Histopathological Image Classification using Discriminative Feature-oriented Dictionary Learning

Tiep Huu Vu[†], *Student Member, IEEE,* Hojjat Seyed Mousavi[†], *Student Member, IEEE,*
Vishal Monga[†], *Senior Member, IEEE,* Ganesh Rao[*] and UK Arvind Rao[*]

Abstract— In histopathological image analysis, feature extraction for classification is a challenging task due to the diversity of histology features suitable for each problem as well as presence of rich geometrical structures. In this paper, we propose an automatic feature discovery framework via learning class-specific dictionaries and present a low-complexity method for classification and disease grading in histopathology. Essentially, our Discriminative Feature-oriented Dictionary Learning (DFDL) method learns class-specific dictionaries such that under a sparsity constraint, the learned dictionaries allow representing a new image sample parsimoniously via the dictionary corresponding to the class identity of the sample. At the same time, the dictionary is designed to be poorly capable of representing samples from other classes. Experiments on three challenging real-world image databases: 1) histopathological images of intraductal breast lesions, 2) mammalian kidney, lung and spleen images provided by the Animal Diagnostics Lab (ADL) at Pennsylvania State University, and 3) brain tumor images from The Cancer Genome Atlas (TCGA) database, reveal the merits of our proposal over state-of-the-art alternatives. Moreover, we demonstrate that DFDL exhibits a more graceful decay in classification accuracy against the number of training images which is highly desirable in practice where generous training is often not available.

Index terms—Histopathological image classification, Sparse coding, Dictionary learning, Feature extraction, Cancer grading.

I. INTRODUCTION

Automated histopathological image analysis has recently become a significant research problem in medical imaging and there is an increasing need for developing quantitative image analysis methods as a complement to the effort of pathologists in diagnosis process. Consequently, an emerging class of problems in medical imaging focuses on the the development of computerized frameworks to classify histopathological images [1]–[5]. These advanced image analysis methods have been developed with three main purposes of (i) relieving the workload on pathologists by sieving out obviously diseased and also healthy cases, which allows specialists to spend more time on more sophisticated cases; (ii) reducing inter-expert variability; and (iii) understanding the underlying reasons for a specific diagnosis that pathologists might not realize.

In the diagnosis process, pathologists often look for problem-specific visual cues, or features, in histopathological images in order to categorize a tissue image as one of the possible categories. These features might come from the distinguishable characteristics of cells or nuclei, for example, size, shape or texture [1], [6]. They could also come from spatially related structures of cells [3], [5], [7], [8]. In some cancer grading problems, features might include the presence of particular regions [5], [9]. Consequently, different customized feature extraction techniques for a variety of problems have been developed based on these observed features [10]–[14]. Morphological image features have been utilized in medical image segmentation [15] for detection of vessel-like patterns. Wavelet features and histograms are also a popular choice of features for medical imaging [16], [17]. Graph-based features such as Delaunay triangulation, Vonoroi diagram, minimum spanning tree [8], query graphs [18] have been also used to exploit spatial structures. Orlov *et al.* [10], [11] have proposed a multi-purpose framework that collects texture information, image statistics and transforms domain coefficients to be set of features. For classification purposes, these features are combined with powerful classifiers such as neural networks or support vector machines (SVMs). Gurcan *et al.* [1] provided detailed discussion of feature and classifier selection for histopathological analysis.

Sparse representation frameworks have also been proposed for medical applications recently [3], [4], [19]. Specifically, Srinivas *et al.* [2], [3] presented a multi-channel histopathological image as a sparse linear combination of training examples under channel-wise constraints and proposed a residual-based classification technique. Yu *et al.* [20] proposed a method for cervigram segmentation based on sparsity and group clustering priors. Song *et al.* [21], [22] proposed a locality-constrained and a large-margin representation method for medical image classification. In addition, Parvin *et al.* [4] combined a dictionary learning framework with an autoencoder to learn sparse features for classification. Chang *et al.* [23] extended this work by adding a spatial pyramid matching to enhance the performance.

