

# Few-shot Learning for Inference in Medical Imaging with Subspace Feature Representations

Jiahui Liu<sup>\*,1</sup>, Keqiang Fan<sup>\*,1</sup>, Xiaohao Cai and Mahesan Niranjan

*School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK*

---

## ARTICLE INFO

**Keywords:**  
 Medical imaging  
 Few-shot learning  
 Classification  
 Discriminant analysis  
 PCA  
 Non-negative matrix factorization  
 Dimensionality reduction

---

## ABSTRACT

Unlike the field of visual scene recognition where tremendous advances have taken place due to the availability of very large datasets to train deep neural networks, inference from medical images is often hampered by the fact that only small amounts of data may be available. When working with very small dataset problems, of the order of a few hundred items of data, the power of deep learning may still be exploited by using a model pre-trained on natural images as a feature extractor and carrying out classic pattern recognition techniques in this feature space, the so-called few-shot learning problem. In regimes where the dimension of this feature space is comparable to or even larger than the number of items of data, dimensionality reduction is a necessity and is often achieved by principal component analysis, i.e., singular value decomposition (SVD). In this paper, noting the inappropriateness of using SVD for this setting, we usher in and explore two alternatives based on discriminant analysis and non-negative matrix factorization (NMF). Using 14 different datasets spanning 11 distinct disease types, we demonstrate that discriminant subspaces at low dimensions achieve significant improvements over SVD-based subspaces and the original feature space. We also show that NMF at modest dimensions is a competitive alternative to SVD in this setting.

---

## 1. Introduction

Impressive empirical performances have been reported in the field of computer vision in recent years, starting from a step improvement reported in the ImageNet challenge [1]. This and subsequent work has used very large neural network architectures, notably their depth, with parameter estimation carried out using equally large datasets. It is common in current computer vision literature to train models with tens of millions of parameters and use datasets of similar sizes. Much algorithmic development to control the complexity of such massive models and to incorporate techniques to handle systematic variability has been developed. Our curiosity about mammalian vision [2, 3] and commercial applications such as self-driving cars and robot navigation [4, 5] has driven the computer vision field. The interest in automatic diagnosis has reached a level of comparing artificial intelligence-based methods against human clinicians [6, 7]. However, compared with natural images, the application of deep learning in the medical domain poses more challenges, such as causality [8], uncertainty [9], and the need to integrate clinical information along with features extracted from images [10]. A particular issue with image-based inference in the medical field is data availability [11]. Often, the number of images available in the medical domain is orders of magnitude smaller than what is state-of-the-art in computer vision. Compared with other domains, due to privacy concerns and the prevalence of adverse medical conditions, most of the medical datasets only contain thousands or even hundreds of images, such as brain imaging [12].

The focus of this paper is on data sparsity/scarcity. Naturally, if we had access to hundreds of thousands of labelled medical images, as might be the case with X-rays and optometry, training a deep neural network from scratch using all the recent methodological advances is the way forward. When the number of images is in the thousands, the strategy of transfer learning is suitable for the medical data by fine-tuning the weights generated from pre-trained natural images. While the scheme is appealing, available empirical evidence for transfer learning is contradictory in the medical field. For example, on a chest X-rays problem, Raghu *et al.* [13] found no significant improvement with the popular ResNet trained on ImageNet as source architecture; more positive results are reported for endoscopy image recognition [14].

---

\*Corresponding author

✉ j14f19@soton.ac.uk (J. Liu); k.fan@soton.ac.uk (K. Fan); x.cai@soton.ac.uk (X. Cai); mn@ecs.soton.ac.uk (M. Niranjan)  
 ORCID(s):

<sup>1</sup>Equal contribution

Another example may be the weakly supervised learning methods [15], whose performance is yet to be seen in medical diagnosis.

Our interest is in a regime of even smaller amounts of data than is needed to fine-tune a pre-trained model with transfer learning. This regime is referred to as “few-shot learning” [16, 17, 18, 19], and is appropriate for dataset sizes of the order of a few hundred or even down to a few tens [20, 21]. Few-shot learning works can be divided into different categories – data, model and algorithm [19]. Most contemporary few-shot learning techniques rely on models and algorithms with fine-tuned parameters based on available data [22, 23]. Data augmentation technology and manifold space [24] have also drawn some attention. Unlike these methods, we in this paper explore few-shot problems from the traditional machine learning perspective by using a pre-trained deep neural network as a feature extractor. In detail, each image is mapped into a fixed dimensional feature space, the dimensions of which, say  $M$ , are defined by the number of neurons in the penultimate fully connected layer of the network, typically 512 or 1024 for the popular architectures. Then we are in a regime where the number of items of data, say  $N$ , is comparable to or even smaller than the dimension of the feature space (i.e., the  $N < M$  problem in statistical inference language [25]), necessitating techniques for dimensionality reduction.

Subspace methods for reducing the dimensionality of data have a long and rich history. They fall under the group of methods known as structured low-rank approximation methods [26, 27, 28]. The basic intuition is a data matrix,  $\mathbf{Y} \in \mathbb{R}^{N \times M}$ , consisting of  $N$  items of data in  $M$  dimensional features, is usually not full rank. This is due to correlations along either of the axes. In the medical context, profiles of patients (i.e. data) may show strong similarities. Along the features axis, some features that have been gathered may be derivable from others. In these situations, we can find low-rank approximations by factorising  $\mathbf{Y}$ , and additionally impose structural constraints on the factors either from prior knowledge or for mathematical convenience. Popular approaches like principal component analysis (PCA) [29] and non-negative matrix factorization (NMF) [30, 31] impose orthogonality and non-negativity constraints on the factors, respectively. Returning to few-shot learning with pre-trained deep neural network as feature extractors and encountering  $N < M$  problems, pattern recognition problems are known to suffer the “curse of dimensionality”. Hence dimensionality reduction techniques are required. The most popular technique used hitherto in the literature is PCA, implemented via singular value decomposition (SVD) [21, 32, 33, 34]. Despite its popularity, PCA has a fundamental weakness in that it is a variance-preserving low-rank approximation technique, more suitable for data that is uni-modal and Gaussian distributed. In the case of classification problems, however, the feature space is necessarily multi-modal with at least as many modes as the number of classes in the problem.

The basic premise of this work is the need for dimensionality reduction in the feature space and that SVD ignores multi-modal data structure. We, for the first time, usher in and explore two alternatives – *discriminant analysis (DA) subspace* and *NMF subspace* – to SVD in few-shot learning on medical imaging with multi-modal structure in the data. The DA subspace introduces the well-known Fisher linear discriminant analysis (FDA) and its multi-dimensional extensions [35]. The NMF [30] and the supervised NMF (SNMF) [36] (where class label information can be injected into the factorization loss function) subspaces focus on the part-based representation with sparsity. A detailed comparison between these subspace representations, including feature selection techniques [37], is conducted. Validating on 14 datasets spanning 11 medical classification tasks with four distinct imaging modalities, we achieve statistically significant improvements in classification accuracy in the subspaces compared to the original high-dimensional feature space, with persuasive results on DA and NMF subspaces as viable alternatives to SVD.

The remainder of this paper is organized as follows. In the next section, we mainly recall the subspace representation methods, i.e., SVD, DA and NMF subspaces. The few-shot learning methodology/scheme on subspace feature representations including the experimental settings in sufficient detail to facilitate reproduction of the work is provided in Section 3. In Section 4, we give succinct descriptions of the datasets used. Section 5 presents the key results of the experimental work. A further discussion is conducted in Section 6, followed by conclusion in Section 7. Some additional details regarding method derivations and extra results are provided in Appendix.

## 2. Subspace Representation

### 2.1. Basic Notations

Given  $N$  samples  $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iM})^\top \in \mathbb{R}^M$ ,  $1 \leq i \leq N$ , we form a data matrix  $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)^\top \in \mathbb{R}^{N \times M}$ , where  $M$  is the number of features of every sample. Suppose that these  $N$  samples belong to  $C$  different classes, namely  $\Lambda_j$ , and their cardinality  $|\Lambda_j| = N_j$ ,  $1 \leq j \leq C$ . Let  $\mathbf{y}_k^j$  represent the  $k$ -th sample in class  $\Lambda_j$ . Clearly,

$N = \sum_{j=1}^C N_j$ ,  $\Lambda_j = \{\mathbf{y}_k^j\}_{k=1}^{N_j}$  and  $\{\mathbf{y}_i\}_{i=1}^N = \bigcup_{j=1}^C \{\mathbf{y}_k^j\}_{k=1}^{N_j}$ . Let  $\bar{\mathbf{y}}$  and  $\bar{\mathbf{y}}_j$  respectively be the mean of the whole samples and the samples in class  $j$ , i.e.,  $\bar{\mathbf{y}} = \frac{1}{N} \sum_{i=1}^N \mathbf{y}_i$ ,  $\bar{\mathbf{y}}_j = \frac{1}{N_j} \sum_{\mathbf{y} \in \Lambda_j} \mathbf{y}$ ,  $1 \leq j \leq C$ .

Let  $\mathbf{S}_W^j$  represent the intra-class scatter for class  $j$ , i.e.,

$$\mathbf{S}_W^j = \sum_{k=1}^{N_j} (\mathbf{y}_k^j - \bar{\mathbf{y}}_j)(\mathbf{y}_k^j - \bar{\mathbf{y}}_j)^\top, \quad 1 \leq j \leq C. \quad (1)$$

Then the inter- and intra-class scatters, denoted as  $\mathbf{S}_B$  and  $\mathbf{S}_W$ , respectively, read

$$\mathbf{S}_B = \sum_{j=1}^C (\bar{\mathbf{y}}_j - \bar{\mathbf{y}})(\bar{\mathbf{y}}_j - \bar{\mathbf{y}})^\top, \quad \mathbf{S}_W = \sum_{j=1}^C \mathbf{S}_W^j. \quad (2)$$

Specifically, for the binary case, i.e.,  $C = 2$ , we also name  $\tilde{\mathbf{S}}_B$  and  $\tilde{\mathbf{S}}_W$  as the inter- and intra-class scatters, i.e.,

$$\tilde{\mathbf{S}}_B = \mathbf{s}_b \mathbf{s}_b^\top, \quad \tilde{\mathbf{S}}_W = \beta \mathbf{S}_W^1 + (1 - \beta) \mathbf{S}_W^2, \quad (3)$$

where  $\mathbf{s}_b = \bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2$  and  $\beta = (N_2 - 1)/(N_1 + N_2 - 2)$ .

## 2.2. Feature Selection

Feature selection has been used in medical imaging [38]. It is the process of extracting a subset of relevant features by eliminating redundant or unnecessary information for model development [37]. There are several types of feature selection techniques, including supervised [39, 40], semi-supervised [41], and unsupervised methods [42]. For example, the Boruta algorithm [40], one of the supervised feature selection methods, selects features by shuffling features of the data and calculating the feature correlations based on classification loss. The approach has also been used to classify medical images [38].

## 2.3. Singular Value Decomposition

SVD is the most common type of matrix decomposition, which can decompose either a square or rectangle matrix. The SVD of the matrix  $\mathbf{Y}$  can be represented as  $\mathbf{Y} = \mathbf{U}\Sigma\mathbf{V}^\top$ , where  $\mathbf{U} \in \mathbb{R}^{N \times N}$  and  $\mathbf{V} \in \mathbb{R}^{M \times M}$  are orthogonal matrices, and  $\Sigma \in \mathbb{R}^{N \times M}$  is a diagonal matrix whose diagonal consists of singular values. The singular values are generally ordered and it is well known that in most real-world problems they reduce quickly to zero, i.e., typically the first 10% or even 1% of the largest singular values could account for more than 99% of the sum of all the singular values. Therefore, the singular vectors corresponding to the top  $p \ll \min\{M, N\}$  largest singular values compose the transformation matrix for the most representative subspace. Meanwhile, the variance preserving property of SVD is extremely effective in data compression and widely employed in deep learning tasks, especially when the data is unimodal. For example, SVD has been used to compress features taken at different layers to compare learning dynamics across layers as well as across models [32].

