, each representing a concrete, independent, and human-understandable concept.

Existing methods for disentangling the latent space can be categorized into two classes: unit-disentanglement methods and group-disentanglement methods. The former methods treat a single latent unit as an independent concept. Following Variational Autoencoders [12], most unit-disentanglement methods [13,14] incorporate KL-divergence into the objective by forcing the latent factors to be statistically independent.

On the contrary, group-disentanglement methods treat a group of latent features as a concept. For example, Attila et al. [11] train their model with input triplets, in which the first two images should differ in the varying factor, but have the same common factor, and the third image is completely different from the first two. Furthermore, they offer solutions and analysis for shortcut problem and reference ambiguity in the disentangled representation learning. As a modification of the variational autoencoder (VAE) framework, Higgins et al. [15] introduce beta-VAE, a framework to automatically discover interpretable factorized latent representations from original image data in a completely unsupervised manner. They further introduce an adjustable hyperparameter to keep balance between latent channel capacity and independence constraints.

Considering rotation transformations harmful to contrastive learning (CL) resulting in failure when the objects show unseen orientations, Bai et al. [16] propose a representation focus shift network (RefosNet), which adds the rotation transformations to CL methods to improve the robustness of representation. Later, Liu et al. [17] carefully review the latest methods, where they motivate the need for disentangled representations, revisit key concepts, and describe practical building blocks and criteria for learning such representations. Person re-identification (Re-ID) in real-world scenarios suffers from various degradations. Follow the idea of disentangled representation learning, Huang et al. [18] propose a degradation invariance learning framework for robust person Re-ID, where they carefully design a content-degradation feature disentanglement strategy to capture and isolate task-irrelevant features.

RGD-Net prefers implicit disease concepts rather than explicit attributes. Moreover, RGD-Net brings capability to be built in a practical medical application by solving the shortcut problem with adversarial constraints.

2.2. Thoracic pathologic prediction

To present ideas for predicting thoracic disease with latest improvement, we focus on related deep learning methods for readers' convenience. Early, Wang et al. [19] propose a weakly supervised framework for multi-label classification of chest diseases, which have done experiments on X-Ray8 dataset for 8 common chest pathologies. Then, Zhou et al. [20] propose a weakly supervised adaptive network, named as DENsenET-169, for chest disease recognition and classification in chest radiography. Specifically, they use different deep learning models for anomaly discovery classification and localization. Subsequently, Li et al. [21] propose a unified method for disease identification and localization with limited labeling data, where they adopt a Multi Instance learning (MIL) formula that improves performance compared with baseline models of ResNet and DenseNet.

Afterwards, Rajpurkar et al. [22] propose a 121-layer convolutional neural network CheXNet, which is trained on “ChestX-ray14”, a expanded dataset of “ChestX-ray8” and containing over 100,000 frontal-view X-ray images with 14 diseases. Then, Wong et al. [23] propose a deep learning-based framework using Inception-ResNet-V2 for abnormal classification of chest X-ray images. Similarly, Wang et al. [24] propose a ChestNet model, which consists of a classification module and an attention module for computer-aided diagnosis of thoracic disease on CXR images.

Building on Information Bottleneck Attribution (IBA) method, Khakzar et al. [25] propose Inverse IBA to identify all informative regions that have high mutual information with the network's output. Thus all predictive cues for pathologies are highlighted on the X-rays, a desirable property for chest X-ray diagnosis. Since input feature attribution methods merely identify the importance of input regions for the prediction, Khakzar et al. [26] try to discover the semantics associated with feature maps extracted by the deep learning network. Later, Calli et al. [27] review recent studies with deep learning on chest radiographs.

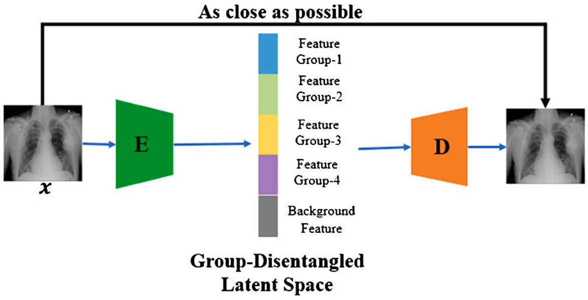


Fig. 2. Group-disentangled latent space encoding with disease-specific feature groups via auto-encoder structure.

Most recently, Wang et al. [28] propose a framework with two steps: the Discrimination-DL and the Localization-DL. The former extracts lung features from chest X-ray radiographs for COVID-19 discrimination, meanwhile the later recognize X-ray radiographs to localize and assign them into the left lung, right lung or bipulmonary. Mao et al. [29] propose a new abnormality detection approach based on an autoencoder, which outputs not only the reconstructed normal version of the input image, but also a pixel-wise uncertainty prediction. Considering most deep learning works suffer from slow convergency and high computing cost, Kong et al. [30] present CT-CAD, context-aware transformers for end-to-end chest abnormality detection on X-Ray images, allowing the transformer to focus on feature subspace and accelerate convergence speed.