A. Challenges and Motivation

While histopathological analysis shares some traits with other image classification problems, there are also principally distinct challenges specific to histopathology. The central challenge comes from the geometric richness of tissue images, resulting in the difficulty of obtaining reliable discriminative features for classification. Tissues from different organs have structural and morphological diversity which often leads to highly customized feature extraction solutions for each problem and hence the techniques lack broad applicability.

[†]T. H. Vu, H. S. Mousavi, and V. Monga are with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA 16802, USA (e-mail: thv102@psu.edu).

[*]Ganesh Rao is with the Department of Neurosurgery, and UK Arvind Rao is with the Department of Bionformatics and Computational Biology, both at the University of Texas MD Anderson Cancer Center, Houston, TX, USA.

Our work aims to produce a more versatile histopathological image classification system through the design of discriminative, class-specific dictionaries which is hence capable of automatic feature discovery using example training image samples. Our proposal evolves from the sparse representation-based classifier (SRC) [24] which has received significant attention recently [25]–[27]. Wright *et al.* [24] proposed SRC with the assumption that given a sufficient collection of training samples from one class, which is referred as a dictionary, any other test sample from the same class can be roughly expressed as a linear combination of these training samples. As a result, any test sample has a *sparse* representation in terms of a big dictionary comprising of class-specific sub-dictionaries. Recent work has shown that learned and data adaptive dictionaries significantly outperform ones constructed by simply stacking training samples together as in [24]. In particular, methods with class-specific constraints [28]–[31] are known to further enhance classification performance.

Being mindful of the aforementioned challenges, we design via optimization, a discriminative dictionary for each class by imposing sparsity constraints that minimizes intra-class differences, while simultaneously emphasizing inter-class differences. On one hand, small intra-class differences encourage the comprehensibility of the set of learned bases, which has ability of representing in-class samples with only few bases (intra class sparsity). This encouragement forces the model to find the representative bases in that class. On the other hand, large inter-class differences prevent bases of a class from sparsely representing samples from other classes. Concretely, given a dictionary from a particular class $\mathbf{D}$ with k bases and a certain sparsity level $L \ll k$, we define an *L-subspace* of $\mathbf{D}$ as a span of a subset of L bases from $\mathbf{D}$. Our proposed Discriminative Feature-oriented Dictionary Learning (DFDL) aims to build dictionaries with this key property: any sample from a class is reasonably close to *an L*-subspace of the associated dictionary while a complementary sample is far from *any L*-subspace of that dictionary. Illustration of the proposed idea is shown in Fig. 1.

B. Contributions

The main contributions of this paper are as follows:

1) **A new discriminative dictionary learning method**[1] for automatic feature discovery in histopathological images is presented to mitigate the generally difficult problem of feature extraction in histopathological images. Our *discriminative* framework learns dictionaries that emphasize inter-class differences while keeping intra-class differences small, resulting in enhanced classification performance. The design is based on solving a sparsity constrained optimization problem, for which we develop a tractable algorithmic solution.

2) **Broad Experimental Validation and Insights**. Experimental validation of DFDL is carried out on three diverse histopathological datasets to show its broad applicability. *The first dataset* is courtesy of the Clarian Pathology Lab and Computer and Information Science Dept., Indiana University-
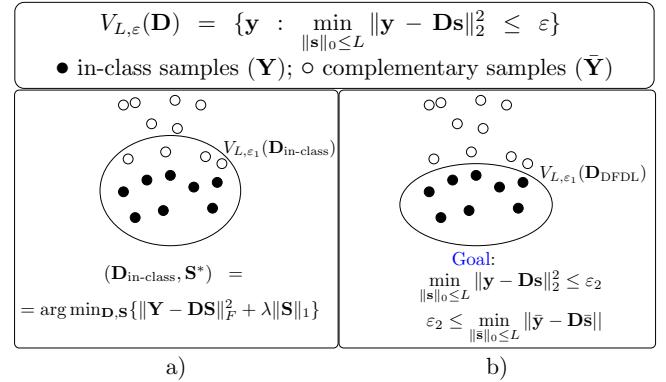
$$V_{L,\varepsilon}(\mathbf{D}) \;=\; \{\mathbf{y} \;:\; \min_{\|\mathbf{s}\|_0 \leq L} \|\mathbf{y} - \mathbf{D}\mathbf{s}\|_2^2 \;\leq\; \varepsilon\}$$

