

HydraViT: Adaptive multi-branch transformer for multi-label disease classification from Chest X-ray images

Şaban Öztürk^{a,b,*}, M. Yiğit Turalı^a, Tolga Çukur^{a,c}

^a Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey

^b Department of Management Information Systems, Ankara Hacı Bayram Veli University, Ankara 06500, Turkey

^c National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara 06800, Turkey

ARTICLE INFO

Keywords:

Multi-label classification
Chest X-ray
Deep learning
Vision transformer
Label co-occurrence

ABSTRACT

Chest X-ray is an essential diagnostic tool in the identification of chest diseases given its high sensitivity to pathological abnormalities in the lungs. However, image-driven diagnosis is still challenging due to heterogeneity in size and location of pathology, as well as visual similarities and co-occurrence of separate pathology. Since disease-related regions often occupy a relatively small portion of diagnostic images, classification models based on traditional convolutional neural networks (CNNs) are adversely affected given their locality bias. While CNNs were previously augmented with attention maps or spatial masks to guide focus on potentially critical regions, learning localization guidance under heterogeneity in the spatial distribution of pathology is challenging. To improve multi-label classification performance, here we propose a novel method, HydraViT, that synergistically combines a transformer backbone with a multi-branch output module with learned weighting. The transformer backbone enhances sensitivity to long-range context in X-ray images, while using the self-attention mechanism to adaptively focus on task-critical regions. The multi-branch output module dedicates an independent branch to each disease label to attain robust learning across separate disease classes, along with an aggregated branch across labels to maintain sensitivity to co-occurrence relationships among pathology. Experiments demonstrate that, on average, HydraViT outperforms competing attention-guided methods by 1.9% AUC and 5.3% MAE, region-guided methods by 2.1% AUC and 8.3% MAE, and semantic-guided methods by 2.0% AUC and 6.5% MAE in multi-label classification performance.

1. Introduction

The prevalence of thoracic diseases is a growing concern that poses a significant threat to human health. Lung cancer, the second most common cancer globally, accounts for 11%–12% of all cancer cases and is responsible for approximately 18% of cancer-related deaths [1]. Aspiration pneumonia, another aggressive thoracic disease, is responsible for about 2%–3% of all deaths in developed countries [2]. A prominent imaging technology for early diagnosis of these deadly conditions is chest X-ray (CXR), which is cost-efficient compared to other common modalities. However, the ever-increasing number of CXR scans, complex pathologies, variable lesion sizes, and subtle texture changes can compromise the accuracy of radiological readings. These challenges are further exacerbated by operator biases in developing countries with relatively limited accumulation of radiological expertise [3]. Therefore, the development of computer-aided diagnosis (CAD) algorithms that can automatically diagnose thoracic diseases from CXR scans can serve to improve efficiency and accuracy in radiological assessments.

The mainstream CAD approach for diagnosing thoracic diseases rests on the extraction of CXR features to help identify and locate pathological regions, followed by classification based on the extracted features to identify the presence of diseases [4,5]. In cases where only a single type of pathology exists per subject, a multi-category classification problem would have to be solved by assigning each CXR image to an exclusive disease category by selecting the label with the highest predicted probability [6]. Yet, separate pathology or multiple instances of a given pathology frequently co-occur in CXR images of individual subjects, rendering diagnosis a multi-label classification problem instead [7]. This significantly increases problem difficulty since a CXR image might have to be assigned simultaneously to multiple disease categories by identifying the subset of labels whose multivariate probability is the highest among all possible subsets. In the face of these challenges, convolutional neural networks (CNNs) have arguably become the de facto standard in extraction of CXR features, given their efficiency in learning visual features for downstream imaging tasks [8,

* Corresponding author at: Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey.
E-mail address: saban.ozturk@hbv.edu.tr (Ş. Öztürk).

9]. However, CNNs often suffer from suboptimal performance during identification of complex pathology in thoracic diseases due to several reasons: (1) The size, location and appearance of pathology in CXR images show high variability across disease classes and across subjects. Since CNN models use compact local filters with static weights for feature extraction, their generalization abilities can be compromised given the anatomical variability of pathology encountered in CXR images. (2) Multiple different pathology can co-occur in CXR images of a single subject, albeit the co-occurrence statistics of simultaneously-present pathology show high heterogeneity across disease classes [7]. CNN models are typically trained conventional softmax output layers, primarily devised for multi-category classification problems relying on exclusivity among classes. However, these output layers can have difficulty in handling multi-label classification of CXR images, as they try to compromise between sensitivity to individual classes versus sensitivity to co-occurrence relationships among classes [10].

Recent studies have considered advanced approaches to improve performance in multi-label CXR classification. A first group of CXR studies have proposed to augment CNN models with attention modules or attention-based masks in order to improve network focus on small-sized pathology [4,11–15]. Despite the efficiency and promising performance of attention-augmented CNNs, they can still show limited capture of the long-range context under multiple distributed or large-sized lesions, and limited generalization performance across subjects. A second group of studies have instead proposed vision transformer (ViT) models based on self-attention mechanisms to improve capture of long-range context and to improve generalization [16,17]. The transformer architecture has been extensively utilized in a myriad of image-based applications, encompassing image segmentation, object detection, classification, and other pertinent tasks, demonstrating its versatility across diverse domains [18–20], albeit often at the expense of elevated computational load. To maintain a desirable trade-off between performance and efficiency, hybrid models that integrate CNN and ViT blocks have also been proposed [5,11,21–25]. While these deep learning models have been adopted to learn representative features in multi-label CXR studies, they often neglect to-occurrence relationships between separate pathology [12,16,26,27]. Pre-defined hierarchical relationships between pathology labels have been considered previously to construct multi-label CXR classifiers [28,29]. However, these previous methods are typically trained to optimize the prediction accuracy for an aggregate output vector across classes, yielding heterogeneous classification performance across individual pathology labels.

Here, we introduce a novel adaptive multi-branch transformer model for multi-label disease classification from CXR images, named HydraViT. Our proposed model leverages a hybrid architecture composed of a convolutional spatial encoder module to efficiently extract feature maps of CXR images, and a transformer-based context encoder module to capture long-range contextual relationships across image patches and co-occurring pathology. To avoid bias due to co-occurrence patterns among disease labels, multi-task learning is performed based on a multi-branch output layer as inspired by recent machine learning studies [30,31]. Yet, differently from previous studies on multi-task learning, we propose a novel loss function that adaptively weights each output branch to further improve the learning of pathology co-occurrences. To our knowledge, HydraViT is the first model in the literature that adopts multi-task learning for each pathology label in multi-label CXR classification. Our main contributions are summarized below:

- HydraViT is a novel hybrid convolutional-transformer model that performs multi-tasking to improve reliability in multi-label disease classification from CXR images.
- HydraViT uses a transformer-based context encoder to capture long-range context and co-occurrence relationships between distinct pathology.

- HydraViT uses distance-based adaptive weights to account for variable co-occurrence statistics between separate pathology and thereby improve the homogeneity of model performance across classes.

The remainder of this paper is structured as follows: Section 2 presents a comprehensive literature review on multi-label CXR classification. Section 3 delineates the theoretical underpinnings and details of the proposed model. Section 4 elaborates on the experimental setup and provides a thorough account of parametric details. Section 5 reports ablation studies on model components and comparison studies against state-of-the-art baselines. Finally, Section 6 concludes with discussions on findings, limitations and future work.

2. Related work

2.1. Deep learning for CXR classification

The traditional framework for automated analysis of CXR images rests on the use of hand-constructed features and manual operator intervention, which undermine classification performance [32,33]. With the advent of deep learning, performance leaps have been attained in CXR analysis based on a variety of different CNN architectures [34–36]. Several lines of improvements in CNN models have been considered including deeper architectures [26], integration of recurrence dependencies [37], pyramidal architectures [17], quantum-classifier-based architectures [38], fusion-based architectures [39] and ensemble architectures to capture a diverse array of features [14,24,35,40–42]. Among recent studies on CNN-based CXR classification, [27] uses deep feature selection to extract most informative image features, [43] utilize CXR images and cough sounds to enhance performance. [7] uses feature selector and integrator branches to learn discriminative features of pathology. [25,44,45] use graph features to capture semantic similarities between image features. [46,47] employ optimization techniques to identify more effective sets of features, which in turn help improve classification performance. Several recent studies have employed diffusion-based methods [48,49] to improve feature reliability based on stable diffusion [50] and latent diffusion models [51].

A primary limitation of conventional CNN models concerns the use of static, local filter weights that can compromise generalization to atypical anatomy that varies in size, location, shape across subjects. Recent CXR classification studies have considered to use attention mechanisms either as augmentation to CNN backbones or as self-attention in ViT backbones to improve generalization and to guide the focus of the model towards disease-relevant regions [4,12,13,15]. Multiple attention mechanisms across different dimensions have been deployed including combination of channel, element, scale attention [5, 52], channel and spatial attention [7,16], class and label attention [6, 22], and multi-head self-attention [17]. Among recent studies, [11] proposes PCAN that uses a pixel-wise attention branch. [21] proposes Thorax-Net with an attention branch to exploit the correlation between class labels and locations of pathology. DualAnet by [24] includes two asymmetric attention networks to extract more discriminative features. [23] introduces PCSANet that uses a shuffle attention module to prioritize features related to pathology. Although these previous methods have focused on architectural improvements to CXR classification models so as to extract task-relevant features by focusing on pathology, they can elicit heterogeneous classification performance when multiple co-occurring pathologies are present in multi-label classification tasks. To address this limitation, HydraViT uniquely uses multi-task learning with a separate output branch for each label and adaptively weights each branch to cope with variable co-occurrence statistics across labels.

