Learning Low-Rank Feature for Thorax Disease Classification

Anonymous authors
Paper under double-blind review

Abstract

In the field of medical imaging, deep neural networks such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) have demonstrated remarkable achievements. In this paper, we focus on classifying thorax diseases based on radiographic images. The key to the success of classification involves effectively extracting features from disease-impacted areas in radiographic images. Although various neural network architectures and training methods, including self-supervised learning through contrastive/restorative techniques, have been utilized for such classification tasks, there remains a lack of systematic approaches to mitigate the negative impacts of noise and non-disease elements in the images. To tackle this issue, we introduce a new Low-Rank Feature Learning (LRFL) technique in this study, which can be implemented in the training processes of different neural networks. The LRFL approach is both empirically inspired by a Low Frequency Property (LFP) and theoretically supported by our precise generalization bounds for neural networks using low-rank features. Notably, LFP is prevalent not only in deep neural networks across general machine learning applications but also across all thorax medical datasets examined in this study. In our empirical evaluation, the LRFL method, when applied to a ViT or CNN that has been pre-trained on unlabeled chest X-rays using Masked Autoencoders (MAE), outperforms existing methods in terms of multi-class area under the receiver operating curve (mAUC) and classification accuracy. The code is available at https://anonymous.4open.science/r/medical_projects-BBFE/.

1 Introduction

Following the huge success of deep learning, recent studies have developed deep neural networks (DNNs) for various tasks in medical imaging, such as disease classification and abnormality detection in chest X-rays (Guendel et al., 2018; Xiao et al., 2023). Accurate clinical decision-making with DNNs heavily relies on learning informative medical feature representation. Early works adopt convolutional neural networks (CNNs) such as U-Net (Ronneberger et al., 2015) for representation learning on radiographic images. Recently, Vision Transformers (ViTs) (Dosovitskiy et al., 2020) are also adopted to learn informative medical representations from radiographic images (Xiao et al., 2023), utilizing their capabilities in capturing long-range feature dependencies. Albeit the success of CNNs and ViTs in analyzing radiographic images, their accuracy heavily relies on the quality and quantity of data and annotations (Feng et al., 2020). However, the collection of large amounts of training data and high-quality annotations in the medical imaging domain is extremely hard (Xiao et al., 2023). To tackle this problem, self-supervised learning (SSL) has been employed as a solution for acquiring representations from unlabeled data. Given the greater availability of unlabeled medical images (Azizi et al., 2022), SSL proves to be an efficient approach for obtaining discriminative representations. SSL employs a range of pretext tasks to acquire transferable representations without manual annotations. Over recent years, numerous variations of self-supervised learning have surfaced using contrastive learning (Chen et al., 2020c) and restorative learning (Xiao et al., 2023).

Building upon the advancements in deep learning, recent studies have pushed forward the development of deep neural networks (DNNs) for applications in medical imaging, such as disease classification and the detection of abnormalities in chest X-rays (Guendel et al., 2018; Xiao et al., 2023). The accuracy of clinical decisions made using DNNs primarily hinges on their capacity to learn robust medical feature representations.

Pioneering efforts utilized convolutional neural networks (CNNs), like U-Net (Ronneberger et al., 2015), to foster representation learning from radiographic images. Lately, Vision Transformers (ViTs) (Dosovitskiy et al., 2020) have also been employed to harvest informative medical representations from these images (Xiao et al., 2023), leveraging their proficiency in handling long-range dependencies among features. While CNNs and ViTs have demonstrated success in processing radiographic images, their effectiveness largely depends on the quality and volume of the available data and annotations (Feng et al., 2020). However, collecting a large dataset of high-quality annotations in medical domains is notably challenging (Xiao et al., 2023). To overcome this issue, self-supervised learning (SSL) has been utilized to procure representations from unlabeled data. Given the increased accessibility of unlabeled medical images (Azizi et al., 2022), SSL has been established as an effective approach for learning discriminative representations. It incorporates a series of pretext tasks designed to learn transferable features without relying on manual annotations. Over the years, a myriad of self-supervised learning methods have emerged, utilizing contrastive learning (Chen et al., 2020c) and restorative learning (Xiao et al., 2023) to enhance the learning process.

Challenges in the Current Literature for Disease Classification. In this paper, we focus on the classification of thorax diseases. Clinical studies indicate that the disease-affected areas in radiographic images are often subtle, showing localized variations, and these challenges are compounded by the pervasive noise present in radiographic imaging as discussed in Section 2.1. It is vital to effectively and robustly extract features from these disease areas for accurate disease classification on radiographic images. While a variety of neural architectures like CNNs and ViTs, along with diverse training methods including self-supervised learning using contrastive and restorative learning, have been applied to this task, there still lacks a principled approach to effectively mitigate the impact of noise and non-disease background on the classification of diseases in radiographic images.

Our Contributions. The contributions of this paper are presented as follows.

First, to address the challenges highlighted earlier, we introduce a novel Low-Rank Feature Learning (LRFL) method, which is adaptable across the training of different neural networks for thorax disease classification. The LRFL strategy utilizes low-rank features to classify diseases. The adoption of low-rank features is inspired by a Low Frequency Property (LFP), as illustrated in Figure 1. LFP suggests that the low-rank projection of the ground truth training labels captures most of the essential information. In fact, LFP is commonly observed in various classification scenarios utilizing deep neural networks, such as (Rahaman et al., 2019; Arora et al., 2019; Cao et al., 2021; Choraria et al., 2022). Inspired by LFP, LRFL integrates the truncated nuclear norm (TNN) as a low-rank regularization term into the training loss of the neural network, promoting the use of low-rank features for classification. Since the features relevant for classification are predominantly low-rank, the high-rank features, which carry most of the noise and background information, are significantly reduced, thereby diminishing their impact on the learning process. *Importantly and significantly different from existing low-rank learning methods reviewed in Section 2.3, we introduce a novel separable approximation for the TNN, enabling the optimization of the LRFL training loss using standard SGD.* Results in Table 4 show that our LRFL method achieves 7×-10× acceleration in the training process compared to the existing augmented Lagrange multiplier based method (Lee & Lam, 2016) for optimizing the TNN. The appropriate feature ranks retained in the LRFL method across various datasets are determined through an efficient cross-validation process, and the optimal ranks are detailed in Table 7 in Section A.2 of the appendix. Extensive experimental results demonstrate that our LRFL method renders new record mAUC on three standard thorax disease datasets, NIH-ChestX-ray (Wang et al., 2017), COVIDx (Pavlova et al., 2022), and CheXpert (Irvin et al., 2019), surpassing the current state-of-the-art (Xiao et al., 2023) with the same pre-training setup.

Second, we provide a theoretical analysis showing a sharp generalization bound for the LRFL method, underscoring the substantial benefits of employing low-rank regularization within LRFL. Given these theoretical insights and the versatility of LRFL across various neural networks, we anticipate broader applications of LRFL in the classification of other diseases beyond thoracic ones, potentially enhancing classification tasks across different radiographic imaging contexts. It is worthwhile to mention that the literature has studied low-rank learning using TNN resembling LRFL as to be reviewed in Section 2.3. Our LRFL method builds upon these foundational principles by incorporating low-rank regularization into the training of neural networks, aiming to improve thorax disease classification by reducing the adverse effects of noise and

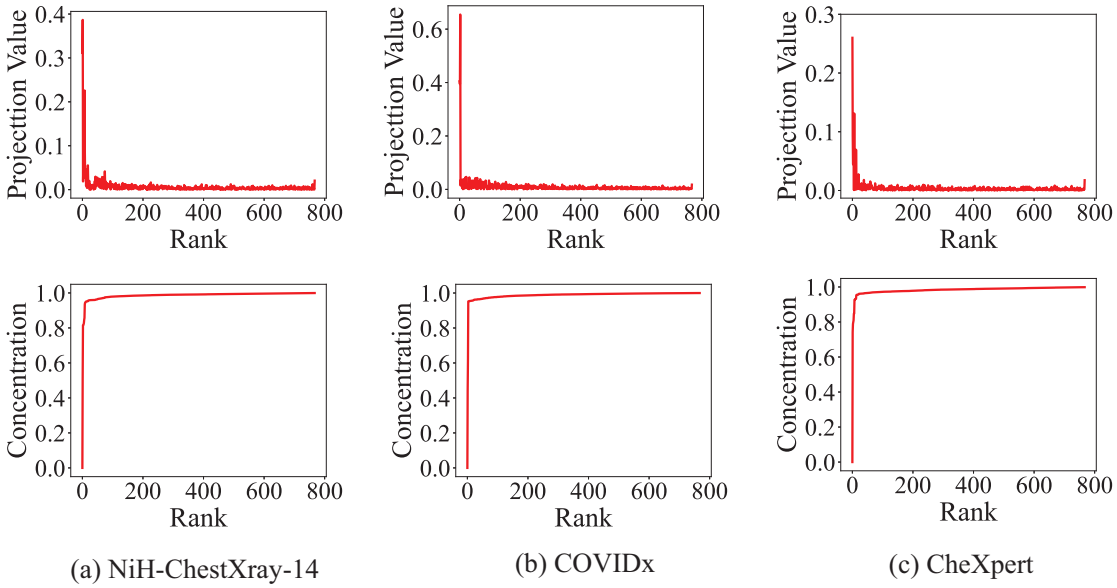


Figure 1: Eigen-projection (first row) and signal concentration ratio (second row) of Vit-Base/16 on NiH-ChestXray-14, COVIDx, and CheXpert. To compute the eigen-projection, we first calculate the eigenvectors $\mathbf{U}$ of the kernel gram matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ computed by a feature matrix $\mathbf{F} \in \mathbb{R}^{n \times d}$, then the projection value is computed by $\mathbf{p} = \frac{1}{C} \sum_{c=1}^{C} \left\| \mathbf{U}^\top \mathbf{Y}^{(c)} \right\|_2^2 / \left\| \mathbf{Y}^{(c)} \right\|_2^2 \in \mathbb{R}^n$, where C is the number of classes, and $\mathbf{Y} \in \{0,1\}^{n \times C}$ is the one-hot labels of all the training data, $\mathbf{Y}^{(c)}$ is the c-th column of $\mathbf{Y}$. The eigen-projection $\mathbf{p}_r$ for $r \in [\min(n,d)]$ reflects the amount of the signal projected onto the r-th eigenvector of $\mathbf{K}$, and the signal concentration ratio of a rank r reflects the proportion of signal projected onto the top r eigenvectors of $\mathbf{K}$. The signal concentration ratio for rank r is computed by $\left\| \mathbf{p}^{(1:r)} \right\|_2$, where $\mathbf{p}^{(1:r)}$ contains the first r elements of $\mathbf{p}$. For example, by the rank $r = 38$, the signal concentration ratio of $\mathbf{Y}$ on NIH ChestX-ray14, COVIDx, and CheXpert are 0.959, 0.964, and 0.962 respectively.

irrelevant background information. **Different from the conventional low-rank learning methods, our approach introduces a separable approximation to the TNN, facilitating the optimization process and enhancing the generalization ability of the model**. Such improved generalization is evidenced by the improved prediction accuracy of LRFL compared to the current state-of-the-art (SOTA) methods in medical image analysis.