3. Problem definition for interpretable prediction

Auto-encoders are often used in deep learning to enhance interpretability. They perform hidden space decomposition, obtaining a quantity of latent vectors that contain vast information corresponding to input chest X-ray images. Through reconstruction training, auto-encoder essentially compresses the input into the hidden space through encoder, and then reconstructs the original image through decoder based on hidden space.

Formally, we define Auto-Encoder: $\mathcal{X} \rightarrow \mathcal{X}$ as a combination of an encoder $E : \mathcal{X} \rightarrow \mathcal{R}^d$; and a decoder $D : \mathcal{R}^d \rightarrow \mathcal{X}$, where d denotes the dimension of the latent space $Z = E(X) \in \mathcal{R}^d$. To enhance interpretability in feature extraction, we wish to divide latent space into several semantic-specific parts as shown in Fig. 2. We define such property with the following formal definition.

Definition *(Group-disentangled latent space)*. A group-disentangled latent space refers to a space consisting of several consecutive, non-overlapping subspaces, each of which is responsible for one specific concept.

Such definition can also be expressed in the view of row-vectors:

$$z^{(1)} = [g_1^{(1)}, g_2^{(1)}, \ldots, g_m^{(1)}, b^{(1)}], \tag{1}$$

where row-vector $z^{(1)}$ is the concatenation of m row-vectors $\{g_i^{(1)} \in \mathcal{R}^{d_i}\}_{i=1}^m$ and a background row-vector $b^{(1)} \in \mathcal{R}^b$. It's noted that $d = \sum_{i=1}^m d_i + b$, where $\{d_i\}_{i=1}^m$ and b are hyper-parameters, and g_i corresponds to the concept c_i.

Although auto-encoder network can compress the image into the latent space and reconstruct it in a proper way, researchers still care about internal structure of latent vectors, since general auto-encoder fails in

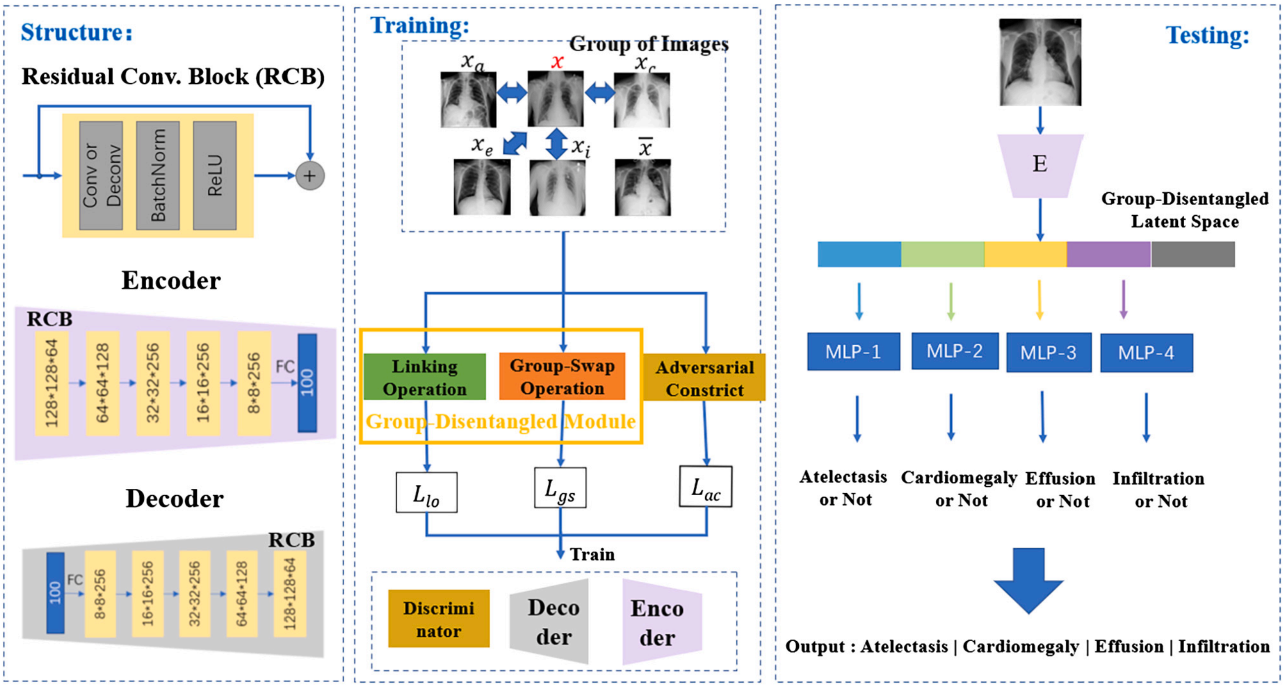


Fig. 3. The overall structure of RGD-Net, which extracts group-disentangled representations of disease through the Group-Disentangled Module and Adversarial Constrict. During testing, we use them to accurately predict corresponding disease labels.

generating separate independent representations of concepts. Without such independent representations, users are hard to be convinced by external and explicit links between concepts and predictions. Therefore, we propose RGD-Net, which can completely disentangle the latent space into different subspaces, each part corresponding to a pathological concept. In that way, RGD-Net provides a robust and reasonable explanation on relationship between feature groups and disease labels.