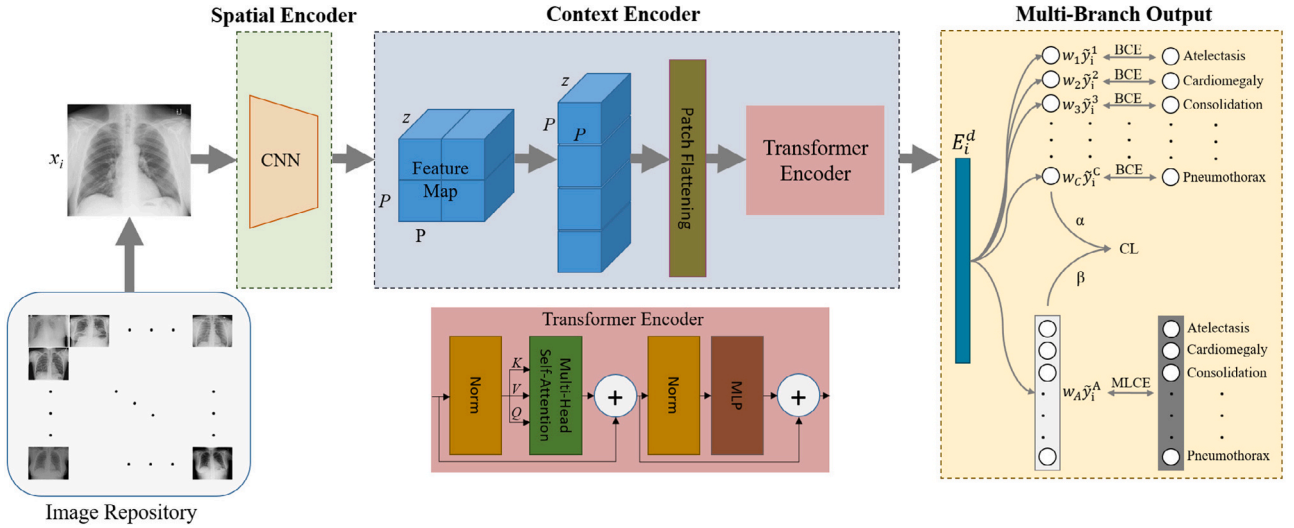


Fig. 1. HydraViT consists of a CNN-based spatial encoder module, a transformer-based context encoder module, and a multi-branch output module to maintain sensitivity to individual labels and to capture label co-occurrence relationships in multi-label CXR classification.

2.2. Multi-label classification

A common approach for multi-label CXR classification is based on model training via a multi-label cross-entropy loss expressed over a multi-dimensional output vector that spans across all examined classes produced by a softmax layer [5,14,26,34,53]. However, this conventional approach can compromise sensitivity to individual labels and can elicit suboptimal capture of co-occurrence relationships between separate labels [13,24]. To improve the learning of co-occurrence statistics between labels, a group of studies propose to refine label predictions via add-on network modules such as a co-occurrence module [54], a graph module [55,56], a recurrent module [37], label correlation guided discriminative learning [57], or a spatial-and-channel encoding module [7] to capture the semantic dependencies between separate labels. As an alternative approach, other studies propose to use modified loss functions such as weighted cross-entropy [58] or multi-label softmax losses [59] to leverage correlations among labels for improved classification performance. Commonly, these previous studies help emphasize label co-occurrence relationships during class predictions. However, since they utilize a single multi-dimensional output vector, they can be suboptimal in preserving sensitivity to individual labels. A recent study aimed to address this issue by an ensemble CNN model, where a separate CNN predicted the label for each class [60]. While this improves sensitivity to individual labels, it ignores label co-occurrence relationships.

To attain sensitivity to both individual labels and their co-occurrence relationships, HydraViT employs a multi-branch output module that maps contextual embeddings extracted by its transformer-based encoder onto class labels. Unlike previous multi-label methods, HydraViT simultaneously uses segregated uni-dimensional output variables for each individual label and an aggregated multi-dimensional output vector across labels. Adaptive weights are assigned to each output variable prior to the calculation of cross-entropy losses, augmented with a consistency loss between the individual and aggregated outputs to maintain consistency between their predicted pathology labels. While several imaging and computer vision studies have considered the use of separate network branches for each individual label in classification models [30,31,60], no previous study has proposed concurrent use of individual and aggregated branches whose predictions are aligned with a consistency loss to our knowledge.

3. Theory

3.1. Problem definition

Let us assume a training set of CXR images and corresponding disease labels $\{x_i, y_i\}_{i=1}^N$, where N is the number of training samples. $x_i \in \mathbb{R}^{H,W}$ is i th image with (H, W) denoting the image size across spatial dimensions. $y_i \in \mathbb{Z}_2^C$ is the C -dim label vector where C is the number of disease classes, and $y_i^c \in \{0, 1\}$ serves as an indicator for the c th class (0: absent, 1: present). To learn the required mapping for multi-label classification, i.e., $f : x_i \rightarrow y_i$, a mainstream approach employs cross-entropy loss [5,12–14,21,45,58]. Yet, conventional cross-entropy loss reflects an aggregate measure across all disease labels, so it does not explicitly consider the co-occurrence relationships among distinct pathology. In turn, a simple adoption of cross-entropy loss in multi-label classification can result in suboptimal performance.

3.2. HydraViT

To address the above-mentioned problems, HydraViT leverages multi-task learning based on a hybrid architecture where a CNN-based spatial encoder extracts lower-dimensional maps of local features followed by a transformer-based context encoder captures contextualized embeddings within and across pathologies in the input CXR image (Fig. 1). Multi-task learning is then exercised via a synergistic combination of dedicated output variables for each individual label and a multi-dimensional output vector aggregated across labels. To maintain sensitivity to both individual labels and label co-occurrence relationships, the learnable weighting of output variables is used in conjunction with a consistency loss between the individual and aggregated output variables. Network components and learning procedures for HydraViT are described below.

3.2.1. Spatial Encoder (SE)

The CNN-based SE module with parameters θ_{SE} is used to extract local spatial features of CXR images and lower dimensionality of feature maps prior to context encoding. Given the input image x_i , a low-dimensional latent representation $m_i \in \mathbb{R}^{H,W,z}$ is derived as $f_{SE} : x_i \rightarrow m_i$, where z is the dimensionality of feature channels:

$$m_i = Pool(\sigma(Conv(\dots Pool(\sigma(Conv(x_i)))\dots))) \quad (1)$$

where $Pool$ denotes a maximum pooling layer across a 2×2 neighborhood for two-fold downsampling, σ is an ReLU activation function, $Conv$ denotes a convolutional block.

3.2.2. Context Encoder (CE)

The transformer-based CE module with parameters θ_{CE} projects spatially-encoded feature maps onto contextualized embedding vectors, $f_{CE} : m_i \rightarrow E_i$, to capture long-range spatial relationships both within and across individual pathologies. For this purpose, m_i from the SE module is split into $N_p = r^2/P^2$ non-overlapping patches of size (P, P) with $P = r/2$, and flattened to zP^2 -dimensional vectors. The transformer encoder first projects the flattened patches onto an N_D -dimensional space through learnable linear projections and positional encodings:

$$E_i^0 = \left[(m_i)^1 P_E; (m_i)^2 P_E; \dots; (m_i)^{N_p} P_E \right] + P_E^{pos} \quad (2)$$

where $E_i^0 \in \mathbb{R}^{N_p \times N_D}$ denote patch embeddings, $(m_i)^p \in \mathbb{R}^{P^2}$ denotes the p th patch, P_E and P_E^{pos} are the linear projections and positional encodings, respectively. Next, path embeddings are processed via L transformer blocks, each comprising a cascade of layer normalization (*Norm*), multi-head self-attention (*MHSA*), and multi-layer perceptron (*MLP*) layers [61]. The l th block performs the following computations:

$$\tilde{E}_i^l = MHSA(Norm(E_i^{l-1})) + E_i^{l-1} \quad (3)$$

$$E_i^l = MLP(Norm(\tilde{E}_i^l)) + \tilde{E}_i^l \quad (4)$$

The output of the CE module E_i^L is taken as the contextualized embedding vector E_i . Please note that the conventional output head in a transformer encoder would map the embedding vector onto activations $o_i^{1 \dots C}$ in C output neurons, and leverage a softmax function (ζ) to compute probabilities for separate classes: $\zeta(o_i^j) = \frac{e^{o_i^j}}{\sum_{k=1}^C e^{o_i^k}}$, for $j = 1, \dots, C$. Because this formulation enforces a single class to dominate over the remaining classes in the output vector, it can be ineffective in capturing co-occurrence relationships between separate labels.

