

Non-negative Subspace Feature Representation for Few-shot Learning in Medical Imaging

Keqiang Fan^{a,*}, Xiaohao Cai^a and Mahesan Niranjan^a

^aElectronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK

ARTICLE INFO

Keywords:
 Few-shot learning
 Principal component analysis
 Non-negative matrix factorization
 Classification
 Subspace
 Medical imaging

ABSTRACT

Unlike typical visual scene recognition domains, in which massive datasets are accessible to deep neural networks, medical image interpretations are often obstructed by the paucity of data. In this paper, we investigate the effectiveness of data-based few-shot learning in medical imaging by exploring different data attribute representations in a low-dimensional space. We introduce different types of non-negative matrix factorization (NMF) in few-shot learning, addressing the data scarcity issue in medical image classification. Extensive empirical studies are conducted in terms of validating the effectiveness of NMF, especially its supervised variants (e.g., discriminative NMF, and supervised and constrained NMF with sparseness), and the comparison with principal component analysis (PCA), i.e., the collaborative representation-based dimensionality reduction technique derived from eigenvectors. With 14 different datasets covering 11 distinct illness categories, thorough experimental results and comparison with related techniques demonstrate that NMF is a competitive alternative to PCA for few-shot learning in medical imaging, and the supervised NMF algorithms are more discriminative in the subspace with greater effectiveness. Furthermore, we show that the part-based representation of NMF, especially its supervised variants, is dramatically impactful in detecting lesion areas in medical imaging with limited samples.

1. Introduction

Recent remarkable advancements in computer vision have prompted an interest in expanding these technologies for diagnostic and prognostic inference based on medical images. In medical imaging, the range of manually extracted image features is often limited. For different diseases, extracting relevant features from complex objects such as lesions and organs can be a challenging task due to their intricate nature, or may be relatively straightforward to achieve. It is advantageous to utilize neural networks, as they provide a consensus way to resolve feature extraction capabilities in medical imaging. According to neural scaling laws [1], the performance of a neural network can be improved consistently given the increase of three factors: model size, dataset size, and the amount of computation engaged in training, illustrating the importance of complex parametric models and large-scale datasets in benefiting the performance of neural networks in computer vision applications. Therefore, research for imaging inference with networks is frequently motivated by very large neural network architectures (especially depth) and equally enormous datasets to address the inability of simple models in parameter estimation of overly complex objects in images.

In contrast to natural images, deep learning techniques in medical imaging encounter multiple challenges [2]. One of the biggest challenges is data scarcity, i.e., the number of images available in the medical field is generally several orders of magnitude lower than that in many other fields;

moreover, data disparities are considerable between different tasks in medical imaging. For instance, participants in the dataset used in the "Leipzig Study for Mind-Body-Emotion Interactions" (LEMON) [3] represent only 10% of those included in DeepLesion [4], which stands as the largest open dataset of clinical CT scans sourced from the NIH Clinical Center. Due to constraints such as expert annotation costs and privacy concerns, constructing large and diverse enough medical datasets for deep learning models, such as disease detection, is extremely difficult. Although the combination of clinical information and medical images prompts several researchers to achieve improvements e.g. by exploring causality [5] and uncertainty [6], the fundamental problem of data scarcity in the medical field is again critical but has not been addressed well.

This paper focuses on medical image classification, particularly in the low data regime. Our interest lies in exploring datasets even smaller than those typically required for the general supervised training paradigm, ranging from a few hundred to even just a few tens of images for each disease. In this case, training a deep learning model from scratch is infeasible due to problems such as over-fitting and poor generalisation. Transfer learning, while fine-tuning the networks' parameters pre-trained from e.g. natural images could be useful to some extent to alleviate the data scarcity issue in the target domain [7]. However, for medical images, it does not considerably enhance classification performance [8]. Though more recent work using cascade learning (i.e., layer-wise pre-training of source models) [9] attempts to overcome this limitation, the work in [8] found that the effects of transfer learning on two large-scale medical imaging datasets (retinal fundus and chest X-ray data) are mainly because of model over-parameterization rather than complicated feature reuse, as widely believed. While alternative methods such as

*Corresponding author

✉ kf1d20@soton.ac.uk (K. Fan); x.cai@soton.ac.uk (X. Cai); mn@ecs.soton.ac.uk (M. Niranjan)
 ORCID(s): 0000-0002-9411-2892 (K. Fan); 0000-0003-0924-2834 (X. Cai); 0000-0001-7021-140X (M. Niranjan)

data augmentation and data synthesis expand data and add features, they may struggle to address the inherent bias of datasets lacking diversity compared to test data, potentially introducing noise, as it is assumed that both training and test data are drawn from the same distribution [10].

“Few-shot learning” techniques have been proven to be quite effective to address the data scarcity challenge [11]. In contrast to current few-shot learning methods based on models and prior information for parameter adjustment (e.g., Reptile [12], MAML [13], prototypical network [14], and matching network [15]), data-based few-shot learning methods can leverage traditional machine learning techniques. Traditional machine learning models focus more on manual prior information, including but not limited to data cleaning, data preprocessing, feature extraction, feature intersection, etc. Influenced by data attributes on pattern classification [16], we intend to explore problems in medical scenarios from the characteristics of the data, e.g., the output from the penultimate layer of a pre-trained network. This will allow us to explore data representation in different subspaces that are helpful for pattern classification.

When analysing data in different spaces, the data quantity will be different from the dimension of the one extracted by pre-trained deep neural networks. Since the limited parameters in traditional machine learning models could pose challenges in representing data with high dimensions, it is crucial to improve the robustness in subspaces and avoid the “curse of dimensionality” with appropriate data representation techniques (including e.g. sparse, collaborative, and non-negative representation), especially in the few-shot learning scenario. For example, effective dimensionality reduction methods can reduce model complexity and enhance generalisation performance by preserving specific characteristics and eliminating collinearity between data attributes.

One dimension reduction method for maintaining data collaboration is principal component analysis (PCA) [17], implemented by singular value decomposition (SVD) [18]. It is one of the most popular dimensionality reduction techniques and has been widely used to explore subspace representations [19, 20]. However, a fundamental weakness of PCA/SVD is that its variance-preserving low-rank approximation properties are mainly suitable for unimodal and Gaussian-distributed data. In the case of classification problems, the feature space is generally multimodal. Therefore, the representation of the principal components can have a certain ambiguity among the classes since the principal components with small eigenvalues may also contain critical divergence information. This implies, in some cases, the results using PCA for dimensionality reduction may not be as interpretable as just using the original sample features.

Pattern classification based on sparse representation and non-negative representation has been widely studied in tasks such as face recognition and object classification [16]. The work in [21] discovered that the effective representation

of homogeneous data samples should be dense and non-negative, which is linked to non-negative matrix factorization (NMF) [22]. In contrast to PCA, there is no subtraction during NMF-based reconstruction, and its non-negative constraint promotes the intuitive notion of combining parts into a whole – the part-based representation. The representation power of homogeneous samples can be boosted with the non-negative property while constraining heterogeneous samples, making the representation sparse and discriminative simultaneously and providing an efficient direction in the subspace for classification.

In this paper, different from the traditional paradigms of few-shot learning, we focus on the data-based few-shot learning paradigm, specifically targeting the utilization of subspaces in medical imaging. We investigate the intrinsic data representation within the features extracted by a pre-trained model, leveraging the benefits of information preservation in the subspace to address challenges in scenarios where the feature dimension exceeds the magnitude of available data. Our main objective is for the first time to explore the innovative application of NMF, particularly its discriminant variants, as alternatives to SVD for dimensionality reduction in low data regimes for multiclass medical inference problems. As a matrix factorization methodology, NMF generates sparse subspaces with part-based representations. Under the premise of sparsity and non-negativity, supervised NMF approaches such as discriminative NMF (DNMF) [23] and supervised and constrained NMF with sparseness (SCNMFS) [24] are also investigated for enhancing the subspace discriminative property by combining labelled samples. Varied combinations of the labelled samples will have different impacts on the subspace throughout the decomposition process. We demonstrate the robustness and generalizability of our method with these subspace representations across 14 datasets, covering 11 distinct medical classification tasks spanning four different imaging modalities. Thorough experimental results and comparison highlight the statistically significant performance improvement of our method in subspace representation, underscoring the benefits of utilizing NMF and supervised NMF subspaces as viable alternatives to SVD. Moreover, the part-based representation of the supervised NMF reveals the potential that SVD lacks in detecting discriminative information in medical imaging, as illustrated by saliency maps.

The rest of the paper is organized as follows. Section 2 briefly reviews the related methods. Section 3 details the few-shot learning framework on subspace feature representations, including the experimental settings and succinct descriptions of the datasets used. Extensive numerical experiments and comparison evaluating NMF and its variants in medical image classification, including detailed comparisons, are presented in Section 4. We conclude and point to some future work in Section 5.

2. Related Work

In this section, we briefly review different subspace representations, including SVD, NMF, DNMF and SCNMFS,

and a correspondence analysis method used for similarity comparison between different subspaces. In the following, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{M \times N}$ represents a data matrix, where $\mathbf{x}_i, 1 \leq i \leq N$, are M -dimensional feature vectors, and N is the number of feature vectors (or data samples).

2.1. Singular Value Decomposition

The SVD of the data matrix \mathbf{X} can be written as

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{P}^\top, \quad (1)$$

where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_M] \in \mathbb{R}^{M \times M}$ and $\mathbf{P} \in \mathbb{R}^{N \times N}$ are unitary matrices, and $\Sigma \in \mathbb{R}^{M \times N}$ is a diagonal matrix formed by r singular values of \mathbf{X} ; note that $r = \text{rank}(\mathbf{X}) \leq \min\{M, N\}$. The columns of \mathbf{U} and \mathbf{P} are the so-called left and right singular vectors, respectively. Let $\mathbf{U}_k = [\mathbf{u}_1, \dots, \mathbf{u}_k], 1 \leq k \leq r$. Utilising \mathbf{U}_k as the projection matrix, the low-dimensional SVD subspace representation of \mathbf{X} can be obtained by computing $\mathbf{V} = \mathbf{X}^\top \mathbf{U}_k$.