4. Methodology

We first introduce overall structure design of RGD-Net. Then, we describe in detail how we perform complete disentanglement via Group-Disentangled Module. Finally, we present the adversarial constrict that mitigates shortcut problem, even facing free-from swap and linking operations existed in former step.

4.1. Workflow design of RGD-Net

As shown in Fig. 3, we propose RGD-Net to obtain group-disentangled latent space by completely disentangling representations of disease concepts (e.g., Atelectasis, Cardiomegaly, Effusion, and Infiltration in our case) based on a group of semantically related images. After training, we use the reconstructed group-disentangled latent space to perform quantity of downstream tasks, such as accurately predicting diseases based on testing images.

Specifically, RGD-Net firstly takes a group of semantically-related X-ray images as inputs. Then, it trains its encoder and decoder through the proposed Group-Disentangled Module and Adversarial Constrict, which forces the encoder structure to reconstruct the input image by a group-disentangled latent space. It's noted that Group-Disentangled Module contains two operations, i.e., Linking and Group-Swap Operation. Linking Operation acts like auto-encoder, which builds direct relationship between semantical concepts of disease and low-level visual features. We calculate its related reconstruction loss L_{lo} for each image. Meanwhile, Group-Swap Operation enforces semantic consistency of disease concepts constrained by the after-swap reconstruction loss L_{gs}. Last but not least, Adversarial Constrict builds on the idea of GAN by involving adversarial loss L_{ac} to solve the model collapse, that may encounter in the process of group-disentanglement and is generally defined as shortcut problem. During training, we combine three kinds of losses as a total loss L:

$$L = \min_{\mathrm{D,E}} \max_{Dis} L_{lo} + \lambda_{gs} L_{gs} + \lambda_{ac} L_{ac}, \tag{2}$$

where L_{lo}, L_{gs} and L_{ac} refer to losses of linking operation, group-swap operation and adversarial constrict part respectively, and scalar coefficients λ_{gs}, λ_{ac} represent the importance factor of different loss terms. It's noted that we optimize L by gradient descent on parameters of encoder (E), decoder (D) and discriminator (Dis).

After training, it's supposed that we can apply RGD-Net on various computer vision tasks to provide convinced results. For example, it's expected to connect a classifier, e.g., MLP, with group of visual features for classification tasks. Alternatively, users can connect a detector with feature groups for detection tasks. In this paper, we demonstrate the effectiveness of RGD-Net to predict four categories of diseases based on chest X-ray images. Guided by the idea to apply on automatical medical application, we use a trained encoder to convert the input image into a group-disentangled latent space during testing phase. Afterwards, we predict thoracic pathologies disease concepts based on the new input X-ray images with an additional classification module with 3 layers of MLPs.

Encoder E is composed of a convolutional layer to generate feature map, four residual convolutional blocks with stride 2 for reshaping feature map to a vector, and a fully-connected layer to output a 100-dimensional vectors as latent feature. Meanwhile, decoder D mirrors the encoder in structure with a fully-connected layer, 4 residual de-conv blocks with stride 2 to reshape into a cuboid, and finally a de-conv layer to compute a synthesized image. Note that our method can be applied to the feature extraction part of any encoder - decoder structure neural network, rather than being limited to residual blocks.

4.2. Group-disentangled representation for learning

While training our RGD-Net, we wish to group-disentangle these latent spaces by E. We use a group-swap operation based on auto-encoder