We also present ablation study results evidencing our contributions. In particular, in Section 4.5, we perform ablation study evidencing the effectiveness of LRFL in reducing the adverse effects of background for thorax disease classification. We also compare the eigenvalues of the kernels and the kernel complexity (Bartlett et al., 2005; Koltchinskii, 2006; Mendelson, 2002) associated with the LRFL models and the corresponding base models in Section A.3.1 of the appendix, and the lower kernel complexity of the LRFL models suggests their lower generalization error (Bartlett et al., 2005; Koltchinskii, 2006; Mendelson, 2002).

Notations. We use bold letters to denote matrices or vectors. $[\mathbf{A}]_i$ stands for the i-th row of a matrix $\mathbf{A}$. $\|\cdot\|_p$ denotes the p-norm of a vector or a matrix. $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm of a matrix. We use $[m \dots n]$ to indicate numbers between m and n inclusively, and $[n]$ denotes the natural numbers between 1 and n inclusively.

2 Related Works

2.1 Radiographic Imaging

Radiographic imaging plays a fundamental role in medical image analysis (Li et al., 2023). Unlike photographic images which exhibit a variety of backgrounds (Deng et al., 2009), radiographic images maintain

consistent backgrounds due to standardized imaging protocols (Zhou, 2021; Li et al., 2022; Shamshad et al., 2022; Xiao et al., 2023). The clinical details are dispersed throughout these images, with regions indicating disease displaying localized variations, adding complexity to their analysis (Xiao et al., 2023; Suetens, 2017; Zhou et al., 2022c). Noise is an inherent feature of radiographic images, arising from several sources such as quantum fluctuations, electronic noise, scatter radiation, motion blur, and overlapping anatomical structures (Siewerdsen et al., 1997; 1998; Manson et al., 2019; Chandra & Verma, 2020). Studies in (Goyal et al., 2018; Hussain & Hyeon Gu, 2024) show that inevitable noise exists in radiographic images and can affect disease detection on them. Particularly, quantum noise, which results from statistical variations in the X-ray photons detected, is often a major concern (Shung et al., 2012b;a; Suetens, 2017; Chandra & Verma, 2020). This type of noise creates a grainy appearance that obscures fine details and reduces image contrast (Shung et al., 2012b). The intensity of quantum noise varies with factors like the X-ray dose, the sensitivity of the detector, and the thickness of the examined object. Although quantum noise is fundamentally modeled as a Poisson process (Suetens, 2017; Chandra & Verma, 2020), at high photon counts, it approximates a Gaussian distribution, which facilitates the application of various noise reduction strategies (Lee et al., 2018; Ding et al., 2018).

2.2 Medical Image Analysis with Deep Learning

Deep learning has achieved impressive advancements in photographic image analysis (He et al., 2016; Lin et al., 2017b;a), leading to a surge of interest in its application within the medical imaging domain due to its capacity to learn discriminative representations. Convolutional Neural Networks (CNNs) such as U-Net (Ronneberger et al., 2015; Falk et al., 2018; Zhou et al., 2018) have been trailblazers in the medical domain, setting new benchmarks across a range of tasks including image classification (Shen & Gao, 2018; Wang et al., 2019; Ma et al., 2020), object detection (Falk et al., 2019; Zhou et al., 2018; Yang & Yu, 2021), and semantic segmentation (Yang & Yu, 2021; Yao et al., 2021; Zhou et al., 2018; Simpson et al., 2019; Sourati et al., 2019). Additionally, methods like Recurrent Neural Networks (RNNs) (Zhou et al., 2019a; Gao et al., 2019) and Reinforcement Learning (RL) techniques (Zhou et al., 2021; Xu et al., 2022; Hu et al., 2023) have been explored for their potential. More recently, vision transformers, inspired by the efficacy of transformers in natural language processing (Vaswani et al., 2017), have surpassed traditional CNNs in various computer vision benchmarks (Yuan et al., 2021; Dosovitskiy et al., 2020; Liu et al., 2021; Zhu et al., 2021; Cai et al., 2023). While there are ongoing discussions regarding the generalization, data requirements, and computational efficiency of transformers versus CNNs (Liu et al., 2022b; Zhou et al., 2022b; Bao et al., 2021; Xiao et al., 2022; Touvron et al., 2021; Ding et al., 2022; Bai et al., 2021; Mao et al., 2022; Zhang et al., 2022; Zhou et al., 2022a; Dosovitskiy et al., 2020; Steiner et al., 2021; Tay et al., 2022; Paul & Chen, 2022), transformers have demonstrated significant promise in the field of medical image analysis (Xiao et al., 2023; Chen et al., 2021a;b), benefiting from the self-attention mechanism that adeptly models long-range dependencies unlike the local convolutions of CNNs (Li et al., 2023). With the limited availability of high-quality annotations, self-supervised contrastive learning approaches (Chen et al., 2020c;b; Grill et al., 2020; Caron et al., 2020; Xiao et al., 2023) have become popular for pre-training networks in medical imaging (Zhou, 2021; Xiao et al., 2023; Chen et al., 2021a). However, the high uniformity in radiographic images due to standardized imaging protocols (Xiang et al., 2021; Haghighi et al., 2022) introduces unique challenges, distinct from those in photographic imaging (He et al., 2020; Chen et al., 2020c). To combat this, recent initiatives employ restorative strategies such as masked autoencoders (MAE) (Alex et al., 2017; Chen et al., 2019; Zhou et al., 2019b; Zhu et al., 2020; Chen et al., 2020a; Xie et al., 2022; Xiao et al., 2023; He et al., 2022) for network pre-training (Xiao et al., 2023). In line with recent developments, we also utilize MAE (Xiao et al., 2023) to pre-train our networks prior to engaging in low-rank feature learning.

2.3 Low-Rank Learning

Low-rank learning has garnered significant attention across various fields for its capacity to reduce dimensionality, suppress noise, and enhance feature extraction. Robust Principal Component Analysis (RPCA) (Candès et al., 2011) serves as a cornerstone in this realm, efficiently separating data matrices into low-rank and sparse components. This technique proves invaluable for vision-related tasks such as image denoising and background subtraction. Building on this foundation, (Yang & Cohen, 2015) introduced singular value

pruning, a method to impose low-rank constraints on neural network layers, thereby boosting both computational efficiency and performance. The concept of TNN regularization (TNNR) has been further refined by researchers like (Hu et al., 2013), who noted that TNNR more accurately approximates the rank function by selectively minimizing singular values, essential for precise low-rank matrix recovery in noisy conditions. Following that, some existing works (Lee & Lam, 2016; Hu et al., 2015; Zhang et al., 2017) propose to perform low-rank feature learning by minimizing the TNN of the feature matrix. Additionally, the use of TNNR in tensor completion has markedly improved the restoration of incomplete visual data, utilizing tensor singular value decomposition (t-SVD) (Liu et al., 2017; Zhang et al., 2020). More contemporary learning-based methods, such as those developed by (Indyk et al., 2019), have optimized low-rank approximations through targeted training, enhancing practical application outcomes. Furthermore, recent studies (Gao et al., 2021; Lu et al., 2015; Ren et al., 2022) also demonstrate that learning low-rank features can significantly enhance the robustness of deep neural networks against noise in input images.

3 Formulation

3.1 Pipeline for Thorax Disease Classification

We utilize the masked MAE technique (He et al., 2022) for the initial pre-training of both CNNs and ViTs following(Xiao et al., 2023), and subsequently fine-tune the pre-trained networks with our Low-Rank Feature Learning (LRFL). The full training pipeline of learning low-rank features for disease classification can be described in three steps. In the first step, which is the **pre-training** step, we pre-train the networks using the self-supervised restorative learning method, masked MAE (He et al., 2022), on a diverse pre-training dataset that includes ImageNet-1k (Krizhevsky et al., 2012) and a collection of X-rays (0.5M) (Xiao et al., 2023). During this phase, we randomly mask patches on input images and drive the networks to optimize pixel-wise image reconstruction for the obscured patches. In the second step, which is the **regular fine-tuning** step, we fine-tune the pre-trained networks employing cross-entropy loss aimed at image classification on specific target datasets, namely NIH-ChestX-ray (Wang et al., 2017), COVIDx (Pavlova et al., 2022), and CheXpert (Irvin et al., 2019). In the last step, which is the **low-rank feature learning** step, we fix the backbones of the networks and fine-tune the linear classifier utilizing our novel LRFL method.

3.2 Problem Setup for LRFL

We now introduce the problem setup for LRFL with training details. Suppose the training data are given as $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ where $\mathbf{x}_i$ and $\mathbf{y}_i \in \mathbb{R}^C$ are the i-th training data point and its corresponding class label vector respectively, and C is the number of classes. Each element $\mathbf{y}_i$ is binary with $[\mathbf{y}_i]_j = 1$ indicating the j-th disease is present in $\mathbf{x}_i$, otherwise $[\mathbf{y}_i]_j = 0$ for $j \in [C]$. Suppose that the neural network trained by step two of our pipeline in Section 3.1 generates a feature vector $f_{\mathbf{W}_1(0)}(\mathbf{x}) \in \mathbb{R}^d$ (the output of the layer preceding the final linear/softmax layer of the network) for any input x, and $f_{\mathbf{W}'}(\cdot)$ is the feature extraction function with $\mathbf{W}'$ being the weights of the feature extraction backbone of the network. Let $\mathbf{W}_1(0)$ denote the weights of feature extraction backbone by step two of the pipeline. We then train a linear neural network by optimizing

$$\min_{\mathbf{W}} L(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}\left(\mathbf{y}_i, \sigma\left(\mathbf{W} f_{\mathbf{W}_1(0)}(\mathbf{x})\right)\right) \tag{1}$$

for the low-rank feature learning step, where $\mathbf{W} \in \mathbb{R}^{C \times d}$ is the weight matrix for the network. Here $\tilde{\sigma}$ is an element-wise sigmoid function, $\tilde{\sigma}(\boldsymbol{a}) \in \mathbb{R}^C$ with $[\tilde{\sigma}(\boldsymbol{a})]_c = 1/(1 + \exp(-\boldsymbol{a}_c))$ for $\boldsymbol{a} \in \mathbb{R}^C$ and $c \in [C]$. KL stands for the element-wise binary cross-entropy function. Given two nonnegative vectors $\mathbf{u} = [u_1, \ldots, u_d] \in \mathbb{R}^d, \mathbf{v} = [v_1, \ldots, v_d] \in \mathbb{R}^d$ where $u_i \in \{0, 1\}$ for all $i \in [d]$ and $\|\mathbf{v}\|_\infty \leq 1$, $\mathrm{KL}(\mathbf{u}, \mathbf{v}) := \sum_{j=1}^d -u_i \log v_i - (1 - u_i) \log(1 - v_i)$ for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. We use $\mathbf{Y} = \left[\mathbf{y}_1^\top; \mathbf{y}_2^\top; \ldots; \mathbf{y}_n^\top\right] \in \mathbb{R}^{n \times C}$ to denote the training label matrix by stacking the label vectors of all the training data. Let the mapping function of the linear neural network used in the loss function $L(\mathbf{W})$ be $\mathrm{NN}_{\mathbf{W}}(\mathbf{x}) = \mathbf{W} f_{\mathbf{W}_1(0)}(\mathbf{x})$.