With the variance preserving property in compressing unimodal data, SVD has been used to interpret the learning dynamics of models across layers and models [18] and to find representative components in few-shot learning [19].

2.2. Non-negative Matrix Factorization

NMF is a well-known matrix factorization method that works by reconstructing a low-rank approximation of the input data matrix under the non-negativity constraint [25]. Given a non-negative data matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$, the purpose of NMF is to seek two non-negative and low-rank matrices $\mathbf{U} \in \mathbb{R}^{M \times k}$ and $\mathbf{V} \in \mathbb{R}^{N \times k}$ under the condition of $\mathbf{X} \approx \mathbf{UV}^\top$, where $k < \min\{M, N\}$. The non-negativity of all entries in \mathbf{U} and \mathbf{V} induces the sparsity of the subspace as well as the part-based representation. NMF can be formulated as the following constrained optimization problem

$$\min_{\mathbf{U}, \mathbf{V}} \|\mathbf{X} - \mathbf{UV}^\top\|_F^2, \quad \text{s.t. } \mathbf{U} \geq 0, \mathbf{V} \geq 0, \quad (2)$$

where $\|\cdot\|_F$ is the Frobenius norm. The work in [25] proposed multiplicative iterative updating rules to find \mathbf{U} and \mathbf{V} for the minimization problem (2), i.e.,

$$u_{ij} \leftarrow u_{ij} \frac{(\mathbf{X}\mathbf{V})_{ij}}{(\mathbf{UV}^\top\mathbf{V})_{ij}}, \quad v_{ij} \leftarrow v_{ij} \frac{(\mathbf{X}^\top\mathbf{U})_{ij}}{(\mathbf{VU}^\top\mathbf{U})_{ij}}. \quad (3)$$

Note that $(\cdot)_{ij}$ represents the (i, j) entry of the given matrix.

NMF has achieved tremendous success in various domains such as signal processing [26], biomedical engineering [27], pattern recognition [22], and image processing [28]. It is unsupervised factorization without fully utilising the label information in classification tasks. Below we briefly recall some supervised NMF methods, which will be investigated in our developed few-shot learning framework in Section 3.

2.3. Discriminative Non-negative Matrix Factorization

Babee et al. [23] proposed the DNMF method, coupling discriminative regularizers generated from the labelled data

with the main NMF objective function. Using the discriminative constraint from labels, each class is dispersed into a separate cluster in the resulting subspace – an appealing property for classification tasks.

For C number of classes, the label matrix $\mathbf{Q} \in \mathbb{R}^{C \times N}$ is introduced with one-hot processing of the labels corresponding to the samples in \mathbf{X} . With an auxiliary matrix $\mathbf{A} \in \mathbb{R}^{C \times k}$, the aim of DNMF is to find \mathbf{U}, \mathbf{V} and \mathbf{A} satisfying

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{V}, \mathbf{A}} & \|\mathbf{X} - \mathbf{UV}^\top\|_F^2 + \alpha \|\mathbf{Q} - \mathbf{AV}^\top\|_F^2, \\ \text{s.t. } & \mathbf{U} \geq 0, \mathbf{V} \geq 0, \end{aligned} \quad (4)$$

where $\alpha > 0$ is a constant used to balance the two terms in (4). Note that \mathbf{A} is allowed to take negative values. Analogous to the iterative formula in (3), problem (4) can be solved by updating \mathbf{U}, \mathbf{V} and \mathbf{A} alternatively, as derived in [23], i.e.,

$$\begin{aligned} u_{ij} & \leftarrow u_{ij} \frac{(\mathbf{X}\mathbf{V})_{ij}}{(\mathbf{UV}^\top\mathbf{V})_{ij}}, \\ v_{ij} & \leftarrow v_{ij} \frac{\left[\mathbf{X}^\top\mathbf{U} + \alpha (\mathbf{VA}^\top\mathbf{A})^- + \alpha (\mathbf{Q}^\top\mathbf{A})^+ \right]_{ij}}{\left[\mathbf{VU}^\top\mathbf{U} + \alpha (\mathbf{VA}^\top\mathbf{A})^+ + \alpha (\mathbf{Q}^\top\mathbf{A})^- \right]_{ij}}, \\ \mathbf{A} & \leftarrow \mathbf{QV} (\mathbf{V}^\top\mathbf{V})^\dagger, \end{aligned} \quad (5)$$

where the notations $(\cdot)^+$ and $(\cdot)^-$ represent the operations of setting the negative entries and positive entries in the given matrix to zero, respectively.

2.4. Supervised and Constrained Non-negative Matrix Factorization with Sparseness

The SCNMFS method proposed in [24] is a variant of the constrained NMF [29]. It takes both the sparsity and discriminative property into account under the condition of integrating the label information into the standard NMF decomposition. The insight into SCNMFS formulation resembles vector quantizing since it forces data with the same label to have the same latent representations [29]. The hard constraints on labels ensure that the latent representations of the data samples from the same class are the same.

After observing the non-negative coefficient matrix $\mathbf{Z} \in \mathbb{R}^{C \times k}$ and the label matrix \mathbf{Q} , the SCNMFS subspace representation is expressed as $\mathbf{V} = \mathbf{Q}^\top \mathbf{Z}$. The reconstruction process is transferred into $\mathbf{X} \approx \mathbf{UV}^\top = \mathbf{UZ}^\top \mathbf{Q}$. In our setting, the matrix \mathbf{U} is subjected to a Frobenius norm constraint, which is a commonly used approach in matrix optimization problems for regularizing the solution and avoiding overfitting. SCNMFS, then, is to find \mathbf{U} and \mathbf{Z} satisfying

$$\min_{\mathbf{U}, \mathbf{Z}} \|\mathbf{X} - \mathbf{UZ}^\top \mathbf{Q}\|_F^2 + \beta \|\mathbf{U}\|_F^2, \quad \text{s.t. } \mathbf{U} \geq 0, \mathbf{Z} \geq 0, \quad (6)$$

where $\beta \in (0, 1)$ balances the two terms in (6). Analogous to the iterative formulas in (3) and (5), problem (6) can be

solved by updating \mathbf{U} and \mathbf{Z} alternatively [24], i.e.,

$$\begin{aligned} u_{ij} &\leftarrow u_{ij} \frac{(\mathbf{X}\mathbf{Q}^\top\mathbf{Z})_{ij}}{(\mathbf{U}\mathbf{Z}^\top\mathbf{Q}\mathbf{Q}^\top\mathbf{Z})_{ij} + \beta u_{ij}}, \\ z_{ij} &\leftarrow z_{ij} \frac{(\mathbf{Q}\mathbf{X}^\top\mathbf{U})_{ij}}{(\mathbf{Q}\mathbf{Q}^\top\mathbf{Z}\mathbf{U}^\top\mathbf{U})_{ij}}. \end{aligned} \quad (7)$$

The final discriminative and sparse subspace is then generated by $\mathbf{V} = \mathbf{Q}^\top\mathbf{Z}$. The learned projection matrix \mathbf{U} with better sparsity ensures the identification capability of the obtained subspace \mathbf{V} .

2.5. Canonical Correlation Analysis

Canonical correlation analysis (CCA) [30] is a multivariate method elucidating the correlation between two datasets by inferring information from a cross-covariance matrix. Given two data matrices $\mathbf{X}_1 \in \mathbb{R}^{M_1 \times N}$ and $\mathbf{X}_2 \in \mathbb{R}^{M_2 \times N}$, the whole process can be expressed as seeking vectors $\mathbf{a} \in \mathbb{R}^{M_1}$ and $\mathbf{b} \in \mathbb{R}^{M_2}$ maximising the correlation ρ , i.e.,

$$\begin{aligned} \rho &= \text{corr}(\mathbf{a}^\top \mathbf{X}_1, \mathbf{b}^\top \mathbf{X}_2) \\ &= \frac{\mathbf{a}^\top \Sigma_{\mathbf{X}_1 \mathbf{X}_2} \mathbf{b}}{\sqrt{\mathbf{a}^\top \Sigma_{\mathbf{X}_1 \mathbf{X}_1} \mathbf{a}} \sqrt{\mathbf{b}^\top \Sigma_{\mathbf{X}_2 \mathbf{X}_2} \mathbf{b}}}, \end{aligned} \quad (8)$$

where $\Sigma_{\mathbf{X}_1 \mathbf{X}_1}$, $\Sigma_{\mathbf{X}_1 \mathbf{X}_2}$, and $\Sigma_{\mathbf{X}_2 \mathbf{X}_2}$ are the cross-covariance matrices.

In this paper, we use CCA to compare the decomposition results of \mathbf{U} and \mathbf{V} obtained by different subspace representation methods.

3. Method

In this section, we present the developed data-driven few-shot learning framework for medical image classification. Different from the existing few-shot learning methods, we exploit the function of NMF and its variants. Moreover, we also describe the data used in this research and the experiment settings.

3.1. Notation

Let $\mathcal{S} = \{\mathbf{D}, \mathcal{Y}\}$ be a given image dataset, where \mathbf{D} and \mathcal{Y} denote the images and their class labels, respectively. They are then separated into training and test sets, i.e., $\mathcal{S}_{\text{train}} = \{\mathbf{D}_{\text{train}}, \mathcal{Y}_{\text{train}}\}$ and $\mathcal{S}_{\text{test}} = \{\mathbf{D}_{\text{test}}, \mathcal{Y}_{\text{test}}\}$. Let $|\mathbf{D}| = L_2$, i.e., the number of images in set \mathbf{D} . Let $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_{L_2}) \in \mathbb{R}^{L_1 \times L_2}$ be a matrix representing the whole images in \mathbf{D} , where $\mathbf{d}_i \in \mathbb{R}^{L_1}$ represents an image by concatenating the image columns and L_1 is therefore the image size.