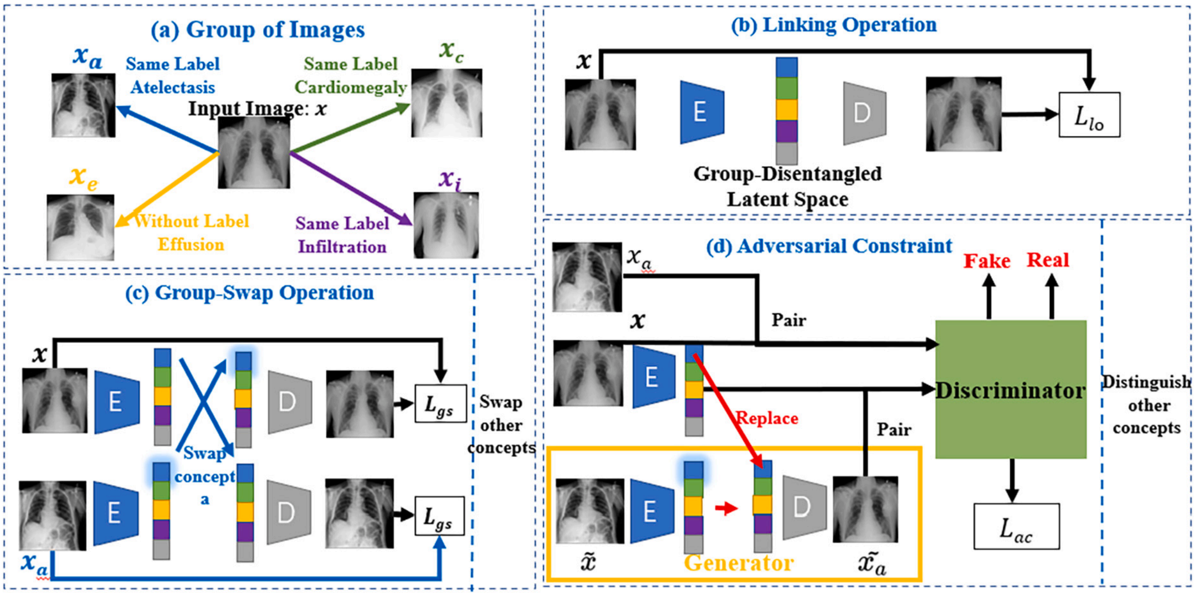


Fig. 4. Four steps in training RGD-Net to learn group-disentangled latent space: (a) **Group Images**, where we input a group of semantically related images to learn their common properties; (b) **Linking Operation**, where we link relationship between semantical concepts of disease and low-level visual features, calculating self-reconstruction loss for each image; (c) **Group-Swap Operation**: where we swap part of the latent representations of their shared concepts to enforce semantic consistency of disease concepts. (d) **Adversarial Constraint** takes triplet images as input to solve the *shortcut problem* caused by trapping in local minimal.

to link low-level visual features with high-level disease concepts in the latent space.

As shown in Fig. 4 (a), we take a group of semantically related images S as the input of RGD-Net. First, we randomly select an image x from the data set, in this case, with atelectasis and cardiomegaly pathology. Based on the same pathology as x, we select images from the data set with Atelectasis and Cardiomegaly and without infiltration and effusion. Therefore, there are five images as input to RGD-Net. As input in this way, common properties between images can be effectively learned.

To retain the information of images in the hidden space, we use a auto-encoder based Linking Scheme, which links relationship between semantical concepts of disease and low-level visual features as shown in Fig. 4 (b). Specifically, for each input X, we embed data in a low-dimensional vector by the encoder. Then we link d_i units of the vector to a specific disease concept c_i. Formally, we select a subset of the latent space $g_i = [\mu_{l_i+1}, \ldots \mu_{l_i+d_i}]$, where l_i is the start position of the subset for concept c_i. Finally, we input this latent vector into the decoder and calculate the reconstruction loss L_{ls} for each image. As shown in Fig. 4 (c), we use the group-swap module to enforce semantic consistency of disease concepts, and extract features of disease concepts by leveraging semantic links between input images.

Taking an image pair sharing a disease as input, the group-swap module exchanges the corresponding part of the disease in the hidden space of the two images, and expects to get same result as the input through the decoder. Formally, for all $x^o \in S,\ x^o \neq x$, with the pair $(x^o,\ x)$ share one concept value j (e.g., Atelectasis), the **Group-Swap** operation is defined as

$$z = E(x)\ ,\ \ z^o = E(x^o) \text{ and } \mathrm{z_s}, \mathrm{z_s^o} = \textbf{swap}(\mathrm{z}, \mathrm{z^o}, \mathrm{j}), \tag{3}$$

where the **swap** operation is defined as

$$\begin{aligned} &\textbf{swap}(z^{(1)}, z^{(2)}, k) \\ &= \textbf{swap}([g_1^{(1)}, \ldots, g_k^{(1)}, \ldots, g_m^{(1)}, b^{(1)}], [g_1^{(2)}, \ldots, g_k^{(2)}, \ldots, g_m^{(2)}, b^{(2)}], k) \\ &= [g_1^{(1)}, \ldots, g_k^{(2)}, \ldots, g_m^{(1)}, b^{(1)}], [g_1^{(2)}, \ldots, g_k^{(1)}, \ldots, g_m^{(2)}, b^{(2)}]. \end{aligned} \tag{4}$$

The group-swap module is subject to the reconstruction loss

$$L_{gs} = ||D(z_s) - x||_2^2 + ||D(z_s^o) - x^o||_2^2. \tag{5}$$

4.3. Adversarial constrict for training

Ideally, if there exist sufficient sample pairs sharing no duplicate concepts, loss of group-swap operation L_{gs} will be zero, so that complete group-disentanglement being logically obtained. However, due to free-form group-swap operation in former group-disentangled module, shortcut problem can occasionally occur with local minimum trap, where RGD-Net may learn degenerate encodings that all information of input images are retained in the group of background features.

It's supposed that shortcut problem can be mitigated by reducing dimensionality of background features, which not only forces the encoder to build a complete representation with several groups other than only one group of features, but also forces both encoder and decoder to properly link feature groups with the corresponding concepts under reasonable space room assumption. However, strategy of reducing dimensionality can only be convenient in practice, nevertheless resulting in time-consuming trial-and-error procedures to guess the proper dimensionality number.