## 2.4. Discriminant Subspaces

It is usually possible to design logic based on the statistics of a design set that achieves a very high recognition rate if the original set of features is well chosen. Discriminant vectors for DA can reduce the error rate and solve the discrimination portion of the task [35]. Since the discriminant vector transformation aims to reduce dimensionality while retaining discriminatory information, sophisticated pattern recognition techniques that were either computationally impractical or statistically insignificant in the original high-dimensional space could be possible in the new and low-dimensional space. The intuitive assumption is that features based on discrimination are better than that based on fitting or describing the data. In what follows, we present different approaches of obtaining discriminant vectors for multiclass and binary classification problems.

### 2.4.1. Multiclass classification Problem

The aim of the multiclass DA is to discover a linear transformation which lowers the dimensionality of an  $M$ -dimensional statistical model with  $C > 2$  classes while keeping as most discriminant information in the lower-dimensional space as possible.

Let  $\mathbf{d} \in \mathbb{R}^M$  serve as the projection direction. In Fisher's discriminant analysis [43], the Fisher criterion reads

$$\max_{\mathbf{d}} \frac{\mathbf{d}^\top \mathbf{S}_B \mathbf{d}}{\mathbf{d}^\top \mathbf{S}_W \mathbf{d}}, \quad (4)$$

which can be addressed by solving

$$\mathbf{S}_B \mathbf{d} = \lambda \mathbf{S}_W \mathbf{d}, \quad (5)$$

where  $\lambda$  is the Lagrange multiplier [44]. This is also known as the generalized eigenvalue problem regarding  $\mathbf{S}_B$  and  $\mathbf{S}_W$ , and  $\mathbf{d}$  is the eigenvector corresponding to the non-zero eigenvalue ( $\lambda$ ) in this situation. Then the transformation matrix can be formed by stacking up the  $(C - 1)$  eigenvectors corresponding to the  $(C - 1)$  largest eigenvalues in Eq. (5). When the number of samples  $N$  is small and/or the dimensionality of the data  $M$  is big,  $\mathbf{S}_W$  is generally singular in practice. This could be dealt with by adding a small perturbation on  $\mathbf{S}_W$ , e.g.,

$$\tilde{\mathbf{S}}_W = \mathbf{S}_W + \delta I, \quad (6)$$

where  $I$  is the identity matrix and  $\delta$  is a relatively small value (e.g.,  $5 \times 10^{-3}$ ) such that  $\tilde{\mathbf{S}}_W$  is therefore invertible. The discriminant directions can then be obtained by conducting the eigenvalue decomposition of  $\tilde{\mathbf{S}}_W^{-1} \mathbf{S}_B$  and finding the  $(C - 1)$  eigenvectors corresponding to the  $(C - 1)$  largest eigenvalues.

#### 2.4.2. Binary Classification Problem

Different from Fisher criterion given in Eq. (4), which can only produce one discriminant direction in the binary classification scenario, the method proposed in [35] can discover more discriminant directions. It is optimal in the sense that a set of projection directions  $\{\mathbf{d}_k\}_{k=1}^n$  is determined under a variety of constraints, see details below.

The Fisher criterion (cf. Eq. (4)) for the binary classification problem reads

$$\mathcal{R}(\mathbf{d}) = \frac{\mathbf{d}^\top \tilde{\mathbf{S}}_B \mathbf{d}}{\mathbf{d}^\top \tilde{\mathbf{S}}_W \mathbf{d}}. \quad (7)$$

Note that  $\mathcal{R}(\mathbf{d})$  is independent of the magnitude of  $\mathbf{d}$ . The first discriminant direction  $\mathbf{d}_1$  is discovered by maximising  $\mathcal{R}(\mathbf{d})$ , and then we have

$$\mathbf{d}_1 = \alpha_1 \tilde{\mathbf{S}}_W^{-1} \mathbf{s}_b, \quad (8)$$

where  $\alpha_1$  (i.e.,  $\alpha_1^2 = (\mathbf{s}_b^\top (\tilde{\mathbf{S}}_W^{-1})^2 \mathbf{s}_b)^{-1}$ ) is the normalising constant such that  $\|\mathbf{d}_1\|_2 = 1$  (and recall  $\mathbf{s}_b$  is the difference of the means of the two classes). The second discriminant direction  $\mathbf{d}_2$  is required to maximise  $\mathcal{R}(\mathbf{d})$  in Eq. (7) and be orthogonal to  $\mathbf{d}_1$ . It can be found by the method of Lagrange multipliers, i.e., finding the stationary points of

$$\mathcal{R}(\mathbf{d}_2) - \lambda [\mathbf{d}_2^\top \mathbf{d}_1], \quad (9)$$

where  $\lambda$  is the Lagrange multiplier. We can then obtain

$$\mathbf{d}_2 = \alpha_2 \left( \tilde{\mathbf{S}}_W^{-1} - \frac{\mathbf{s}_b^\top (\tilde{\mathbf{S}}_W^{-1})^2 \mathbf{s}_b}{\mathbf{s}_b^\top (\tilde{\mathbf{S}}_W^{-1})^3 \mathbf{s}_b} (\tilde{\mathbf{S}}_W^{-1})^2 \right) \mathbf{s}_b, \quad (10)$$

where  $\alpha_2$  is the normalising constant such that  $\|\mathbf{d}_2\|_2 = 1$ .

The above procedure can be extended to any number of directions (until the number of features  $M$ ) recursively as follows. The  $n$ -th discriminant direction  $\mathbf{d}_n$  is required to maximise  $\mathcal{R}(\mathbf{d})$  in Eq. (7) and be orthogonal to  $\mathbf{d}_k, k = 1, 2, \dots, n-1$ . It can be shown that

$$\mathbf{d}_n = \alpha_n \tilde{\mathbf{S}}_W^{-1} \left\{ \mathbf{s}_b - [\mathbf{d}_1 \dots \mathbf{d}_{n-1}] \mathbf{S}_{n-1}^{-1} \begin{bmatrix} 1/\alpha_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\}, \quad (11)$$

**Algorithm 1** LDA for binary classification

**Require:**  $\{\mathbf{y}_i\}_{i=1}^N$  and  $L \leq M$ , i.e., the given samples and the number of discriminant vectors.

**Compute**  $\tilde{\mathbf{S}}_W$  and  $s_b$  in Eq. (3);

**Compute**  $\mathbf{d}_1$  using Eq. (8) and  $\mathbf{S}_1$  using Eq. (12);

$n = 1$ ;

**for**  $n < L$  **do**

$n = n + 1$ ;

**Compute**  $\mathbf{d}_n$  using Eq. (11);

**Compute**  $\mathbf{S}_n$  using Eq. (12);

**end for**

**Return**  $\{\mathbf{d}_n\}_{n=1}^L$

where  $\alpha_n$  is the normalising constant such that  $\|\mathbf{d}_n\|_2 = 1$  and the  $(i, j)$  entries of  $\mathbf{S}_{n-1} \in \mathbb{R}^{(n-1) \times (n-1)}$  are defined as

$$\mathbf{d}_i^\top \tilde{\mathbf{S}}_W^{-1} \mathbf{d}_j, \quad 1 \leq i, j \leq n-1. \quad (12)$$

The whole procedure of finding  $L$  number of discriminant vectors  $\{\mathbf{d}_n\}_{n=1}^L$  is summarised in Algorithm 1.

Similar to how each singular vector correlates to a singular value, each discriminant vector  $\mathbf{d}_n$  corresponds to a “discrim-value” say  $\gamma_n$ , where

$$\gamma_n = \frac{\mathbf{d}_n^\top \tilde{\mathbf{S}}_B \mathbf{d}_n}{\mathbf{d}_n^\top \tilde{\mathbf{S}}_W \mathbf{d}_n}. \quad (13)$$

The discriminant vectors  $\{\mathbf{d}_n\}_{n=1}^L$  are naturally ordered by their discriminant values, following  $\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_L \geq 0$ .

The DA subspace formed by  $\{\mathbf{d}_n\}_{n=1}^L$  offers considerable potential for feature extraction and dimensionality reduction in many fields like pattern recognition [35]. For example, face recognition has been enhanced by LDA [45] outperforming PCA in many cases.

## 2.5. Non-negative Matrix Factorization

In the process of matrix factorization, reconstructing a low-rank approximation for the data matrix  $\mathbf{Y}$  is of great importance. NMF is a technique dealing with  $\mathbf{Y} \geq 0$  whose entries are all non-negative [46], with great achievements in many fields such as signal processing [47], biomedical engineering [36], pattern recognition [30] and image processing [48]. The sparsity of the NMF subspace has also received extensive attention. In genomics, for example, the work in [49] factorized gene expression matrices across different experimental conditions, showing that the sparsity of NMF contributes to decreasing noise and extracting biologically meaningful features. The purpose of NMF is to find two non-negative and low-rank matrices, i.e., one base matrix  $\mathbf{X} \in \mathbb{R}^{p \times M}$  and one coefficient matrix  $\mathbf{K} \in \mathbb{R}^{N \times p}$ , satisfying

$$\mathbf{Y} \approx \mathbf{K}\mathbf{X}, \quad (14)$$

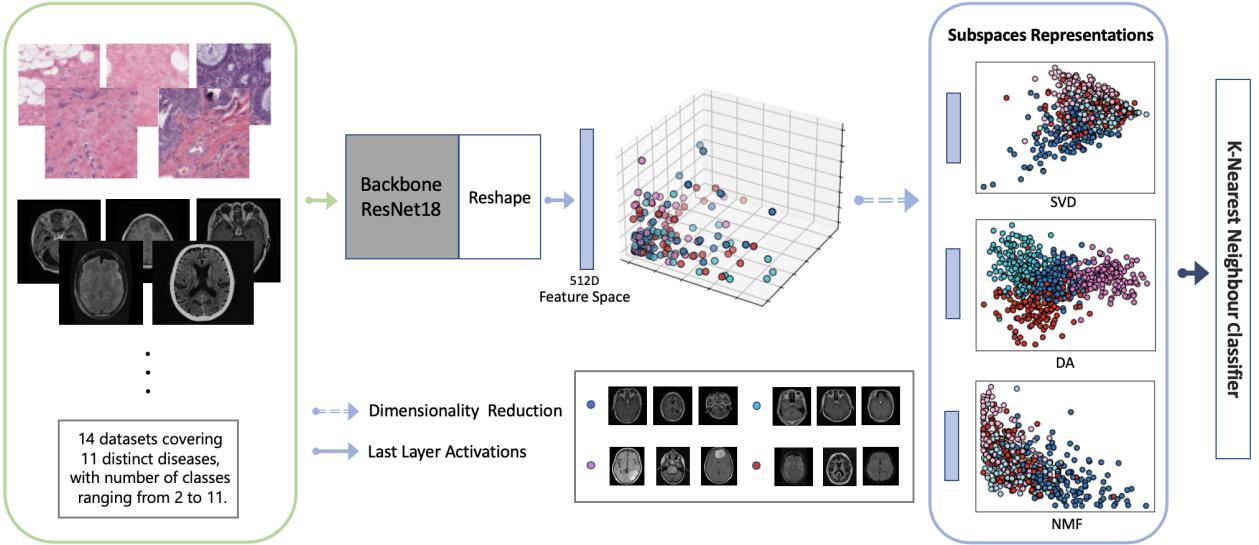
where  $p < \min\{M, N\}$ . Let  $\mathbf{K} = (\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_N)^\top$ . We have  $\mathbf{y}_i^\top \approx \mathbf{k}_i^\top \mathbf{X}$ ,  $1 \leq i \leq N$ . In other words, every sample  $\mathbf{y}_i$  can be represented by a linear combination of the rows of  $\mathbf{X}$  with the components in  $\mathbf{k}_i$  serving as weights. Therefore,  $\mathbf{X}$  is also known as consisting of basis vectors which can project the data matrix  $\mathbf{Y}$  into a low-dimensional subspace. The number of basis vectors  $p$  will affect the degree of approximation to the data matrix  $\mathbf{Y}$ .