3.2.3. Multi-Branch Output (MBO)

Given E_i , the MBO module computes C uni-dimensional output variables for each label $\tilde{y}_i^1, \tilde{y}_i^2, \dots, \tilde{y}_i^C$ such that $\{\tilde{y}_i^c \in \mathbb{R}^1 : 0 \leq \tilde{y}_i^c \leq 1\}$, along with a multi-dimensional output vector aggregated across labels $\tilde{y}_i^A \in \mathbb{R}^C$ such that $\{[\tilde{y}_i^A]^c \in \mathbb{R}^1 : 0 \leq [\tilde{y}_i^A]^c \leq 1\}$. The resultant mapping is given as $f_{MHO} : E_i^d \rightarrow [\tilde{y}_i^1], [\tilde{y}_i^2], \dots, [\tilde{y}_i^C], [\tilde{y}_i^A]$. A learnable weight is employed for each output variable $w_1, w_2, \dots, w_C, w_A \in \mathbb{R}^1$ to account for label co-occurrence. These weights are initialized based on the observed ratios of samples in the training set (i.e., $w_c = N / (C * N_c)$ for the c th class with N_c training samples; $w_A = 1/(C+1)$). The weights are incorporated in the loss terms for model training:

$$\text{Loss}(x_i, y_i) = \frac{1}{C} \sum_{c=1}^C \text{BCE}(y_i^c, w_c \tilde{y}_i^c) + \text{MLCE}(y_i, w_A \tilde{y}_i^A) + \text{CL}(\alpha ([w_1 \tilde{y}_i^1, \dots, w_C \tilde{y}_i^C]), \beta (w_A \tilde{y}_i^A)) \quad (5)$$

The proposed loss terms employ binary cross-entropy (BCE) loss terms for each label in conjunction with a multi-label cross-entropy (MLCE) loss term across labels to maintain sensitivity to both individual labels and their co-occurrence statistics. The first term in Eq. (5) denotes BCE loss expressed for each uni-dimensional output variable:

$$\text{BCE}(y, \tilde{y}) = -[y \log(\tilde{y}) + (1 - y) \log(1 - \tilde{y})] \quad (6)$$

where y and \tilde{y} denote binary scalars that reflect the true and predicted labels for a given class. The second term in Eq. (5) denotes MLCE loss expressed for the multi-dimensional output vector as:

$$\text{MLCE}(y^A, \tilde{y}^A) = \frac{1}{C} \sum_{c=1}^C (\text{BCE}([y^A]^c, [\tilde{y}^A]^c)) \quad (7)$$

where y^A and \tilde{y}^A denote binary vectors that reflect the true and predicted labels across C classes. The final term in Eq. (5) is consistency loss that enforces the predictions from the individual and aggregated output variables to be consistent with each other:

$$\text{CL}(\tilde{y}^{A,1}, \tilde{y}^{A,2}) = \|\tilde{y}^{A,1} - \tilde{y}^{A,2}\|_2 \quad (8)$$

where $\tilde{y}^{A,1}$ is formed by concatenating individual output variables across the label dimension and scaling the resultant vector by a factor of α , and $\tilde{y}^{A,2}$ is derived from the aggregated output vector via scaling by a factor of β . The scaling factors are taken as learnable parameters. The training procedures for HydraViT based on the loss given in Eq. (5) are described in Alg. 1. During inference on a test CXR image, x_{qi} , the predictions from the individual output variables, i.e., $[w_1 \tilde{y}_{qi}^1, \dots, w_C \tilde{y}_{qi}^C]$, are used to generate the class predictions.

Algorithm 1: Training procedure for HydraViT

Input: Dataset: $\{x_i, y_i\}_{i=1}^N$, x_i : CXR image, y_i : label
 f_{SE} : Spatial encoder with param. θ_{SE}
 f_{CE} : Context encoder with param. θ_{CE}
 $\tilde{y}_i^{1 \dots C}$: Individual output variables
 \tilde{y}_i^A : Aggregated output vector
 $Opt()$: Optimizer for computing param. updates
Output: ψ : $\{w_1, \dots, w_C, w_A, \alpha, \beta, \theta_{SE, CE}\}$
Initialize parameters.
for $i = 1:N$ **do**
 Compute $m_i, f_{SE}(\theta_{SE}) : x_i \rightarrow m_i$
 Compute $E_i, f_{CE}(\theta_{CE}) : m_i \rightarrow E_i$
 Project onto individual branches, $[\tilde{y}_i^1], [\tilde{y}_i^2], \dots, [\tilde{y}_i^C]$
 Project onto an aggregate branch, \tilde{y}_i^A
 Compute BCE for individual branches via Eq. (6)
 Compute MLCE for the aggregate branch via Eq. (7)
 Compute Loss based on Eq. (5)
 Update model parameters: $\psi \leftarrow \psi - Opt(\nabla_{\psi} \text{Loss})$
return ψ

4. Experimental setup

4.1. Dataset

Demonstrations were performed on the ChestX-ray14 dataset [62] with 112,120 frontal-view images from 30,805 unique patients with ages 1–95 years. Of these patients, 56.49% are male, and 43.51% are female. The dataset includes labels for 15 classes for each CXR image, including the ‘No Finding’ class for healthy individuals, and 14 different pathologies (atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia). Note that the average size of pathological regions is approximately 7.5% of the image size. The classes except for ‘No Finding’ can be simultaneously present in a given patient, yielding a multi-label classification problem. The ‘No Finding’ class accounts for 53.83% of the total dataset with 60,412 samples. Major pathological abnormalities such as ‘Infiltration’ and ‘Effusion’ have sample sizes of 19,894 and 13,317, respectively, while minor pathological abnormalities such as ‘Hernia’ and ‘Pneumonia’ have sample sizes of 227 and 1,431, respectively.

Prior to modeling, all CXR images were spatially downsampled to a 224×224 grid for computational efficiency. Data were split into training and test sets without any patient-level overlap, while preserving the ratios between the number of samples for separate classes. The training split contained 86,524 images, whereas the test split contained 25,596 images.

4.2. Implementation details

HydraViT leverages a hybrid network architecture with a spatial encoder module, a context encoder module and a multi-branch output layer. The spatial encoder module was implemented based on a pre-trained VGG16 architecture [63] with $z = 512$ and $r = 7$. This architecture consists of 13 convolutional layers with a 3×3 kernel size, ReLU layers, 5 max-pooling layers with a 2×2 kernel size, and 3 fully connected layers (FCL). The context encoder module used a pre-trained ViT architecture [64] with 12 transformer blocks with a projection dimension of 512, 20 attention heads, $P = 4$ resulting in

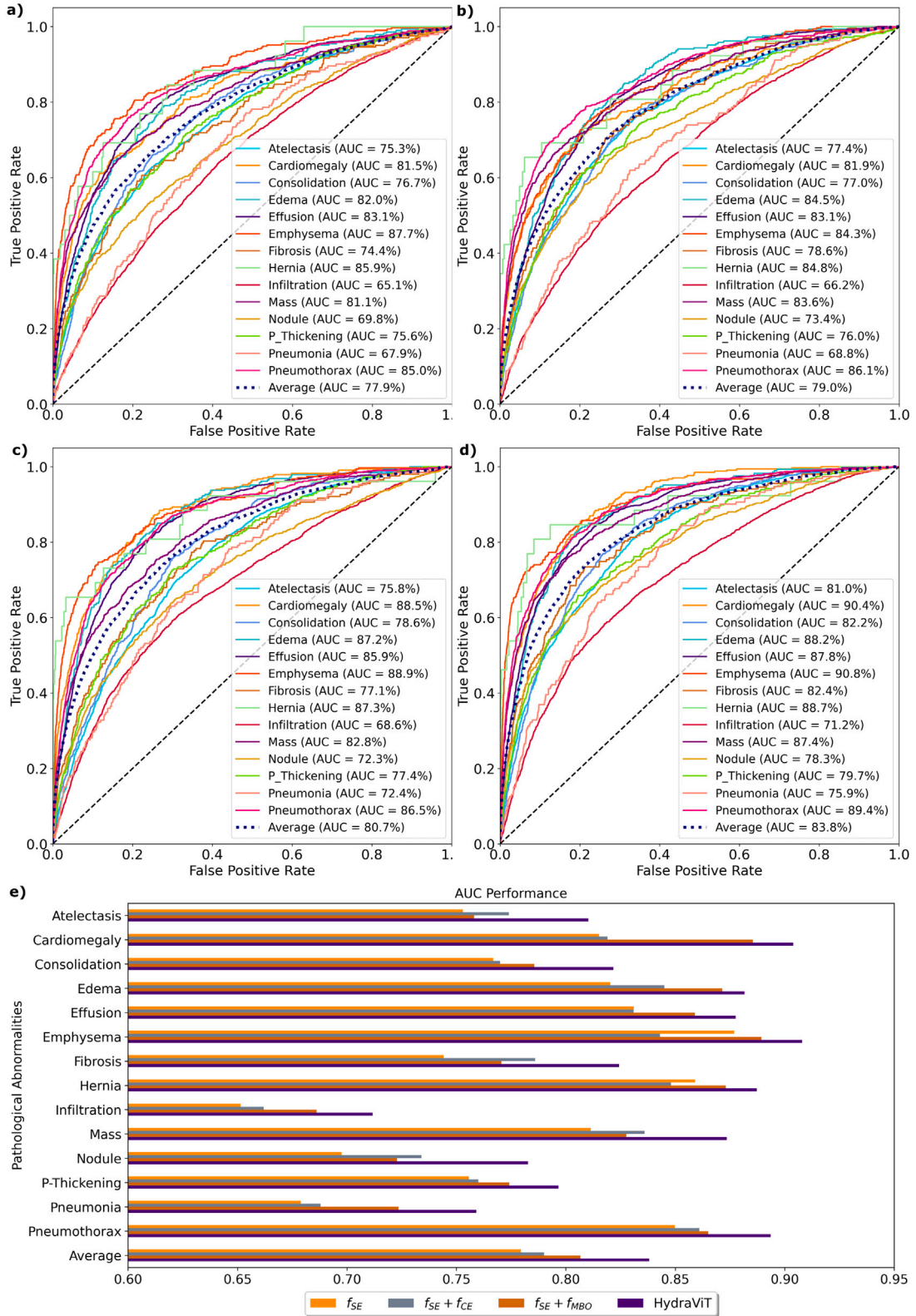


Fig. 2. Multi-label CXR classification performance of HydraViT and variant models. Results are shown for a variant model that contained only the SE module f_{SE} that was augmented with a softmax output layer for multi-label classification, a variant model that contained the SE and CE modules $f_{SE} + f_{CE}$ augmented with a softmax output layer, and a variant model that contained the SE and MBO modules $f_{SE} + f_{MBO}$. Class-wise and class-average ROC curves for (a) f_{SE} , (b) $f_{SE} + f_{CE}$, (c) $f_{SE} + f_{MBO}$, and (d) HydraViT; and (e) class-wise and class-average AUC metrics for all models.