Let f_{Θ_1} represent the pre-trained deep neural network which will be used to extract the prior information (i.e., features of our interest) from the given training/test images. For example, $\forall \mathbf{d}_i \in \mathbf{D}, f_{\Theta_1}(\mathbf{d}_i)$ is the obtained feature vector of image \mathbf{d}_i . Then all the feature vectors obtained by f_{Θ_1} on set \mathbf{D} can form a feature space, being represented by matrix

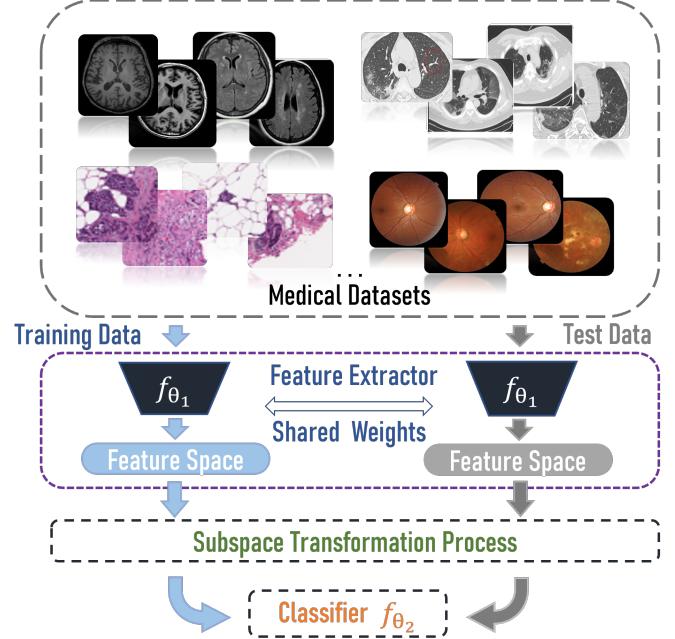


Figure 1: Few-shot learning framework on medical imaging; see the main text for the detailed description. Note that there is no fine-tuning or training process for feature extraction (i.e., the purple block).

$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{L_2}) = (f_{\Theta_1}(\mathbf{d}_1), f_{\Theta_1}(\mathbf{d}_2), \dots, f_{\Theta_1}(\mathbf{d}_{L_2}))$. For simplicity, we define the process as $\mathbf{X} \leftarrow f_{\Theta_1}(\mathbf{D})$. Analogously, the feature spaces corresponding to the training and test sets can be obtained by $\mathbf{X}_{\text{train}} \leftarrow f_{\Theta_1}(\mathbf{D}_{\text{train}})$ and $\mathbf{X}_{\text{test}} \leftarrow f_{\Theta_1}(\mathbf{D}_{\text{test}})$, respectively.

Let Δ represent the method SVD, NMF, DNMF or SC-NMFS described in Section 2. Denote $\mathbf{U}_{\text{train}}^\Delta$ and $\mathbf{V}_{\text{train}}^\Delta$ (and $\mathbf{U}_{\text{test}}^\Delta$ and $\mathbf{V}_{\text{test}}^\Delta$) as the projection matrix and the subspace representation with k columns, respectively, generated by method Δ at the training stage (and the test stage). Let f_{Θ_2} be the classifier trained on $\{\mathbf{V}_{\text{train}}^\Delta, \mathcal{Y}_{\text{train}}\}$ and tested on $\{\mathbf{V}_{\text{test}}^\Delta, \mathcal{Y}_{\text{test}}\}$.

3.2. Framework

Fig. 1 shows our proposed framework utilizing a pre-trained backbone network and an NMF dimensionality reduction scheme. The original medical dataset is randomly divided into a training set and a test set. The corresponding feature space is then extracted by the pre-trained network without fine-tuning steps. Following the dimensionality reduction of the feature space with the introduced different types of NMF methods, the connection between the subspaces of the training set and the test set is formed by implicit operations in the subspace transformation process. Finally, one classifier is trained from the generated subspace and performs the prediction task.

To better understand the framework in Fig. 1, the procedures are summarised in Algorithm 1. Firstly, the feature spaces corresponding to the training and test image sets are generated by $\mathbf{X}_{\text{train}} \leftarrow f_{\Theta_1}(\mathbf{D}_{\text{train}})$ and $\mathbf{X}_{\text{test}} \leftarrow$

Table 1

Details of the 14 medical datasets used in the experiments.

Data	Brief introduction	# Classes	Modality	Scale
BreastCancer[31]	Invasive ductal carcinoma (IDC), the most prevalent subtype of all breast cancers, is collected, containing breast histopathology images.	2	Histopathology	277,524
BrainTumor[32]	Brain tumor dataset provides human brain images including the intricate abnormalities in brain tumor size and location with 4 categories: glioma, meningioma, no tumor and pituitary.	4	MRI	7,022
CovidCT[33]	CovidCT contains hundreds of scans for COVID-19 with high diagnostic accuracy.	2	CT	812
DeepDRiD [34]	Deep Diabetic Retinopathy dataset provides dual-view fundus images of eyes with discernible quality levels, containing information from 500 patients.	5	Regular refund image	2,000
BloodMNIST[35]	BloodMNIST dataset is built on a dataset of individual normal blood cells from 8 categories without infection during collection.	8	Blood Cell Microscope	17,092
BreastMNIST[36]	BreastMNIST dataset is composed of breast ultrasound images. The original categorized classes (normal, benign, and malignant) are simplified into binary classification by combining normal and benign as positive and classifying them against malignant as negative. We also explore the deep features class activation map with its provided segmentation masks.	2	Breast Ultrasound	780
DermaMNIST[37, 38]	DermaMNIST dataset is regarding common pigmented skin lesions with 7 distinct illnesses.	7	Dermatoscope	10,015
OCTMNIST[39]	OCTMNIST is based on valid optical coherence tomography (OCT) images for retinal diseases.	4	Retinal OCT	109,309
OrganAMNIST[40]	OrganA/C/SMNIST datasets are generated from 3D CT images of Liver			58,850
OrganCMNIST[40]	Tumor Segmentation Benchmark[40], based on cropping from the centre slices of 3D bounding boxes in axial, coronal and sagittal views, respectively.	11	Abdominal CT	23,660
OrganSMNIST[40]				25,221
PathMNIST[41]	PathMNIST dataset is based on two previous studies (NCT-CRC-HE-100K and CRC-VAL-HE-7K) that used colorectal cancer histology slides to predict survival.	9	Colon Pathology	107,180
PneumoniaMNIST[39]	PneumoniaMNIST is a dataset of pediatric chest X-Ray images.	2	Chest X-Ray	5,856
TissueMNIST[42]	TissueMNIST dataset is obtained from the Broad Bioimage Benchmark Collection [42, 43] which focuses on the human kidney cortex cells, segmented from 3 reference tissue specimens and organized into 8 categories.	8	Kidney Cortex Microscope	236,386

$f_{\Theta_1}(\mathbf{D}_{\text{test}})$, respectively. With the selected method Δ and the subspace dimension k , the projection matrix $\mathbf{U}_{\text{train}}^{\Delta}$ and the subspace representation $\mathbf{V}_{\text{train}}^{\Delta}$ are derived from the feature space $\mathbf{X}_{\text{train}}$. The projection matrix $\mathbf{U}_{\text{test}}^{\Delta}$ for the test data is yielded by the established implicit relationship between the training set and the test set (i.e., step 10 in Algorithm 1). The subspace representation $\mathbf{V}_{\text{test}}^{\Delta}$ for the test data is then obtained by decomposing \mathbf{X}_{test} by $\mathbf{U}_{\text{test}}^{\Delta}$. Classifier f_{Θ_2} trained on the subspace representation $\mathbf{V}_{\text{train}}^{\Delta}$ is finally applied on the subspace representation $\mathbf{V}_{\text{test}}^{\Delta}$ to predict the final classification results.

3.3. Data Description

A total of 14 different datasets covering a range of problems in diagnostics are employed for validation. The detailed description of these datasets is given in Table 1. The four

selected medical datasets (BreastCancer, BrainTumor, CovidCT and DeepDRiD) cover diseases such as breast cancer [31, 44], brain tumor [32], COVID-19 [33] and diabetic retinopathy [34]. Early identification and classification of these diseases is an important research topic in medical imaging since it aids in the selection of the best treatment choices for patients. Ten datasets with MNIST in their names are part of the MedMNIST [45] benchmark collection. They cover primary data modalities (i.e., X-Ray, OCT, Ultrasound, CT and Electron Microscope) and diverse classification tasks (i.e., binary/multiclass with classes ranging from 2 to 11).

Each image in our experiments is resized to 50×50 pixels and we randomly sample each of those datasets into two sizes (i.e., including 300 and 600 images) instead of using the entire datasets. The sizes of datasets are chosen to be on either side of the dimensionality of the feature space, which for ResNet is 512. The greyscale images are

Algorithm 1 Few-shot learning framework

Require: S_{train} , S_{test} , f_{Θ_1} , k (i.e., subspace dimension), and Metype (i.e., method type)

Ensure: The predicted labels and accuracy of $\mathcal{D}_{\text{test}}$

- 1: Generate feature spaces X_{train} and X_{test} , i.e., $X_{\text{train}} \leftarrow f_{\Theta_1}(\mathcal{D}_{\text{train}})$ and $X_{\text{test}} \leftarrow f_{\Theta_1}(\mathcal{D}_{\text{test}})$;
- 2: Calculate $U_{\text{train}}^{\Delta}$ and $V_{\text{train}}^{\Delta}$ in the **Switch** function below;
- 3: **Switch** (MeType)
 - 4: **case** SVD: using (1);
 - 5: **case** NMF: using (3);
 - 6: **case** DNMF: using (5);
 - 7: **case** SCNMFS: using (7);
- 8: **end switch**
- 9: Train the classifier f_{Θ_2} with $V_{\text{train}}^{\Delta}$ and $\mathcal{Y}_{\text{train}}$;
- 10: Let $U_{\text{test}}^{\Delta} = U_{\text{train}}^{\Delta}$;
- 11: **Switch** (MeType)
 - 12: **case** SVD: compute V_{test}^{Δ} with U_{test}^{Δ} using (1);
 - 13: **case** NMF / DNMF / SCNMFS:
 - fix U_{test}^{Δ} and compute V_{test}^{Δ} using (3);
- 14: **end switch**
- 15: Compute the predicted labels of $\mathcal{D}_{\text{test}}$ using f_{Θ_2} with V_{test}^{Δ} , and compute the accuracy using $\mathcal{Y}_{\text{test}}$.

converted into RGB images via the strategy in [45] to ensure the compatibility of the pre-trained network. It is noteworthy that the purpose of selecting these datasets is to illustrate the subspace mechanism of few-shot learning, rather than to compare with previous results on these datasets.