Finding  $\mathbf{K}$  and  $\mathbf{X}$  satisfying Eq. (14) can be addressed by solving the following minimisation problem:

$$\min_{\mathbf{K}, \mathbf{X}} \|\mathbf{Y} - \mathbf{K}\mathbf{X}\|_F^2, \quad \text{s.t. } \mathbf{K} \geq 0, \mathbf{X} \geq 0, \quad (15)$$

where  $\|\cdot\|_F$  is the Frobenius norm. To solve problem (15), a common technique is to update  $\mathbf{K}$  and  $\mathbf{X}$  alternatively, i.e.,

$$\mathbf{K} \leftarrow \mathbf{K} \odot \frac{\mathbf{Y}\mathbf{X}^\top}{\mathbf{K}\mathbf{X}\mathbf{X}^\top}, \quad \mathbf{X} \leftarrow \mathbf{X} \odot \frac{\mathbf{K}^\top \mathbf{Y}}{\mathbf{K}^\top \mathbf{K}\mathbf{X}}, \quad (16)$$



**Figure 1:** Few-shot learning schematic diagram on different subspaces. From left to right: A pre-trained deep neural network (e.g. ResNet18) to solve a large natural image classification problem is exploited to extract features of medical images (i.e., inputs in the green box), and then the extracted features are projected to subspace representations (i.e., outputs in the blue box), followed by a classifier (e.g. KNN) delivering the classification results. The extracted features for individual images are visualised as dots with different colours representing different classes (i.e., the middle of the diagram).

where  $\odot$  denotes the pointwise product. For more algorithmic details please refer to e.g. [46].

NMF is an unsupervised method that decomposes the data matrix without utilising the class label information. Regarding the binary classification problem, the SNMF (supervised NMF) proposed in [36] extends the standard unsupervised NMF approach by exploiting feature extraction and integrating the cost function of the classification method into NMF. In SNMF, the classification labels are incorporated into the algorithms to extract the specific data patterns relevant to the respective classes. The whole algorithm of SNMF is provided in Appendix A.

### 3. Few-shot Learning on Subspace Representations

We deploy few-shot learning techniques in investigating medical imaging particularly in the data scarcity scenario. We consider problems in which the feature space dimensionality is usually high in comparison to the number of images we have; hence subspace representations are sought. The adopted few-shot learning scheme on subspace feature representations and experimental settings are presented in what follows.

#### 3.1. Framework

The deployed and enhanced few-shot learning schematic diagram on different subspaces is shown in Figure 1. Firstly, a pre-trained deep neural network (e.g. ResNet18) to solve a large natural image classification problem is prepared and then used to extract features of medical images in the given datasets (i.e., the green box in Figure 1). After that the extracted features are projected to subspace representations (i.e., the blue box in Figure 1). In this paper, we consider three different methods (i.e., SVD, DA and NMF) described in Section 2 to achieve this. Finally, a classifier (e.g. the K-Nearest Neighbour (KNN) or Support Vector Machine (SVM)) is employed to perform few-shot learning – predicting the final classification results. Extensive exploration in terms of the benefits of different subspace representations and insightful suggestions and comparisons in the regime of few-shot learning in medical imaging will be conducted in Section 5.

#### 3.2. Experimental Settings

We explore 14 datasets covering 11 distinct diseases, with the number of classes ranging from 2 to 11, see Section 4 for more detail. The pre-trained deep model, ResNet18, is used as the source model in our experiment. Each input is pre-processed and pixel-wised by subtracting its mean and being divided by the standard deviation without data

augmentation. The feature space is from the features in the penultimate layer of the pre-trained model (ResNet-18) extracted by PyTorch hooks [50], yielding a 512-dimensional feature vector for each image. The low-dimensional representations are then generated from the introduced methods. The number of iterations related to NMF and SNMF is set to 3000 to ensure convergence. The mean result of the KNN classifier with selected  $K$  (with values of 1, 5, 10 and 15) nearest neighbours is used to evaluate the final performance. Except for KNN, we also implement SVM as the classifier for comparison. The detailed experimental setting and results of the SVM classifier are shown in Appendix B.3. To quantify the uncertainty of the classification accuracy and produce more reliable quantitative results, we present averages and standard deviations across 10 distinct times of random samplings in each dataset. In addition to the accuracy, the reconstruction error of NMF at different random initialization is conducted to demonstrate its convergence. Moreover, we also compare our method with other well-known few-shot learning algorithms like the prototypical network [51]. The experimental setup and results are presented in Appendix B.5.

## 4. Data

A total of 14 different datasets covering a range of problems in diagnostics are employed for our empirical work. The number of classes ranges from 2 to 11 and the imaging modalities include X-rays, CT scans, MRI and Microscope. The datasets with MNIST within their name come from a benchmark family referred to as MedMNIST<sup>2</sup>. In order to illustrate the regime of few-shot learning, randomly sampled subsets of the whole individual datasets are used for our training and test. The corresponding data split for each class in the training and test sets for all the datasets is presented in Table 1. It is worth noting that our intention is not to compare with previously published results which have used the whole individual datasets. For ease of reference, brief descriptions of these individual datasets together with our implementations are given below.

- 1) BreastCancer (IDC) data [52, 53] is a binary classification problem sampled from digitised whole slide histopathology images. The source of the data is from 162 women diagnosed with Invasive Ductal Carcinoma (IDC), the most common phenotypic subtype in breast cancers. From these annotated images 277,524 patches had been segmented. An accuracy of 84.23% using the whole dataset is reported in [53].
- 2) BrainTumor data [54, 55] is a four-category problem, consisting of 7,022 images of human brain MRI images, three types of tumours (i.e., glioma, meningioma and pituitary), and a control group.
- 3) CovidCT data [56] is a binary classification problem, which is of great interest due to the COVID-19 pandemic. It contains 349 CT scans that are positive for COVID-19 and 397 negatives that are normal or contain other types of diseases. Two-dimensional slices from the scans are used in the study.
- 4) DeepDRiD data [57] is a five-category problem. Diabetic retinopathy is a prevalent eyesight condition in eye care. With early detection and treatment, the majority of these disorders may be controlled or cured. In this dataset, a total of 2,000 regular fundus images were acquired using Topcon digital fundus camera from 500 patients.
- 5) BloodMNIST data [58] is an eight-category problem, including a total of 17,092 images. It consists of individual normal cells, captured from individuals without infection, hematologic or oncologic disease and free of any pharmacologic treatment at the time of blood collection.
- 6) BreastMNIST data [59] is a binary classification problem, including a total of 780 breast ultrasound images. An accuracy of 94.62% is claimed in [60] in the computer-aided diagnostic (CAD) setting on the whole dataset. The grayscale images are replicated in order to match the pre-trained model.
- 7) DermaMNIST data [61, 62] is a multi-source dermatoscopic image collection of common pigmented skin lesions. It contains 10,015 dermatoscopic images, which are classified into seven diseases.
- 8) OCTMNIST data [63] is for retinal diseases, including a total of 109,309 valid optical coherence tomography images, with four diagnostic categories.
- 9) OrganAMNIST, OrganCMNIST and OrganSMNIST datasets [64] are eleven-category problem. They are benchmarks for segmenting liver tumours from 3D computed tomography images (LiTS). Organ labels were obtained using boundary box annotations of the 11 bodily organs studied, which are renamed from Axial, Coronal and Sagittal for simplicity. Grayscale images were converted into RGB images through the instruction in [65].
- 10) PathMNIST data [66] is based on the study of using colorectal cancer histology slides to predict survival, including a total of 107,180 images and nine different types of tissues. An accuracy of 94% was achieved in [66] by training a CNN using transfer learning on a set of 7,180 images from 25 CRC patients.

<sup>2</sup><https://medmnist.com/>

**Table 1**

Data split for each class in the training and test sets of each dataset.

Datasets	#Classes	#Samples for each class	
		Training	Test
BreastCancer[52]	2	300	40
BrainTumor[54]	4	160	40
CovidCT[56]	2	300	40
DeepDRID [57]	5	118	29
BloodMNIST[58]	8	75	25
BreastMNIST[59]	2	263	88
DermaMNIST[61]	7	75	25
OCTMNIST[63]	4	150	50
OrganAMNIST[64]	11	50	15
OrganCMNIST[64]	11	50	15
OrganSMNIST[64]	11	50	15
PathMNIST[66]	9	60	20
PneumoniaMNIST[63]	2	262	87
TissueMNIST[67]	8	65	20

11) PneumoniaMNIST data [63] is to classify pneumonia into two categories – severe and mild. It consists of 5,856 paediatric chest X-ray images. The source images are grayscale, which are converted to RGB for training in the same manner as the OrganAMNIST dataset.

12) TissueMNIST data [67] is derived from the Broad Bioimage Benchmark Collection. It consists of 236,386 human kidney cortex cells, segmented and labelled into eight categories. An accuracy of 80.26% was achieved in [67] using a custom 3D CNN on the whole dataset.

## 5. Experimental Results

In this section, we investigate the performance of the few-shot learning scheme described in Section 3 on subspace representations using SVD, DA and NMF. Note, importantly, that our main interest is to introduce DA and NMF as alternative subspace representations to SVD in the regime of few-shot learning in medical imaging. In addition to the comparison between the SVD, DA and NMF subspaces, we also compare them with other relevant feature selection, dimensionality reduction, and few-shot learning methods. For visual inspection, we visualise the subspace distributions of SVD, DA and NMF by T-SNE built-in function in Python (see the results in Appendix B.4).

### 5.1. Discriminant versus Principal Component Subspaces

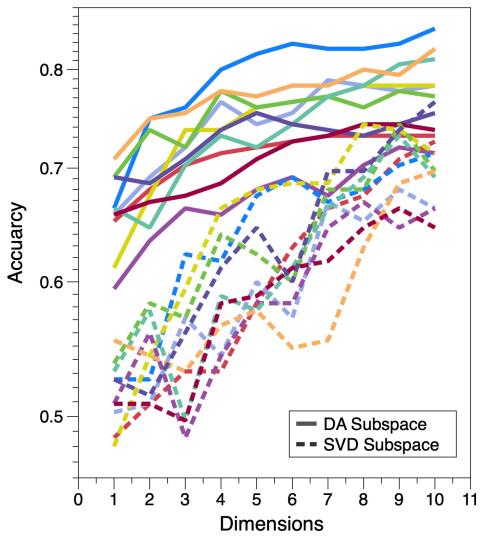
We first conduct comparison between DA and PCA. Table 2 shows the few-shot learning classification accuracy on the 14 datasets/problems, comparing the feature space in its original dimension of the ResNet18 with the PCA and DA subspaces. The accuracy results are the average of  $K$  values of KNN classifier chosen to be 1, 5, 10 and 15. We note that with a single exception of the CovidCT dataset, principal component dimensionality reduction loses information about class separation, whereas the discriminant subspace representation maintains the separation extremely well, thereby showing significant improvement over the original feature space. In detail, in 11 of the 14 problems, the SVD subspace performs worse than the original feature space. In contrast, the DA subspace shows significant improvement over the corresponding SVD subspace in all the 14 problems; and in 13 of the 14 problems, the DA subspace shows significant improvement over the original feature space. Furthermore, Z-test was also carried out and it is confirmed that the results are statistically significant at  $P$  values smaller than  $10^{-3}$ .