● in-class samples ($\mathbf{Y}$); ○ complementary samples ($\bar{\mathbf{Y}}$)

$V_{L,\varepsilon_1}(\mathbf{D}_{\text{in-class}})$

$V_{L,\varepsilon_1}(\mathbf{D}_{\text{DFDL}})$

$(\mathbf{D}_{\text{in-class}}, \mathbf{S}^*) \;=\;$

$= \arg\min_{\mathbf{D},\mathbf{S}}\{\|\mathbf{Y} - \mathbf{D}\mathbf{S}\|_F^2 + \lambda\|\mathbf{S}\|_1\}$

Goal:
$$\min_{\|\mathbf{s}\|_0 \leq L} \|\mathbf{y} - \mathbf{D}\mathbf{s}\|_2^2 \leq \varepsilon_2$$
$$\varepsilon_2 \leq \min_{\|\bar{\mathbf{s}}\|_0 \leq L} \|\bar{\mathbf{y}} - \mathbf{D}\bar{\mathbf{s}}\|$$

a)   b)

Figure 1: Main idea: a) The sparse representation space of learned dictionary using in-class samples only, e.g. KSVD [33] or ODL [34]($V_{L,\varepsilon_1}(\mathbf{D}_{\text{in-class}})$ may also cover some complementary samples), and b) desired DFDL ($V_{L,\varepsilon_2}(\mathbf{D}_{\text{DFDL}})$) cover in-class samples only.

Purdue University Indianapolis (IUPUI). The images acquired by the process described in [6] correspond to human Intraductal Breast Lesions (IBL). Two well-defined categories will be classified: Usual Ductal Hyperplasia (UDH)–benign, and Ductal Carcinoma In Situ (DCIS)–actionable. *The second dataset* contains images of brain cancer (glioblastoma or GBM) obtaind from The Cancer Genome Atlas (TCGA) [35] provided by the National Institute of Health, and will henceforth be referred as the TCGA dataset. For this dataset, we address the problem of detecting MicroVascular Proliferation (MVP) regions, which is an important indicator of a high grade glioma (HGG) [5]. *The third dataset* is provided by the Animal Diagnostics Lab (ADL), The Pennsylvania State University. It contains tissue images from three mammalian organs - kidney, lung and spleen. For each organ, images will be assigned into one of two categories–healthy or inflammatory. The samples of these three datasets are given in Figs. 2, 3, and 4, respectively. Extensive experimental results show that our method outperforms many competing methods, particularly in low training scenarios. In addition, Receiver Operating Characteristic (ROC) curves are provided that facilitate a trade-off between false alarm and miss rates.

3) **Complexity analysis.** We derive the computational complexity of DFDL as well as competing dictionary learning methods in terms of approximate number of operations needed. We also report experimental running time of DFDL and three other dictionary learning methods.

4) **Reproducibility**. All results in the manuscript are reproducible via a user-friendly software[2]. The software (MATLAB toolbox) is also provided with the hope of usage in future research and comparisons via peer researchers.

The remainder of this paper is organized as follows. Our proposed DFDL via a sparsity constrained optimization and the solution for the said optimization problem are detailed in Section II. Section II-D also presents our algorithmic classification procedures for the three diverse histopathological problems stated above. Section III presents classification accuracy as well as run-time complexity comparisons with existing methods in the literature to reveal merits of the proposed
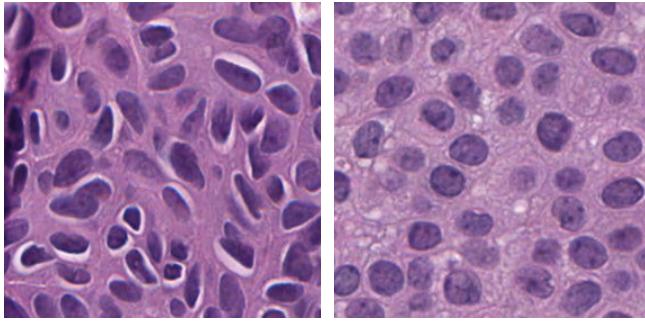
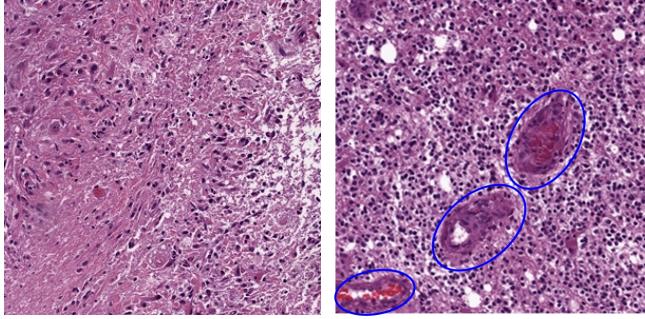Figure 2: Samples form IBL dataset: left-UDH, right-DCIS