Motivation for Low-Rank Regularization The Low Frequency Property is illustrated in Figure 1, that is, the low-rank projection of the ground truth class labels possesses the majority of the information of

the class labels. Inspired by this observation, our LRFL encourages the low-rank part of the feature to participate in the classification process. In this way, the noise and non-disease areas in the high-rank part of the feature are mostly not learned by LRFL so as to improve the classification accuracy. Using notations in Section 3.2, the truncated nuclear norm (TNN) of $\mathbf{F}$ is $\|\mathbf{F}\|_T \coloneqq \sum_{i=T+1}^{d} \sigma_i$ where $T \in [0, d]$ and we use the convention that $\sum_{i=d'}^{d} \cdot = 0$ for $d' > d$. It will be shown by the generalization error bound to be discussed in Section 3.3 that a smaller $\|\mathbf{F}\|_T$ renders a tighter upper bound for the generalization error of the linear neural network used for LRFL. This observation gives a strong theoretical motivation for us to add the TNN $\|\mathbf{F}\|_T$ to the training loss $L(\mathbf{W})$ so as to reduce $\|\mathbf{F}\|_T$.

3.3 Generalization Bound for Low-Rank Feature Learning

We define the loss function $\ell(\mathrm{NN}_{\mathbf{W}}(\mathbf{x}), \mathbf{y}) \coloneqq \|\mathrm{NN}_{\mathbf{W}}(\mathbf{x}) - \mathbf{y}\|_2^2$, and the generalization error of the network NN is the expected risk of the loss ℓ, which is denoted by $L_{\mathcal{D}}(\mathrm{NN}_{\mathbf{W}}) \coloneqq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell(\mathrm{NN}_{\mathbf{W}}(\mathbf{x}), \mathbf{y})\right]$, with $\mathcal{D}$ being the distribution of the data $\mathbf{x}$ and its class label $\mathbf{y}$. The network $\mathrm{NN}_{\mathbf{W}}$ generates a feature $\mathbf{F} \in \mathbb{R}^{n\times d}$ of all the training data with $\mathbf{F}_i = f_{\mathbf{W}_1(0)}^{\top}(\mathbf{x}_i)$ for $i \in [n]$. The kernel gram matrix for the feature $\mathbf{F}$ is $\mathbf{K}_n = \frac{1}{n}\mathbf{F}\mathbf{F}^{\top}$. We let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots \geq \widehat{\lambda}_{\bar{r}} > 0$ be the eigenvalues of $\mathbf{K}_n$ where $\bar{r} \leq \min\{n, d\}$ is the rank of $\mathbf{K}_n$. Suppose the Singular Value Decomposition of $\mathbf{F}$ is $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$, where $\mathbf{U} \in \mathbb{R}^{n\times d}$ has orthogonal columns, $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with diagonal elements being the singular values of $\mathbf{F}$, and $\mathbf{V} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. The columns of $\mathbf{U}$ and $\mathbf{V}$ are also called the left eigenvectors and the right eigenvectors of $\mathbf{F}$ respectively. Let $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_d$ be the singular values of $\mathbf{F}$, and $\bar{\mathbf{Y}} = \mathbf{U}^{(\bar{r})}{\mathbf{U}^{(\bar{r})}}^{\top}\mathbf{Y}$ be the projection of the training label matrix $\mathbf{Y}$ onto the subspace spanned by the top-$\bar{r}$ left eigenvectors of $\mathbf{F}$, where $\mathbf{U}^{(\bar{r})} \in \mathbb{R}^{n\times\bar{r}}$ is formed by the top $\bar{r}$ eigenvectors in $\mathbf{U}$. Then we have the following theorem giving the sharp generalization error bound for the linear neural network $\mathrm{NN}_{\mathbf{W}}$ in (1). The proof this theorem is deferred to Section B of the appendix.

Theorem 3.1. For every $x > 0$, with probability at least $1 - \exp(-x)$, after the t-th iteration of gradient descent for all $t \geq 1$, we have

$$L_{\mathcal{D}}(\mathrm{NN}_{\mathbf{W}}) \leq c_1\left\|\mathbf{Y} - \bar{\mathbf{Y}}\right\|_{\mathrm{F}}^2 + c_1\left(1 - \eta\widehat{\lambda}_r\right)^{2t}\|\mathbf{Y}\|_{\mathrm{F}}^2 + c_2 \min_{h\in[0,r]}\left(\frac{h}{n} + \sqrt{\frac{1}{n}\sum_{i=h+1}^{r}\widehat{\lambda}_i}\right) + \frac{c_3 x}{n}, \tag{2}$$

where c_1, c_2, c_3 are positive constants.

Remark 3.2. The RHS of (2) is the generalization error bound for the linear neural network used in LRFL as step three of the pipeline in Section 3.1. Moreover, let $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_d$ be the singular values of $\mathbf{F}$. Due to the fact that $\sqrt{\frac{1}{n}\sum_{i=h+1}^{r}\widehat{\lambda}_i} \leq \frac{1}{n}\sum_{i=h+1}^{r}\sigma_i$, it follows by (2) that

$$L_{\mathcal{D}}(\mathrm{NN}_{\mathbf{W}}) \leq c_1\left\|\mathbf{Y} - \bar{\mathbf{Y}}\right\|_{\mathrm{F}}^2 + c_1\left(1 - \eta\widehat{\lambda}_r\right)^{2t}\|\mathbf{Y}\|_{\mathrm{F}}^2 + c_2\left(\frac{h}{n} + \frac{1}{n}\sum_{i=T+1}^{d}\sigma_i\right) + \frac{c_3 x}{n}, \tag{3}$$

which holds for all $T \in [0, d]$. (3) motivates the reduction of the TNN of the feature $\mathbf{F}$, as detailed in the next subsection.

3.4 Optimization of the TNN in SGD

The TNN $\|\mathbf{F}\|_T$ is not separable, so the training loss with $\|\mathbf{F}\|_T$ cannot be directly optimized by the standard SGD. To address this problem, we propose an approximation $\overline{\|\mathbf{F}\|_T}$ to $\|\mathbf{F}\|_T$ which is separable so that $\overline{\|\mathbf{F}\|_T}$ can be optimized by standard SGD.

First, we note that if $\mathbf{U}, \mathbf{V}$ are known, then $\mathbf{\Sigma} = \mathbf{U}^{\top}\mathbf{F}\mathbf{V}$. If we have an approximation $\overline{\mathbf{U}}$ to $\mathbf{U}$ and an approximation $\overline{\mathbf{V}}$ to $\mathbf{V}$, then $\mathbf{\Sigma}$ can be approximated by $\overline{\mathbf{\Sigma}} = \overline{\mathbf{U}}^{\top}\mathbf{F}\overline{\mathbf{V}}$. As a result, the approximation $\overline{\|\mathbf{F}\|_T}$ to the TNN is $\overline{\|\mathbf{F}\|_T} = \sum_{i=1}^{n}\left(\sum_{s=T+1}^{d}\sum_{k=1}^{d}\overline{\mathbf{U}}_{si}^{\top}\mathbf{F}_{ik}\overline{\mathbf{V}}_{ks}\right)$. Due to the above discussions, the loss function

of LRFL with the approximate TNN $\overline{\|\mathbf{F}\|_T}$ is $\mathcal{L}_{\mathrm{LRFL}}(\mathbf{W}) = \frac{1}{m}\sum_{v_i \in \mathcal{V}_{\mathcal{L}}} \mathrm{KL}\left(\mathbf{y}_i, [\sigma(\mathbf{F}\mathbf{W})]_i\right) + \eta\overline{\|\mathbf{F}\|_T}$, which is separable, so that it can be trained by the standard SGD. $\eta > 0$ is the weighting parameter for the TNN. Because $\mathcal{L}_{\mathrm{LRFL}}(\mathbf{W})$ is to be optimized by the standard SGD, we have the loss function of LRFL for the j-th minibatch $\mathcal{B}_j \subseteq [n]$ as

$$\mathcal{L}_j(\mathbf{W}) = \frac{1}{|\mathcal{B}_j|}\sum_{i \in \mathcal{B}_j} \mathrm{KL}\left(\mathbf{y}_i, [\sigma(\mathbf{F}\mathbf{W})]_i\right) + \frac{\eta}{|\mathcal{B}_j|}\sum_{i \in \mathcal{B}_j}\left(\sum_{s=T+1}^{d}\sum_{k=1}^{d}\overline{\mathbf{U}}_{si}^{\top}\mathbf{F}_{ik}\overline{\mathbf{V}}_{ks}\right). \tag{4}$$

The approximation $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ can be computed as the left and right eigenvectors of the feature $\mathbf{F}$ computed at earlier epochs. In order to save computation and avoiding performing SVD for $\mathbf{F}$ at every epoch, we propose to update $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ only after certain epochs. Algorithm 1 describes the training algorithm for the neural network trained with LRFL, which uses the standard SGD to optimize the loss function $\mathcal{L}_{\mathrm{LRFL}}(\mathbf{W})$, as step three of our pipeline in Section 3.1. Before the first epoch, we compute $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ as the left and right eigenvectors of the feature $\mathbf{F}$ at the initialization of the neural network. After every t_0 epochs for t_0 being a constant integer, we update $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ as the left and right eigenvectors of the feature $\mathbf{F}$ produced by the neural network right after t_0-th epoch, with t_0 being a constant integer.