3.4. Experiment Setting

In our experiments, each dataset takes two sizes (i.e., 300 and 600 images) as the training sets, together with test sets with scales of 80 and 160 images, respectively. All the data are further preprocessed by subtracting the mean and being divided by the standard deviation before being used. Therefore, each image is distributed between 0 and 1 to eliminate the effect caused by abnormal data. ResNet18 pre-trained on ImageNet is used to generate the corresponding feature space. The feature space represents the output of the penultimate layer of ResNet18 implemented by PyTorch hooks [46], yielding a 512-dimensional feature vector for each image.

The subspace representations of the extracted feature space are generated by the methods introduced in Section 2. The number of iterations related to NMF, DNMF and SCNMFS is set to 3000, ensuring the convergence. The main classifier used is the K -nearest neighbours (KNN) algorithm, where K is set to 5 unless otherwise specified. The reported average classification accuracy with standard deviation is achieved by repeating random samplings of the data 10 times. Besides, we localise the distinct medical image regions (saliency map) from the subspace using the

class activation map (CAM) method [47]. The feature maps in the last convolutional layer are extracted using the same method as the feature space. To obtain visualizations maps, we take a fully connected layer as the classifier instead of KNN to find the weights and gradients. The detailed steps of implementing the CAM method are shown in Appendix A.

Comparisons are firstly conducted between the subspaces generated from SVD, NMF, DNMF and SCNMFS, with quantitative and qualitative metrics like classification accuracy and saliency maps. As widely believed that data augmentation techniques could bring positive impacts for the low data regime, A comparison with data augmentation techniques [48] and their impact on the feature space are also conducted in our experiments. Each image is subjected to standard rotation and cropping to increase the dataset size. Ablation research involves comparing the feature space with different subspaces through dimensionality reduction, as conducted in Section 4.1.1. Moreover, we also compare our method with a well-known few-shot learning method – prototypical network [14]. The network is composed of four convolution blocks. Each block is composed of a 64-filter 3×3 convolution, batch normalization layer, a ReLU nonlinearity and a 2×2 max-pooling layer. In our experiments, we pre-trained the network on the omniglot dataset [49] via SGD with Adam [50] and obtained 99% accuracy in the 10-shot scenario.

In the experiments, the “ C -way l -shot” setting means l samples are given for each class in the support set, and l query images per class are provided to validate the final performance. The original feature space represents the features extracted by the pre-trained network without dimension reduction. Different number of dimensions for the SVD/NMF/DNMF/SCNMFS subspaces are tested. The final classification accuracy is computed by averaging over 10 randomly generated episodes from each medical dataset.

4. Experimental Results and Discussion

In this section, we evaluate the effectiveness of using subspaces in data-based few-shot learning for medical images with limited samples, and evaluate the part-based representation of NMF and its variants in detecting the distinct lesion area in medical imaging.

4.1. Classification Results and Discussion

4.1.1. Effect of subspace

Table 2 shows the classification results of using the original feature space and four subspaces obtained by SVD, NMF, DNMF and SCNMFS, including data augmentation. Both the original feature space and the feature space with data augmentation have 512 dimensions and are derived directly from the pre-trained network. The dimension of the subspaces derived by SVD, NMF, DNMF and SCNMFS is kept at 30. As indicated before, experiments are conducted at two different training set sizes of 300 and 600 to explore the cases that the data size is smaller and larger than the dimension of the feature space (i.e., 512). In general, an increase in the amount of data affects the classification

Table 2

Few-shot learning classification accuracy with KNN classifier. The original feature space and 30-dimensional subspaces obtained by SVD, NMF, DNMF and SCNMFS, including data augmentation, are evaluated on 14 medical datasets.

Datasets	Size	Augmentation	Feature space	Accuracy(%)				
				SVD	NMF	DNMF	SCNMFS	Classes
BreastCancer[31]	300	40.75±6.25	65.00±2.37	70.25±4.43	62.00±4.37	69.00±6.63	77.75±3.48	2
	600	42.75±4.30	69.00±2.87	72.25±1.51	65.25±2.64	70.38±2.19	75.32±2.12	
BrainTumor[32]	300	58.00±4.51	63.75±6.66	65.00±8.02	64.50±3.67	64.00±4.06	66.00±5.67	4
	600	59.40±3.50	69.62±4.36	69.38±1.81	66.75±1.08	65.25±2.26	67.88±2.42	
CovidCT[33]	300	62.70±4.70	78.00±4.00	75.25±2.00	71.00±5.39	72.25±6.19	71.75±6.20	2
	600	62.12±3.72	81.50±1.46	79.25±2.81	77.50±1.53	76.12±2.28	70.75±1.74	
DeepDRiD [34]	300	24.85±3.53	45.00±4.03	43.50±5.67	41.75±4.58	45.50±6.35	47.72±5.67	5
	600	30.91±5.21	54.50±2.35	54.50±2.83	52.38±3.59	53.25±2.48	49.50±2.03	
BloodMNIST[35]	300	18.50±4.64	34.75±5.83	42.25±6.19	32.25±2.15	39.75±4.21	46.75±4.59	8
	600	15.12±2.45	37.62±2.78	45.00±3.10	39.75±2.64	41.50±2.70	48.88±2.78	
BreastMNIST[36]	300	71.75±8.82	72.74±4.36	73.75±5.30	70.50±6.50	71.75±5.62	68.75±5.18	2
	600	71.44±4.04	74.50±1.74	73.12±3.51	70.87±3.41	72.62±3.76	68.62±4.72	
DermaMNIST[37]	300	18.00±6.83	23.75±1.77	29.50±3.92	23.75±4.18	27.25±5.88	37.75±4.70	7
	600	20.62±2.23	28.62±2.35	32.38±3.36	27.38±1.39	31.87±2.71	38.50±2.42	
OCTMNIST[39]	300	23.00±5.45	29.00±7.59	30.75±3.22	29.75±4.64	30.50±6.05	31.75±4.23	4
	600	25.38±3.25	34.62±4.41	34.25±5.43	29.88±3.52	29.25±2.94	37.38±3.43	
OrganAMNIST[40]	300	17.25±6.24	23.50±5.09	35.00±6.17	27.75±4.14	30.25±5.09	42.75±4.57	11
	600	17.12±4.50	34.62±2.58	40.88±2.64	34.63±3.15	39.12±4.36	46.12±1.50	
OrganCMNIST[40]	300	20.00±1.11	21.50±5.50	29.50±4.91	20.25±1.66	28.00±1.27	35.75±3.59	11
	600	12.75±2.87	28.75±4.15	34.25±3.61	28.12±2.34	30.87±3.32	42.25±3.10	
OrganSMNIST[40]	300	13.75±3.44	22.50±3.71	24.25±3.12	18.50±2.15	24.75±3.66	30.25±3.74	11
	600	13.88±1.33	25.25±2.26	28.75±1.63	26.25±2.90	27.00±2.22	32.38±1.00	
PathMNIST[41]	300	18.75±6.02	28.75±4.33	45.75±4.44	34.75±6.24	44.00±6.19	54.75±6.04	9
	600	17.00±2.94	33.25±2.18	45.13±2.54	39.62±6.37	44.50±1.74	54.62±1.79	
PneumoniaMNIST[39]	300	54.50±7.27	66.75±4.08	72.75±3.98	63.25±6.25	74.25±5.28	80.75±3.22	2
	600	50.37±4.26	71.00±1.70	76.25±2.17	71.12±2.86	74.12±3.55	81.12±2.48	
TissueMNIST[42]	300	10.63±15.30	17.25±4.57	20.00±2.85	15.50±3.02	18.75±3.79	21.75±3.76	8
	600	13.13±14.80	19.38±3.35	19.88±1.50	19.38±3.19	18.88±2.66	23.62±1.00	

results, as more various features are brought in. Interestingly, the results in Table 2 show that the subspace does not appear to have the same trend of improvement as the original feature space. The subspace representations (i.e., using SVD, NMF, DNMF and SCNMFS) overall yield better results than the original feature space (except for the datasets CovidCT and BrainTumor), indicating that dimensionality reduction is an important step for working in low data regimes. Table 2 also shows that data augmentation does not increase but decrease the classification performance, indicating its negative impact on the classification tasks here in medical imaging.

As shown in Fig. 2, we see that using subspaces is robust with different dimensions starting from 10. Table 3 show that subspaces with appropriate dimensions are undoubtedly a better way with even only 10 samples per class available. Our proposed method also outperforms the prototypical network in all the datasets.

4.1.2. Performance of SVD and NMF

Regarding the performance of subspaces in few-shot learning on both the binary class and multiclass problems, the classification result discrepancies between the standard NMF and SVD are modest as shown in Table 2, with SVD performing slightly better. The selected rank during decomposition has an impact on the efficacy of the subspace when NMF converts data into sparse and part-based subspaces. As shown in Fig. 2, NMF yields stable and reliable results when the subspace dimension is not too small (e.g. > 10). SVD's performance is also limited by the number of dimensions, with features from its lower dimensions yielding poor performance in classification tasks.