We now evaluate the impact of the subspace dimensions on the classification accuracy for DA and SVD. Figure 2(a) shows how the classification accuracy varies as the subspace dimensions increase on the PneumoniaMNIST dataset (consistent results are observed for other datasets). In particular, ten different random partitions of the training-test set are utilised to shuffle the data (which will make the results more credible) and dimensions from one to ten are investigated in Figure 2(a). We observe that the performance of both the DA and SVD methods increases monotonically

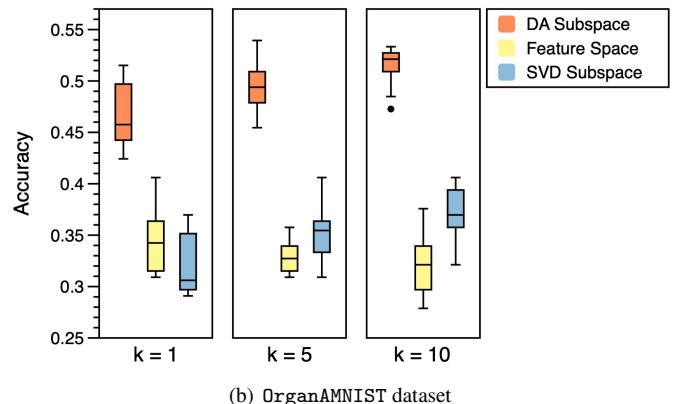
**Table 2**

Few-shot learning classification accuracy on 14 medical datasets with the KNN classifier. The original feature space and the subspaces obtained by SVD and DA are tested. The number of classes ranging from 2–11 shows the multi-modal structure in each dataset.

Datasets	Feature Space	SVD	DA	Classes
BreastCancer[52]	63.25±4.80	60.85±8.44	<b>66.76±5.39</b>	2
BrainTumor[54]	68.63±4.17	54.63±4.73	<b>73.19±2.98</b>	4
CovidCT[56]	<b>77.11±2.89</b>	66.68±6.82	70.58±3.39	2
DeepDRiD [57]	48.25±6.94	36.19±4.98	<b>55.11±4.19</b>	5
BloodMNIST[58]	37.49±3.88	37.10±3.91	<b>54.33±3.56</b>	8
BreastMNIST[59]	69.78±3.79	68.45±4.19	<b>70.08±3.58</b>	2
DermaMNIST[61]	25.03±4.64	21.16±3.29	<b>33.52±3.13</b>	7
OCTMNIST[63]	31.61±4.44	28.25±3.60	<b>34.85±3.13</b>	4
OrganAMNIST[64]	32.65±2.58	35.30±3.26	<b>49.67±2.98</b>	11
OrganCMNIST[64]	25.80±3.09	26.88±3.13	<b>45.93±4.14</b>	11
OrganSMNIST[64]	24.80±2.37	24.18±2.54	<b>39.59±2.64</b>	11
PathMNIST[66]	33.97±2.37	38.47±4.59	<b>58.68±3.75</b>	9
PneumoniaMNIST[63]	70.43±3.70	61.60±7.48	<b>73.76±5.22</b>	2
TissueMNIST[67]	18.89±2.80	16.88±2.42	<b>21.86±2.15</b>	8



(a) PneumoniaMNIST dataset

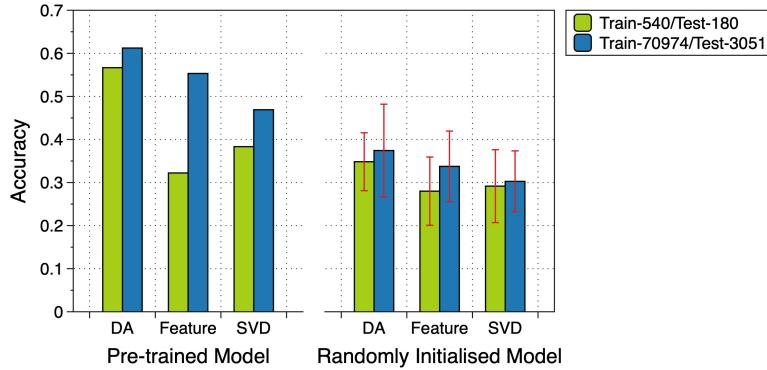


(b) OrganAMNIST dataset

**Figure 2:** Comparison between DA and PCA subspaces in terms of classification accuracy corresponding to different dimensions and different neighbourhood size  $K$  in the KNN classifier. Figure (a) shows the DA subspace taken at different dimensions consistently outperform the SVD subspace (*cf.* Table 2 for the performance on the full 512 dimensional feature space). Figure (b) shows the excellent performance of the DA subspace against PCA and the original feature space, irrespective of the choice of  $K$  in the classifier.

corresponding to the number of dimensions, with the DA subspace consistently outperforming SVD. Given the performance achieved using the full set of features is  $70.43 \pm 3.70$  in Table 2, hence the increase for SVD is not sustainable beyond this point.

The effect of different neighbourhood size  $K$  of the KNN classifier is reported in Figure 2(b), where the eleven-class dataset OrganAMNIST (consistent results are observed for other datasets) is used. Moreover, the performance of the



**Figure 3:** Comparison between DA and PCA subspaces and the original feature space in terms of classification accuracy corresponding to different dataset sizes. Dataset PathMNIST with nine classes is used. The left and right three pairs of bars in the panel are the results of the pre-trained model and the model with randomly initialised weights, respectively. The results reveal that the performance of the DA subspace always outperforms the SVD and the original feature space, irrespective of the choice of the data size. Moreover, the results achieved using the DA subspace are highly comparable to those obtained by using the entire dataset, whereas the results of SVD fall short.

**Table 3**

Binary classification accuracy comparison between subspaces (i.e., SVD, NMF and SNMF) and the original feature space.

Datasets	Feature Space	SVD	NMF	SNMF
CovidCT	<b>77.11±2.89</b>	75.75±2.82	74.96±1.98	76.47±2.25
BreastCancer	63.25±4.80	67.15±4.18	68.75±4.45	<b>69.53±4.98</b>
PneumoniaMNIST	70.43±3.70	73.32±1.21	75.08±2.07	<b>75.28±2.02</b>
BreastMNIST	69.78±3.79	70.92±2.88	71.54±3.54	<b>73.81±2.63</b>

SVD and DA subspaces with dimension equal to ten against the original feature space corresponding to  $K = 1, 5$  and 10 is evaluated in Figure 2(b). Uncertainty in results is evaluated over 10 random partitions of the training-test set, with 550 and 165 images for training and test, respectively. Figure 2(b) shows substantial improvement in DA subspace representation over both the original feature space and the SVD reduced subspace irrespective of the choice of  $K$  in the KNN classifier.

Finally, we investigate the effect of the dataset size on the performance of the methods compared. Figure 3 shows the results regarding the DA and PCA subspaces and the original feature space on a small subset (i.e., 540 and 180 images for training and test, respectively) of the dataset as well as the entire dataset (i.e., 70,974 and 3,051 images for training and test, respectively), where nine-class dataset PathMNIST (consistent results are observed for other datasets) is used for illustration. The value  $K$  in the KNN classifier is set to 5. In Figure 3, we also evaluate the effect of the pre-trained model on ImageNet versus the model whose weights are defined by random initialisation. The findings reveal that the performance of the DA subspace always outperforms the SVD and the original feature space, irrespective of the choice of the data size. Particularly, it also shows that, although utilising only 0.7% of the entire dataset, the results achieved using the DA subspace are highly comparable to those obtained using the entire dataset, whereas the results of SVD fall short. This confirms that the DA subspace is more stable than the SVD subspace, providing a discriminative subspace ideal for classification problems. In passing, we also see that the performance of the pre-trained model is better than that of the model with randomly initialised weights, which fits our expectations. More results – the comparison between DA and the manifold learning method Isomap (a non-linear dimensionality reduction process) – on all the datasets are given in Appendix B.1.

## 5.2. Non-negative Matrix Factorization Subspace

The classification accuracy of the NMF subspace (including NMF and SNMF) and the comparison with the SVD subspace and the original feature space on the binary class and multiclass problems are shown in Tables 3 and 4

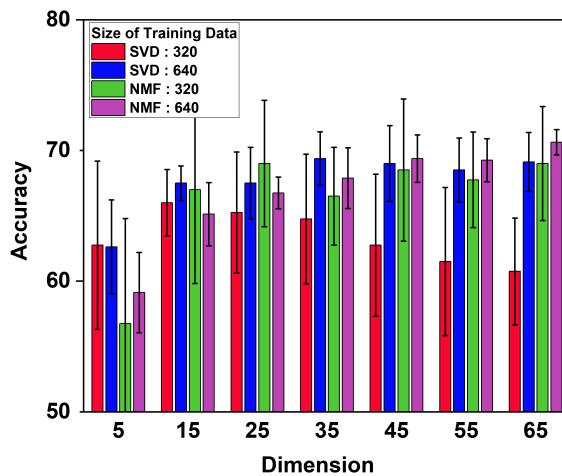
**Table 4**

Multiclass classification accuracy comparison between subspaces (i.e., SVD and NMF) and the original feature space.

Datasets	Feature Space	SVD	NMF
DeepDRid	48.25±6.94	48.59±3.90	<b>50.88±3.97</b>
BrainTumor	68.63±4.17	70.01±2.94	<b>70.10±2.52</b>
BloodMNIST	37.49±3.88	44.88±1.59	<b>45.18±2.30</b>
DermaMNIST	25.03±4.64	28.96±2.44	<b>29.79±2.01</b>
OCTMNIST	31.61±4.14	31.47±2.50	<b>31.80±2.43</b>
OrganAMNIST	32.65±2.58	38.62±1.29	<b>39.00±2.23</b>
OrganCMNIST	25.80±3.09	<b>32.92±2.77</b>	32.11±2.07
OrganSMNIST	24.80±2.37	<b>29.12±1.46</b>	29.01±1.51
PathMINST	33.97±2.37	43.35±1.23	<b>44.74±1.55</b>
TissueMNIST	18.89±2.80	<b>21.03±1.58</b>	20.75±1.52

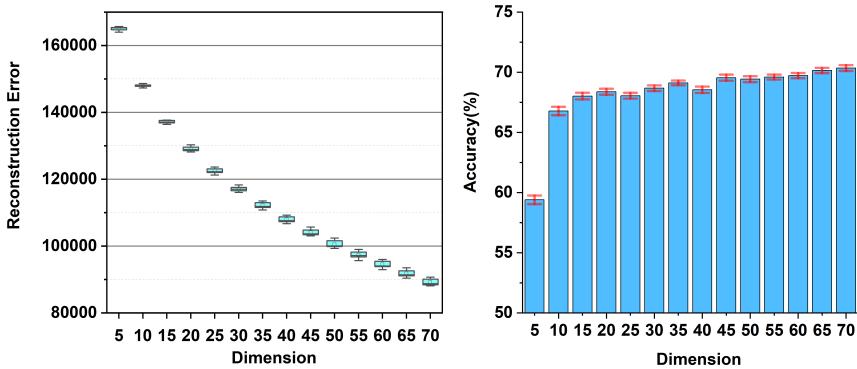
respectively. The SNMF subspace is only limited to the binary class problem and the dimension of related subspaces is kept as 30. It shows that, generally, the subspace representations (either SVD or NMF) deliver better performance than the original feature space. With SNMF marginally outperforming NMF in binary classification tests, NMF and SVD subspace both perform comparably and the trend is also preserved in multiclass classification problems. This prompts NMF can be a viable alternative to SVD, particularly when sparse representation is of great interest.

Different from the dimension selected in the DA subspace, the dimension of the NMF/SVD subspaces is retained as 30. Mainly because the NMF (including SNMF) approximates the original data with the product of two matrices and is affected by the selected rank during decomposition. While for the DA subspace, the dimensions are determined by the number of classes for the multiclass problems. Our results show that the performance of NMF is stable only after reaching a specific dimension, which is similar to the selection of the number of eigenvectors in SVD. Detailed trends regarding the performance of NMF and SVD subspaces on the 14 datasets against the changes in dimension are presented in Appendix B.1, including the comparison with the non-linear dimensionality reduction method Isomap in Figure 8.