Algorithm 1 Training Algorithm with the Approximate TNN by SGD

```
Initialize the weights W_1 by W_1 = W_1(0), and initialize W_2 randomly
Compute feature F by the neural network, and its SVD as F = UΣV
Update Ū = U, V̄ = V
for t = 1, 2, ..., t_max do
    if t ≡ 0 (mod t_0) then
        Compute feature F of the neural network, and its SVD F = UΣV.
        Update Ū = U, V̄ = V
    end if
    for b = 1, 2, ..., B do
        Update W by applying gradient descent on batch B_j ⊆ [n] using the gradient of the loss L_j in Eq.(4)
    end for
end for
return The trained weights W of the network
```

4 Experimental Results

In this section, we conduct experiments on medical datasets to demonstrate the effectiveness of the proposed LRFL method. The experiments section is organized as follows. In Section 4.1, we discuss our experimental setup and implementation details. In Sections 4.2, Section 4.3, and Section 4.4, we evaluate LRFL models for thorax disease classification on CheXpert, NIH ChestX-ray 14, and COVIDx, respectively. Comprehensive ablation studies on LRFL are performed in Section 4.5. Additional experimental results are deferred to Section A of the appendix. In Section A.1, we evaluate the training time of LRFL models compared with the baseline models. In Section A.2, we show the cross-validation results of hyper-parameters for different models and different datasets. Additional ablation study results are presented in Section A.3 of the appendix.

4.1 Implementation Details

In this section, we assess the effectiveness of the proposed Low-Rank Feature Learning (LRFL) method for classifying thoracic diseases. We employ networks that were previously trained on ImageNet (Russakovsky et al., 2015) or chest X-rays detailed in (Xiao et al., 2023) using MAE, a self-supervised strategy that involves reconstructing absent pixels from masked regions of the input images. When ImageNet-1k and X-rays (0.5M) are used for the pre-training of models in our paper, all the images will be reshaped to $224 \times 224 \times 3$ following the settings in (Xiao et al., 2023). We subsequently fine-tune these pre-trained networks employing low-rank regularization on three publicly available X-ray datasets: (1) NIH ChestX-ray14 (Wang et al., 2017), (2) Stanford CheXpert (Irvin et al., 2019), and (3) COVIDx (Pavlova et al., 2022). Fine-tuning involves using

the ADAM optimizer, with a batch size set at 1024 for all datasets. The fine-tuning process consists of two phases. Initially, we fine-tune the entire network for 75 epochs as per the protocols described in (Xiao et al., 2023). This is followed by an additional 75 epochs of fine-tuning that incorporates low-rank regularization. The learning rate follows a cosine schedule, with the initial rate, denoted by μ, determined through cross-validation for each model and dataset combination. Standard values for momentum and weight decay are maintained at 0.9 and 0, respectively. We also implement typical data augmentation techniques, which include random-resize cropping, random rotation, and random horizontal flipping. To ensure a balanced evaluation, all baseline models are also subjected to an additional 150 epochs of fine-tuning, which typically results in negligible improvements. A thorough analysis of this extended fine-tuning phase is provided in Table 11 in Section A.3.3 of the appendix. Our evaluation of the LRFL approach spans both CNN and vision transformer architectures, including models like ResNet-50, DenseNet, ViT-S, and ViT-B. We denote our LRFL models as 'X-LR', where 'X' represents the base model such as ResNet-50.

Tuning the T, η, and μ by Cross-Validation. We search for the optimal values of feature rank T, the weighting parameter for the TNN η, and the learning rate μ on each dataset. Let $T = \lceil \gamma \min(n, d) \rceil$, where γ is the rank ratio. The selection process for γ and η involves 5-fold cross-validation, utilizing 20% of the training data for each dataset. We test γ values from the set $\{0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.2\}$ and η from $\{5 \times 10^{-4}, 1 \times 10^{-3}, 2.5 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}\}$. Additionally, μ is evaluated from $\{5 \times 10^{-4}, 2.5 \times 10^{-4}, 1 \times 10^{-4}, 5 \times 10^{-5}, 2.5 \times 10^{-5}, 1 \times 10^{-5}\}$. We determine the optimal values for η, γ, and μ with a sequential greedy search strategy. Initially, we fix η and μ to ascertain the optimal γ through cross-validation. With the best γ identified, we then keep μ constant and search for the ideal η. Finally, having established the optimal values for both γ and η, we adjust to find the best μ through cross-validation. The selected optimal values for η, γ, and μ confirmed through cross-validation are detailed in Table 7 in Section A.2 of the appendix. Additionally, Table 8 in Section A.2 the appendix shows the time of the entire cross-validation process, affirming its efficiency and minimal impact on the overall computational burden of the training procedure.

4.2 Stanford CheXpert

Experimental setup. The CheXpert dataset (Irvin et al., 2019) encompasses a collection of 224,316 chest X-rays obtained from 65,240 patients, with 191,028 of these images designated for training purposes. Each X-ray in the dataset is accompanied by radiology reports that identify the presence of 14 distinct diseases. Following the protocol in (Xiao et al., 2023), all images are uniformly resized to a resolution of 224×224. Furthermore, we calculate and report the mean Area Under the Curve (AUC) for five specific disease classes and undertake a thorough comparison with state-of-the-art baseline methods to evaluate the effectiveness of our approach.

Results and analysis. Table 1 presents the performance comparisons between the baseline models and the models enhanced with LRFL. Throughout this section, we use the postfix "-LR" to indicate a neural network trained with our LRFL. For instance, we employ a Vision Transformer (ViT-B) model pre-trained on $489,090$ chest X-rays and a ViT-S model pre-trained on $266,340$ chest X-rays using Masked Autoencoders (MAE) (Xiao et al., 2023). After fine-tuning the ViT-B network on the CheXpert dataset, it records a mean AUC of 89.3%. Notably, the ViT-B-LR, trained with our LRFL method, attains state-of-the-art performance with an mAUC of 89.8%, representing an enhancement of 0.5% over the standard ViT-B. Similarly, the ViT-S-LR model shows a 0.4% improvement in mAUC over the original ViT-S model, underlining the efficacy of the LRFL approach. We also show the classification accuracy for the five diseases in Table 1, where our method performs much better than baseline methods. For example, the ViT-S-LR model achieves an mAUC of 86.3% on Cardiomegaly, marking a 4.5% improvement compared to the ViT-S model trained with MAE. These improvements underscore the potential of LRFL in enhancing disease detection capabilities.

4.3 NIH ChestX-ray14

Experimental setup. NIH ChestX-ray14 (Wang et al., 2017) consists of $112,120$ X-rays collected from $30,805$ unique patients. Each X-ray can have up to 14 associated labels, with the possibility of multiple labels per image. Following the official data split in (Wang et al., 2017), we use $75,312$ images for training

Table 1: The table shows the performance of various state-of-the-art (SOTA) baseline methods on CheXpert. DN represents DenseNet.

Method	Architecture	Rank	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	mAUC (%)
Allaouzi et al.(Allaouzi & Ahmed, 2019)	DN121	-	72.0	**88.0**	77.0	87.0	90.0	82.8
Irvin et al.(Irvin et al., 2019)		-	81.8	82.8	93.8	93.4	92.8	88.9
Seyyedkalantari et al.(Seyyed-Kalantari et al., 2020)		-	81.2	83.0	90.0	88.3	93.8	87.3
Pham et al.(Pham et al., 2021)		-	82.5	85.5	93.7	93.0	92.3	89.4
Hosseinzadeh et al.(Hosseinzadeh Taher et al., 2021)		-	-	-	-	-	-	87.1
Haghighi et al.(Haghighi et al., 2022)		-	-	-	-	-	-	87.6
Kang et al.(Kang et al., 2021)		-	82.1	85.9	**94.4**	89.2	93.6	89.0
DN121 (MoCo v2) (Xiao et al., 2023)		-	78.5	77.9	92.5	92.8	92.7	88.7
DN121 (Xiao et al., 2023)		-	81.5	77.6	89.4	92.3	92.0	88.7
ViT-S (Xiao et al., 2023)	ViT-S/16	-	83.5	81.8	93.5	94.0	93.2	89.2
ViT-S-LR (Ours)	ViT-S/16	0.05r	**83.7**	86.3	90.9	93.7	93.1	89.6
ViT-B (Xiao et al., 2023)	ViT-B/16	-	82.7	83.5	92.5	93.8	**94.1**	89.3
ViT-B-LR (Ours)	ViT-B/16	0.05r	81.6	85.4	93.4	**94.6**	93.9	**89.8**

and 25,596 images for testing. The raw images from the dataset are sized 1024×1024. In our experiments, we scale down the input images to 224×224. We report the mean AUC (Area Under the Curve) for 14 distinct classes and conduct a comprehensive comparison with 18 widely recognized and influential baseline methods.

Table 2: The table shows the performance of various state-of-the-art (SOTA) CNN-based and Transformer-based methods on ChestX-ray14. RN, DN, and SwinT represent ResNet, DenseNet, and Swin Transformer.

Method	Architecture	Pre-training	Rank	mAUC
Hermoza et al.(Hermoza et al., 2020)	DN121	ImageNet-1K	-	82.1
Kim et al.(Kim et al., 2021)	DN121		-	82.2
Haghighi et al.(Haghighi et al., 2022)	DN121		-	81.7
Liu et al.(Liu et al., 2022a)	DN121		-	81.8
Taslimi et al.(Taslimi et al., 2022)	SwinT		-	81.0
MoCo v2 (Xiao et al., 2023)	DN121	X-rays (0.3M)	-	80.6
MAE (Xiao et al., 2023)	DN121		-	81.2
RN-50 (Xiao et al., 2023)	RN50	ImageNet-1K	-	81.8
RN-50-LR (Ours)	RN50		0.05r	82.2
DN-121 (Xiao et al., 2023)	DN121	ImageNet-1K	-	82.0
DN-121-LR (Ours)	DN121		0.05r	82.4
ViT-S (Xiao et al., 2023)	ViT-S/16	X-rays (0.3M)	-	82.3
ViT-S-LR (Ours)	ViT-S/16		0.05r	82.7
ViT-B (Xiao et al., 2023)	ViT-B/16	X-rays (0.5M)	-	83.0
ViT-B-LR (Ours)	ViT-B/16		0.05r	**83.4**

Results and Analysis. Table 2 presents the performance comparisons between the LRFL models and state-of-the-art (SOTA) baselines on the NIH ChestX-ray14 dataset. For example, we use ViT-B model pre-trained on 266,340 chest X-rays with Masked Autoencoders (MAE) (Xiao et al., 2023). It is observed that all LRFL models achieve improvements in mean AUC compared to the corresponding base models. For instance, the pre-trained ViT-B network is fine-tuned on the NIH ChestX-ray14 dataset and achieves a mean AUC of 83.0. The corresponding LRFL model, denoted as ViT-B-LR, achieves a mean AUC of 83.4. ViT-S-LR improves its base model, ViT-S, by a mean AUC of 0.4%. Similar improvements are observed for CNN-based models as well. For example, ResNet-50-LR improves its base model by a mean AUC of 0.4%. It is important to highlight that the research community dedicated four years to enhancing the mAUC score for CNN-type architectures, advancing it from 74.5% to 82.2%, which was primarily attributed to the challenging nature of the classification with the NIH ChestX-ray14 dataset.