$N_p = 4$, and $d = 25,088$. The w_A value in the multi-head output block was initialized as $1/(C + 1)$. The α and β parameters were initialized randomly in the range of $[0, 5]$. HydraViT was implemented using the

TensorFlow framework and executed on an NVidia RTX 3090 GPU. Models were trained via the Adam algorithm with a batch size of 35, a learning rate of 10^{-4} , and 120 epochs.

4.3. Competing methods

HydraViT was compared against several state-of-the-art deep-learning models for multi-label CXR classification. Three main groups of competing methods were considered: attention-guided, region-guided, and semantic-guided methods.

4.3.1. Attention-guided methods

PCAN: Pixel-wise classification and attention network (PCAN) [11] extracts mid-level CXR image features via a CNN, and uses pixel-wise branches for classification.

A³Net: Triple-attention learning (A³Net) [5] extracts features via DenseNet121, and uses channel, element, scale attention.

CBAtt: Class-based attention (CBAtt) [22] extracts features via ResNet50, and learns class-specific attention maps.

ConsultNet: ConsultNet [7] uses a two-branch architecture based on DenseNet121, spatial and channel attention to learn discriminative features.

DualAnet: Dual lesion attention network (DualAnet) [24] consists of two asymmetric attention networks based on DenseNet169 and ResNet152.

C-Tran: Classification transformer (C-Tran) [65] employs a transformer backbone for multi-label image classification by capturing intricate relationships between visual features and labels.

4.3.2. Region-guided methods

TSCN: Two-stream collaborative network (TSCN) [14] creates a segmentation mask via U-Net and performs feature extraction on the masked region via DenseNet169.

WSLM: Weakly supervised localization method (WSLM) [12] generates masks for pathology-containing regions and performs feature extraction via ResNet50.

RpSal: RpSal [42] extracts features via a pyramid network followed by region proposal and saliency detection for simultaneous localization and classification.

LLAGnet: Lesion location attention guided network (LLAGnet) [13] extracts features via DenseNet169, and uses weakly supervised attention to localize lesions in CXR images.

4.3.3. Semantic-guided methods

SEMM: SEMM [52] extracts semantic features via DenseNet121 that are split into three branches using multi-map transfer learning. Features are concatenated across branches following class-wise pooling.

CheXGCN: Label co-occurrence learning framework based on graph convolution networks (CheXGCN) [45] extracts features via DenseNet 169, and captures co-occurrence relationships via a GCN.

SSGE: Semantic similarity graph embedding (SSGE) [44] constructs a similarity graph from learned image features, and uses knowledge distillation to capture semantic similarities.

TNELF: Triple network ensemble learning framework (TNELF) [41] uses ensemble learning based on DenseNet169, ResNet50, and Efficient Net-B4 backbones, and performs feature-wise concatenation for classification.

4.4. Performance evaluation

Model performance was characterized via the Area Under the Curve (AUC) metric, which is commonly preferred for quantitative evaluation of multi-label CXR classification results. To do this, the area under the receiver operating characteristic (ROC) curve was first computed for each class. These areas were then averaged across classes. A higher AUC score indicates improved classification performance. Given a test set of CXR image $x_q = \{x_{q1}, x_{q2}, \dots, x_{qN}\}$, the AUC of a classification model for the c th class is computed as:

$$AUC = \frac{\sum_{qi=1}^N \sum_{qj=1}^N \xi(y_{qi}^c < y_{qj}^c) \xi(\tilde{y}_{qi}^c < \tilde{y}_{qj}^c)}{\sum_{qi=1}^N \sum_{qj=1}^N \xi(y_{qi}^c < y_{qj}^c)} \quad (9)$$

Table 1

Performance of HydraViT and ablated variant models are listed as average \pm std AUC and MAE across labels. Results are shown for the subset of test samples where only a single label is present, the subset of test samples where only multiple co-occurring labels are present, and all test samples.

		w/o f_{MBO}	w/o f_{CE}	w/o \tilde{y}_i^A	w/o init.	HydraViT
Aggregate AUC	Single	77.0 \pm 8.5	76.8 \pm 8.8	79.7 \pm 8.8	78.0 \pm 8.5	79.8 \pm 8.6
	Multiple	80.1 \pm 4.4	83.7 \pm 5.0	85.5 \pm 4.1	85.8 \pm 4.1	86.3 \pm 3.6
	All	79.0 \pm 6.0	80.7 \pm 6.3	83.3 \pm 5.9	82.8 \pm 6.0	83.8 \pm 5.8
Aggregate MAE	Single	0.086	0.090	0.084	0.080	0.084
	Multiple	0.083	0.077	0.071	0.075	0.069
	All	0.085	0.082	0.075	0.079	0.076

where $\xi(\cdot)$ denotes an indicator function for the condition expressed via its input argument. The AUC metric reflects the discriminative ability of a classification model by quantifying the proportion of cases in which the ranking of predicted labels is aligned with the ranking of true labels.

We also utilize the mean absolute error (MAE) metric to further analyze the performance of the proposed technique. The MAE of a classification model across classes is computed as:

$$MAE = \frac{1}{N} \sum_{qi=1}^N |y_{qi} - \tilde{y}_{qi}| \quad (10)$$

where y_{qi} is the vector of ground-truth labels across classes, and \tilde{y}_{qi} is the output vector across classes for the qi th test image. Note that MAE is reported as an aggregate performance measure across classes.

5. Results

5.1. Ablation studies

Several ablation studies were conducted on the ChestX-ray14 dataset to demonstrate the contribution of the individual components in HydraViT to method performance. First, we examined the effect of the spatial encoder, the context encoder, and the multi-branch output modules. For this purpose, HydraViT that comprises all three modules was compared against a variant that contained only the SE module f_{SE} that was augmented with a softmax output layer for multi-label classification, a variant that contained the SE and CE modules $f_{SE} + f_{CE}$ augmented with a softmax output layer, and a variant that contained the SE and MBO modules $f_{SE} + f_{MBO}$. Fig. 2 displays ROC curves and respective AUC metrics for the compared models, separately for each pathology label and on average across labels. On average across labels, HydraViT outperforms f_{SE} by 5.9%, $f_{SE} + f_{CE}$ by 4.8%, and $f_{SE} + f_{MBO}$ by 3.1% AUC. We also find that $f_{SE} + f_{MBO}$ consistently outperforms f_{SE} and that HydraViT consistently outperforms $f_{SE} + f_{CE}$ across labels. For these cases, the most notable improvements due to the introduction of the MBO module are observed for relatively rare labels such as ‘Cardiomegaly’, ‘Infiltration’, and ‘Pneumonia’ that often co-occur with other pathology. This finding indicates the importance of the MBO module over a conventional softmax classification layer in multi-label classification. We also observe that HydraViT consistently outperforms $f_{SE} + f_{MBO}$ across labels, with more notable improvements for labels such as ‘Nodule’, ‘Mass’, ‘Fibrosis’, and ‘Atelectasis’ where pathology can manifest with an atypical intensity distribution at both local and global scale. This result indicates the importance of the self-attention mechanism in CE to capture local and global contextual features of CXR images.

Next, we examined the benefits of the multi-task training of the MBO module in HydraViT for multi-label classification. HydraViT was compared against an aggregated variant that contained the SE and CE modules $f_{SE} + f_{CE}$ augmented with a softmax output layer based on an aggregated output vector, and an ensemble variant that utilized multiple $f_{SE} + f_{CE}$ models with uni-dimensional output variables trained separately for each individual label. Fig. 3 illustrates classification performance for the compared models. Among the variants, either the aggregated or the ensemble variant yields better performance

Table 2

Classification performance of competing methods on the ChestX-ray14 dataset. Results are shown for attention-guided, region-guided, and semantic-guided baselines along with HydraViT. AUC for each pathology label is listed on separate rows; aggregate AUC is given as average±std across labels; aggregate MAE is given as average across labels. Bold font marks the top performing method in each task.