Figure 3: Samples form TCGA dataset. Left: regions without MVP. Right: regions with MVP are inside blue ovals.

DFDL. A detailed analytical comparison of complexity against competing dictionary learning methods is provided in the Appendix. Section IV concludes the paper.

II. CONTRIBUTIONS

A. Notation

The vectorization of a small block (or patch)[3] extracted from an image is denoted as a column vector $\mathbf{y} \in \mathbb{R}^d$ which will be referred as a sample. In a classification problem where we have c different categories, collection of all data samples from class i (i can vary between 1 to c) forms the matrix $\mathbf{Y}_i \in \mathbb{R}^{d \times N_i}$ and let $\bar{\mathbf{Y}}_i \in \mathbb{R}^{d \times \bar{N}_i}$ be the matrix containing all complementary data samples i.e. those that are not in class i. We denote by $\mathbf{D}_i \in \mathbb{R}^{d \times k_i}$ the dictionary of class i that is desired to be learned through our DFDL method.

For a vector $\mathbf{s} \in \mathbb{R}^k$, we denote by $\|\mathbf{s}\|_0$ the number of its non-zero elements. The sparsity constraint of $\mathbf{s}$ can be formulated as $\|\mathbf{s}\|_0 \leq L$. For a matrix $\mathbf{S}$, $\|\mathbf{S}\|_0 \leq L$ means that *each* column of $\mathbf{S}$ has no more than L non-zero elements.

B. Discriminative Feature-oriented Dictionary Learning

We aim to build class-specific dictionaries $\mathbf{D}_i$ such that each $\mathbf{D}_i$ can sparsely represent samples from class i but is *poorly* capable of representing its complementary samples with small number of bases. Concretely, for the learned dictionaries we need:

$$\min_{\|\mathbf{s}_l\|_0 \leq L_i} \|\mathbf{y}_l - \mathbf{D}_i \mathbf{s}_l\|_2^2, \quad \forall l = 1, 2, \ldots, N_i \quad \text{to be small}$$

and

$$\min_{\|\bar{\mathbf{s}}_m\|_0 \leq L_i} \|\bar{\mathbf{y}}_m - \mathbf{D}_i \bar{\mathbf{s}}_m\|_2^2, \quad \forall m = 1, 2, \ldots, \bar{N}_i \quad \text{to be large.}$$

---

[3]In our work, a training vector is obtained by vectorizing all three RGB channels followed by concatenating them together to have a long vector.
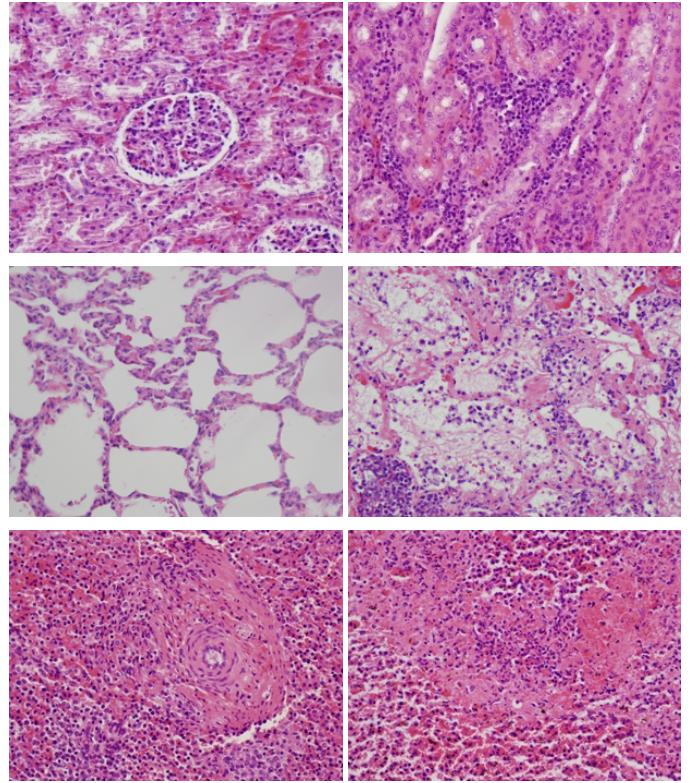