Unlike adding constraints on feature space with setting a hyper-parameter, we propose an adversarial constrict to solve the shortcut problem without additional magic number to determine. As shown in Fig. 4 (d), we take triplet images, i.e., x_a, x and $\tilde{x_a}$, as input and introduce an adversarial training style. Specifically, the generator uses encoder-decoder structure to replace one specific feature group (represented as concept a) from x to $\tilde{x}$, thus generating new image $\tilde{x_a}$. In other words, the generator learns to generate an image containing pathological features shared by x and x_a, trying to fool the discriminator with new and fake image pair $[x, \tilde{x_a}]$. Meanwhile, the discriminator is designed as neural network to distinguish between original/real image pair $[x, x_a]$ and new/fake image pair $[x, \tilde{x_a}]$. Formally, $\tilde{x_a}$ can be defined as

$$\tilde{z} = E(\tilde{x})\ ,\ \ z = E(x) \text{ and } \tilde{\mathrm{x}_\mathrm{a}} = \textbf{swap}(\tilde{z}, z, \mathrm{a}). \tag{6}$$

In Fig. 4 (d), we show an example in adversarial training style by swapping the first feature group. Similarly, we construct image triples with

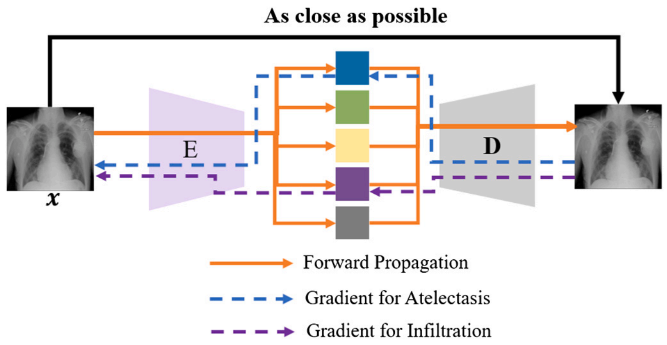


Fig. 5. Local tuning principle. When the model produces bad results in the medical automated decision-making process, the model can be tuned locally by limiting the gradient propagation, rather than tuning the whole model.

different feature groups, i.e., concepts, and calculate total adversarial losses with all image triples:

$$L_{ac} = \sum_{a=1,...,5} log(Dis(x, x_a)) + log(1 - Dis(x, \bar{x_a})), \tag{7}$$

where the total number of disease concepts is 5 in our medical diagnosis application, and function $Dis()$ represents the discriminator to judge real or fake pair.

Essentially, the key idea in adversarial constricts is to achieve minimal adversarial loss, thus forcing informative low-level features belong to disease concepts to keep remained during adversarial training process. In other words, global minimum can be achieved only if original image pair $[x, x_a]$ is real and new image pair $[x, \bar{x_a}]$ is fake, so that replacing any concept can't beat the original concept pairing. In that way, we prove the best matching performance of the original image pair $[x, x_a]$, thus forbidding all image pairs collapse to wrong pairs. With global optimum of first adversarial loss L_{ac} and then total loss L, latent space could be completely and stably group-disentangled.

4.4. Local tuning medical application

Based on RGD-Net, we can obtain a group-disentangled feature space, which not only brings interpretability and purer features, but also benefits medical automated decision-making processes. Consider yourself as a medical practitioner, while automating decisions with deep learning algorithms, has found that a few diseases are diagnosed with major errors, and that in most cases it works well. In this case, you should correct the part that caused the error and leave the rest unchanged.

Based on this idea, we propose a local tuning method for fine-tuning specific parts of the model in medical automated decision making, rather than retraining the entire model. As shown in Fig. 5, we use solid and dashed lines to represent the forward propagation and gradient back-propagation of the network respectively.

The forward propagation of the network is no different from other neural networks, but the specific diseases can be considered in the gradient back-propagation, and the gradient is calculated only for the corresponding part of the group-disentangled feature space, while the other parts are frozen. For example, when it is found that the model has a large error in the diagnosis of Atelectasis in the process of medical automated decision making, we can only make local tuning to the network, i.e., only allow the gradient of the corresponding features of Atelectasis to be transmitted, while gating the gradient of other features.

5. Experiments

We first evaluate the performance of our method on thoracic pathologic prediction and compare it with other non-disentangled DL methods. Then, to demonstrate the effectiveness of the module, we performed ablation experiments. Further, We evaluate our performance on learning group-disentangled representations and compare it with some partially disentangled and non-disentangled methods. Finally, we introduce the local tuning method to help clinicians improve performance in medical automated decision making.

Table 1
Comparison experiments on ChestXray-14 dataset. For each pathology, the highest AUROC scores are bolded.

Methods	Atel	Card	Effu	Infi
RGD-Net (ours)	**0.8630**	0.8980	**0.9269**	**0.8653**
CheXNet [22]	0.8094	**0.9248**	0.8638	0.7345
Yao et al. [31]	0.7720	0.9040	0.8590	0.6950
Wang et al. [19]	0.7160	0.8070	0.7840	0.6090
ChestNet [24]	0.7433	0.8748	0.8114	0.6772
Li et al. [21]	0.8000	0.8700	0.8700	0.7000
Zhou et al. [20]	0.8121	0.9066	0.8786	0.7065

Table 2
Comparison experiments on ChestXpert dataset. For each pathology, the highest AUROC scores are bolded.