**Figure 4:** Comparison between NMF and SVD subspaces in terms of classification accuracy corresponding to different dataset size as the subspace dimension changes. Dataset BrainTumor with four classes is used. Uncertainty is evaluated over 5 random partitions of the training-test set; and two types of training datasets with 320 and 640 images are created. The value  $K = 5$  is used in the KNN classifier. It shows that the performance of the NMF is stable for both types of datasets, whereas SVD suffers dimensional issues in the small dataset (with 320 images).

Additionally, we investigate the stability and uncertainty of NMF from the viewpoints of dataset size and the effects of random NMF initialization in various dimensions, respectively. Figure 4 describes how the volume of datasets influences the classification performance as the subspace dimensions ranging from 5 to 65 on the BrainTumor



**Figure 5:** Performance of the NMF subspace corresponding to different initialisation as the dimension of the subspace increases. The left and right panels respectively show the reconstruction error and the classification performance with 20 random NMF initializations on the BrainTumor dataset, indicating that the performance of NMF is quite stable corresponding to different dimensions with random initialisation except for the 5-dimensional subspace.

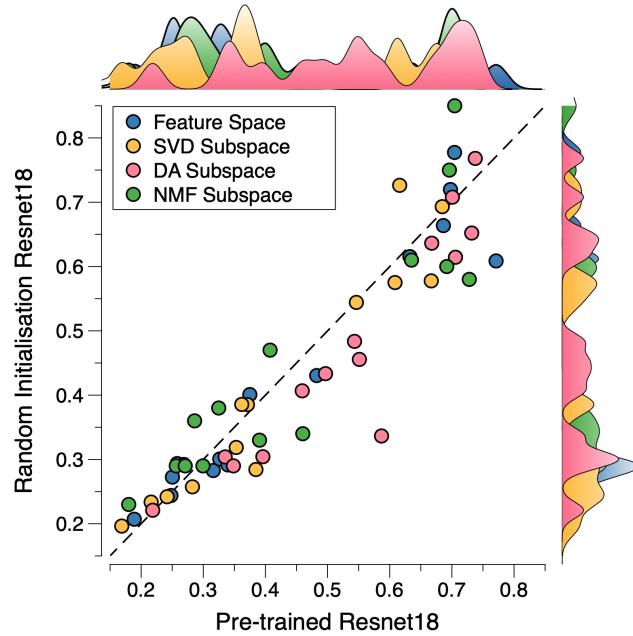
dataset. Two training datasets with the size of 320 and 640 images are created for the SVD and NMF subspaces, represented by different colour bars. It shows that on the big dataset (with 640 images), SVD and NMF are quite similar (see the blue and purple bars). On the small dataset (with 320 images), the NMF subspace outperforms the SVD subspace (see the red and green bars). SVD suffers from dimension issues in the small dataset since it performs gradually worse rather than better when the dimension becomes higher (e.g. when the dimension increases from 15 to 65). In contrast, the results of the NMF subspace are relatively stable in different dimensions and have similar accuracy. Although NMF behaves not good in extremely low dimensions (such as 5 dimensions), it gets improved as the dimension increases, which is consistent with the statement mentioned before. The uncertainty of NMF is evaluated by randomly initialising the NMF corresponding to different dimensions. In Figure 5, the left and right images show the reconstruction error and the classification performance with 20 random NMF initializations on the BrainTumor dataset. It reveals that the reconstruction error decreases as the dimensionality increases and the performance of NMF is quite stable corresponding to different dimensions with random initialisation.

### 5.3. Role of the Feature Extractor

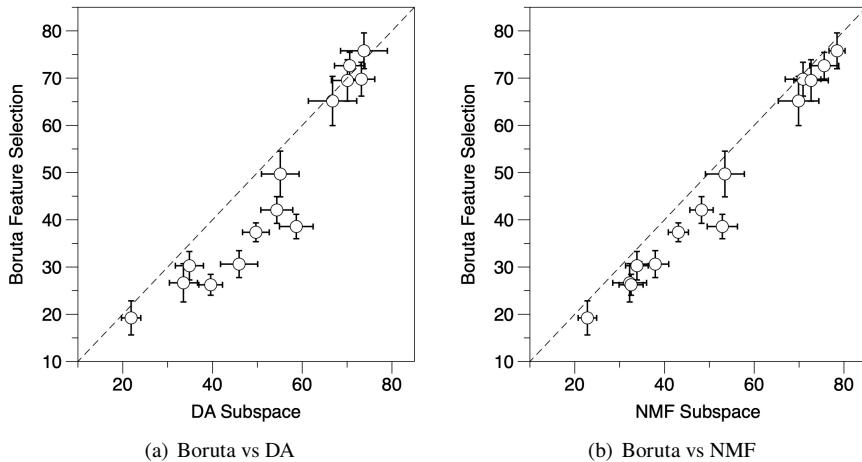
In the few-shot learning paradigm considered, the pre-trained source model serves as a feature extractor, mapping the medical images into a high dimensional space. To explore the impact of parameters in the model, we compare the classification accuracy from the related subspaces (i.e., feature space, PCA, DA and NMF) in random initialization and pre-trained models. Figure 6 shows the performance of the pre-trained model and the average of ten random initialization models on all the 14 datasets. ResNet18 is used as the base feature extractor with various parameters in this experiment. As we expected, the features extracted by the pre-trained model retain the good discriminant properties. Surprisingly, the performance of the features extracted by the randomly initialized model and the corresponding subspaces is not significantly degraded, indicating that the same discriminative properties are properly preserved in its extracted features. The DA results in the figure further illustrate this point and prove that subspace perspective provides directions for solving the few-shot learning on medical imaging.

### 5.4. Boruta Subspace

To investigate the performance of feature selection techniques in the few-shot learning framework, we below compare the subspace extracted from the Boruta feature selection method with the dimensionality reduction methods (i.e., SVD, DA and NMF). We follow the Boruta method and extract the related features on the 14 medical datasets (see results in Appendix B.2). Figure 7 presents the classification results comparing the Boruta feature selection method against DA and NMF. It shows that feature selection, like the Boruta method which only selects a subset from the 512 dimensions based on the voting results of a wrapper algorithm around a random forest, generally is not a good choice for the few-shot learning architecture we present. Instead of selecting features randomly like Boruta, we prefer to conditionally maintain the original attributes (e.g. discriminability, sparsity and non-negativity) of the data in the subspace. In addition, in terms of the computation time, DA and NMF is dramatically faster than Boruta (needing a high number of iterations), showing the efficiency of the introduced subspace representations.



**Figure 6:** Comparison of the use of features derived from pre-trained models against models with random initializations in the few-shot learning framework on the 14 datasets. Information extracted from the pre-trained source models helps in downstream medical tasks, although the fixed random transformations also retain discriminant information.



**Figure 7:** Few-shot learning classification accuracy comparison between Boruta feature selection and the approaches using DA (left panel) and NMF (right panel) subspaces on the 14 distinct medical datasets.

## 6. Discussion

For few-shot learning with only hundreds of images, comparable in order of magnitude to the feature dimensions (typically 512 or 1024 of popular models), dimensionality reduction is essential. While popular method of dimensionality reduction is PCA/SVD, its limitations as a variance preserving approximation suitable for uni-modal data need to be considered. We have addressed this by exploring DA and NMF as alternatives to SVD for few-shot learning in medical imaging.

By presenting the results in the experiment section, we discovered that the subspace obtained by DA is more useful for classification problems than the variance-preserving dimensionality reduction PCA/SVD. DA performs well on multiple disease datasets and effectively distinguishes the classes of disease in the low-dimensional space. However, DA also has some limitations, e.g. the maximum dimension of its subspace is one less than the number of classes for multiclass problems. This limitation is related to the rank of the covariance generated by the dataset. Moreover, DA may not perform ideally with classification when the data information depends on variance rather than the mean.

We also restricted our work on SNMF (supervised NMF) to binary classification problems for which the derivation is readily available. While for multiclass problems, more attempts will be necessary. This is mainly due to the fact that NMF is an inherently unsupervised matrix factorization algorithm and how to properly combine label signals and generate discriminate subspaces remain to be discussed. These, however, do not limit the scope of the conclusions we reach regarding the desirability of alternatives to the widely used SVD. Future work could be focusing on deriving the solutions to these cases. Additionally, it is also interesting to explore automatic rank selection using information theoretic concepts such as minimum description length considered in [68].

The comparison between feature selection techniques e.g. [37, 38] and the dimensionality reduction (i.e., SVD, DA and NMF) reveals that just selecting some specific features is less effective than eliminating less relevant information via dimensionality reduction. Moreover, plain feature selection can be quite unstable and may also be time-consuming. In comparison, since our few-shot learning architecture uses a pre-trained network for feature extraction, it is quite efficient. Most of the time consumed by our few-shot learning architecture is the dimensionality reduction and classification with a simple classifier. Benefiting from the dimensionality reduction, the final classification step is also quite economical.

Finally, it is worth mentioning that in clinical settings the validation and accuracy evaluation of the developed technique in medical imaging are extremely challenging (which is also true for all the related techniques). This is far beyond the lack of data challenge since clinical settings may require the involvement of clinicians, hospitals, patients and even the government, which are all difficult to reach out for individual academics or research groups. Collective effort from all interests is essential to validate/evaluate the practical use of any new method in medical imaging.

## 7. Conclusion

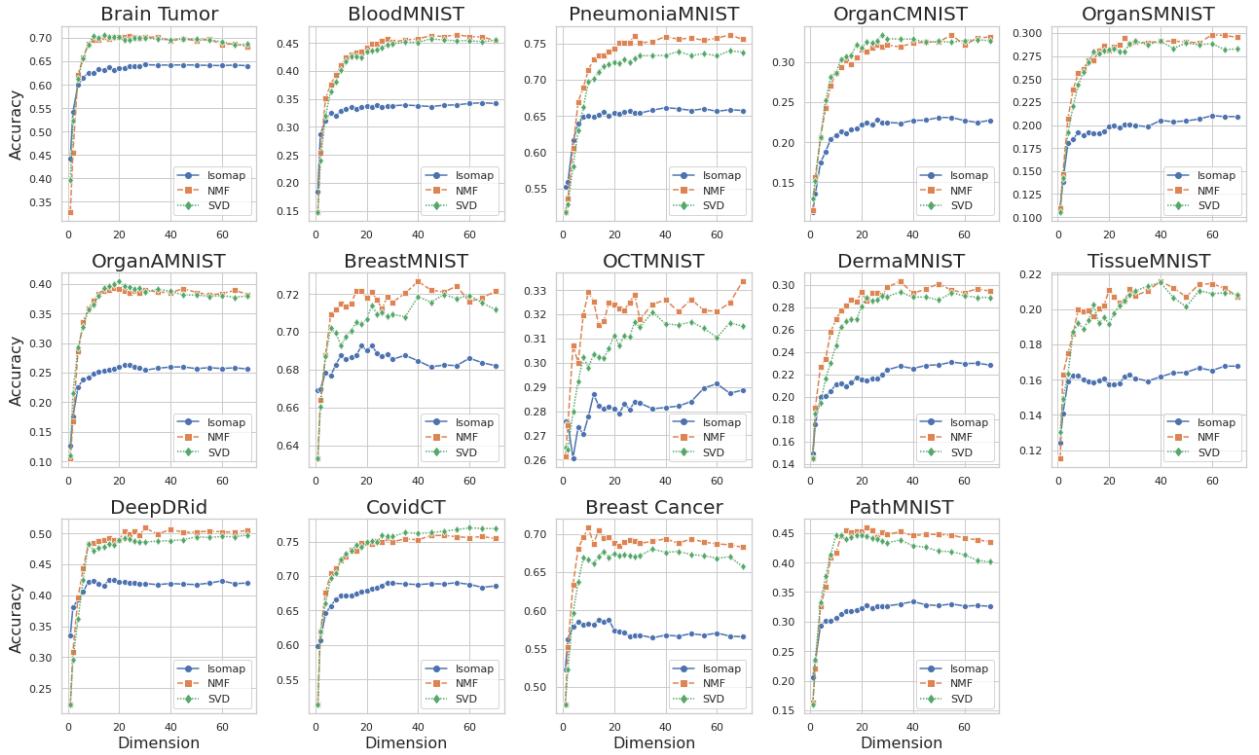
In this paper, we explored two different subspace representations – DA and NMF – of features learned from deep neural networks pre-trained on large computer vision datasets, adopted for few-shot learning on small medical imaging datasets. Our empirical work is carried out on 14 different datasets spanning 11 distinct diseases and four image acquisition modalities. Across these, we demonstrate the following: I) there is a consistent performance advantage on dimensionality reduction in the few-shot learning on medical imaging; II) working with DA derived subspaces gives significant performance gains over PCA/SVD based variance preserving dimensionality reductions, and even when taken at very low dimensions, these gains are statistically significant; and III) NMF-based representation, including its supervised variation, is a viable alternative to SVD-based low dimensional subspaces. NMF also shows a comparable advantage on part-based representation in moderate low dimensions. Overall, the developed few-shot learning framework with the newly introduced subspace representations is a very powerful approach in tackling medical imaging multiclass classification problems. One of important future avenues could be extending the developed approaches in this work in other fields.