4.4 COVIDx

Experimental setup. The COVIDx dataset (Version 9A) (Pavlova et al., 2022) consists of 30,386 chest X-rays sourced from 17,026 unique patients. Consistent with prior studies (Pavlova et al., 2022; Xiao et al., 2023), we partition this dataset into 29,986 images for training across four distinct classes, and 400 images for testing, which are categorized into three classes. For objective evaluations and to ensure fair comparisons

with previous methodologies, we report the Top-1 accuracy for the test set, which encompasses the three classes.

Results and Analysis. Table 3 compares the performance of SOTA transformer-based models and the LRFL models on the COVIDx dataset. Similar to Section 4.2, the base Vision Transformers (ViTs), namely ViT-S and ViT-B, are initially pre-trained on chest X-rays using Masked Autoencoders (MAE). Subsequently, these pre-trained models are fine-tuned on the COVIDx dataset. Results in Table 3 show that both ViT-S-LR and ViT-B-LR models surpass their respective base models. Specifically, ViT-S-LR and ViT-B-LR demonstrate an improvement in accuracy by 1.6% and 1.7%, respectively. Table 3 also presents a performance comparison between our LRFL models and other leading models on the COVIDx dataset. Notably, the LRFL models significantly outperform CNN-based models such as DenseNet-121. For instance, the ViT-B-LR model achieves a new SOTA top-1 accuracy of 97% at an input resolution of 224×224. This represents a substantial increase of 1.7% over the previous SOTA performance as reported in (Xiao et al., 2023), highlighting the effectiveness of integrating LRFL into transformer-based models for medical image analysis on the COVIDx dataset.

Table 3: The table shows the performance of various state-of-the-art (SOTA) CNN-based and Transformer-based methods on COVIDx. With the same pre-training settings as ViT-S and ViT-B in (Xiao et al., 2023), our ViT-S-LR and ViT-B-LR achieve 97.0% Accuracy. DN represents DenseNet.

Method	Architecture	Pre-training	Rank	mAUC
Wang et al. (Wang et al., 2017)	RN50	ImageNet-1K	-	74.5
Li et al.(Li et al., 2018)	RN50		-	75.5
Yao et al. (Yao et al., 2018)	RN&DN		-	76.1
Wang et al.(Wang et al., 2019)	R152		-	78.8
Ma et al.(Ma et al., 2019)	R101		-	79.4
Tang et al.(Tang et al., 2018)	RN50		-	80.3
Baltruschat et al.(Baltruschat et al., 2019)	RN50		-	80.6
Guendel et al.(Guendel et al., 2018)	DN121		-	80.7
Guan et al.(Guan & Huang, 2018)	DN121		-	81.6
Seyyed et al.(Seyyed-Kalantari et al., 2020)	DN121		-	81.2
Ma et al.(Ma et al., 2020)	DN121(×2)		-	81.7
Hermoza et al.(Hermoza et al., 2020)	DN121		-	82.1
Kim et al.(Kim et al., 2021)	DN121		-	82.2
Haghighi et al.(Haghighi et al., 2022)	DN121		-	81.7
Liu et al.(Liu et al., 2022a)	DN121		-	81.8
Taslimi et al.(Taslimi et al., 2022)	SwinT		-	81.0
MoCo v2 (Xiao et al., 2023)	DN121	X-rays (0.3M)	-	80.6
MAE (Xiao et al., 2023)	DN121		-	81.2
RN-50 (Xiao et al., 2023)	RN50	ImageNet-1K	-	81.8
RN-50-LR (Ours)	RN50		0.05r	82.2
DN-121 (Xiao et al., 2023)	DN121	ImageNet-1K	-	82.0
DN-121-LR (Ours)	DN121		0.05r	82.4
ViT-S (Xiao et al., 2023)	ViT-S/16	X-rays (0.3M)	-	82.3
ViT-S-LR (Ours)	ViT-S/16		0.05r	82.7
ViT-B (Xiao et al., 2023)	ViT-B/16	X-rays (0.5M)	-	<u>83.0</u>
ViT-B-LR (Ours)	ViT-B/16		0.05r	**83.4**

4.5 Ablation Study

Efficiency Analysis of the Separable Approximation to the TNN. To verify the efficiency of the novel training algorithm of LRFL with the separable approximation to the TNN in Algorithm 1, we compare the training time of our LRFL models with an existing method for optimizing the TNN, TNNM-ALM (Lee & Lam, 2016), on NIH ChestX-ray14, CheXpert, and COVIDx. The results in Table 4 show that our LRFL method achieves 7×-10× acceleration in the training process on the three datasets, demonstrating the effectiveness and efficiency of the separable approximation to the TNN proposed in our paper.

Ablation Study on LRFL in Reducing the Adverse Effects of Background. To demonstrate that LRFL models are more robust to the background than the baselines, we perform an ablation study on LRFL

Table 4: Training time (minues) comparisons on NIH ChestX-ray14, CheXpert, and COVIDx.

Methods	Training Time (minutes)		
	NIH ChestX-ray14	CheXpert	COVIDx
ViT-S	54	90	23
ViT-S (TNM-ALM) (Lee & Lam, 2016)	804	854	342
ViT-S-LR	98	117	38
ViT-B	72	162	32
ViT-B (TNM-ALM) (Lee & Lam, 2016)	915	1461	418
ViT-B-LR	113	185	45

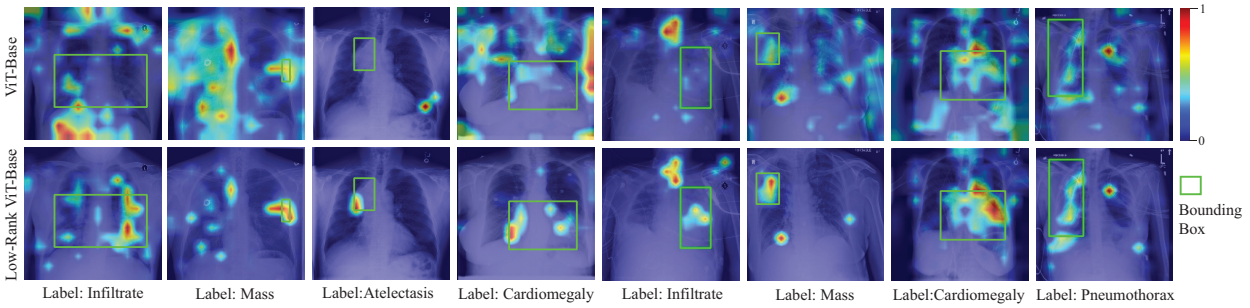


Figure 2: Grad-CAM (Selvaraju et al., 2017) visualization results on NIH ChestX-ray 14 are shown. The figures in the first row depict the visualization results of ViT-Base, while the figures in the second row show the visualization results of Low-Rank ViT-Base. The ground-truth bounding box for each disease is indicated in green. Additional Robust Grad-CAM (Selvaraju et al., 2017) visualization results of Low-Rank ViT-Base and Low-Rank ResNet-50 can be found in Figure 3 and Figure 4 in the appendix.

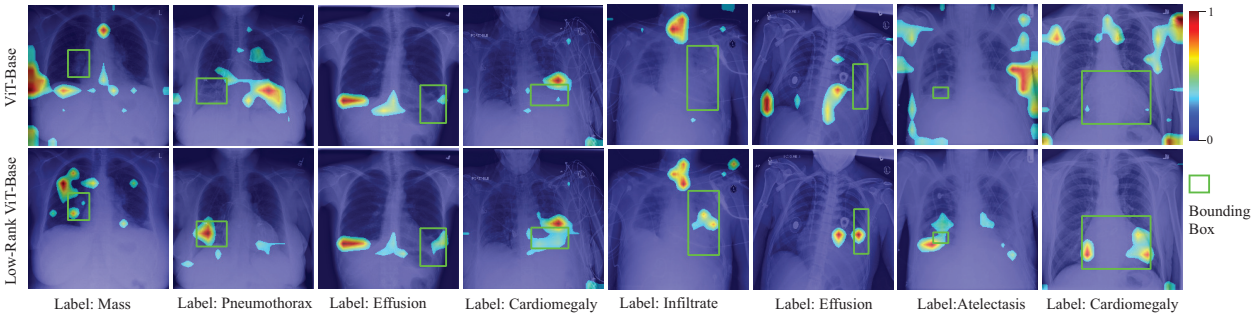


Figure 3: Robust Grad-CAM (Selvaraju et al., 2017) visualization results on NIH ChestX-ray 14. The figures in the first row are the visualization results of ViT-Base, and the figures in the second row are the visualization results of Low-Rank ViT-Base.

to reduce the adverse effects of background. In this study, we create a mask for the disease area for each original image, then decompose the original image, which has a bounding box for the disease, into a disease image and a background image. Both the disease image and the background image are of the same size as the original image. The background image has grayscale 0 in the masked disease area, and the disease image has grayscale 0 in the non-disease area. We feed the three images, which are the original image, the disease image, and the background image, to an LRFL model and obtain the original features, disease features, and background features for the LRFL model, respectively. We also feed these three images to a baseline model and obtain the original features, disease features, and background features for the baseline model. For each original image, we measure the distance between the disease features and original features using KL-divergence on the softmaxed features for the LRFL model and the baseline model. We then compute the average feature distance for each model, which is the average distance between the disease features and original features over the images with a ground-truth bounding box for the disease in the NIH ChestX-ray

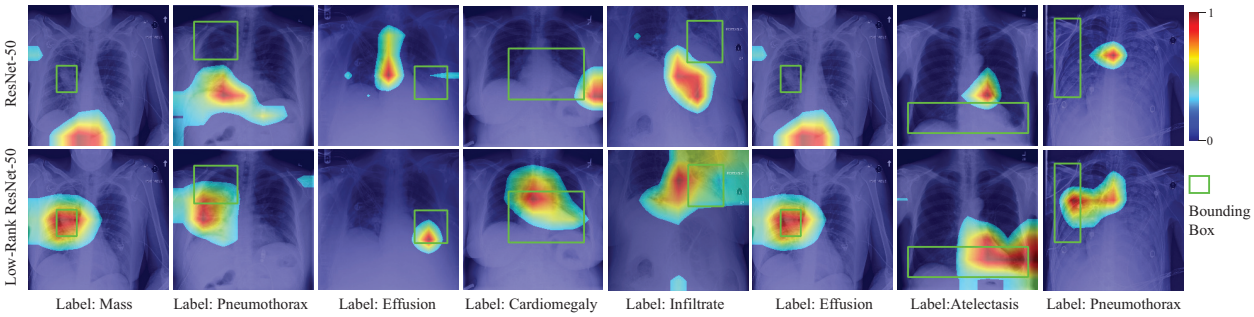


Figure 4: Robust Grad-CAM (Selvaraju et al., 2017) visualization results on NIH ChestX-ray 14. The figures in the first row are the visualization results of ViT-Base, and the figures in the second row are the visualization results of Low-Rank ResNet-50.