Methods	Effu	Edema	Card
RGD-Net (ours)	0.900	0.9023	0.8871
Ye et al. [32]	0.9166	0.9436	0.8703
Pham et al. [33]	**0.9640**	**0.9580**	**0.910**
Irvin et al. [34]	0.9360	0.9280	0.8540

5.1. Datasets and measurements

We adopt two datasets to conduct thoracic pathologic prediction, i.e., chestxray-14 and ChestXpert. For the former dataset, we select a subset for experiments, which contains 36764 training images and 7353 testing images with 4 pathology labels (Atelectasis, Cardiomegaly, Effusion and Infiltration), which are extracting from the associated radiological reports using natural language processing. For the latter one, ChestXpert is a much larger data set, which demonstrates the performance of our method under large data volumes. We also select a subset in ChestXpert, which contains 162188 training images and 32437 testing images with 3 pathology labels (Pleural Effusion, Edema and Cardiomegaly).

To evaluate the performance of prediction, we follow the evaluation rules of both datasets, and adopt the area under receiver operating characteristic curve (AUROC) as our evaluation metric.

5.2. Accuracy analysis of thoracic pathologic prediction

Table 1 shows that our RGD-Net, which has significantly improved on ChestXray-14 dataset by prediction with group-disentangled latent representation compared with the existing methods.

The AUROC values of the network on A, B and C reached 1, 2, and 3 respectively The AUROC values of RGD-Net on Atelectasis, Cardiomegaly, Effusion and Infiltration reached 86.30%, 89.80%, 92.69%, 86.53% respectively, being 5.36%, -2.68%, 6.31% and 13.08% higher than the second-highest achieved by CheXNet. Considering the reason for the decline of AUROC in predicting Cardiomegaly, we explored the ChestXray-14 and found that there was an extreme imbalance of the label of Cardiomegaly. This may be a weakness of interpretable models, making it difficult to learn concepts from these imbalanced datasets.

To prove the performance of the proposed method on large medical datasets, we test the prediction performance of the proposed model on ChestXpert, one of the largest datasets currently available. As shown in Table 2, the accuracy of the proposed network is slightly lower on ChestXpert than the two latest networks, that is because our method considers not only the categories of predicted pathology, but also the

Table 3
Ablation and Division Experiments. AUROC of Thoracic Pathologic Prediction by RGD-Net with different structures on ChestXray-14.

Methods	Atel	Card	Effu	Infi
$L_{ls} + L_{gsm} + L_{GAN}$ (ours)	**0.8630**	0.8980	**0.9269**	**0.8653**
$L_{ls} + L_{gsm}$	0.6076	0.7048	0.7444	0.6332
$L_{ls} + L_{GAN}$	0.5263	0.5141	0.5365	0.5468
L_{ls} (Auto-Encoder)	0.5065	0.4713	0.5032	0.5289
Less Background	0.8497	0.8749	0.9013	0.8633
More Background	0.8263	**0.9048**	0.8883	0.8445

Table 4
Group-disentangled representation analysis. We use the row disease features to predict the AUROC of column diseases on the test set by a simple 3-MLP. Diagonals are bolded and '-' means that Esther et al. [8] fail to disentangle concept of background.

Disease	RGD-Net (ours) (completely disentangled)				Esther et al. [8] (partly disentangled)				AutoEncoder (without disentangled)			
	Atel	Card	Effu	Infi	Atel	Card	Effu	Infi	Atel	Card	Effu	Infi
Atelectasis	**0.8630**	0.4855	0.5094	0.5005	**0.6136**	0.4960	0.4816	0.5050	**0.6076**	0.4990	0.4802	0.5297
Cardiomegaly	0.4822	**0.8980**	0.4836	0.5063	0.5062	**0.6610**	0.4968	0.4758	0.5067	**0.7048**	0.5183	0.4864
Effusion	0.4893	0.5061	**0.9269**	0.5229	0.5153	0.5038	**0.6688**	0.5099	0.4884	0.4985	**0.7444**	0.5292
Infiltration	0.4986	0.4900	0.4985	**0.8653**	0.4863	0.5230	0.5315	**0.5910**	0.4996	0.4955	0.4911	**0.6332**
Background	0.4983	0.5200	0.4926	0.4926	-	-	-	-	0.5045	0.5029	0.5087	0.4887

interpretability of the network. It is useful in promoting clinicians' and patients' confidence and expanding the usage of DL in automated disease diagnosis. Moreover, our method can be easily migrated to other medical computing tasks. For example, we can replace the classification head with the detection head to detect the part related to the case.

5.3. Ablation experiments

To verify the effectiveness of each module in the proposed method, we conduct ablation studies on ChestXray-14, where performance is listed in Table 3. The AUROC will decrease by a large percentage without the help of L_{gsm} module (row 3), since group-swap implies that swapping one attribute does not destroy latent information for other attributes. Moreover, the group-swap module can enforce semantic consistency of disease concepts, and extract features of disease concepts by leveraging semantic links between input images. When we remove this module, the model's understanding of the concept of disease is reduced, and it is naturally difficult to accurately predict the pathology.

As removing the L_{GAN} module (shown in row 4), the model will degenerate to auto-encoder, the AUROC decreases again, indicating that the adversarial training module also has a certain group-disentanglement effect. As removing the L_{GAN} module from the RGD-Net (shown in row 2), AUROC drops to an approximate value of Auto-Encoder, which implies the fact that the model is collapsing. All the features of the image are kept in the background features, which makes the features not discriminative.