## Acknowledgements

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

## A. Supervised NMF

The supervised NMF suggested by [36], introducing a logistic regression model into the cost function of NMF in Eq. (15). Let  $\mathbf{Z} := [\mathbf{1} \mid \mathbf{Y}\mathbf{X}^\top] = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N)^\top \in \mathbb{R}^{N \times (p+1)}$ . Note that  $\mathbf{z}_i \in \mathbb{R}^{p+1}, 1 \leq i \leq N$ . Considering the binary classification problem, let  $\mathbf{u} \in \{0, 1\}^N$  be the vector representing the labels (i.e., 0 or 1) of the given  $N$  samples.



**Figure 8:** Classification accuracy comparison between the SVD, Isomap and NMF subspaces on 14 datasets with subspace dimensions ranging from 1 to 70. Ten random partitions of the training-test set on each of the 14 datasets are conducted. It shows that in many cases NMF subspace exhibits slightly better performance in different dimensions than the SVD subspace. Moreover, both SVD and NMF outperform the Isomap subspace by a large margin.

Let  $\beta \in \mathbb{R}^{p+1}$ . The total loss function of the supervised NMF is defined as

$$\min_{\mathbf{K}, \mathbf{X} \geq 0, \beta} \frac{1}{2} \|\mathbf{Y} - \mathbf{K}\mathbf{X}\|_F^2 + \frac{\tilde{\lambda}}{N} \left( \sum_{i=1}^N \log (1 + \exp(\mathbf{z}_i^\top \beta)) - \mathbf{u}^\top \mathbf{Z}\beta \right), \quad (17)$$

where  $\tilde{\lambda} \geq 0$  is the regularization parameter. Problem (17) can be minimised corresponding to  $\mathbf{K}$ ,  $\mathbf{X}$  and  $\beta$  alternatively. Minimisation with respect to  $\mathbf{K}$  has the same update rules with (16). Because of the additional included logistic regression term in the loss function, we follow the stochastic gradient descent ADADELTA used in [36] to update  $\mathbf{X}$  and  $\beta$ . Method ADADELTA leverages the knowledge from previously computed gradients to determine the required gradients  $\Delta\mathbf{X}$  and  $\Delta\beta$  in the next step. To guarantee non-negativity of  $\mathbf{X}$ , one projection step, denoted by  $\text{proj}(\mathbf{X})$ , is added to replace all entries of the  $\mathbf{X}$  less than 0 with a small positive number. The whole update rules for supervised NMF is

$$\mathbf{K} \leftarrow \mathbf{K} \circ \frac{\mathbf{Y}\mathbf{X}^\top}{\mathbf{K}\mathbf{X}\mathbf{X}^\top}, \quad \mathbf{X} \leftarrow \text{proj}(\mathbf{X}) + \Delta\mathbf{X}, \quad \beta \leftarrow \beta + \Delta\beta. \quad (18)$$

For more details please see [36].

## B. Experiment Supplement

### B.1. Comparison with Manifold Learning

A manifold in mathematics is a topological space that resembles Euclidean space. One of the most commonly used manifold learning methods is the nonlinear dimensionality reduction technique Isomap [69]. Isomap is used for computing a quasi-isometric and low-dimensional embedding of a set of high-dimensional data points. Comparison

**Table 5**

Few-shot learning classification accuracy by Boruta feature selection on the 14 medical datasets.

Datasets	Accuracy	Selected Features	Classes
BreastCancer[52]	65.16±5.21	32	2
BrainTumor[54]	69.77±3.58	284	4
CovidCT[56]	72.63±2.82	58	2
DeepDRiD [57]	49.70±4.84	106	5
BloodMNIST[58]	42.08±2.83	183	8
BreastMNIST[59]	69.50±4.39	13	2
DermaMNIST[61]	26.66±4.06	41	7
OCTMNIST[63]	30.28±3.01	19	4
OrganAMNIST[64]	37.36±1.99	170	11
OrganCMNIST[64]	30.62±2.85	125	11
OrganSMNIST[64]	26.24±2.21	100	11
PathMNIST[66]	38.58±2.58	197	9
PneumoniaMNIST[63]	75.78±3.79	71	2
TissueMNIST[67]	19.23±3.61	26	8

between NMF, SVD and Isomap subspaces in terms of classification accuracy on the 14 datasets as the subspace dimensions ranging from 1 to 70 is given in Figure 8. Ten random partitions of the training-test set on each of the 14 datasets are conducted. It shows that in many cases NMF subspace exhibits slightly better performance in different dimensions than the SVD subspace. Moreover, both SVD and NMF outperform the Isomap subspace by a large margin. The comparison between DA, SVD and Isomap subspaces in terms of classification accuracy on all the datasets is given in Figure 9. It shows that Isomap and SVD produce similar results in low dimensions, while DA is able to achieve better results than both Isomap and SVD by a large margin. On the whole, the above results again suggest the viable alternatives of using NMF and DA to the SVD-based low dimensional subspaces.

## B.2. Boruta Results

Table 5 gives the classification results based on Boruta feature selection and the average number of selected features across ten runs for each of the 14 medical datasets, from which we see that the Boruta approach is highly unstable (i.e., yielding large deviation).

## B.3. Classification Results via SVM

Figure 10 depicts the performance of applying SVM as a classifier in the developed few-shot learning framework using the same setup as KNN on the 14 medical datasets. In particular, we implemented both the rbf kernel and the linear kernel for SVM. The C and gamma parameters were also fine-tuned for different datasets. The same dimensions as NMF and DA were applied to SVD to ensure fair comparison. We discovered that the rbf kernel worked better in comparisons between SVD and DA in low dimensions, and the linear kernel performed better in comparisons between SVD and NMF in medium dimensions. On the whole, consistent results were obtained by using SVM and KNN as classifiers on the 14 distinct medical datasets, indicating that the developed few-shot learning architecture is robust to the choice of classifiers.

## B.4. T-SNE Visualization

For better visual validation, Figure 11 shows an example of visualising the subspaces utilising the T-SNE visualization technique (built-in function in Python) on the brain tumour dataset. Figure 11(a) and Figure 11(b) are the 2D feature visualization of the original feature space (i.e., features extracted by the pre-trained network) and the DA subspace, respectively. Figure 11(c) and Figure 11(d) show the features that are projected to the 30-dimensional subspace by SVD and NMF, respectively. These plots visually prove that the DA and NMF subspaces are indeed viable alternatives to SVD.

**Table 6**

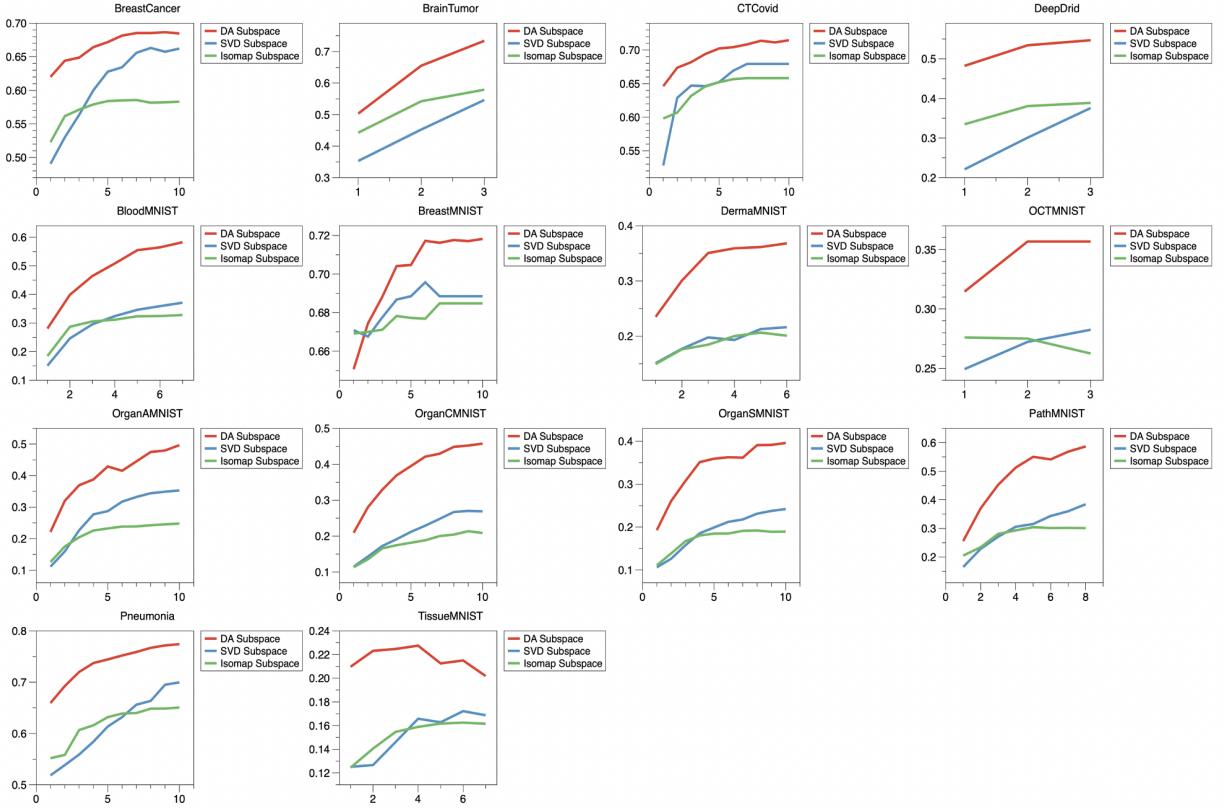
Classification accuracy comparison between the prototypical network and the few-shot learning with subspace feature representations. In particular, 5 samples from each class in each dataset are used for training, i.e., forming the “ $C$ -way 5-shot” setting (recall  $C$  is the number of classes in each dataset). Dim stands for dimensions.