Figure 4: Samples form ADL dataset. First row: kidney. Second row: lung. Last row: spleen. Left: healthy. Right: inflammatory.

where L_i controls the sparsity level. These two sets of conditions could be simplified in the matrix form:

$$\text{intra-class differences:} \quad \frac{1}{N_i} \min_{\|\mathbf{S}_i\|_0 \leq L_i} \|\mathbf{Y}_i - \mathbf{D}_i \mathbf{S}_i\|_F^2 \quad \text{small,} \quad (1)$$

$$\text{inter-class differences:} \quad \frac{1}{\bar{N}_i} \min_{\|\bar{\mathbf{S}}_i\|_0 \leq L_i} \|\bar{\mathbf{Y}}_i - \mathbf{D}_i \bar{\mathbf{S}}_i\|_F^2 \quad \text{large.} \quad (2)$$

The averaging operations $\left(\dfrac{1}{N_i} \text{ and } \dfrac{1}{\bar{N}_i} \right)$ are taken here for avoiding the case where the largeness of inter-class differences is solely resulting from $\bar{N}_i \gg N_i$.

For simplicity, from now on, we consider only one class and drop the class index in each notion, i.e., using $\mathbf{Y}, \mathbf{D}, \mathbf{S}, \bar{\mathbf{S}}, N, \bar{N}, L$ instead of $\mathbf{Y}_i, \mathbf{D}_i, \mathbf{S}_i, \bar{\mathbf{S}}_i, N_i, \bar{N}_i$ and L_i. Based on the argument above, we formulate the optimization problem for each dictionary:

$$\mathbf{D}^* = \arg\min_{\mathbf{D}} \left(\frac{1}{N} \min_{\|\mathbf{S}\|_0 \leq L} \|\mathbf{Y} - \mathbf{D}\mathbf{S}\|_F^2 - \frac{\rho}{\bar{N}} \min_{\|\bar{\mathbf{S}}\|_0 \leq L} \|\bar{\mathbf{Y}} - \mathbf{D}\bar{\mathbf{S}}\|_F^2 \right), \quad (3)$$

where ρ is a positive regularization parameter. The first term in the above optimization problem encourages intra-class differences to be small, while the second term, with minus sign, emphasizes inter-class differences. By solving the above problem, we can jointly find the appropriate dictionaries as we desire in (1) and (2).

How to choose L: The sparsity level L for classes might be different. For one class, if L is too small, the dictionary might not appropriately express in-class samples, while if it is too large, the dictionary might be able to represent complementary

samples as well. In both cases, the classifier might fail to determine identity of one new test sample. We propose a method for estimating L as follows. First, a dictionary is learned using ODL [34] using in-class samples $\mathbf{Y}$ only:

$$(\mathbf{D}^0, \mathbf{S}^0) = \arg\min_{\mathbf{D},\mathbf{S}}\{\|\mathbf{Y} - \mathbf{D}\mathbf{S}\|_F^2 + \lambda\|\mathbf{S}\|_1\}, \qquad (4)$$

where λ is a positive regularization parameter controlling the sparsity level. Note that the same λ can still lead to different L for different classes, depending on the intra-class variablity of each class. Without prior knowledge of those variablities, we choose the same λ for every class. After $\mathbf{D}^0$ and $\mathbf{S}^0$ have been computed, $\mathbf{D}^0$ could be utilized as a warm initialization of $\mathbf{D}$ in our algorithm, $\mathbf{S}^0$ could be used to estimate the sparsity level L:

$$L \approx \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{s}_i^0\|_0. \qquad (5)$$