In the last two rows of Table 3, assuming that these thoracic pathologies are independent of each other, we distribute their corresponding latent subspace with same size. But these pathologies are not necessarily independent of each other and contain different amounts of information. Therefore, we modify the size of the latent space so that it is no longer equally divided. It's noted that less background of RGD-Net represents $g_i = 22, i = 1, 2, 3, 4$ and $b = 12$, and more background represents $g_i = 15, i = 1, 2, 3, 4$ and $b = 40$.

As we allocate less (as 12 in our paper) dimensions of latent space to represent background, the AUROC decreases by 1.33%, 2.31%, 2.56% and 0.2% for each pathology. If more (as 40 in our paper) latent space is used for background, the AUROC change by a percentage of -6.37%, +0.68%, -3.86% and -2.08%. This experiment shows that division equally is the most effective for this task.

5.4. Capability of group-disentangled representation analysis

To see the effect of group-disentanglement of our RGD-Net, we use the subspaces of disease concepts to predict four thoracic pathologies by a simple 3-MLP. If the hidden subspace contains all the information about the disease, the predicted result should be a matrix with 1 on the diagonal and 0.5 on the rest.

We use Esther et al. [8] and standard auto-encoder with classification head as comparison methods. The former partly disentangles the latent space, and the latter is not a disentangled method. Table 4 shows that RGD-Net successfully decomposes the image into a group-disentangled latent space and uses each subspace to accurately predict the corresponding concept, but not to predict other concepts. Results of two comparison methods, whose latent space is not completely group-disentangled, show that each subspace doesn't know what it corresponds to, so their AUROCs are nearly 0.5.

This result shows that our method can effectively learn to group-disentangle representation and decompose the feature space into several independent parts, each of which represents a certain disease concept. However, other methods do not enforce the semantic consistency between the latent space and the concept of diseases, which leads to unsatisfactory results.

6. Conclusion

This paper proposes a Representation Group-Disentangling Network (RGD-Net), which tries to explain the process of feature extraction and decision making inside deep learning framework. In fact, RGD-Net completely extracts group-disentangled disease representations with fine-grained and non-overlapping features. Specifically, RGD-Net extracts group-disentangled representations from input X-ray images with Group-Disentangle Module, which constructs semantic latent space by enforcing semantic consistency of attributes. To avoid possible model collapse problem, we propose adversarial constricts on mapping from features to diseases for robustness. Experiments demonstrate that RGD-Net can significantly improve classification accuracy, when comparing with partly disentangled interpretable. Our future work includes to extend such framework into other applications of medical image analysis.

CRediT authorship contribution statement

Hao Li: Conceptualization, Data curation, Methodology, Writing – original draft. **Yirui Wu:** Formal analysis, Funding acquisition, Writing – review & editing. **Hexuan Hu:** Investigation, Writing – review & editing. **Hu Lu:** Project administration, Supervision, Writing – review & editing. **Qian Huang:** Resources, Software, Writing – review & editing. **Shaohua Wan:** Validation, Visualization, Writing – review & editing.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Data availability

The data used in the manuscript is publicly available and can be downloaded directly from the Web.

Acknowledgement

This work was supported in part by a grant from National Key R&D Program of China under Grant No. 2021YFB3900601, the Fundamental Research Funds for the Central Universities under Grant B220202074, the Fundamental Research Funds for the Central Universities, JLU, the Natural Science Foundation of China under Grant 61702160, and the National Natural Science Foundation of China under Grant 61976050.

References

[1] D. Bouchacourt, R. Tomioka, S. Nowozin, Multi-level variational autoencoder: learning disentangled representations from grouped observations, in: Proceedings of AAAI, 2018, pp. 2095–2102.

[2] Y. Ge, S. Abu-El-Haija, G. Xin, L. Itti, Zero-shot synthesis with group-supervised learning, in: Proceedings of ICLR, 2021.

[3] T. Wang, Z. Yue, J. Huang, Q. Sun, H. Zhang, Self-supervised learning disentangled group representation as feature, in: Proceedings of Advances in Neural Information Processing Systems, 2021, pp. 18225–18240.

[4] L.K. Tam, X. Wang, E. Turkbey, K. Lu, Y. Wen, D. Xu, Weakly supervised one-stage vision and language disease detection using large scale pneumonia and pneumothorax studies, in: Proceedings of MICCAI, vol. 12264, 2020, pp. 45–55.

[5] F. Karim, M.A. Shah, H.A. Khattak, Z. Ameer, U. Shoaib, H.T. Rauf, F. Al-Turjman, Towards an effective model for lung disease classification: using dense capsule nets for early classification of lung diseases, Appl. Soft Comput. 124 (2022) 109077.

[6] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of CVPR, 2016, pp. 2921–2929.

[7] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in: Proceedings of ICCV, 2017, pp. 618–626.

[8] P.-A. Esther, C. Chen, J.R. Clough, Interpretable deep models for cardiac resynchronisation therapy response prediction, in: Proceedings of MICCAI, vol. 12261, 2020, pp. 284–293.