The ROC results from the KNN classifier in multiple subspaces are shown in Fig. 3. In this experiment, the PneumoniaMNIST dataset of size 300 is used, and two different dimensions (i.e., 5 and 15) are reported as an example. Different colour lines represent the average results of the

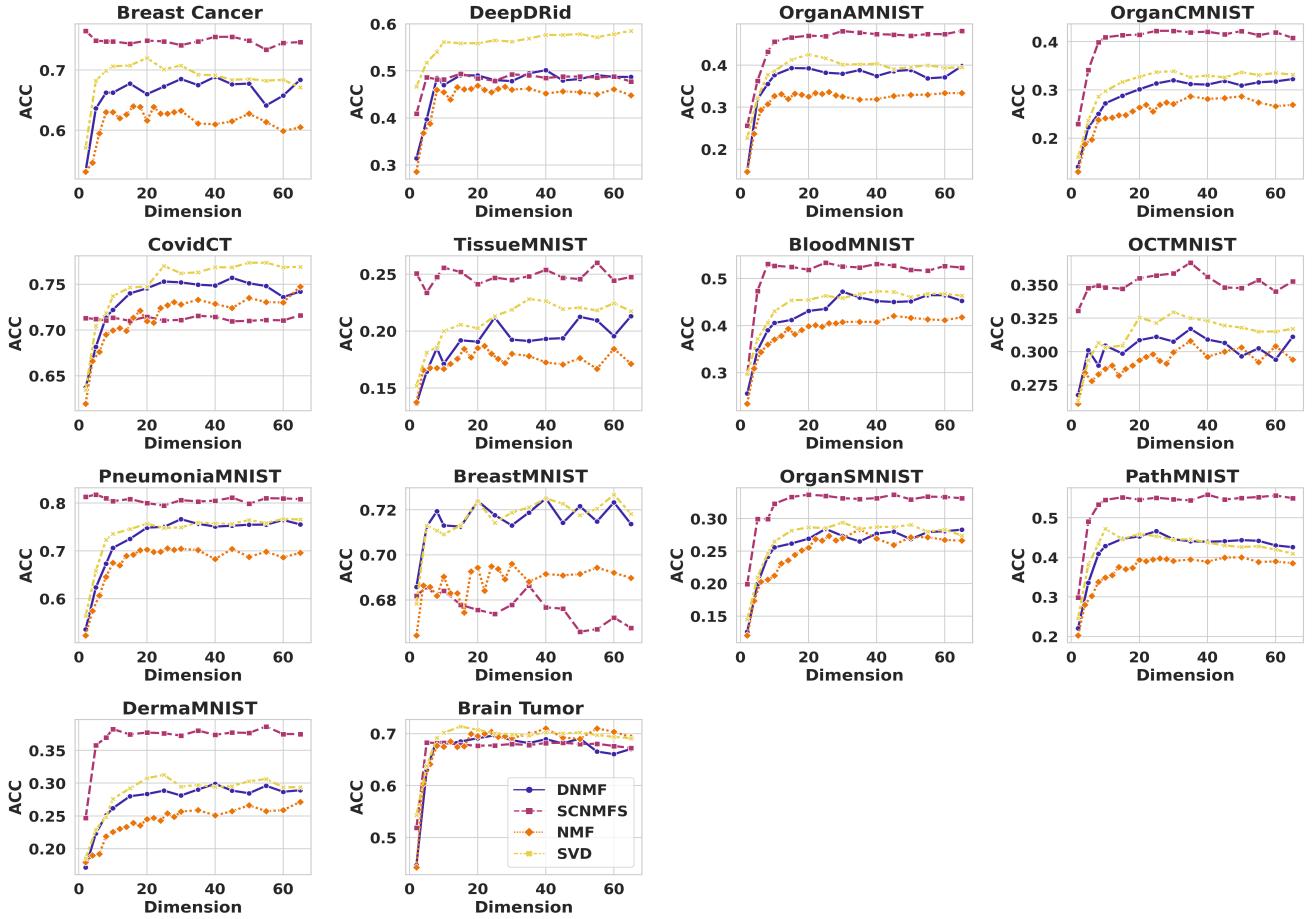


Figure 2: Classification accuracy of the NMF, DNMF, SCNMFS and SVD subspaces on 14 medical datasets with subspace dimensions ranging from 2 to 70. The dataset size is chosen as 600. In the plots, different colours correspond to different methods. Ten random partitions of the training-test set on each of the 14 datasets are conducted. It shows that the supervised NMF, especially SCNMFS, achieves significant improvements over the SVD-based subspace.

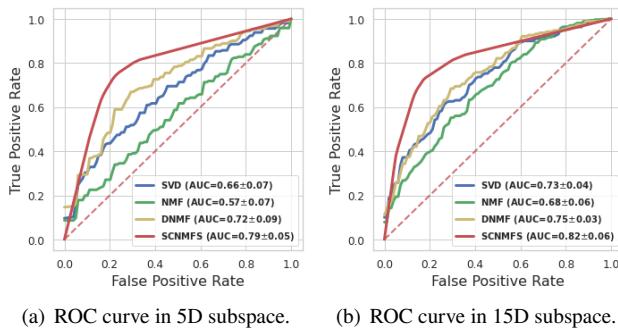


Figure 3: ROC curve of different subspaces (i.e., SVD, NMF, DNMF and SCNMFS) in 5D and 15D subspace representations on the PneumoniaMNIST dataset with the size of 300. The blue, green, yellow and red lines represent the KNN results on SVD, NMF, DNMF and SCNMFS subspaces, respectively. It shows that the performance of SCNMFS is much more stable than others including SVD in both low and high dimensions (i.e., 5D and 15D).

different subspace methods. It shows that the performance of the SVD is better than the NMF subspaces (see the blue

and green lines in Fig. 3). Interestingly, in Table 3, NMF shows its advantage compared to SVD when fewer data is available (e.g., 10 images per class) on the 14 medical datasets. This may indicate that the subspace obtained by SVD may not be as meaningful as NMF until the data size reaches a certain scale. Under appropriate dimensions, NMF can perform better than SVD. The sparse and part-based representation obtained by NMF can effectively preserve the original information in the subspace of the few-shot learning mechanism. Therefore, standard NMF can be a viable alternative to SVD when sparse subspace representation is of great interest.

4.1.3. Performance of the supervised NMF

The results of using the subspaces obtained by the supervised NMF methods (i.e., DNMF and SCNMFS) in few-shot learning are also given in Tables 2 and 3. In Table 2, compared with NMF, the performance of the supervised NMF is significantly improved, indicating the effectiveness of incorporating the label information into the decomposition process. Moreover, it also shows that the subspace generated from DNMF exhibits a slight advantage over the standard NMF and the performance of SCNMFS is more

Table 3

Comparison between the prototypical network and the few-shot learning with subspace feature representations. Note that C is the number of classes in each dataset and Dim stands for the subspace dimensions.