	Attention-guided methods					Region-guided methods					Semantic-guided methods				
	PCAN	A ³ Net	CBAtt	ConsultNet	DualLanet	C-Tran	TSCN	WSLM	RpSal	LLAGnet	SEMM	CheXGCN	SSGE	TNELF	HydraViT
Atelectasis	79.1	77.9	79.0	79.7	78.3	80.2	78.5	79.0	77.5	78.3	79.2	78.6	79.2	78.8	81.0
Cardiomegaly	88.7	89.5	91.0	90.9	88.4	90.4	88.7	91.0	88.1	88.5	88.1	89.3	89.2	87.5	90.4
Consolidation	75.9	75.9	76.0	77.9	74.6	82.0	75.4	74.0	74.7	75.4	76.0	75.1	75.3	75.6	82.2
Edema	85.4	85.5	86.0	85.8	84.1	87.1	84.9	86.0	84.6	85.1	84.8	85.0	84.8	85.4	88.2
Effusion	84.1	83.6	83.0	84.8	83.2	84.6	83.1	84.0	83.1	83.4	84.1	83.2	84.0	83.7	87.8
Emphysema	94.4	93.3	93.0	92.9	93.7	93.1	93.0	95.0	93.6	93.9	92.2	94.4	94.8	93.4	90.8
Fibrosis	81.9	83.8	82.0	83.4	82.0	82.6	83.3	84.0	83.3	83.2	83.3	83.4	84.0	85.1	84.5
Hernia	88.6	89.8	87.0	88.3	89.0	88.4	88.6	88.0	89.1	89.5	89.0	88.8	88.0	89.0	90.8
Infiltration	73.2	72.8	74.0	75.1	72.1	74.0	72.2	73.0	73.1	72.5	73.0	72.0	73.0	72.9	75.5
Mass	82.4	81.0	82.0	81.5	82.0	83.4	82.3	81.0	81.2	82.6	82.0	82.0	82.4	82.8	84.0
Nodule	78.4	78.5	79.0	78.4	78.6	79.4	78.2	79.0	78.6	78.8	79.1	79.2	79.0	78.7	80.0
Pleural Thickening	81.6	81.8	82.0	81.9	81.0	82.1	81.3	82.0	81.0	81.7	82.1	81.5	82.0	81.4	83.0
Pneumonia	74.3	74.5	73.0	75.0	74.0	74.1	73.6	73.0	74.0	74.2	74.0	74.3	74.0	74.0	75.8
Pneumothorax	85.3	86.0	86.0	85.8	85.0	86.0	85.7	85.0	85.2	85.4	85.3	85.6	85.2	85.4	87.6
Aggregate AUC	82.2 ± 5.8	82.1 ± 5.5	82.1 ± 5.9	82.6 ± 5.8	81.4 ± 5.4	83.3 ± 5.8	81.9 ± 5.6	82.2 ± 6.0	81.8 ± 5.5	81.9 ± 5.5	81.9 ± 5.5	82.2 ± 5.8	82.2 ± 5.8	82.1 ± 5.5	84.1 ± 5.4
Aggregate MAE	0.078	0.082	0.078	0.078	0.080	0.079	0.083	0.081	0.082	0.081	0.079	0.077	0.081	0.084	0.075

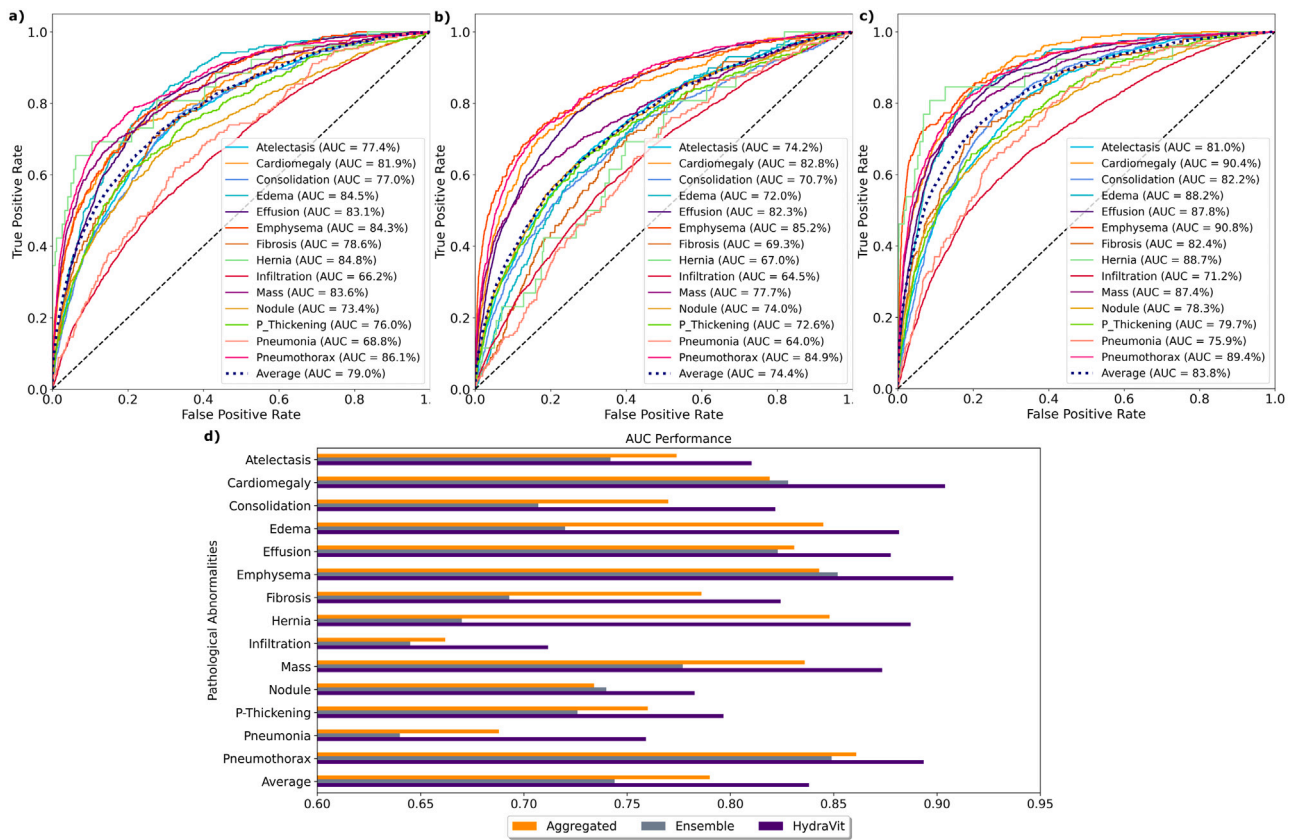


Fig. 3. Multi-label CXR classification performance of HydraViT and variant models. Results are shown for a variant model that trained a single $f_{SE} + f_{CE}$ architecture with an aggregated output vector across labels, and a variant model that ensembles separate $f_{SE} + f_{CE}$ architectures trained for each individual label. Class-wise and class-average ROC curves for (a) the aggregated variant, (b) the ensemble variant, and (c) HydraViT; and (d) class-wise and class-average AUC metrics for all models.

in some labels, and the two variants perform similarly in remaining labels. That said, HydraViT consistently outperforms the two variants across all labels. On average, HydraViT outperforms the aggregated variant by 4.8% suggesting that multi-task training based on separate output branches helps improve sensitivity to individual labels. HydraViT also outperforms the ensemble variant by 9.4% suggesting that the aggregated output branch in HydraViT helps improve capture of co-occurrence relationships among pathology.

We also assessed the benefits of the MBO module, CE module, aggregated output vector and MBO weight initialization in HydraViT on classification performance for single versus multiple labels. A variant that excluded the MBO module from HydraViT (w/o f_{MBO}), a variant that excluded the CE module (w/o f_{CE}), a variant that excluded the aggregated output vector and the associated multi-label cross-entropy

and consistency loss terms (w/o \tilde{y}^A), and a variant with zero initialization of weights in the MBO module (w/o $init.$) were considered. Table 1 lists the classification performance of the compared models on the subset of test samples that contain only a single label, on the subset of the test sample that contains multiple co-occurring labels, and on the entire test set. We find that HydraViT outperforms the variants in all cases, albeit performance benefits are more notable for the multiple label case. In the single-label case, HydraViT outperforms w/o f_{MBO} by 2.8% AUC and 2.3% MAE, w/o f_{CE} by 3.0% AUC and 6.7% MAE, w/o \tilde{y}^A by 0.1% AUC (albeit yields on par MAE), and w/o $init.$ by 1.8% AUC (albeit yields lower MAE). In the multiple-label case, HydraViT outperforms w/o f_{MBO} by 6.2% AUC and 16.9% MAE, w/o f_{CE} by 2.6% AUC and 10.4% MAE, w/o \tilde{y}^A by 0.8% AUC and 2.8% MAE, and w/o $init.$ by 0.5% AUC and 8.0% MAE. Across the entire validation set, HydraViT outperforms w/o f_{MBO} by 4.8% AUC and

Table 3

The number of model parameters, training time per image, and inference time per image for HydraViT variants.

	w/o f_{MBO}	w/o f_{CE}	w/o ($f_{MBO}+f_{CE}$)	HydraViT
Parameters	409 M	17.8 M	17.1 M	409.7 M
Training time	273 ms	160 ms	156 ms	277 ms
Inference time	7.6 ms	5.0 ms	4.9 ms	7.9 ms

10.6% MAE, w/o f_{CE} by 3.1% AUC and 7.3% MAE, w/o \tilde{y}^A by 0.5% AUC (albeit yields on par MAE), and w/o $init.$ by 1.0% AUC and 3.8% MAE. Note also that HydraViT yields the lowest standard deviation in AUC across labels, indicating an improvement in homogeneity of classification performance across distinct pathology.

A practical concern regarding CXR classification methods pertains to the computational burden encountered during model training and inference. Table 3 lists the number of model parameters, the training time per image and the inference time per image for HydraViT and its ablated variants. As expected, the CE module that contains transformer components carries a notably higher proportion of model parameters in comparison to the MBO module that is relatively light weight. Analogously, the CE module also has a relatively higher influence on the training and inference times.