Data	Prototypical Network	Feature Space	C-way 5-shot Accuracy(%)							
			Methods							
			Few-shot Learning with Subspace Feature Representations (Ours)							
			2 Dim	5 Dim	10 Dim	20 Dim	30 Dim	40 Dim	50 Dim	
CovidCT (2 classes)	53.33±6.99	52.22±5.88	SVD	49.78±10.95	54.56±5.80	52.11±5.32	52.11±5.32	52.11±5.32	52.11±5.32	52.11±5.32
			NMF	53.00±10.05	<b>56.89±8.07</b>	56.44±4.89	55.89±6.61	55.78±6.36	56.33±6.09	56.44±6.08
			DA				51.91±3.41 (10 Dim)			
BreastCancer (2 classes)	72.33±8.68	70.22±8.68	SVD	66.89±5.83	70.33±8.61	70.11±9.07	70.11±9.07	70.11±9.07	70.11±9.07	70.11±9.07
			NMF	68.78±7.83	72.44±7.71	72.78±9.69	72.11±7.82	<b>73.00±8.93</b>	71.33±9.27	72.44±8.37
			DA				62.69±4.43 (10 Dim)			
PneumoniaMNIST (2 classes)	64.33±8.17	66.56±9.79	SVD	61.56±9.34	66.11±7.32	66.56±8.92	66.56±8.92	66.56±8.92	66.56±8.92	66.56±8.92
			NMF	62.67±9.44	70.33±6.38	72.00±6.46	72.89±6.03	72.56±5.75	73.33±7.42	<b>73.78±6.78</b>
			DA				68.45±5.79 (10 Dim)			
BreastMNIST (2 classes)	54.00±9.40	54.11±6.43	SVD	54.00±7.14	54.11±6.67	54.67±7.47	54.67±7.47	54.67±7.47	54.67±7.47	54.67±7.47
			NMF	53.00±7.70	59.11±4.63	60.33±5.17	60.56±3.66	59.00±6.84	60.00±6.19	<b>62.33±5.45</b>
			DA				59.98±8.08 (Dim = 10)			
DeepDRid (5 classes)	30.07±7.87	29.47±4.91	SVD	27.58±4.53	28.03±5.59	29.42±4.75	29.28±4.95	29.47±4.98	29.47±4.98	29.47±4.98
			NMF	28.92±4.51	30.03±5.59	31.07±4.48	31.27±4.59	31.02±4.54	<b>31.47±4.17</b>	31.37±4.07
			DA				<b>31.63±7.07</b> (4 Dim)			
BrainTumor (4 classes)	34.67±4.82	35.42±5.29	SVD	33.12±4.69	34.54±4.84	35.00±5.29	35.42±5.23	35.42±5.23	35.42±5.23	35.42±5.23
			NMF	33.96±5.45	36.04±4.92	36.88±4.73	37.33±5.20	<b>37.87±5.10</b>	37.50±5.10	37.46±4.87
			DA				<b>61.88±5.50</b> (3 Dim)			
BloodMNIST (8 classes)	47.58±5.24	48.29±2.52	SVD	36.29±2.98	46.42±4.65	47.62±2.49	48.06±3.15	<b>48.33±2.45</b>	48.33±2.63	48.33±2.63
			NMF	36.60±3.77	47.33±4.96	46.94±4.42	47.02±3.87	46.23±2.71	45.90±3.65	46.58±3.56
			DA				<b>54.33±8.08</b> (7 Dim)			
DermaMNIST (7 classes)	26.57±4.69	25.83±3.57	SVD	21.43±4.61	25.07±3.63	25.19±3.28	25.33±3.67	25.71±3.57	25.76±3.58	25.76±3.58
			NMF	22.71±3.49	27.45±4.02	26.19±3.36	26.55±3.16	27.86±3.40	26.83±2.92	<b>27.79±2.98</b>
			DA				<b>31.14±5.54</b> (6 Dim)			
OCTMNIST (4 classes)	28.64±3.95	29.79±3.02	SVD	26.08±3.80	28.50±4.31	29.42±2.94	29.83±3.43	29.83±3.43	29.83±3.43	29.83±3.43
			NMF	26.12±3.63	31.17±2.52	<b>32.17±3.12</b>	31.75±3.20	32.13±3.49	31.04±3.16	32.12±3.57
			DA				<b>32.92±6.68</b> (3 Dim)			
OrganAMNIST (11 classes)	47.94±2.39	52.50±3.03	SVD	34.26±3.28	45.52±3.92	50.03±3.36	52.73±3.29	52.74±3.48	52.89±3.15	52.71±3.22
			NMF	33.70±3.38	45.39±2.67	53.00±3.31	53.88±3.38	<b>53.88±3.26</b>	52.97±2.82	53.58±3.99
			DA				<b>60.94±3.86</b> (10 Dim)			
OrganCMNIST (11 classes)	48.55±4.11	50.23±3.84	SVD	30.61±3.48	42.12±4.15	48.68±3.71	49.70±4.23	50.18±4.46	50.18±4.02	50.08±4.25
			NMF	29.86±2.77	40.55±3.83	49.55±3.67	51.92±4.72	<b>52.29±4.40</b>	51.26±4.55	50.91±3.55
			DA				<b>60.62±3.10</b> (10 Dim)			
OrganSMNIST (11 classes)	34.67±4.21	36.95±3.69	SVD	25.74±2.64	33.79±3.32	35.70±4.01	36.52±3.86	<b>37.18±3.43</b>	37.05±3.45	36.94±3.58
			NMF	24.67±2.41	33.97±3.04	36.06±3.24	36.45±3.11	35.48±2.97	35.53±3.34	34.67±3.80
			DA				<b>41.23±2.62</b> (10 Dim)			
PathMINST (9 classes)	36.22±4.77	37.69±3.53	SVD	28.83±5.01	36.17±4.08	37.72±3.45	37.96±3.72	37.50±3.53	37.76±3.64	37.74±3.59
			NMF	28.67±4.26	36.69±3.04	<b>39.76±3.10</b>	39.07±3.33	38.67±2.42	37.54±2.95	37.04±3.29
			DA				<b>41.43±4.19</b> (8 Dim)			
TissueMNIST (8 classes)	24.42±3.67	23.65±2.22	SVD	19.02±3.15	20.94±2.74	22.60±2.39	23.21±2.53	23.83±2.44	23.75±2.21	23.75±2.21
			NMF	18.50±3.63	22.40±2.43	<b>24.58±2.22</b>	24.23±2.39	23.31±2.58	22.96±2.63	23.02±1.72
			DA				<b>38.35±5.59</b> (7 Dim)			

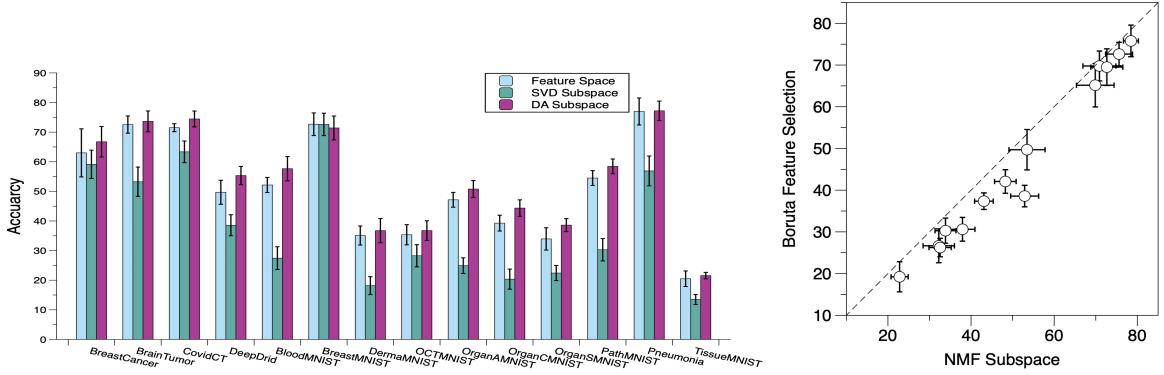
## B.5. Comparison with the Prototypical Network

Below we compare our techniques with the well-known few-shot learning algorithm, the prototypical network [51], on all the 14 medical datasets, see Table 6. The architecture of the prototypical network used in this experiment is the same as the one in [51], which is composed of four convolution blocks and has been trained on the omniglot dataset [70] via SGD with Adam optimiser [71] and obtained 99% accuracy in the 5-shot scenario. Each block in this network is comprised of a 64-filter  $3 \times 3$  convolution, batch normalisation layer, a ReLU nonlinearity and a  $2 \times 2$  max-pooling layer. The classification accuracy reported in Table 6 is averaging over 10 randomly generated episodes from the test set. The setting for the test experiment is “ $C$ -way 5-shot,” where 5 samples are given for each class in the support set and  $C$  is the number of classes in each dataset. To validate the final performance, 15 query images per class are provided. Recall that the original feature space represents the features extracted by the network without dimensionality reduction. The dimensions for the NMF/SVD subspaces are chosen as 2, 5, 10, 20, 30, 40 and 50, separately. The dimensions of the DA subspace are chosen as 10 for the binary classification problem and  $(C - 1)$  for the multiclass

## Few-shot Learning for Inference in Medical Imaging



**Figure 9:** Classification accuracy comparison between the SVD, Isomap and DA subspaces on the 14 datasets. It shows that DA outperforms both SVD and Isomap by a large margin.

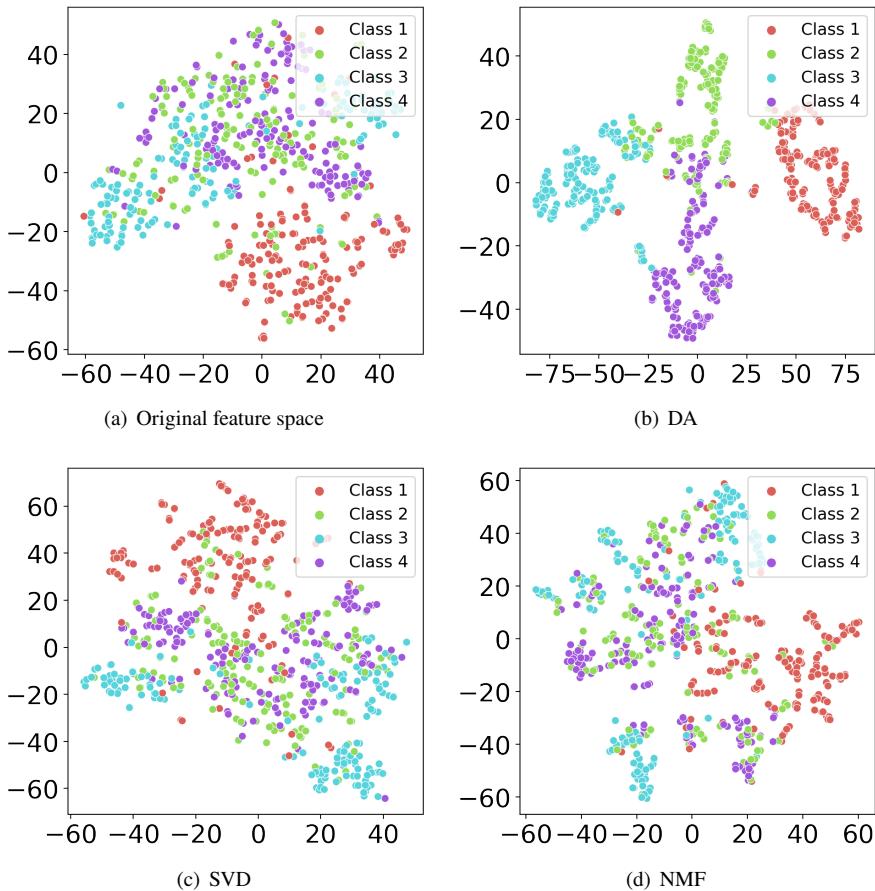


**Figure 10:** Few-shot learning performance using SVM as classifier on the 14 medical datasets. Left panel: results obtained by using the original feature space, the SVD subspace and the DA subspace. Right panel: results obtained by using the features in 30-dimensional subspace derived by SVD and NMF.