14. The results in Table 5 indicate that the original features are closer to the disease features by the LRFL models compared to the baseline models, evidencing the effectiveness of the LRFL models in reducing the adverse effect of the background area. We also remark that since only the low-rank part of the original features participates in the classification process, the noise and non-disease areas in the high-rank part of the features are mostly not learned by LRFL, and in this manner, LRFL is robust to both noise and background.

Table 5: Average feature distance between original features and disease features of images with a ground-truth bounding box for the disease in the NIH ChestX-ray 14.

Methods	mAUC (%)	Average Feature Distance
ViT-S	82.3	0.7030
ViT-S-LR	**82.7**	**0.6352**
ViT-B	83.0	0.5642
ViT-B-LR	**83.4**	**0.6628**

Grad-CAM Visualization. To study how LRFL improves the performance of base models in disease detection, we use the Grad-CAM (Selvaraju et al., 2017) to visualize the parts in the input images that are responsible for the predictions of the base models and low-rank models. Grad-CAM (Selvaraju et al., 2017) visualization results of Low-Rank ViT-Base are illustrated in Figure 2. Robust Grad-CAM (Selvaraju et al., 2017) visualization results of Low-Rank ViT-Base and Low-Rank ResNet-50 are illustrated in Figure 3 and Figure 4. All Grad-CAM visualization results illustrate that our LRFL models usually focus more on the areas inside the bounding box associated with the labeled disease. In contrast, the base models also focus on the areas outside the bounding box or even areas in the background. Additional Grad-CAM visualization results can be found in Figure 6 in Section A.3.4 of the appendix.

In addition, we compare the kernel eigenvalues and kernel complexities LRFL models and base models in Table 9 and Figure 5 in Section A.3.1 of the appendix. We perform an ablation study on LRFL models in scenarios with limited data availability in Table 10 in Section A.3.2 of the appendix. In Table 11 in Section A.3.3 of the appendix, we compare LRFL with other fine-tuning strategies.

5 Conclusion

In this study, we propose a novel Low-Rank Feature Learning (LRFL) method designed for the classification of thorax diseases. LRFL aims to mitigate the negative impacts of noise and non-disease areas in radiographic images, enhancing disease classification accuracy. The LRFL method is universally adaptable across various neural network architectures, drawing empirical support from the low frequency property and theoretical support from a sharp generalization bound developed for neural networks utilizing low-rank features. Our comprehensive experimental evaluations across several thorax disease datasets including NIH-ChestX-ray, COVIDx, and CheXpert—highlight robust performance improvements of LRFL in terms of both multi-class Area Under the Curve (mAUC) and overall classification accuracy.

References

Varghese Alex, Kiran Vaidhya, Subramaniam Thirunavukkarasu, Chandrasekharan Kesavadas, and Ganapathy Krishnamurthi. Semisupervised learning using denoising autoencoders for brain lesion detection and segmentation. *Journal of Medical Imaging*, 4(4):041311, 2017.

Imane Allaouzi and Mohamed Ben Ahmed. A novel approach for multi-label chest x-ray classification of common thorax diseases. *IEEE Access*, 7:64279–64288, 2019.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 322–332. PMLR, 2019.

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S Sara Mahdavi, Ellery Wulczyn, et al. Robust and efficient medical imaging with self-supervision. *arXiv preprint arXiv:2205.09723*, 2022.

Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843, 2021.

Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9(1):1–10, 2019.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005.

Han Cai, Chuang Gan, and Song Han. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In Zhi-Hua Zhou (ed.), *International Joint Conference on Artificial Intelligence*, pp. 2205–2211. ijcai.org, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Tej Bahadur Chandra and Kesari Verma. Analysis of quantum noise-reducing filters on chest x-ray images: A review. *Measurement*, 153:107426, 2020.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12299–12310, 2021a.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021b.

Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019.

Mark Chen, Alec Radford, Rewon Child, Jeff Wu, and Heewoo Jun. Generative pretraining from pixels. *Advances in Neural Information Processing Systems*, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020b.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Moulik Choraria, Leello Tadesse Dadi, Grigorios Chrysos, Julien Mairal, and Volkan Cevher. The spectral bias of polynomial neural networks. In *International Conference on Learning Representations*. OpenReview.net, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Qiaoqiao Ding, Yong Long, Xiaoqun Zhang, and Jeffrey A Fessler. Statistical image reconstruction using mixed poisson-gaussian noise model for x-ray ct. *arXiv preprint arXiv:1801.09533*, 2018.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11963–11975, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, pp. 1, 2018.

Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019.

Ruibin Feng, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Parts2whole: Self-supervised contrastive learning via reconstruction. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pp. 85–95. Springer, 2020.

Ming Gao, Runmin Liu, and Jie Mao. Noise robustness low-rank learning algorithm for electroencephalogram signal classification. *Frontiers in Neuroscience*, 15:797378, 2021.

Riqiang Gao, Yuankai Huo, Shunxing Bao, Yucheng Tang, Sanja L Antic, Emily S Epstein, Aneri B Balar, Steve Deppen, Alexis B Paulson, Kim L Sandler, et al. Distanced lstm: time-distanced gates in long short-term memory models for lung cancer detection. In *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*, pp. 310–318. Springer, 2019.

Bhawna Goyal, Sunil Agrawal, and BS Sohi. Noise issues prevailing in various types of medical images. *Biomedical & Pharmacology Journal*, 11(3):1227, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Qingji Guan and Yaping Huang. Multi-label chest x-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters*, 2018.

Sebastian Guendel, Sasa Grbic, Bogdan Georgescu, Siqi Liu, Andreas Maier, and Dorin Comaniciu. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In *Iberoamerican Congress on Pattern Recognition*, pp. 757–765. Springer, 2018.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20824–20834, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Renato Hermoza, Gabriel Maicas, Jacinto C Nascimento, and Gustavo Carneiro. Region proposals for saliency map refinement for weakly-supervised disease localisation and classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 539–549. Springer, 2020.

Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pp. 3–13. Springer, 2021.

Mingzhe Hu, Jiahan Zhang, Luke Matkovic, Tian Liu, and Xiaofeng Yang. Reinforcement learning in medical image analysis: Concepts, applications, challenges, and future directions. *Journal of Applied Clinical Medical Physics*, 24(2):e13898, 2023.

Qingsong Hu, Deyu Zhang, Wei Zhang, and Xuelong Li. Truncated nuclear norm regularization for tensor completion. *arXiv preprint arXiv:1308.0737*, 2013.

Yao Hu, Zhongming Jin, Yi Shi, Debing Zhang, Deng Cai, and Xiaofei He. Large scale multi-class classification with truncated nuclear norm regularization. *Neurocomputing*, 148:310–317, 2015.

Dildar Hussain and Yeong Hyeon Gu. Exploring the impact of noise and image quality on deep learning performance in dxa images. *Diagnostics*, 14(13):1328, 2024.

Piotr Indyk, Ali Vakilian, and Yang Yuan. Learning-based low-rank approximations. In *Advances in Neural Information Processing Systems*, 2019.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Mintong Kang, Yongyi Lu, Alan L Yuille, and Zongwei Zhou. Label-assemble: Leveraging multiple datasets with partial labels. *In Submission: Thirty-Sixth Conference on Neural Information Processing Systems*, 2021. URL https://arxiv.org/pdf/2109.12265.pdf.

Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. Xprotonet: Diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15719–15728, June 2021.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 12 2006.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Chul Lee and Edmund Y Lam. Computationally efficient truncated nuclear norm minimization for high dynamic range imaging. *IEEE Transactions on Image Processing*, 25(9):4145–4157, 2016.

Sangyoon Lee, Min Seok Lee, and Moon Gi Kang. Poisson–gaussian noise analysis and estimation for low-dose x-ray images in the nsct domain. *Sensors*, 18(4):1019, 2018.

Jun Li, Junyu Chen, Yucheng Tang, Bennett A Landman, and S Kevin Zhou. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *arXiv preprint arXiv:2206.01136*, 2022.

Jun Li, Junyu Chen, Yucheng Tang, Ce Wang, Bennett A Landman, and S Kevin Zhou. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *Medical image analysis*, pp. 102762, 2023.

Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8290–8299, 2018.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017b.

Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20706, 2022a.

Jinyu Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Low-rank tensor completion by truncated nuclear norm regularization. *arXiv preprint arXiv:1712.00704*, 2017.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022b.

Yuwu Lu, Zhihui Lai, Yong Xu, Xuelong Li, David Zhang, and Chun Yuan. Low-rank preserving projections. *IEEE transactions on cybernetics*, 46(8):1900–1913, 2015.

Congbo Ma, Hu Wang, and Steven C. H. Hoi. Multi-label thoracic disease image classification with cross-attention networks, 2020.

Yanbo Ma, Qiuhao Zhou, Xuesong Chen, Haihua Lu, and Yong Zhao. Multi-attention network for thoracic disease classification and localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1378–1382. IEEE, 2019.

EN Manson, V Atuwo Ampoh, E Fiagbedzi, JH Amuasi, JJ Flether, and C Schandorf. Image noise in radiography and tomography: Causes, effects and reduction techniques. *Curr. Trends Clin. Med. Imaging*, 2(5):555620, 2019.

Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12042–12051, 2022.

Shahar Mendelson. Geometric parameters of kernel machines. In Jyrki Kivinen and Robert H. Sloan (eds.), *Conference on Learning Theory (COLT)*, volume 2375 of *Lecture Notes in Computer Science*, pp. 29–43. Springer, 2002.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2071–2081, 2022.