5.2. Comparison studies

We comparatively demonstrated the performance of HydraViT in multi-label CXR classification against several state-of-the-art methods including attention-guided (PCAN [11], A³Net [5], CBAAtt [22], ConsultNet [7], DualAnet [24], C-Tran [65]), region-guided (TSCN [14], WSLM [12], RpSal [42], LLAGnet [13]), and semantic guided models (SEMM [52], CheXGCN [45], SSGE [44], TNELF [41]). Table 2 lists the classification performance of the competing models separately for each pathology label, and on average across labels. In terms of pathology labels, the performance improvements offered by HydraViT are most notable for ‘Atelectasis’, ‘Consolidation’, ‘Edema’, ‘Effusion’, ‘Mass’, ‘Pneumonia’, ‘Pneumothorax’ that can show a relatively broad spatial distribution across CXR images. HydraViT yields comparable performance to most baselines for labels such as ‘P-Thickening’, ‘Infiltration’, ‘Fibrosis’, and ‘Cardiomegaly’. Meanwhile, several baselines can yield higher performance than HydraViT for ‘Emphysema’, ‘Hernia’, ‘Nodule’. On average, HydraViT outperforms competing attention-guided methods by 1.9% AUC and 5.3% MAE, region-guided methods by 2.1% AUC and 8.3% MAE, and semantic-guided methods by 2.0% AUC and 6.5% MAE. Furthermore, HydraViT also achieves the lowest standard deviation in AUC across labels among competing methods, indicating an improvement in homogeneity of classification performance across distinct pathology. These findings suggest that HydraViT improves performance and reliability in multi-label CXR classification.

Finally, we examined the performance of HydraViT in multi-label classification qualitatively by inspecting the predicted labels in representative CXR images. Fig. 4 displays the CXR images, corresponding heatmaps extracted via the GradCAM method [66] that highlight salient regions relevant to pathology, and the scores for top-5 predicted labels. The ground-truth labels are annotated in green font. We observe that the top-ranked labels by HydraViT are closely aligned with the ground-truth labels. In each case, the average score estimated by HydraViT for the ground-truth labels is significantly higher than the average score for non-present labels within the list of top-5 (e.g., 0.4 versus 0.1, 0.7 versus 0.1, 0.6 versus 0.1, 0.6 versus 0.2 for the representative samples presented in Fig. 4). These results indicate that HydraViT yields an accurate estimation of pathology in multi-label CXR classification.

6. Discussion and conclusion

In this study, we proposed a novel deep learning method to improve performance in multi-label CXR classification of thoracic diseases. The proposed HydraViT model uses a hybrid convolutional-transformer backbone to extract contextualized embeddings of CXR images, and a multi-branch output module with adaptive weights to improve capture of co-occurring pathology. While branched output modules have been previously considered for multi-task learning problems in the machine learning literature, to our knowledge, HydraViT is the first method to devise a multi-branch architecture that comprises output heads for individual and aggregated labels in multi-label classification.

A set of ablation studies were conducted to demonstrate the contribution of individual design elements in HydraViT. These studies indicate that the introduction of the transformer-based context encoder helps significantly boost classification performance. They also indicate that the multi-branch output module in HydraViT yields elevated performance over training separate network models for each label and training a single model with only separate heads for individual labels. Note that there is also a computational benefit for training a single multi-branch architecture as in HydraViT, which is nearly 14 times faster compared to sequential training of single-branch architectures for each pathology label separately.

HydraViT was comparatively demonstrated against state-of-the-art deep learning methods for multi-label CXR classification. Attention-guided, region-guided, and semantic-guided baselines were considered. While there were occasional cases where a competing baseline yielded comparable or higher scores for 1–3 pathology labels out of 14, HydraViT generally outperformed baselines in the majority of labels. Across all labels, HydraViT yielded the highest average performance in multi-label classification. The success of HydraViT in identifying the presence of multiple pathologies was also corroborated via visual inspections. Therefore, our results suggest that HydraViT is a promising approach for CXR-based classification of pathology in thoracic diseases.

Here, HydraViT was demonstrated on the ChestX-ray14 dataset that contains over hundred thousand CXR images and associated pathology labels across a diverse patient cohort. Naturally, the reliability of data-driven deep-learning classifiers depend critically on the use of such large, diverse training sets that accurately reflect the anatomical variability and co-occurrence statistics of pathology. When trained on limited CXR datasets with intrinsic biases in the data distribution, multi-label classification models might suffer from poor generalization performance. A potential remedy that can help enhance generalizability under limited-data settings could be to use generative modeling approaches that synthesize CXR data with a high degree of anatomical and label variability [51]. Future work is warranted to examine the performance of HydraViT on other datasets possessing differences in the CXR data distribution in order to comprehensively evaluate generalization capabilities.

Despite its promising performance, several practical challenges might hamper the rapid development and clinical adoption of HydraViT. To facilitate deployment of the CXR classification model by lowering computational demands, here we implemented HydraViT based on a hybrid CNN-transformer architecture. The proposed architecture improves computational efficiency by using light-weight convolutional modules at relatively high spatial resolutions to extract compact representations of input images, and by using parameter-dense transformer modules on these representations to extract contextual embeddings [67]. Still, transformer-based architectures elevate training and inference times compared to simpler models (e.g., based solely on a CNN), which can introduce challenges for users operating under resource-constrained settings such as rural or underfunded healthcare facilities. In such cases, knowledge distillation methods could be employed to transfer the information captured by a pre-trained HydraViT model onto more compact models that can be run under limited compute resources [68]. Hybrid CNN-transformer architectures have

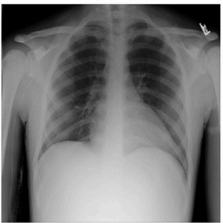

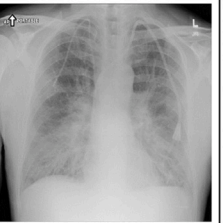
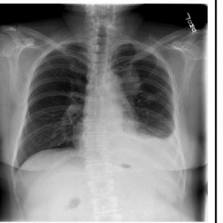
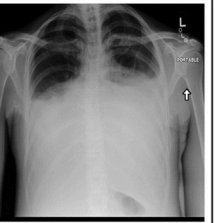
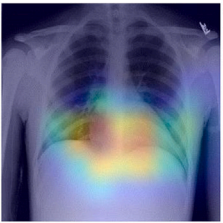

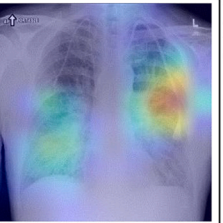

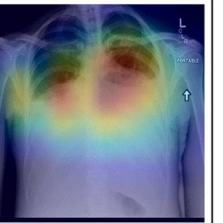
Images					
Heatmaps					
Scores	Pneumothorax : 0.5175 Emphysema : 0.4428 Infiltration : 0.2278 Effusion : 0.1163 P-Thickening : 0.0711	Pneumothorax : 0.7220 Infiltration : 0.1819 P-Thickening : 0.1180 Effusion : 0.0795 Nodule : 0.0426	Edema : 0.5828 Infiltration : 0.5342 Pneumonia : 0.1617 Nodule : 0.1114 Mass : 0.0956	Mass : 0.6619 Effusion : 0.5988 Atelectasis : 0.1701 Infiltration : 0.1662 P-Thickening : 0.1141	Effusion : 0.7430 Atelectasis : 0.1426 Infiltration : 0.1248 Mass : 0.0744 Nodule : 0.0617

Fig. 4. Representative CXR images from the ChestX-ray14 dataset, and respective multi-label predictions generated by HydraViT. The top-5 predicted pathologies and probability scores are listed, and the ground truth labels are marked in green. GradCAM-derived heatmaps for the CXR images are also given to highlight pathological regions.

become pervasive in recent years, and the backbones in HydraViT were based on commonly available architectures in the literature. Furthermore, the multi-branch output module used learnable weights that do not require any tuning. Yet, a degree of expertise could still be helpful in tuning of optimization parameters in conjunction with specific architectural choices. For users with limited expertise, establishment of pre-trained models that manifest extensive transfer learning capabilities might be a potential path to facilitate adoption of advanced models such as HydraViT [37,50]. Meanwhile, since transformers rely on relatively complex, non-local attention mechanisms, they may also face additional challenges in terms of interpreting classification decisions as an important component of model validation. Recent explanatory techniques devised specifically for transformer-based architectures might help alleviate this problem by building user trust [68,69].

Several technical limitations can be addressed in future work to further improve the performance of HydraViT. Sensitivity for long-range context can be further boosted by adopting a pure transformer architecture at the expense of elevated model complexity. In those cases, low-rank approximations on the self-attention matrix or pyramidal transformer architectures with limited attention windows can help improve efficiency [69]. Note that the pathology annotations for the CXR dataset analyzed in this study were mined from radiological reports via language models, so they contain an inherent level of noise [70]. Such noisy levels can introduce a degree of bias in trained models that take annotations as ground truth. Learning procedures that take into account the possibility of erroneous labels can help boost classification performance. Here, a context encoder module equipped with self-attention mechanisms was used to enable the model to focus on pathological regions in CXR images. For improved localization, anomaly detection on CXR images based on generative approaches such as diffusion models could be utilized [48,49]. Finally, sensitivity for co-occurrence relationships can be enhanced by adopting focal modulation networks or state space sequences models instead of transformers with self-attention filtering [71,72].