Data	C-way 10-shot Accuracy (%)							
	Prototypical Network		Feature Space	Subspaces	Few-shot Learning with Subspace Feature Representations (Ours)			
	Network	Feature Space			2 Dim	5 Dim	10 Dim	20 Dim
CovidCT ($C=2$)	51.00 \pm 9.78	51.00 \pm 8.17	SVD	52.67 \pm 7.72	52.33 \pm 6.51	51.33 \pm 8.97	51.33 \pm 8.97	51.33 \pm 8.97
			NMF	49.67 \pm 8.23	56.67 \pm 7.45	56.67 \pm 8.56	58.00 \pm 8.06	57.00 \pm 5.68
			DNMF	48.33 \pm 7.92	53.00 \pm 5.47	53.00 \pm 8.23	49.00 \pm 6.84	51.33 \pm 11.18
			SCNMFS	49.00 \pm 9.32	51.00 \pm 9.78	51.33 \pm 8.33	52.70 \pm 9.04	51.33 \pm 8.33
BreastCancer ($C=2$)	70.00 \pm 12.45	74.00 \pm 11.14	SVD	66.00 \pm 15.94	71.5 \pm 11.84	72.50 \pm 10.31	74.50 \pm 9.86	74.50 \pm 9.86
			NMF	66.00 \pm 14.46	74.50 \pm 9.60	76.00\pm8.60	77.00 \pm 7.14	77.00 \pm 6.40
			DNMF	66.00 \pm 11.14	74.50 \pm 9.07	70.00 \pm 12.25	72.00 \pm 9.27	71.00 \pm 7.35
			SCNMFS	69.50 \pm 12.13	70.50 \pm 12.34	70.50 \pm 12.34	71.00 \pm 12.41	70.50 \pm 12.34
PneumoniaMNIST ($C=2$)	70.0 \pm 10.25	73.5 \pm 7.09	SVD	65.00 \pm 10.72	73.00 \pm 8.12	75.00 \pm 7.75	74.00 \pm 7.68	74.00 \pm 7.68
			NMF	61.00 \pm 11.14	82.50 \pm 6.80	82.50 \pm 6.42	82.00 \pm 7.14	82.50 \pm 5.59
			DNMF	62.00 \pm 12.69	77.50 \pm 9.01	75.00 \pm 5.00	72.50 \pm 8.73	77.00 \pm 6.00
			SCNMFS	70.00 \pm 11.4	71.50 \pm 10.50	75.00 \pm 8.37	73.50 \pm 9.76	73.50 \pm 10.50
BreastMNIST ($C=2$)	59.5 \pm 13.68	62.5 \pm 10.31	SVD	53.50 \pm 11.19	58.00 \pm 10.30	60.00 \pm 14.49	62.00 \pm 9.54	62.00 \pm 9.54
			NMF	56.00 \pm 11.58	66.50 \pm 9.76	69.00 \pm 10.91	69.00\pm9.95	65.00 \pm 13.23
			DNMF	56.00 \pm 10.44	60.00 \pm 12.04	56.50 \pm 16.44	61.00 \pm 14.11	61.50 \pm 12.66
			SCNMFS	62.00 \pm 10.30	62.50 \pm 9.81	64.50 \pm 9.34	64.50 \pm 7.57	64.50 \pm 8.20
DeepDRid ($C=5$)	33.00 \pm 6.02	29.00 \pm 5.95	SVD	28.40 \pm 6.05	29.60 \pm 4.45	29.00 \pm 5.31	28.40 \pm 6.97	28.60 \pm 6.26
			NMF	30.40 \pm 6.44	32.40 \pm 3.98	33.20 \pm 4.21	34.00 \pm 4.38	34.20 \pm 5.90
			DNMF	28.40 \pm 6.05	29.20 \pm 4.40	29.00 \pm 2.72	30.00 \pm 6.07	29.40 \pm 4.90
			SCNMFS	34.00 \pm 5.29	37.00 \pm 6.08	36.60 \pm 5.59	38.00 \pm 5.87	38.40\pm7.36
BrainTumor ($C=4$)	59.75 \pm 4.25	56.75 \pm 8.95	SVD	52.50 \pm 10.19	56.00 \pm 10.26	57.75 \pm 9.78	56.75 \pm 10.25	56.75 \pm 9.75
			NMF	52.75 \pm 8.02	60.00 \pm 8.94	64.00 \pm 7.76	61.25 \pm 4.37	62.25 \pm 6.37
			DNMF	51.50 \pm 9.23	56.75 \pm 9.22	59.25 \pm 7.08	58.50 \pm 6.34	58.25 \pm 8.22
			SCNMFS	44.75 \pm 5.75	62.50 \pm 4.61	62.00 \pm 5.57	62.00 \pm 5.22	62.75\pm5.53
BloodMNIST ($C=8$)	4.88 \pm 5.38	55.5 \pm 6.94	SVD	40.00 \pm 3.58	51.12 \pm 6.62	50.62 \pm 5.65	53.62 \pm 6.43	55.12 \pm 6.97
			NMF	40.25 \pm 5.53	51.88 \pm 6.36	53.50 \pm 5.09	56.50 \pm 5.83	57.25 \pm 4.14
			DNMF	39.50 \pm 6.30	48.87 \pm 6.41	51.12 \pm 5.08	51.87 \pm 5.51	49.50 \pm 5.48
			SCNMFS	39.38 \pm 6.13	54.38 \pm 6.60	56.12 \pm 4.89	56.12 \pm 5.17	56.75\pm6.35
DermaMNIST ($C=7$)	29.14 \pm 5.83	33.00 \pm 6.31	SVD	21.86 \pm 5.61	28.57 \pm 6.23	31.14 \pm 6.38	33.00 \pm 7.65	33.57 \pm 6.00
			NMF	24.43 \pm 5.36	31.43 \pm 4.69	35.43 \pm 5.78	34.14 \pm 4.58	35.86 \pm 6.04
			DNMF	23.00 \pm 5.17	30.71 \pm 6.55	32.14 \pm 4.88	34.00 \pm 5.14	31.00 \pm 8.31
			SCNMFS	23.57 \pm 5.93	33.29 \pm 4.70	35.86\pm5.05	35.14 \pm 5.20	35.57 \pm 3.86
OCTMNIST ($C=4$)	25.0 \pm 9.22	28.25 \pm 6.32	SVD	27.50 \pm 5.00	25.25 \pm 6.84	26.75 \pm 5.25	27.75 \pm 6.84	28.00 \pm 5.6
			NMF	28.25 \pm 6.90	31.50 \pm 6.04	33.00 \pm 7.81	34.00 \pm 5.15	36.50\pm3.74
			DNMF	26.50 \pm 7.67	26.50 \pm 7.84	27.25 \pm 7.78	33.00 \pm 6.87	31.25 \pm 7.35
			SCNMFS	26.00 \pm 5.94	28.75 \pm 7.35	29.25 \pm 8.44	28.75 \pm 8.46	30.00 \pm 7.83
OrganAMNIST ($C=11$)	55.18 \pm 3.57	63.73 \pm 3.91	SVD	36.64 \pm 3.81	52.55 \pm 4.02	61.18 \pm 3.90	62.82 \pm 3.76	63.27 \pm 4.15
			NMF	36.36 \pm 2.73	55.36 \pm 4.31	62.27 \pm 4.01	65.27 \pm 3.94	65.18 \pm 5.18
			DNMF	35.18 \pm 3.20	51.73 \pm 5.98	61.09 \pm 3.94	63.09 \pm 3.62	64.27 \pm 5.32
			SCNMFS	30.18 \pm 3.44	48.91 \pm 2.93	58.91 \pm 3.14	61.00 \pm 2.92	61.45 \pm 1.92
OrganCMNIST ($C=11$)	49.64 \pm 3.28	60.18 \pm 6.27	SVD	27.82 \pm 2.88	43.82 \pm 4.67	53.27 \pm 4.49	58.73 \pm 6.09	59.55 \pm 6.05
			NMF	28.09 \pm 3.51	43.73 \pm 5.55	56.64 \pm 2.76	61.36 \pm 4.25	63.18\pm4.69
			DNMF	27.27 \pm 3.98	41.27 \pm 5.84	53.91 \pm 4.26	60.55 \pm 4.38	58.64 \pm 4.92
			SCNMFS	27.18 \pm 3.89	43.82 \pm 3.72	53.82 \pm 4.06	56.36 \pm 4.21	56.09 \pm 4.05
OrganSMNIST ($C=11$)	38.0 \pm 4.51	43.82 \pm 4.37	SVD	24.45 \pm 2.92	38.36 \pm 6.73	41.18 \pm 5.46	42.09 \pm 4.64	42.91 \pm 4.44
			NMF	26.00 \pm 3.73	40.36 \pm 3.82	42.27 \pm 3.69	45.73 \pm 3.02	45.18 \pm 2.41
			DNMF	25.00 \pm 3.28	37.64 \pm 4.31	40.36 \pm 3.75	40.82 \pm 4.90	41.27 \pm 5.91
			SCNMFS	24.27 \pm 3.07	37.91 \pm 4.86	41.09 \pm 4.74	41.91 \pm 4.70	42.18 \pm 4.71
PathMINST ($C=9$)	36.78 \pm 4.02	42.56 \pm 3.88	SVD	26.11 \pm 2.91	36.67 \pm 3.65	40.44 \pm 3.52	42.67 \pm 4.16	41.89 \pm 4.87
			NMF	28.00 \pm 4.89	39.89 \pm 3.28	44.22 \pm 3.74	45.78 \pm 2.85	45.56 \pm 2.58
			DNMF	27.33 \pm 4.51	35.00 \pm 4.01	40.67 \pm 4.13	41.33 \pm 4.73	42.11 \pm 4.70
			SCNMFS	26.22 \pm 3.19	40.56 \pm 3.11	41.89 \pm 3.52	43.89\pm3.58	42.56 \pm 4.28
TissueMNIST ($C=8$)	25.5 \pm 3.67	24.5 \pm 3.63	SVD	20.50 \pm 2.32	22.50 \pm 3.58	22.37 \pm 3.86	24.12 \pm 3.92	23.75 \pm 3.01
			NMF	20.88 \pm 4.03	26.88 \pm 4.55	26.25 \pm 3.62	25.25 \pm 3.82	26.88 \pm 3.17
			DNMF	19.25 \pm 3.84	22.62 \pm 4.73	23.12 \pm 3.55	23.75 \pm 2.37	22.25 \pm 4.96
			SCNMFS	23.38 \pm 4.58	28.00 \pm 3.54	28.50 \pm 3.44	28.75\pm3.31	27.50 \pm 3.16

robust compared with the original feature space and other subspaces. This is also in line with the results of DNMF (yellow line) and SCNMFS (red line) shown in Fig 3. Both DNMF and SCNMFS perform better than NMF in different

subspace dimensions. Compared to SVD, the effect of SCNMFS is more stable. In Table 3, the performance of the supervised NMF method is not that impressive compared to the standard NMF. This tells us that incorporating label

signals into the decomposition process of NMF will be affected by the data size. Considering the real case where the number of images in most medical datasets are hundreds available, the subspace generated by the supervised NMF method is quite a competitive solution.

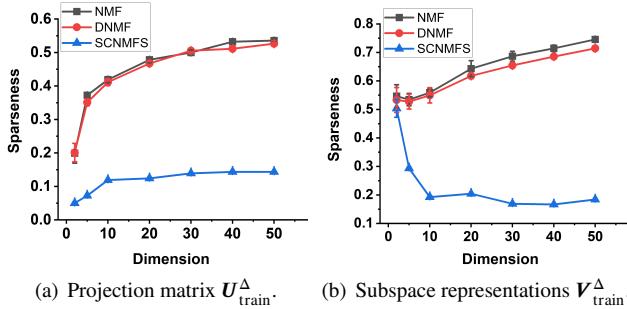


Figure 4: Sparsity analysis of NMF, DNMF and SCNMFS by investigating the projection matrices $U_{\text{train}}^{\Delta}$ and the subspace representations $V_{\text{train}}^{\Delta}$ with different subspace dimensions. Dataset PneumoniaMNIST of size 300 is used in this experiment.

Fig. 4 presents a further comparison between NMF and the supervised NMF methods in terms of sparsity of the generated subspaces. In detail, the sparsity comparison is conducted by investigating the projection matrix $U_{\text{train}}^{\Delta}$ and the subspace representation $V_{\text{train}}^{\Delta}$ generated by NMF, DNMF and SCNMFS on the PneumoniaMNIST dataset. Note that compared to the standard NMF, which can generate sparse representation in subspace, DNMF and SCNMFS are different from it due to the way of combining labels. Fig. 4 shows that the subspace generated by SCNMFS is less sparse. This occurs because SCNMFS consolidates data with identical labels into a single point (signifying that all elements with the same label in the subspace exhibit roughly equal activity), while DNMF organizes data from each class into distinct clusters along the axis. The limited correlation regarding the matrices $U_{\text{train}}^{\Delta}$ and $V_{\text{train}}^{\Delta}$ produced by these methods further reinforces this trend, as depicted in Fig. 5.