Maya Pavlova, Tia Tuinstra, Hossein Aboutalebi, Andy Zhao, Hayden Gunraj, and Alexander Wong. Covidx cxr-3: a large-scale, open-source benchmark dataset of chest x-ray images for computer-aided covid-19 diagnostics. *arXiv preprint arXiv:2206.03671*, 2022.

Hieu H Pham, Tung T Le, Dat Q Tran, Dat T Ngo, and Ha Q Nguyen. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing*, 437:186–194, 2021.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 09–15 Jun 2019.

Jiahuan Ren, Zhao Zhang, Richang Hong, Mingliang Xu, Haijun Zhang, Mingbo Zhao, and Meng Wang. Robust low-rank convolution network for image denoising. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6211–6219, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pp. 232–243. World Scientific, 2020.

Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*, 2022.

Yan Shen and Mingchen Gao. Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization. In *International Workshop on Machine Learning in Medical Imaging*, pp. 389–397. Springer, 2018.

K Kirk Shung, Michael Smith, and Benjamin MW Tsui. *Principles of medical imaging*. Academic Press, 2012a.

K Kirk Shung, Michael Smith, and Benjamin MW Tsui. *Principles of medical imaging*. Academic Press, 2012b.

JH Siewerdsen, LE Antonuk, Y El-Mohri, J Yorkston, W Huang, JM Boudry, and IA Cunningham. Empirical and theoretical investigation of the noise performance of indirect detection, active matrix flat-panel imagers (amfpis) for diagnostic radiology. *Medical physics*, 24(1):71–89, 1997.

JH Siewerdsen, LE Antonuk, Y El-Mohri, J Yorkston, W Huang, and IA Cunningham. Signal, noise power spectrum, and detective quantum efficiency of indirect-detection flat-panel imagers for diagnostic radiology. *Medical physics*, 25(5):614–628, 1998.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

Jamshid Sourati, Ali Gholipour, Jennifer G Dy, Xavier Tomas-Fernandez, Sila Kurugol, and Simon K Warfield. Intelligent labeling based on fisher information for medical image segmentation using deep learning. *IEEE transactions on medical imaging*, 38(11):2642–2653, 2019.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Paul Suetens. *Fundamentals of medical imaging*. Cambridge university press, 2017.

Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald M Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In *International Workshop on Machine Learning in Medical Imaging*, pp. 249–258. Springer, 2018.

Sina Taslimi, Soroush Taslimi, Nima Fathi, Mohammadreza Salehi, and Mohammad Hossein Rohban. Swinchex: Multi-label classification on chest x-ray images with transformers. *arXiv preprint arXiv:2206.04246*, 2022.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q Tran, Dani Yogatama, and Donald Metzler. Scaling laws vs model architectures: How does inductive bias influence scaling? *arXiv preprint arXiv:2207.10551*, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Hongyu Wang, Haozhe Jia, Le Lu, and Yong Xia. Thorax-net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography. *IEEE journal of biomedical and health informatics*, 24(2):475–485, 2019.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Tiange Xiang, Yongyi Liu, Alan L Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. In-painting radiography images for unsupervised anomaly detection. *arXiv preprint arXiv:2111.13495*, 2021.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Transforming radiograph imaging with transformers: Comparing vision transformers with convolutional neural networks in multi-label thorax disease classification. In *Radiological Society of North America (RSNA)*, 2022.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3588–3600, 2023.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9653–9663, June 2022.

Lanyu Xu, Simeng Zhu, and Ning Wen. Deep reinforcement learning and its applications in medical imaging and radiation therapy: a survey. *Physics in Medicine & Biology*, 67(22):22TR02, 2022.

Ruixin Yang and Yingyan Yu. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in oncology*, 11:638182, 2021.

Yiming Yang and William W Cohen. Singular value pruning of deep neural networks with application to dialogue response selection. *arXiv preprint arXiv:1509.08865*, 2015.

Li Yao, Jordan Prosky, Eric Poblenz, Ben Covington, and Kevin Lyman. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*, 2018.

Yuan Yao, Fengze Liu, Zongwei Zhou, Yan Wang, Wei Shen, Alan Yuille, and Yongyi Lu. Unsupervised domain adaptation through shape modeling for medical image segmentation. In *Medical Imaging with Deep Learning*, 2021.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 558–567, 2021.

Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7277–7286, 2022.

Fanlong Zhang, Heyou Chang, Guowei Yang, Zhangjing Yang, and Minghua Wan. Truncated nuclear norm based low rank embedding. In *Biometric Recognition: 12th Chinese Conference, CCBR 2017, Shenzhen, China, October 28-29, 2017, Proceedings 12*, pp. 708–715. Springer, 2017.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Xinyu Zhang, Guanghui Wang, Xiangzhao Li, Xuelin Liu, and Wei Liu. An efficient tensor completion method via truncated nuclear norm. *ScienceDirect*, 2020.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pp. 27378–27394. PMLR, 2022a.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *ICLR*, 2022b.

S Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger. *Handbook of medical image computing and computer assisted intervention*. Academic Press, 2019a.

S Kevin Zhou, Hoang Ngan Le, Khoa Luu, Hien V Nguyen, and Nicholas Ayache. Deep reinforcement learning in medical imaging: A literature review. *Medical image analysis*, 73:102193, 2021.

Zongwei Zhou. *Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis*. PhD thesis, Arizona State University, 2021.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11. Springer, 2018.

Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pp. 384–393. Springer, 2019b.

Zongwei Zhou, Michael Gotway, and Jianming Liang. Interpreting medical images. In *Intelligent Systems in Medicine and Health: The Role of AI*. Springer, 2022c.

Jiuwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S Kevin Zhou, and Yefeng Zheng. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical Image Analysis*, 64:101746, 2020.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2021.

A More Experimental Results

A.1 Training Time Analysis

We evaluate the training time of LRFL models and compare them with the training time of the baseline models. The evaluation of LRFL models and baseline models is performed on 4 Nvidia A100 GPUs. It is observed from the results in Table 6 that the training time of LRFL models is comparable to the training time of LRFL models. The main computational overhead of LRFL models is the computation of the eigenvectors of the feature matrix $\mathbf{F}$ and the TNN. However, the computation overhead is largely reduced by avoiding performing SVD for the feature matrix $\mathbf{F}$ at every epoch, benefiting from the approximation algorithm we designed in Algorithm 1.

Table 6: Training time comparison between LRFL models and baseline models on NIH ChestX-ray14, CheXpert, and CovidX. All the results are reported in minutes.

Datasets	NIH ChestX-ray14	CheXpert	CovidX
ViT-S	54	90	23
ViT-S-LR	98	117	38
ViT-B	72	162	32
ViT-B-LR	113	185	45

A.2 Cross-Validation Results

The optimal values of the rank ratio γ, weighting parameter η, and learning rate μ decided by cross-validation for different models on different datasets are shown in Table 7.

Table 7: Optimal values of rank ratio γ, weighting parameter η, and learning rate μ decided by cross-validation for different models on different datasets.

Models	Parameters	NIH-ChestX-ray	COVIDx	CheXpert
ViT-S	γ	0.05	0.01	0.05
	η	5×10^{-4}	1×10^{-3}	1×10^{-3}
	μ	5×10^{-5}	2.5×10^{-5}	1×10^{-5}
ViT-B	γ	0.05	0.003	0.05
	η	5×10^{-4}	1×10^{-3}	1×10^{-3}
	μ	5×10^{-5}	2.5×10^{-5}	2.5×10^{-5}

In addition, the time for the entire cross-validation process in searching for the optimal values of the rank ratio γ, weighting parameter η, and learning rate μ are shown in Table 8. The evaluation is performed on

4 Nvidia A100 GPUs. As we use only 20% of the training data for cross-validation and train the models with each option for only 40% of the entire number of training epochs, the entire cross-validation process is efficient and does not largely increase the computation cost of the training process.

Table 8: Time Spent for cross-validation on NIH ChestX-ray14, CheXpert, and CovidX. All the results are reported in minutes.

Datasets	NIH ChestX-ray14	CheXpert	CovidX
ViT-S-LR	149	178	57
ViT-B-LR	172	285	69

A.3 Additional Ablation Study

A.3.1 Study on the Kernel Eigenvalues and Kernel Complexity

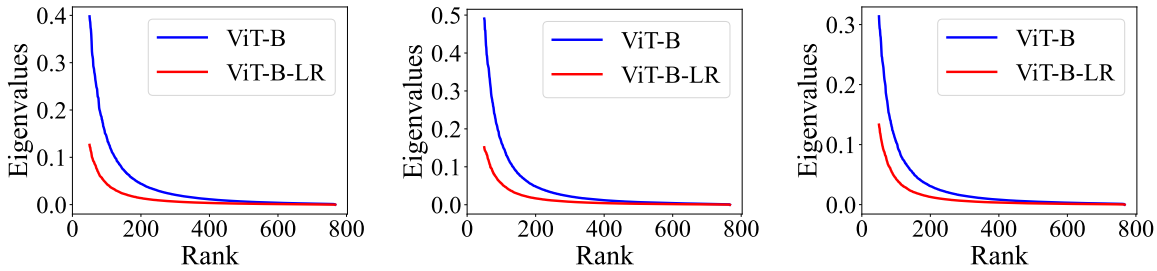
Kernel complexity (Bartlett et al., 2005; Koltchinskii, 2006; Mendelson, 2002) is a widely-studied complexity measure for the generalization capability of kernel-based learning algorithms. In this section, we compare the eigenvalues of the kernel and kernel complexity of ViT-B-LR and ViT-B on ChestX-ray14, COVIDx, and CheXpert. Given the representations of all the training images $\mathbf{F}$ learned by ViT-B or ViT-B-LR, the kernel complexity of the gram matrix $\mathbf{K}_n = \frac{1}{n}\mathbf{F}\mathbf{F}^\top$, which is also defined in Section 3.3, can be computed by

$$\text{Kernel Complexity of } \mathbf{K}_n = \min_{h \in [0,n]} \left(\frac{h}{n} + \sqrt{\frac{\sum_{i=h+1}^{n} \widehat{\lambda}_i}{n}} \right). \tag{5}$$

The eigenvalues of ViT-B-LR and ViT-B on ChestX-ray14, COVIDx, and CheXpert are illustrated in Figure 5. The computed kernel complexities of ViT-B-LR and ViT-B on ChestX-ray14, COVIDx, and CheXpert are shown in Table 9. It is observed that LRFL significantly reduces the kernel complexity of the image representations, which suggests that the LRFL models have lower generalization errors (Bartlett et al., 2005; Koltchinskii, 2006; Mendelson, 2002).