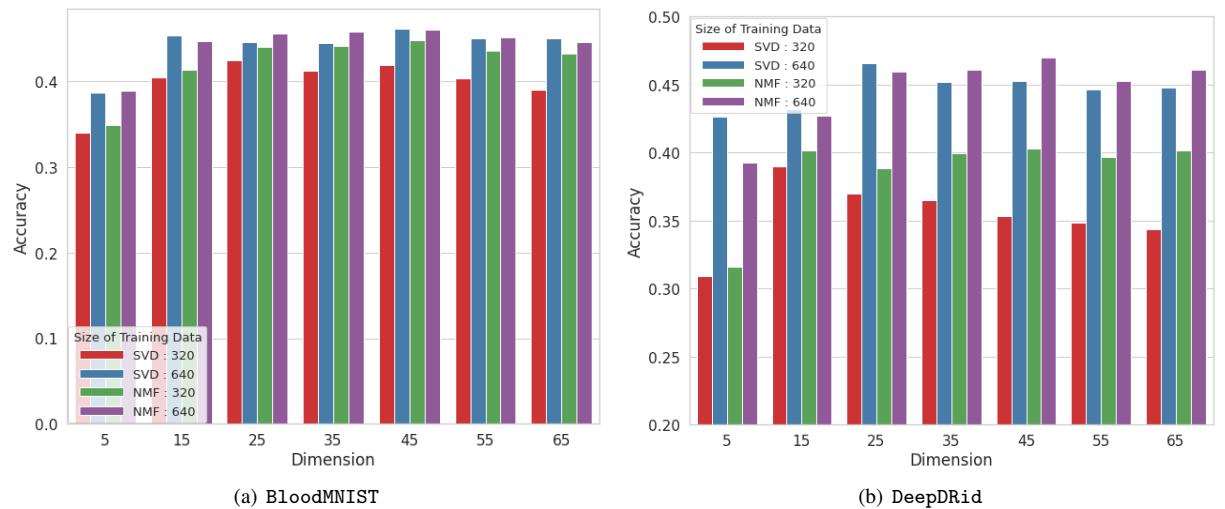
classification problem. The results in Table 6 demonstrate that our method outperforms all the other methods, i.e., the prototypical network and the ones with the original feature space and SVD.

### B.6. Impact of the Dataset Size and Dimensionality

Figure 12 shows the impact of the dataset size on the classification accuracy of NMF and SVD as the subspace dimension changes, where datasets BloodMNIST with eight classes and DeepDRid with five classes are used. The setting in Figure 12 is the same as the one used in Figure 4. It shows that, in the multiclass classification problem,



**Figure 11:** Subspace visualization by T-SNE on the brain tumour dataset. (a)–(d): the results regarding the originally feature space, DA subspace, 30-dimensional subspace derived by SVD and NMF, respectively.



**Figure 12:** Comparison between NMF and SVD subspaces in terms of classification accuracy corresponding to different dataset size as the subspace dimension changes. Datasets BloodMNIST with eight classes and DeepDRid with five classes are used in panels (a) and (b), respectively.

fewer categories will result in higher accuracy, and SVD suffers from dimension changes more in the datasets with small size. Without enough data, SVD could not extract specific precise features that match the target categories or may extract insignificant features as the dimension increases, resulting in low classification performance (e.g. see Figure 12(b)). In contrast, the results of NMF are robust to dimension changes in datasets with different size. This great performance of NMF benefits from its part-based representation maximising the features preserved in the subspace.

## References

- [1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [2] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.
- [3] Nicolas Pinto, David D Cox, and James J DiCarlo. Why is real-world visual object recognition hard? *PLoS computational biology*, 4(1):e27, 2008.
- [4] Abhishek Gupta, Alagan Anpalagan, Ling Guan, and Ahmed Shaharyar Khwaja. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, 10:100057, 2021.
- [5] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017.
- [6] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019.
- [7] Filippo Pesapane, Marina Codari, and Francesco Sardanelli. Artificial intelligence in medical imaging: threat or opportunity? radiologists again at the forefront of innovation in medicine. *European radiology experimental*, 2(1):1–10, 2018.
- [8] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.
- [9] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.
- [10] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019. Special Issue: Deep Learning in Medical Physics.
- [11] Mijung Kim, Jasper Zullaert, and Wesley De Neve. Few-shot learning using a small-sized dataset of high-resolution fundus images for glaucoma diagnosis. In *Proceedings of the 2nd international workshop on multimedia for personal health and health care*, pages 89–92, 2017.
- [12] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences*, 11(2):796, 2021.
- [13] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.
- [14] Andrea Caroppo, Alessandro Leone, and Pietro Siciliano. Deep transfer learning approaches for bleeding detection in endoscopy images. *Computerized Medical Imaging and Graphics*, 88:101852, 2021.
- [15] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5866–5885, 2021.
- [16] Hsin-Ping Huang, Krishna C Puvvada, Ming Sun, and Chao Wang. Unsupervised and semi-supervised few-shot acoustic event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 331–335. IEEE, 2021.
- [17] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 441–449, 2019.
- [18] Hao Tang, Zechao Li, Zhimao Peng, and Jinhui Tang. Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning. In *Proceedings of the 28th ACM international conference on multimedia*, pages 610–618, 2020.
- [19] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.
- [20] David Argüeso, Artzai Picon, Unai Irusta, Alfonso Medela, Miguel G San-Emeterio, Arantza Bereciartua, and Aitor Alvarez-Gila. Few-shot learning approach for plant disease classification using images taken in the field. *Computers and Electronics in Agriculture*, 175:105542, 2020.
- [21] Gwenolé Quellec, Mathieu Lamard, Pierre-Henri Conze, Pascale Massin, and Béatrice Cochener. Automatic detection of rare pathologies in fundus photographs using few-shot learning. *Medical image analysis*, 61:101660, 2020.
- [22] Debasmit Das and C. S. George Lee. A two-stage approach to few-shot learning for image recognition. *IEEE Transactions on Image Processing*, 29:3336–3350, 2020.
- [23] Xiaokang Zhou, Wei Liang, Shohei Shimizu, Jianhua Ma, and Qun Jin. Siamese neural network based few-shot learning for anomaly detection in industrial cyber-physical systems. *IEEE Transactions on Industrial Informatics*, 17(8):5790–5798, 2021.
- [24] Debasmit Das, JH Moon, and George Lee. Few-shot image recognition with manifolds. In *International Symposium on Visual Computing*, pages 3–14. Springer, 2020.
- [25] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- [26] Ivan Markovsky. *Low rank approximation: algorithms, implementation, applications*, volume 906. Springer, 2012.
- [27] Omar Shetta, Mahesan Niranjan, and Srinandan Dasmahapatra. Convex multi-view clustering via robust low rank approximation with application to multi-omic data. *IEEE/ACM transactions on computational biology and bioinformatics*, 2021.

- [28] Ivan Markovsky and Mahesan Niranjan. Approximate low-rank factorization with structured factors. *Computational Statistics & Data Analysis*, 54(12):3411–3420, 2010.
- [29] Dimitris Papailiopoulos, Alexandros Dimakis, and Stavros Korokythakis. Sparse pca through low-rank approximations. In *International Conference on Machine Learning*, pages 747–755. PMLR, 2013.
- [30] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- [31] Zechao Li, Jinhui Tang, and Xiaofei He. Robust structured nonnegative matrix factorization for image representation. *IEEE transactions on neural networks and learning systems*, 29(5):1947–1960, 2017.
- [32] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.
- [33] Aming Wu, Suqi Zhao, Cheng Deng, and Wei Liu. Generalized and discriminative few-shot object detection via svd-dictionary enhancement. *Advances in Neural Information Processing Systems*, 34, 2021.
- [34] He Zhang and Lili Liang. Res-svndt: A metric learning method for few-shot image classification. In *2021 40th Chinese Control Conference (CCC)*, pages 7400–7405, 2021.
- [35] Donald H. Foley and John W Sammon. An optimal set of discriminant vectors. *IEEE Transactions on computers*, 100(3):281–289, 1975.
- [36] Johannes Leuschner, Maximilian Schmidt, Pascal Fernsel, Delf Lachmund, Tobias Boskamp, and Peter Maass. Supervised non-negative matrix factorization methods for maldi imaging applications. *Bioinformatics*, 35(11):1940–1947, 2019.
- [37] Zechao Li, Jing Liu, Jinhui Tang, and Hanqing Lu. Robust structured subspace learning for data representation. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):2085–2098, 2015.
- [38] Rong Tang and Xiaojun Zhang. Cart decision tree combined with boruta feature selection for medical data classification. In *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, pages 80–84. IEEE, 2020.
- [39] Samuel H Huang. Supervised feature selection: A tutorial. *Artif. Intell. Res.*, 4(2):22–37, 2015.
- [40] Miron B Kursa and Witold R Rudnicki. Feature selection with the boruta package. *Journal of statistical software*, 36:1–13, 2010.
- [41] Zechao Li and Jinhui Tang. Semi-supervised local feature selection for data classification. *Science China Information Sciences*, 64(9):1–12, 2021.
- [42] Zechao Li and Jinhui Tang. Unsupervised feature selection via nonnegative spectral analysis and redundancy control. *IEEE Transactions on Image Processing*, 24(12):5343–5355, 2015.
- [43] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.
- [44] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.
- [45] Fatma Zohra Chelali, A Djeradi, and R Djeradi. Linear discriminant analysis for face recognition. In *2009 International Conference on Multimedia Computing and Systems*, pages 1–10. IEEE, 2009.
- [46] Daniel Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001.
- [47] Aimei Dong, Zhigang Li, and Qiuyu Zheng. Transferred subspace learning based on non-negative matrix factorization for eeg signal classification. *Frontiers in Neuroscience*, 15, 2021.
- [48] Zhikui Chen, Shan Jin, Runze Liu, and Jianing Zhang. A deep non-negative matrix factorization model for big data representation learning. *Frontiers in Neurorobotics*, page 93, 2021.
- [49] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.
- [50] Pytorch, forward and backward function hooks—pytorch documentation.
- [51] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.
- [52] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.
- [53] Angel Cruz-Roa, Ajay Basavanhalli, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganeshan, Natalie Shih, John Tomaszewski, and Anant Madabhushi. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *Medical Imaging 2014: Digital Pathology*, volume 9041, page 904103. SPIE, 2014.
- [54] Jun Cheng. brain tumor dataset, Apr 2017.
- [55] Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS one*, 10(10):e0140381, 2015.
- [56] Xuehai He, Xingyi Yang, Shanghang Zhang, Jinyu Zhao, Yichen Zhang, Eric Xing, and Pengtao Xie. Sample-efficient deep learning for covid-19 diagnosis based on ct scans. *medrxiv*, 2020.
- [57] The 1st diabetic retinopathy – classification of fundus images according to the severity level of diabetic retinopathy.
- [58] Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, ISSN: 23523409, Vol. 30,(2020), 2020.
- [59] Walid Al-Dhabayani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.
- [60] Woo Kyung Moon, Yan-Wei Lee, Hao-Hsiang Ke, Su Hyun Lee, Chiun-Sheng Huang, and Ruey-Feng Chang. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Computer methods and programs in biomedicine*, 190:105361, 2020.
- [61] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.
- [62] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging

- collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.
- [63] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.
  - [64] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (LiTS). *arXiv preprint arXiv:1901.04056*, 2019.
  - [65] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *arXiv preprint arXiv:2110.14795*, 2021.
  - [66] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019.
  - [67] Andre Woloshuk, Suraj Khochare, Aljohara F Almulhim, Andrew T McNutt, Dawson Dean, Daria Barwinska, Michael J Ferkowicz, Michael T Eadon, Katherine J Kelly, Kenneth W Dunn, et al. In situ classification of cell types in human kidney tissue using 3d nuclear staining. *Cytometry Part A*, 99(7):707–721, 2021.
  - [68] Steven Squires, Adam Prügel-Bennett, and Mahesan Niranjan. Rank selection in nonnegative matrix factorization using minimum description length. *Neural computation*, 29(8):2164–2176, 2017.
  - [69] Mukund Balasubramanian and Eric L Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.
  - [70] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. The omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences*, 29:97–104, 2019.
  - [71] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.