HydraViT contains a multi-branch output (MBO) module to improve reliability in multi-label classification. Loss functions computed based on the outputs of this module assign learnable weights to different labels so as to balance their contributions, and maintain a degree of reliability against potential class imbalances. That said, we acknowledge

that HydraViT can still show susceptibility to class imbalance particularly in the context of rare pathology, where scarce representation of such pathology in the training data can prohibit adequate learning regardless of loss term weightings. It might be possible to address this challenge by performing discrimination among labels after embedding them in a dedicated latent space according to their semantic similarities with the intent to mitigate apparent biases due to a one-hot representation. In this manner, the intrinsic relationships between different labels could be leveraged rather than relying solely on their frequency within the training dataset [73,74]. Another limitation of the MBO module concerns its growing complexity with the number of distinct pathology labels, which might necessitate prolonged training on relatively larger datasets in order to accurately capture label co-occurrence statistics. In such cases, a hierarchical classification model can be built following known disease taxonomies, and a separate HydraViT instance can be used to run the decisions at each stage of the hierarchy [28].

In sum, here we introduced a novel multi-label classification model for characterizing thoracic diseases from CXR images. Naturally, the spatial encoder, context encoder and MBO modules in HydraViT were designed to optimize performance for the X-ray modality. The general architecture of HydraViT can be adopted for multi-label classification tasks in other modalities such as MRI or CT assuming that adjustments are performed to cope with varying image dimensionality. Yet, it remains to be demonstrated whether the specific backbone choices in HydraViT modules are preferable in other modalities given differences in the distribution of images and pathology labels.

CRediT authorship contribution statement

Şaban Öztürk: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **M. Yiğit Turalı:** Visualization, Software, Methodology, Conceptualization. **Tolga Çukur:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Saban Ozturk reports financial support was provided by TUBITAK. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by a TUBITAK 118C543 grant awarded to Şaban Öztürk.

References

- [1] A. Saha, P. Dickinson, R. Shrimali, A. Salem, S. Agarwal, Is thoracic radiotherapy an absolute contraindication for treatment of lung cancer patients with interstitial lung disease? A systematic review, *Clin. Oncol.* 34 (12) (2022) e493–e504.
- [2] T. Gupte, A. Knack, J.D. Cramer, Mortality from aspiration pneumonia: incidence, trends, and risk factors, *Dysphagia* 37 (6) (2022) 1493–1500.
- [3] M.E. Kruk, A.D. Gage, N.T. Joseph, G. Danaei, S. García-Saisó, J.A. Salomon, Mortality due to low-quality health systems in the universal health coverage era: a systematic analysis of amenable deaths in 137 countries, *Lancet* 392 (10160) (2018) 2203–2212.
- [4] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, Y. Yang, Thorax disease classification with attention guided convolutional neural network 131, 2020, pp. 38–45.
- [5] H. Wang, S. Wang, Z. Qin, Y. Zhang, R. Li, Y. Xia, Triple attention learning for classification of 14 thoracic diseases using chest radiography, 67, 2021, 101846.
- [6] J. Su, Z. Luo, S. Li, Consistent response for automated multilabel thoracic disease classification, *Concurr. Comput.* 34 (23) (2022) e7201.
- [7] Q. Guan, Y. Huang, Y. Luo, P. Liu, M. Xu, Y. Yang, Discriminative feature learning for thorax disease classification in chest X-ray images, 30, 2021, pp. 2476–2487, <http://dx.doi.org/10.1109/TIP.2021.3052711>.
- [8] H.E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M.E. Maros, T. Ganslandt, Transfer learning for medical image classification: A literature review, *BMC Med. Imaging* 22 (1) (2022) 69.
- [9] P. Celard, E. Iglesias, J. Sorribes-Fdez, R. Romero, A.S. Vieira, L. Borrajo, A survey on deep learning applied to medical images: From simple artificial neural networks to generative models, *Neural Comput. App.* 35 (3) (2023) 2291–2323.
- [10] H. Wang, Y. Xia, Domain-ensemble learning with cross-domain mixup for thoracic disease classification in unseen domains, *Biomed. Signal Process. Control* 81 (2023) 104488.
- [11] X. Zhu, S. Pang, X. Zhang, J. Huang, L. Zhao, K. Tang, Q. Feng, PCAN: Pixel-wise classification and attention network for thoracic disease classification and weakly supervised localization, *Comput. Med. Imaging Graph.* 102 (2022) 102137.
- [12] H.-G. Jung, W.-J. Nam, H.-W. Kim, S.-W. Lee, Weakly supervised thoracic disease localization via disease masks, *Neurocomput* 517 (2023) 34–43.
- [13] B. Chen, J. Li, G. Lu, D. Zhang, Lesion location attention guided network for multi-label thoracic disease classification in chest X-rays, 24(7), 2019, pp. 2016–2027.
- [14] B. Chen, Z. Zhang, J. Lin, Y. Chen, G. Lu, Two-stream collaborative network for multi-label chest X-ray image classification with lung segmentation, 135, 2020, pp. 221–227.
- [15] M. Hossain, S. Hossain, M. Zunaed, T. Hasan, et al., A novel attention mechanism using anatomical prior probability maps for thoracic disease classification from X-Ray images, 2022, [arXiv:2210.02998](https://arxiv.org/abs/2210.02998).
- [16] M.S. Lee, S.W. Han, DuETNet: Dual encoder based transfer network for thoracic disease classification, 161, 2022, pp. 143–153.
- [17] F. Li, L. Zhou, Y. Wang, C. Chen, S. Yang, F. Shan, L. Liu, Modeling long-range dependencies for weakly supervised disease classification and localization on chest X-ray, *Quant. Imaging Med. Surg.* 12 (6) (2022) 3364.
- [18] B. Cheng, I. Misra, A.G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 1280–1289.
- [19] C. Zhao, S. Xiang, Y. Wang, Z. Cai, J. Shen, S. Zhou, D. Zhao, W. Su, S. Guo, S. Li, Context-aware network fusing transformer and V-Net for semi-supervised segmentation of 3D left atrium, *Expert Syst. Appl.* 214 (2023) 119105.
- [20] S. Xiang, N. Li, Y. Wang, S. Zhou, J. Wei, S. Li, Automatic delineation of the 3D left atrium from LGE-mri: Actor-critic based detection and semi-supervised segmentation, *IEEE J. Biomed. Health Inf.* 28 (6) (2024) 3545–3556.
- [21] H. Wang, H. Jia, L. Lu, Y. Xia, Thorax-net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography, 24(2), 2019, pp. 475–485.
- [22] D. Sriker, H. Greenspan, J. Goldberger, Class-Based Attention Mechanism for Chest Radiograph Multi-Label Categorization, *IEEE*, 2022, pp. 1–5.
- [23] K. Chen, X. Wang, S. Zhang, Thorax disease classification based on pyramidal convolution shuffle attention neural network, *IEEE Access* 10 (2022) 85571–85581.
- [24] V. Teixeira, L. Braz, H. Pedrini, Z. Dias, Dualnet: dual lesion attention network for thoracic disease classification in chest X-rays, in: IWSSIP, *IEEE*, 2020, pp. 69–74.
- [25] B. Jung, L. Gu, T. Harada, Graph interaction for automated diagnosis of thoracic disease using x-ray images, in: SPIE Med. Imaging, 12032, SPIE, 2022, pp. 135–147.
- [26] H. Wang, Y.-Y. Yang, Y. Pan, P. Han, Z.-X. Li, H.-G. Huang, S.-Z. Zhu, Detecting thoracic diseases via representation learning with adaptive sampling, *Neurocomput* 406 (2020) 354–360.
- [27] S. Kabir, L. Farrokhar, A. Dabouei, A weakly supervised approach for thoracic diseases detection, *Expert Syst. Appl.* 213 (2023) 118942.
- [28] H. Chen, S. Miao, D. Xu, G.D. Hager, A.P. Harrison, Deep hierarchical multi-label classification applied to chest X-ray abnormality taxonomies, *Med. Image Anal.* 66 (2020) 101811.
- [29] 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, *Neurocomput* 437 (2021) 186–194.
- [30] R.T. Mullanpudi, W.R. Mark, N. Shazeer, K. Fatahalian, Hydranets: Specialized dynamic architectures for efficient inference, 2018, pp. 8080–8089.
- [31] B. Velasco, J. Cerquides, J.L. Arcos, Hydranet: A neural network for the estimation of multi-valued treatment effects, in: *NeurIPS 2022 Workshop on Causality for Real-World Impact*.
- [32] A. Kumar, Y.-Y. Wang, K.-C. Liu, I.-C. Tsai, C.-C. Huang, N. Hung, Distinguishing normal and pulmonary edema chest x-ray using gabor filter and SVM, in: *IEEE ISBB 2014*, *IEEE*, 2014, pp. 1–4.
- [33] F.H.O. Alfdhli, A.A. Mand, M.S. Sayeed, K.S. Sim, M. Al-Shabi, Classification of tuberculosis with SURF spatial pyramid features, in: *ICORAS, IEEE*, 2017, pp. 1–5.
- [34] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017, [arXiv:1711.05225](https://arxiv.org/abs/1711.05225).
- [35] J. Rubin, D. Sanghavi, C. Zhao, K. Lee, A. Qadir, M. Xu-Wilson, Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks, 2018, [arXiv:1804.07839](https://arxiv.org/abs/1804.07839).
- [36] R. Dharmesh Ishwerlari, R. Agarwal, K. Sujatha, Lung disease classification using chest x ray image: An optimal ensemble of classification with hybrid training, *Biomed. Signal Process. Control* 91 (2024) 105941.
- [37] L. Yao, E. Poblentz, D. Dagunts, B. Covington, D. Bernard, K. Lyman, Learning to diagnose from scratch by exploiting dependencies among labels, 2017, [arXiv:1710.10501](https://arxiv.org/abs/1710.10501).
- [38] G. Eswara Rao, R. B., P.N. Srinivasu, M.F. Ijaz, M. Woźniak, Hybrid framework for respiratory lung diseases detection based on classical CNN and quantum classifiers from chest X-rays, *Biomed. Signal Process. Control* 88 (2024) 105567.
- [39] U. Hasanah, C. Avian, J.T. Darmawan, N. Bachroin, M. Faisal, S.W. Prakosa, J.-S. Leu, C.-T. Tsai, CheXNet and feature pyramid network: a fusion deep learning architecture for multilabel chest X-Ray clinical diagnoses classification, *Int. J. Cardiovascular Imag.* 40 (4) (2024) 709–722.
- [40] S.U. Amin, S. Taj, A. Hussain, S. Seo, An automated chest X-ray analysis for COVID-19, tuberculosis, and pneumonia employing ensemble learning approach, *Biomed. Signal Process. Control* 87 (2024) 105408.
- [41] M. Yang, H. Tanaka, T. Ishida, Performance improvement in multi-label thoracic abnormality classification of chest X-rays with noisy labels, *Int. J. Comput. Assist. Radiol. Surg.* 18 (1) (2023) 181–189.
- [42] R. Hermoza, G. Maicas, J.C. Nascimento, G. Carneiro, Region proposals for saliency map refinement for weakly-supervised disease localisation and classification, 2020, pp. 539–549.
- [43] H. Malik, T. Anees, Multi-modal deep learning methods for classification of chest diseases using different medical imaging and cough sounds, *Plos one* 19 (3) (2024) e0296352.
- [44] B. Chen, Z. Zhang, Y. Li, G. Lu, D. Zhang, Multi-label chest X-Ray image classification via semantic similarity graph embedding, *IEEE Trans. Circuits Syst. Video Technol.* 32 (4) (2022) 2455–2468.
- [45] B. Chen, J. Li, G. Lu, H. Yu, D. Zhang, Label co-occurrence learning with graph convolutional networks for multi-label chest X-Ray image classification, 24(8), 2020, pp. 2292–2302, <http://dx.doi.org/10.1109/JBHI.2020.2967084>.
- [46] V. Parthasarathy, S. Saravanan, Chaotic sea horse optimization with deep learning model for lung disease pneumonia detection and classification on chest X-ray images, *Multimedia Tools Appl.* (2024) 1–23.
- [47] Z. Ali, M.A. Khan, A. Hamza, A.I. Alzahrani, N. Alalwan, M. Shabaz, F. Khan, A deep learning-based x-ray imaging diagnosis system for classification of tuberculosis, COVID-19, and pneumonia traits using evolutionary algorithm, *Int. J. Imaging Syst. Technol.* 34 (1) (2024) e23014.