Classification scheme: In the same manner with SRC [24], a new patch $\mathbf{y}$ is classified as follows. Firstly, the sparse codes $\hat{\mathbf{s}}$ are calculated via l_1-norm minimization:

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}}\left\{\|\mathbf{y} - \mathbf{D}_{total}\mathbf{s}\|_2^2 + \gamma\|\mathbf{s}\|_1\right\}, \qquad (6)$$

where $\mathbf{D}_{total} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_c]$ is the collection of all dictionaries and γ is a scalar constant. Secondly, the identity of $\mathbf{y}$ is determined as: $\arg\min_{i\in\{1,\ldots,c\}}\{r_i(\mathbf{y})\}$ where

$$r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{D}_i\delta_i(\hat{\mathbf{s}})\|_2 \qquad (7)$$

and $\delta_i(\hat{\mathbf{s}})$ is part of $\hat{\mathbf{s}}$ associated with class i.

C. Proposed solution

We use an iterative method to find the optimal solution for the problem in (3). Specifically, the process is iterative by fixing $\mathbf{D}$ while optimizing $\mathbf{S}$ and $\bar{\mathbf{S}}$ and vice versa.

In the sparse coding step, with fixed $\mathbf{D}$, optimal sparse codes $\mathbf{S}^*, \bar{\mathbf{S}}^*$ can be found by solving:

$$\mathbf{S}^* = \arg\min_{\|\mathbf{S}\|_0\leq L}\|\mathbf{Y} - \mathbf{D}\mathbf{S}\|_F^2; \quad \bar{\mathbf{S}}^* = \arg\min_{\|\bar{\mathbf{S}}\|_0\leq L}\|\bar{\mathbf{Y}} - \mathbf{D}\bar{\mathbf{S}}\|_F^2.$$

With the same dictionary $\mathbf{D}$, these two sparse coding problems can be combined into the following one:

$$\hat{\mathbf{S}}^* = \arg\min_{\|\hat{\mathbf{S}}\|_0\leq L}\left\|\hat{\mathbf{Y}} - \mathbf{D}\hat{\mathbf{S}}\right\|_F^2. \qquad (8)$$

with $\hat{\mathbf{Y}} = [\mathbf{Y}, \bar{\mathbf{Y}}]$ being the matrix of all training samples and $\hat{\mathbf{S}} = [\mathbf{S}, \bar{\mathbf{S}}]$. This sparse coding problem can be solved effectively by OMP [36] using SPAMS toolbox [37].

For the bases update stage, $\mathbf{D}^*$ is found by solving:

$$\mathbf{D}^* = \quad \arg\min_{\mathbf{D}}\left\{\frac{1}{N}\|\mathbf{Y} - \mathbf{D}\mathbf{S}\|_F^2 - \frac{\rho}{\bar{N}}\|\bar{\mathbf{Y}} - \mathbf{D}\bar{\mathbf{S}}\|_F^2\right\}, \qquad (9)$$

$$= \quad \arg\min_{\mathbf{D}}\left\{-2\text{trace}(\mathbf{E}\mathbf{D}^\top) + \text{trace}(\mathbf{D}\mathbf{F}\mathbf{D}^\top)\right\}. \qquad (10)$$

We have used the equation $\|\mathbf{M}\|_F^2 = \text{trace}(\mathbf{M}\mathbf{M}^\top)$ for any matrix $\mathbf{M}$ to derive (10) from (9) and denoted:

$$\mathbf{E} = \frac{1}{N}\mathbf{Y}\mathbf{S}^\top - \frac{\rho}{\bar{N}}\bar{\mathbf{Y}}\bar{\mathbf{S}}^\top; \quad \mathbf{F} = \frac{1}{N}\mathbf{S}\mathbf{S}^\top - \frac{\rho}{\bar{N}}\bar{\mathbf{S}}\bar{\mathbf{S}}^\top. \qquad (11)$$