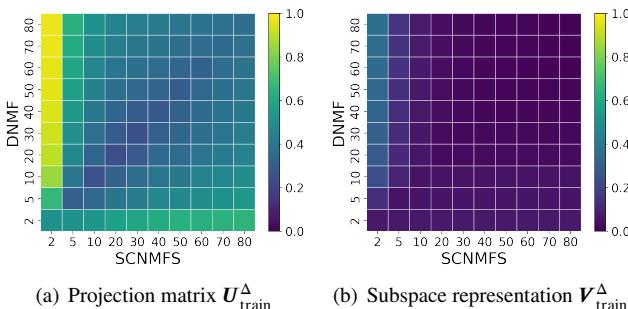


Figure 5: Correlation analysis regarding projection matrices $U_{\text{train}}^{\Delta}$ and subspace representations $V_{\text{train}}^{\Delta}$ generated by DNMF and SCNMFS with different subspace dimensions. Dataset PneumoniaMNIST is used.

4.1.4. Discussion

The results reported above support the initial finding that subspace-based few-shot learning is a promising direction in data-driven few-shot learning in medical imaging. More importantly, NMF and its variations can be a viable alternative to SVD. The basic vectors in the projection matrix of NMF preserve more suitable information than the principal components in SVD. For the few-shot learning model, e.g., the prototypical network [14], the effect is not as obvious as expected in medical imaging. Moreover, data augmentation could even introduce negative effect in medical imaging.

For generally hundreds of data that are available in medical scenarios, the supervised NMF methods (e.g., DNMF and SCNMFS) are superior to the unsupervised subspace methods (e.g., SVD and NMF) in classification tasks, and the results of SCNMFS are more robust than DNMF in different subspace dimensions. Besides, we found that, from the data characteristics point of view, the non-negative representation in NMF will bring sparsity and enhance the ability of data representation in subspaces. Appropriate label signal combination (e.g., DNMF and SCNMFS) will further boost the discriminative ability of the subspace, which is particularly important for classification tasks with limited data.

Another feature of NMF is the part-based representation in the subspace. The non-negativity constraints are compatible with the intuitive notion of combining parts to form a whole. In the next subsection we further show how part-based representations of NMF can be improved by the supervised NMF methods via locating the discriminative image regions in medical images.

4.2. Validation by Class Activation Map and Discussion

Finally, to further understand/validate the properties of the above tested different subspace methods, we develop a CAM method, see Appendix A for the detail, based on the one proposed in [47] to illustrate the discriminative image regions these subspace methods generated.

Fig. 6 gives the CAM results for different subspace methods (i.e., SVD, NMF, DNMF and SCNMFS) on the breast ultrasound image dataset [36] with a training set size of 300. In Fig. 6, the result for feature space in the first column of each panel is the one obtained without dimensionality reduction, while the rest of the columns in each panel are the results of the subspace representations in 3 selected dimensions (i.e., 5D, 10D and 15D for different rows in Fig. 6). The discriminant areas are highlighted in red. The probability score under each CAM is the predication result (to be cancer) of each subspace method.

It is observed that when the prediction is correct, there is no significant disparity in the CAMs between the results of the subspaces and the feature space, but the discriminant regions inferred from the subspaces are more centralised than the feature space, indicating the virtue of exploiting subspaces. For the incorrect predictions, the CAMs by NMF vary as the dimension increases, and the same tendency is also appeared in the CAMs by DNMF, see the second and

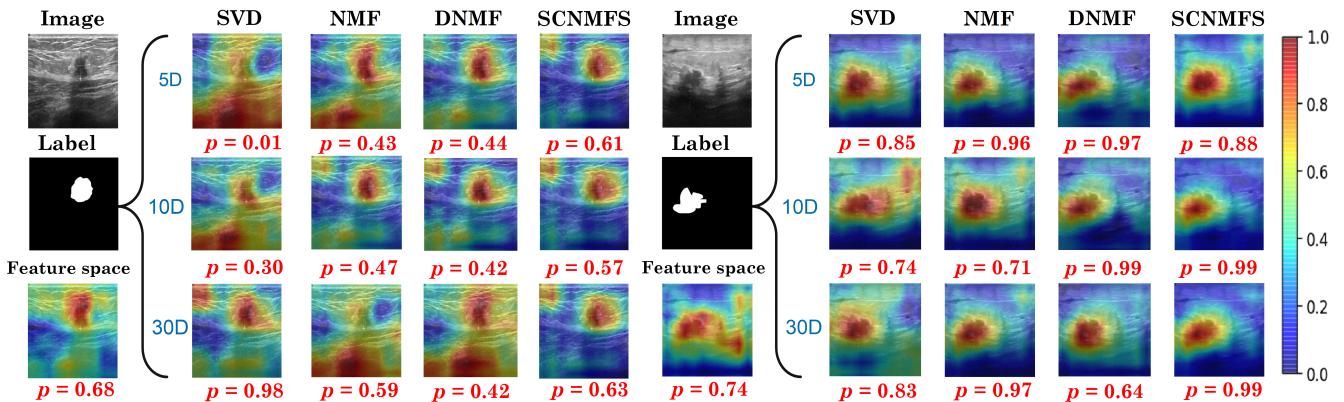


Figure 6: CAM validation for different subspace methods (i.e., SVD, NMF, DNMF and SCNMFS) on the breast ultrasound image dataset. Results of each method for 3 different dimensions (i.e., 5D, 10D and 15D) are considered. The discriminant regions, i.e., those regions to which the output decision is most sensitive to, appear as red. Results under ‘Feature space’ are the ones obtained without dimensionality reduction. Predicted cancer class scores by each method are shown below the individual CAMs; particularly, $p > 0.5$ indicates the prediction is correct. Compared with the feature space and other subspaces, we see that SCNMFS achieves the best in positioning the lesion area and this localisation of discriminative features is consistent across dimensions.

third columns in the left panel in Fig 6. This demonstrates that although DNMF is a supervised variation of NMF, it does not fundamentally change the prediction performance compared to NMF. In contrast, the supervised NMF method SCNMFS performs excellently, i.e., the discriminative regions in its CAMs are stable and invariant to different dimensions. In contrast, SVD suffers from incorrect prediction from its eigenvectors; moreover, as shown in Fig. 6, the discriminant regions in its CAMs shift as the dimension increases, indicating that the key information it uses for classification prediction is preserved in its eigenvectors, including the ones corresponding to small eigenvalues. This tendency is more obvious when the model with SVD makes incorrect predictions (i.e., the CAMs with $p < 0.5$); see the left panel in Fig. 6.

The results from the subspaces experiments empirically demonstrate that the part-based representation can be enhanced in supervised NMF, thereby contributing to lesion detection. The vector quantization property in SCNMFS ensures stability across different dimensions. Even with limited data, the discriminant areas achieved by SCNMFS closely align with the ground truth compared to other methods. Furthermore, the localized diagnosis by the CAM method underscores the effectiveness of the part-based representation provided by supervised NMF in identifying discriminative information in medical imaging.

5. Conclusion

In this paper, we explored the innovative application of NMF (non-negative matrix factorization) and its supervised variations (i.e., DNMF and SCNMFS) as tools to gain insights regarding the properties of subspace features extracted from deep neural networks pre-trained on large natural image datasets, adapted for medical imaging in few-shot learning. In our experiments, we found that for few-shot learning with limited datasets (e.g., a few hundred images),

data augmentation methods are not as useful as widely believed, while reducing dimensionality could bring a suitable solution in the case that the magnitude of the data is smaller than the feature dimension. Using this insight, we introduced NMF and supervised NMF as alternatives to replace the common dimension reduction method PCA/SVD. Our suggestion will alleviate the limitations of using PCA/SVD in multimodal data and shed lights on the link between non-negativity and sparsity in few-shot learning.

By carrying out the experiments on 14 medical datasets including 11 distinct disease and 4 image acquisition modalities, our work exposes considerable additional fundamental and surprising findings as follows. I) Dimensionality reduction yields a constant performance advantage in the few-shot learning regime. II) The NMF-based representation, including its supervised variants (i.e., DNMF and SCNMFS), serves as a feasible alternative to SVD-based subspaces. SVD suffers from incorrect predictions from its eigenvectors for classification tasks with insufficient data. Moreover, utilizing the SCNMFS subspace instead of PCA/SVD-based variance-preserving dimensionality reduction yields significant performance improvements, even at extremely low dimensions. III) The combination of label information in supervised NMF greatly impacts the interpretation of the subspace, e.g. SCNMFS gives a way of constructing a discriminative subspace with vector quantization rather than preserving the distribution as DNMF does. IV) The subspace sparsity does not considerably improve classification performance due to data scarcity and may even decrease the accuracy in the few-shot learning framework, yet non-negativity is more worthwhile to pursue given its part-based representation capacity. V) Part-based representations obtained from NMF-based subspaces can facilitate localization diagnosis in medical imaging, especially with limited medical images.

In this study, we adopted two supervised NMF methods, i.e., DNMF and SCNMFS, to investigate the effect of label information in generating discriminative subspaces. We acknowledge the significant contributions of NMF and its supervised variants, which encourage us to approach the concept of subspace from a different perspective. This also drives our investigation into the possibilities of a novel data-based few-shot learning approach, leveraging the features preserved within the NMF subspace for few-shot learning in medical imaging. During our study, we encountered several limitations that warrant consideration. In addition to choosing a modest subspace dimension, our experiments focused solely on exploring subspaces using image data. However, in medical scenarios, multi-modal data (such as images and text) are often associated with the same disease. Future research could explore integrating multi-modal data into the analysis. This could involve investigating novel NMF variants and exploring alternative loss functions to enhance the stability and effectiveness of subspace representation in medical imaging analysis, thereby improving classification performance.

6. Acknowledgments

MN's contribution to this work is funded by Grant EP/S000356/1, Artificial and Augmented Intelligence for Automated Scientific Discovery, Engineering and Physical Sciences Research Council (EPSRC), UK.