Table 9: Kernel complexity comparison between ViT-B-LR and ViT-B on ChestX-ray14, COVIDx, and CheXpert.

Method	ChestX-ray14		COVIDx		CheXpert	
	Kernel Complexity	h	Kernel Complexity	h	Kernel Complexity	h
ViT-B	0.0101	465	0.0207	303	0.0040	766
ViT-B-LR	0.0076	262	0.0155	187	0.0038	389



(a) ChestX-ray14 (b) COVIDx (c) CheXpert

Figure 5: Eigenvalues comparison between ViT-B-LR and ViT-B on ChestX-ray14, COVIDx, and CheXpert.

A.3.2 Experiments in Small Data Regimes

Experimental setup. We explore the effectiveness of low-rank features learned in scenarios with limited data availability, which is particularly significant given the challenges in acquiring high-quality data annotations in the medical imaging domain. We expect that LRFL models can demonstrate improved performance in such situations due to our theoretical guarantee of the better generalization capability of LRFL. We randomly select 5%, 10%, 15%, 20%, 25%, and 50% of the training data from the NIH ChestX-ray14 dataset and then fine-tune the base model using its default training configurations. We then train LRFL models for 20 epochs.

Results and analysis. As depicted in Table 10, our LRFL models consistently outperform their corresponding base methods across all data subsets, including 5%, 10%, 15%, 20%, 25%, and 50% on the NIH ChestX-ray14 dataset. Notably, the average improvement in performance is more substantial for the 5% data subset compared to the remaining subsets. For instance, ViT-B-LR exhibits a remarkable improvement of 1.05% for the 5% data subset, which significantly surpasses the improvements of 0.15%, 0.06%, 0.06%, 0.09%, and 0.11% observed for the 10%, 15%, 20%, 25%, and 50% training data subsets, respectively. These findings are consistent with our expectations, showcasing the strong generalization capability of LRFL models in mitigating over-fitting issues with limited data. In conclusion, our findings in the low-data regimes demonstrate the superiority of our LRFL in delivering more generalizable and robust representations for tasks with limited data availability, thereby contributing to the reduction of annotation costs.

Table 10: The table evaluates the performance of various models under low data regimes on the NIH ChestX-rays14 dataset. Models trained with low-rank features effectively combat overfitting in scenarios with limited data availability, thereby enhancing the quality of representations for downstream tasks.

Pre-training Dataset	Model	Label Fractions											
		5%		10%		15%		20%		25%		50%	
		Rank	mAUC	Rank	mAUC	Rank	mAUC	Rank	mAUC	Rank	mAUC	Rank	mAUC
X-rays(0.3M)	ViT-S	-	61.22	-	73.19	-	76.99	-	78.65	-	79.57	-	81.20
	ViT-S-LR(Ours)	$0.05r$	61.81	$0.2r$	73.84	$0.04r$	77.21	$0.04r$	78.86	$0.05r$	79.65	$0.05r$	81.35
X-rays(0.5M)	ViT-B	-	70.71	-	78.67	-	79.99	-	80.59	-	81.13	-	82.19
	ViT-B-LR (Ours)	$0.05r$	71.76	$0.2r$	78.82	$0.2r$	80.05	0.1r	80.65	$0.05r$	81.22	$0.05r$	82.30

A.3.3 Exploring Fine-tuning Strategies

Our LRFL method learns low-rank features by leveraging models pre-trained on the target dataset. In this section, we conduct an ablation study to investigate the significance of low-rank regularization in the fine-tuning process. A detailed comparative analysis of low-rank regularization against several performance-enhancing techniques, including mix-up (Zhang et al., 2018), label smoothing (Müller et al., 2019), and EMA (Wightman, 2019), is presented in Table 11. We performed an experiment by fine-tuning without low-rank regularization and other tricks, which serves as a baseline for studying the effects of fine-tuning strategies. All models underwent equivalent training epochs to ensure a fair comparison. The results demonstrate that LRFL models achieve the highest performance improvement compared to all other approaches. Notably, unlike natural images, applying mix-up, label smoothing, or EMA to the NIH ChestX-ray dataset leads to performance drops (see Table 11). Fine-tuning models pre-trained on the target dataset without low-rank regularization does not lead to performance improvements compared to fine-tuning with low-rank regularization. For example, the original ViT-S (Xiao et al., 2023) achieves a mean AUC of 82.27% on NIH Chest Xray-14. Fine-tuning this model for 20 epochs without low-rank regularization leads to a mean AUC of 82.26%, whereas fine-tuning with low-rank regularization for 75 epochs results in a mean AUC of 83.40%. We observe similar results for all models based on low-rank features, demonstrating the significance of LRFL.

A.3.4 Additional Grad-CAM Visualization Results

More grad-cam visualization results of Low-Rank ViT-Base on NIH ChestX-ray 14 are illustrated in Figure 6. We visualize the parts in the input images that are responsible for the predictions of the ground-truth disease label for base models and low-rank models. The visualization results show that our low-rank models usually

Table 11: Comparison of fine-tuning strategies on NIH ChestX-ray14.

Model	mAUC					
	Base Model	Fine-tuning	Mix-up (Zhang et al., 2018)	Label Smoothing (Müller et al., 2019)	EMA (Wightman, 2019)	LRFL
ViT-S	82.27	82.26	82.09	82.24	82.26	82.70
ViT-B	<u>83.00</u>	<u>83.00</u>	82.37	82.99	82.98	**83.40**

focus more on the areas inside the bounding box associated with the labeled disease. In contrast, the base models also focus on the areas outside the bounding box or even areas in the background.

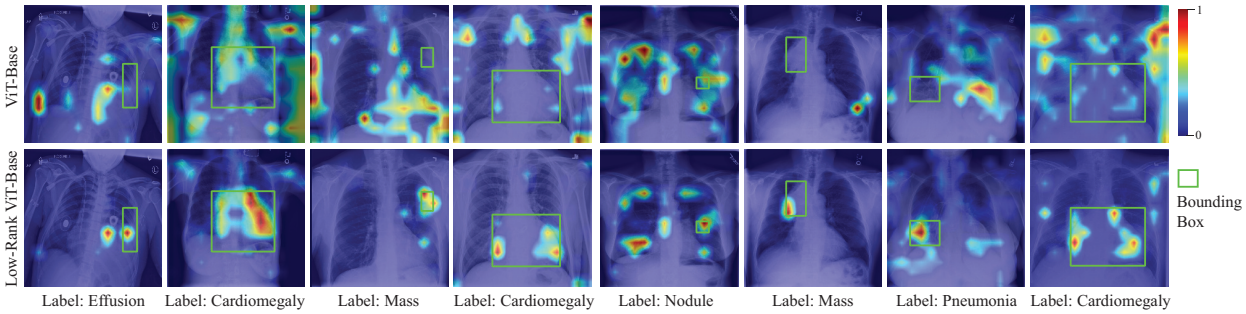


Figure 6: Grad-CAM visualization results on NIH ChestX-ray 14. The figures in the first row are the visualization results of ViT-Base, and the figures in the second row are the visualization results of Low-Rank ViT-Base.

B Proofs

Proof of Theorem 3.1. It can be verified that at the t-th iteration of gradient descent for $t \geq 1$, we have

$$\mathbf{W}^{(t)} = \mathbf{W}^{(t-1)} - \frac{\eta}{n}\mathbf{F}^\top\left(\mathbf{F}\mathbf{W}^{(t-1)} - \mathbf{Y}\right). \tag{6}$$

It follows by (6) that

$$\begin{aligned}\mathbf{F}\mathbf{W}^{(t)} &= \mathbf{F}\mathbf{W}^{(t-1)} - \eta\mathbf{K}_n\left(\mathbf{F}\mathbf{W}^{(t-1)} - \mathbf{Y}\right)\\ &= \mathbf{F}\mathbf{W}^{(t-1)} - \eta\mathbf{K}_n\left(\mathbf{F}\mathbf{W}^{(t-1)} - \bar{\mathbf{Y}}\right),\end{aligned} \tag{7}$$

where $\mathbf{K}_n = 1/n \cdot \mathbf{F}\mathbf{F}^\top$, $\bar{\mathbf{Y}} = \mathbf{U}^{(\bar{r})}{\mathbf{U}^{(\bar{r})}}^\top\mathbf{Y}$.

We define $\mathbf{F}(\mathbf{W}, t) \coloneqq \mathbf{F}\mathbf{W}^{(t)}$, then it follows by (7) that

$$\mathbf{F}(\mathbf{W}, t) - \bar{\mathbf{Y}} = (\mathbf{I}_n - \eta\mathbf{K}_n)\left(\mathbf{F}(\mathbf{W}, t) - \bar{\mathbf{Y}}\right),$$

which indicates that

$$\begin{aligned}\mathbf{F}(\mathbf{W}, t) - \bar{\mathbf{Y}} &= (\mathbf{I}_n - \eta\mathbf{K}_n)^t\left(\mathbf{F}(\mathbf{W}, 0) - \bar{\mathbf{Y}}\right)\\ &= -(\mathbf{I}_n - \eta\mathbf{K}_n)^t\bar{\mathbf{Y}},\end{aligned}$$

and

$$\begin{aligned}\|\mathbf{F}(\mathbf{W}, t) - \mathbf{Y}\|_{\mathrm{F}} &\leq \left\|\mathbf{Y} - \bar{\mathbf{Y}}\right\|_{\mathrm{F}} + \left(1 - \eta\widehat{\lambda}_r\right)^t\left\|\bar{\mathbf{Y}}\right\|_{\mathrm{F}}\\ &\leq \left\|\mathbf{Y} - \bar{\mathbf{Y}}\right\|_{\mathrm{F}} + \left(1 - \eta\widehat{\lambda}_r\right)^t\|\mathbf{Y}\|_{\mathrm{F}}.\end{aligned} \tag{8}$$

As a result of (8), by using the proof of (Bartlett et al., 2005, Theorem 3.3,Corollary 6.7), for every $x > 0$, with probability at least $1 - \exp(-x)$,

$$L_{\mathcal{D}}(\mathrm{NN}_{\mathbf{W}}) \leq c_1 \left\| \mathbf{Y} - \bar{\mathbf{Y}} \right\|_{\mathrm{F}}^2 + c_1 \left(1 - \eta \widehat{\lambda}_r\right)^{2t} \|\mathbf{Y}\|_{\mathrm{F}}^2 + c_2 \min_{h \in [0,r]} \left(\frac{h}{n} + \sqrt{\frac{1}{n} \sum_{i=h+1}^{r} \widehat{\lambda}_i} \right) + \frac{c_3 x}{n}. \tag{9}$$

□