- [48] A. Güngör, S.U. Dar, Şaban. Öztürk, Y. Korkmaz, H.A. Bedel, G. Elmas, M. Özbey, T. Çukur, Adaptive diffusion priors for accelerated MRI reconstruction, *Med. Image Anal.* (2023) 102872.
- [49] M. Özbey, O. Dalmaz, S.U. Dar, H.A. Bedel, Ş. Öztürk, A. Güngör, T. Çukur, Unsupervised medical image translation with adversarial diffusion models, *IEEE Trans. Med. Imaging* 42 (12) (2023) 3524–3539.
- [50] P. Chambon, C. Bluethgen, C.P. Langlotz, A. Chaudhari, Adapting pretrained vision-language foundational models to medical imaging domains, 2022, [arXiv:2210.04133](https://arxiv.org/abs/2210.04133).
- [51] K. Packhäuser, L. Folle, F. Thamm, A. Maier, Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems, 2022, [arXiv:2211.01323](https://arxiv.org/abs/2211.01323).
- [52] C. Yan, J. Yao, R. Li, Z. Xu, J. Huang, Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays, in: *ACM Int Conf Bioinf Comput Biol Health Inf*, 2018, pp. 103–110.
- [53] S.E. Sorour, A.A. Wafa, A.A. Abohany, R.M. Hussien, A deep learning system for detecting cardiomegaly disease based on cxr image, *Int. J. Intell. Syst.* 2024 (1) (2024) 8997093.
- [54] B. Chen, Y. Lu, G. Lu, Multi-label chest X-ray image classification via label co-occurrence learning, in: *PRCV*, Springer, 2019, pp. 682–693.
- [55] A.I. Aviles-Rivero, N. Papadakis, R. Li, P. Sellars, Q. Fan, R.T. Tan, C.-B. Schönlieb, GraphX^{small} NET-NET-Chest X-Ray Classification Under Extreme Minimal Supervision, Springer, 2019, pp. 504–512.
- [56] J. Zhao, K. Yan, Y. Zhao, X. Guo, F. Huang, J. Li, Transformer-based dual relation graph for multi-label image recognition, in: *2021 IEEE/CVF International Conference on Computer Vision, ICCV*, 2021, pp. 163–172.
- [57] K. Zhang, W. Liang, P. Cao, X. Liu, J. Yang, O. Zaiane, Label correlation guided discriminative label feature learning for multi-label chest image classification, *Comput. Methods Programs Biomed.* 245 (2024) 108032.
- [58] B. Zhou, Y. Li, J. Wang, A weakly supervised adaptive densenet for classifying thoracic diseases and identifying abnormalities, 2018, [arXiv:1807.01257](https://arxiv.org/abs/1807.01257).
- [59] Z. Ge, D. Mahapatra, S. Sedai, R. Garnavi, R. Chakravorty, Chest x-rays classification: A multi-label and fine-grained problem, 2018, [arXiv:1807.07247](https://arxiv.org/abs/1807.07247).
- [60] R. López-González, J. Sánchez-García, B. Fos-Guarinos, F. García-Castro, Á. Alberich-Bayarri, E. Soria-Olivas, C. Muñoz-Núñez, L. Martí-Bonmatí, Automated chest radiographs triage reading by a deep learning referee network, 2021, pp. 1–9, [medRxiv](https://medrxiv.org/).
- [61] Y. Korkmaz, T. Çukur, V.M. Patel, Self-supervised MRI reconstruction with unrolled diffusion models, in: *MICCAI*, 2023, pp. 491–501.
- [62] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-Ray8: Hospital-scale chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, 2017, pp. 3462–3471, [http://dx.doi.org/10.1109/CVPR.2017.369](https://doi.org/10.1109/CVPR.2017.369).
- [63] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- [64] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020, [http://dx.doi.org/10.48550/arXiv.2010.11929](https://dx.doi.org/10.48550/arXiv.2010.11929), [arXiv e-prints, arXiv:2010.11929](https://arxiv.org/abs/2010.11929).
- [65] J. Lanchantin, T. Wang, V. Ordonez, Y. Qi, General multi-label image classification with transformers, in: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2021, pp. 16473–16483.
- [66] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, 2017, pp. 618–626, [http://dx.doi.org/10.1109/ICCV.2017.74](https://dx.doi.org/10.1109/ICCV.2017.74).
- [67] O. Dalmaz, M. Yurt, T. Çukur, ResViT: Residual vision transformers for multi-modal medical image synthesis, 41(10), 2022, pp. 2598–2614, [http://dx.doi.org/10.1109/TMI.2022.3167808](https://dx.doi.org/10.1109/TMI.2022.3167808).
- [68] H.A. Bedel, T. Çukur, DreaMR: Diffusion-driven counterfactual explanation for functional MRI, 2023, [arXiv:2307.09547](https://arxiv.org/abs/2307.09547).
- [69] H.A. Bedel, I. Sivgin, O. Dalmaz, S.U. Dar, T. Çukur, BoIT: Fused window transformers for fMRI time series analysis, *Med. Image Anal.* 88 (2023) 102841.
- [70] A. Majkowska, S. Mittal, D.F. Steiner, J.J. Reicher, S.M. McKinney, G.E. Duggan, K. Eswaran, P.-H. Cameron Chen, Y. Liu, S.R. Kalidindi, A. Ding, G.S. Corrado, D. Tse, S. Shetty, Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation, *Radiology* 294 (2) (2020) 421–431.
- [71] J. Yang, C. Li, X. Dai, J. Gao, Focal modulation networks, 35, 2022, pp. 4203–4217.
- [72] O.F. Atli, B. Kabas, F. Arslan, M. Yurt, O. Dalmaz, T. Çukur, I2I-Mamba: Multi-modal medical image synthesis via selective state space modeling, 2024, [arXiv:2405.14022](https://arxiv.org/abs/2405.14022).
- [73] Ş. Öztürk, T. Çukur, Deep clustering via center-oriented margin free-triplet loss for skin lesion detection in highly imbalanced datasets, 26, 2022, pp. 4679–4690, [http://dx.doi.org/10.1109/JBHI.2022.3187215](https://dx.doi.org/10.1109/JBHI.2022.3187215).
- [74] Ş. Öztürk, E. Çelik, T. Çukur, Content-based medical image retrieval with opponent class adaptive margin loss, 2023, 118938.