A. Class Activation Map Generation

Algorithm 2 CAM generation with subspaces

Require: \mathbf{d} , f_{Θ_1} , f_{Θ_2} and $\mathbf{U}_{\text{train}}^{\Delta}$

Ensure: Class activation map \mathbf{R}

- 1: Obtain the feature map \mathbf{M} and form $\bar{\mathbf{M}}$;
 - 2: Obtain f_{Θ_2} weight matrix $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_C) \in \mathbb{R}^{k \times C}$
 - 3: Obtain the predict label i for image \mathbf{d} from f_{Θ_2} and the corresponding vector \mathbf{w}_i ;
 - 4: Compute the weight vector $\mathbf{x}' = \mathbf{U}_{\text{train}}^{\Delta} \mathbf{w}_i$;
 - 5: Form $\mathbf{R}' \in \mathbb{R}^{h \times w}$ by reshaping $(\mathbf{x}')^\top \bar{\mathbf{M}} \in \mathbb{R}^{h \cdot w}$;
 - 6: Resize \mathbf{R}' to the size of the test image and generate \mathbf{R} .
-

This Appendix shows the steps of calculating CAMs for our studied few-shot learning framework in medical imaging. Let $\mathbf{M} \in \mathbb{R}^{c \times h \times w}$ be the feature map produced by the last convolutional layer of the pre-trained network f_{Θ_1} corresponding to the input image $\mathbf{d} \in \mathcal{D}$, where c represents the number of channels, and h, w are the size of the feature map in each channel. We flatten the channel-wise feature maps, and then \mathbf{M} is changed into a matrix $\bar{\mathbf{M}}, \bar{\mathbf{M}} \in \mathbb{R}^{c \times (h \cdot w)}$. The pre-trained model f_{Θ_1} transfers the input image \mathbf{d} into a vector $\mathbf{x} \in \mathbb{R}^c$ in the feature space, i.e., $\mathbf{x} \leftarrow f_{\Theta_1}(\mathbf{d})$. With the selected rank k and subspace method Δ , the corresponding subspace representation $\mathbf{v}^{\Delta} \in \mathbb{R}^k$ is generated by using the projection matrix $\mathbf{U}_{\text{train}}^{\Delta} \in \mathbb{R}^{c \times k}$. Let

$\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_C) \in \mathbb{R}^{k \times C}$ be the weight matrix of f_{Θ_2} . Since f_{Θ_2} is a fully connected layer, the predicted process for the input image \mathbf{d} can be regarded as $\mathbf{z} \approx (\mathbf{v}^{\Delta})^\top \mathbf{W}$. Assume class i is predicted (i.e., \mathbf{z} 's largest component is the i -th entry) for image \mathbf{d} . Then the weight vector \mathbf{w}_i is used to generate $\mathbf{x}' = \mathbf{U}_{\text{train}}^{\Delta} \mathbf{w}_i \in \mathbb{R}^c$, which will be used as a weight vector for $\bar{\mathbf{M}}$. The initial CAM result is formed by reshaping $(\mathbf{x}')^\top \bar{\mathbf{M}} \in \mathbb{R}^{h \cdot w}$ to a matrix \mathbf{R}' , $\mathbf{R}' \in \mathbb{R}^{h \times w}$. The final CAM result \mathbf{R} is drawn by resizing \mathbf{R}' to the size of the test image by interpolation. The above steps are summarised in Algorithm 2.

References

- [1] J. Kaplan, S. McCandlish, T. J. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *ArXiv*, vol. abs/2001.08361, 2020.
- [2] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," *Classification in BioApps*, pp. 323–350, 2018.
- [3] A. Babayan, M. Erbey, D. Kumral, J. D. Reinelt, A. M. Reiter, J. Röbbig, H. L. Schaare, M. Uhlig, A. Anwander, P.-L. Bazin *et al.*, "A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults," *Scientific data*, vol. 6, no. 1, pp. 1–21, 2019.
- [4] K. Yan, X. Wang, L. Lu, and R. M. Summers, "Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations," *arXiv preprint arXiv:1710.01766*, 2017.
- [5] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, pp. 1–10, 2020.
- [6] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier, "Well-calibrated regression uncertainty in medical imaging with deep learning," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 393–412.
- [7] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, pp. 1–40, 2016.
- [8] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [9] J. Wang, X. Du, K. Farrahi, and M. Niranjan, "Deep cascade learning for optimal medical image feature representation," *Machine Learning for Healthcare (MLHC)*, 2022.
- [10] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.
- [11] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.
- [12] A. Nichol and J. Schulman, "Reptile: a scalable metalearning algorithm," *arXiv preprint arXiv:1803.02999*, vol. 2, no. 3, p. 4, 2018.
- [13] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.
- [14] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [15] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [16] J. Xu, W. An, L. Zhang, and D. Zhang, "Sparse, collaborative, or nonnegative representation: which helps pattern classification?" *Pattern Recognition*, vol. 88, pp. 679–688, 2019.
- [17] D. Papailiopoulos, A. Dimakis, and S. Korokythakis, "Sparse pca through low-rank approximations," in *International Conference on Machine Learning*. PMLR, 2013, pp. 747–755.

- [18] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, “Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [19] C. Simon, P. Koniusz, R. Nock, and M. Harandi, “Adaptive subspaces for few-shot learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4136–4145.
- [20] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, “Pcanet: A simple deep learning baseline for image classification?” *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2015.
- [21] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
- [22] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
- [23] M. Babaee, S. Tsoukalas, M. Babaee, G. Rigoll, and M. Datcu, “Discriminative nonnegative matrix factorization for dimensionality reduction,” *Neurocomputing*, vol. 173, pp. 212–223, 2016.
- [24] X. Cai and F. Sun, “Supervised and constrained nonnegative matrix factorization with sparseness for image representation,” *Wireless Personal Communications*, vol. 102, no. 4, pp. 3055–3066, 2018.
- [25] D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. MIT Press, 2001. [Online]. Available: <https://proceedings.neurips.cc/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf>
- [26] A. Dong, Z. Li, and Q. Zheng, “Transferred subspace learning based on non-negative matrix factorization for eeg signal classification,” *Frontiers in Neuroscience*, vol. 15, 2021.
- [27] J. Leuschner, M. Schmidt, P. Fernsel, D. Lachmund, T. Boskamp, and P. Maass, “Supervised non-negative matrix factorization methods for maldi imaging applications,” *Bioinformatics*, vol. 35, no. 11, pp. 1940–1947, 2019.
- [28] Z. Chen, S. Jin, R. Liu, and J. Zhang, “A deep non-negative matrix factorization model for big data representation learning,” *Frontiers in Neurorobotics*, p. 93, 2021.
- [29] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang, “Constrained nonnegative matrix factorization for image representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1299–1311, 2011.
- [30] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.
- [31] A. Janowczyk and A. Madabhushi, “Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases,” *Journal of Pathology Informatics*, vol. 7, 2016.
- [32] J. Cheng, “brain tumor dataset,” Apr 2017. [Online]. Available: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427/5
- [33] J. Zhao, Y. Zhang, X. He, and P. Xie, “Covid-ct-dataset: a ct scan dataset about covid-19,” *arXiv preprint arXiv:2003.13865*, 2020.
- [34] R. Liu, X. Wang, Q. Wu, L. Dai, X. Fang, T. Yan, J. Son, S. Tang, J. Li, Z. Gao *et al.*, “Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge,” *Patterns*, p. 100512, 2022.
- [35] A. Acevedo, A. Merino, S. Alférez, Á. Molina, L. Boldú, and J. Rodellar, “A dataset of microscopic peripheral blood cell images for development of automatic recognition systems,” *Data in Brief*, ISSN: 23523409, Vol. 30,(2020), 2020.
- [36] W. Al-Dhabayani, M. Gomaa, H. Khaled, and A. Fahmy, “Dataset of breast ultrasound images,” *Data in Brief*, vol. 28, p. 104863, 2020.
- [37] P. Tschandl, C. Rosendahl, and H. Kittler, “The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Scientific Data*, vol. 5, no. 1, pp. 1–9, 2018.
- [38] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic),” in *2018 IEEE 15th International symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.
- [39] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, “Identifying medical diagnoses and treatable diseases by image-based deep learning,” *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
- [40] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C. W. Fu, X. Han, P.-A. Heng, J. Hesser *et al.*, “The liver tumor segmentation benchmark (LiTS),” *arXiv preprint arXiv:1901.04056*, 2019.
- [41] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber *et al.*, “Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study,” *PLoS Medicine*, vol. 16, no. 1, p. e1002730, 2019.
- [42] A. Woloshuk, S. Khochare, A. F. Almulhim, A. T. McNutt, D. Dean, D. Barwinska, M. J. Ferkowicz, M. T. Eadon, K. J. Kelly, K. W. Dunn *et al.*, “In situ classification of cell types in human kidney tissue using 3d nuclear staining,” *Cytometry Part A*, vol. 99, no. 7, pp. 707–721, 2021.
- [43] V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter, “Annotated high-throughput microscopy image sets for validation,” *Nature Methods*, vol. 9, no. 7, pp. 637–637, 2012.
- [44] A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, “Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks,” in *Medical Imaging 2014: Digital Pathology*, vol. 9041. SPIE, 2014, p. 904103.
- [45] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, “Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification,” *arXiv preprint arXiv:2110.14795*, 2021.
- [46] Pytorch, “Pytorch, forward and backward function hooks—pytorch documentation.” [Online]. Available: https://pytorch.org/tutorials/beginner/former_torchies/nnft_tutorial.html#forward-and-backward-function-hooks
- [47] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.
- [48] K. Alomar, H. I. Aysel, and X. Cai, “Data augmentation in classification and segmentation: a survey and new strategies,” *Journal of Imaging*, vol. 9, 2023. [Online]. Available: <https://doi.org/10.3390/jimaging9020046>
- [49] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, “The omniglot challenge: a 3-year progress report,” *Current Opinion in Behavioral Sciences*, vol. 29, pp. 97–104, 2019.
- [50] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.