[9] M.J. Vowels, N.C. Camgöz, R. Bowden, Gated variational autoencoders: incorporating weak supervision to encourage disentanglement, in: Proceedings of 15th IEEE International Conference on Automatic Face and Gesture Recognition, 2020, pp. 125–132.

[10] J. Ouyang, Q. Zhao, E. Adeli, G. Zaharchuk, K.M. Pohl, Self-supervised learning of neighborhood embedding for longitudinal MRI, Med. Image Anal. 82 (2022) 102571.

[11] A. Szabó, Q. Hu, T. Portenier, M. Zwicker, P. Favaro, Understanding degeneracies and ambiguities in attribute transfer, in: Proceedings of ECCV, vol. 11209, 2018, pp. 721–736.

[12] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, in: Proceedings of ICLR, 2014.

[13] R.T.Q. Chen, X. Li, R. Grosse, D. Duvenaud, Isolating sources of disentanglement in variational autoencoders, in: Proceedings of ICLR, 2018.

[14] F. Locatello, S. Bauer, et al., Challenging common assumptions in the unsupervised learning of disentangled representations, in: Proceedings of ICLR, 2019.

[15] I. Higgins, L. Matthey, A. Pal, C.P. Burgess, X. Glorot, M.M. Botvinick, S. Mohamed, A. Lerchner, beta-VAE: learning basic visual concepts with a constrained variational framework, in: Proceedings of International Conference on Learning Representations, 2017.

[16] G. Bai, W. Xi, X. Hong, X. Liu, Y. Yue, S. Zhao, Robust and rotation-equivariant contrastive learning, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–14, https://doi.org/10.1109/TNNLS.2023.3243258.

[17] X. Liu, P. Sanchez, S. Thermos, A.Q. O'Neil, S.A. Tsaftaris, Learning disentangled representations in the imaging domain, Med. Image Anal. 80 (2022) 102516.

[18] Y. Huang, X. Fu, L. Li, Z. Zha, Learning degradation-invariant representation for robust real-world person re-identification, Int. J. Comput. Vis. 130 (11) (2022) 2770–2796.

[19] X. Wang, Y. Peng, et al., ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of CVPR, 2017, pp. 3462–3471.

[20] B. Zhou, Y. Li, J. Wang, A weakly supervised adaptive densenet for classifying thoracic diseases and identifying abnormalities, CoRR, arXiv:1807.01257 [abs], 2018.

[21] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L. Li, L. Fei-Fei, Thoracic disease identification and localization with limited supervision, in: Proceedings of CVPR, 2018, pp. 8290–8299.

[22] P. Rajpurkar, J. Irvin, et al., CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning, CoRR, arXiv:1711.05225 [abs], 2017.

[23] K.C.L. Wong, M. Moradi, J.T. Wu, T.F. Syeda-Mahmood, Identifying disease-free chest X-ray images with deep transfer learning, in: Medical Imaging: Computer-Aided Diagnosis, vol. 10950, 2019, 109500P.

[24] H. Wang, Y. Xia, ChestNet: a deep neural network for classification of thoracic diseases on chest radiography, CoRR, arXiv:1807.03058 [abs], 2018.

[25] A. Khakzar, Y. Zhang, W. Mansour, Y. Cai, Y. Li, Y. Zhang, S.T. Kim, N. Navab, Explaining COVID-19 and thoracic pathology model predictions by identifying informative input features, in: Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention, vol. 12903, 2021, pp. 391–401.

[26] A. Khakzar, S. Musatian, J. Buchberger, I.V. Quiroz, N. Pinger, S. Baselizadeh, S.T. Kim, N. Navab, Towards semantic interpretation of thoracic disease and COVID-19 diagnosis models, in: Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention, vol. 12903, 2021, pp. 499–508.

[27] E. Çalli, E. Sogancioglu, B. van Ginneken, K.G. van Leeuwen, K. Murphy, Deep learning for chest X-ray analysis: a survey, Med. Image Anal. 72 (2021) 102125.

[28] Z. Wang, Y. Xiao, Y. Li, J. Zhang, F. Lu, M. Hou, X. Liu, Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays, Pattern Recognit. 110 (2021) 107613.

[29] Y. Mao, F. Xue, R. Wang, J. Zhang, W. Zheng, H. Liu, Abnormality detection in chest X-ray images using uncertainty prediction autoencoders, in: Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention, vol. 12266, 2020, pp. 529–538.

[30] Q. Kong, Y. Wu, C. Yuan, Y. Wang, CT-CAD: context-aware transformers for end-to-end chest abnormality detection on X-rays, in: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, 2021, pp. 1385–1388.

[31] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, K. Lyman, Learning to diagnose from scratch by exploiting dependencies among labels, CoRR, arXiv:1710.10501 [abs], 2017.

[32] W. Ye, J. Yao, H. Xue, Y. Li, Weakly supervised lesion localization with probabilistic-cam pooling, CoRR, arXiv:2005.14480 [abs], 2020.

[33] H.H. Pham, T.T. Le, D.Q. Tran, D.T. Ngo, H.Q. Nguyen, Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels, Neurocomputing 437 (2021) 186–194.

[34] J. Irvin, P. Rajpurkar, et al., Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of AAAI, 2019, pp. 590–597.