Algorithm 1 Discriminative Feature-oriented Dictionary Learning

---

function $\mathbf{D}^* = \text{DFDL}(\mathbf{Y}, \bar{\mathbf{Y}}, k, \rho)$
    INPUT: $\mathbf{Y}, \bar{\mathbf{Y}}$: collection of all in-class samples and complementary samples. k: number of bases in the dictionary. ρ: the regularization parameter.
    1. Choose initial $\mathbf{D}^*$ and L as in (4) and (5).
    while not converged **do**
        2. Fix $\mathbf{D} = \mathbf{D}^*$ and update $\mathbf{S}, \bar{\mathbf{S}}$ by solving (8);
        3. Fix $\mathbf{S}, \bar{\mathbf{S}}$, calculate:

$$\mathbf{E} = \frac{1}{N}\mathbf{Y}\mathbf{S}^\top - \frac{\rho}{\bar{N}}\bar{\mathbf{Y}}\bar{\mathbf{S}}^\top; \quad \mathbf{F} = \frac{1}{N}\mathbf{S}\mathbf{S}^\top - \frac{\rho}{\bar{N}}\bar{\mathbf{S}}\bar{\mathbf{S}}^\top.$$

        4. Update $\mathbf{D}$ from:

$$\mathbf{D}^* = \arg\min_{\mathbf{D}}\left\{-2\text{trace}(\mathbf{E}\mathbf{D}^\top) + \text{trace}\left(\mathbf{D}\left(\mathbf{F} - \lambda_{\min}(\mathbf{F})\mathbf{I}\right)\mathbf{D}^\top\right)\right\}$$

$$\text{subject to:}\|\mathbf{d}_i\|_2^2 = 1, i = 1, 2, \ldots, k.$$

    end while
    RETURN: $\mathbf{D}^*$
end function

The objective function in (10) is very similar to the objective function in the dictionary update stage problem in [34] except that it is not guaranteed to be convex. It is convex if and only if $\mathbf{F}$ is positive semidefinite. For the discriminative dictionary learning problem, the symmetric matrix $\mathbf{F}$ is *not* guaranteed to be positive semidefinite, even all of its eigenvalues are real. In the worst case, where $\mathbf{F}$ is negative semidefinite, the objective function in (10) becomes concave; if we apply the same dictionary update algorithm as in [34], we will obtain its maximum solution instead of the minimum.

To deal with this situation, we propose a technique which convexifies the objective function based on the following observation.

If we look back to the main optimization problem stated in (3):

$$\mathbf{D}^* = \arg\min_{\mathbf{D}}\left(\frac{1}{N}\min_{\|\mathbf{S}\|_0\leq L}\|\mathbf{Y} - \mathbf{D}\mathbf{S}\|_F^2 - \frac{\rho}{\bar{N}}\min_{\|\bar{\mathbf{S}}\|_0\leq L}\|\bar{\mathbf{Y}} - \mathbf{D}\bar{\mathbf{S}}\|_F^2\right),$$

we can see that if $\mathbf{D} = \begin{bmatrix}\mathbf{d}_1 & \mathbf{d}_2 & \ldots & \mathbf{d}_k\end{bmatrix}$ is an optimal solution, then $\mathbf{D} = \begin{bmatrix}\frac{\mathbf{d}_1}{a_1} & \frac{\mathbf{d}_2}{a_2} & \ldots & \frac{\mathbf{d}_k}{a_k}\end{bmatrix}$ is also an optimal solution as we multiply j-th rows of optimal $\mathbf{S}$ and $\bar{\mathbf{S}}$ by a_j, where $a_j, j = 1, 2, \ldots, k$, are arbitrary nonzero scalars. Consequently, we can introduce constraints: $\|\mathbf{d}_i\|_2^2 = 1, j = 1, 2, \ldots, k$, without affecting optimal value of (10). With these constraints, $\text{trace}(\mathbf{D}\lambda_{\min}(\mathbf{F})\mathbf{I}_k\mathbf{D}^\top) = \lambda_{\min}(\mathbf{F})\text{trace}(\mathbf{D}^\top\mathbf{D}) = \lambda_{\min}(\mathbf{F})\sum_{i=1}^{k}\mathbf{d}_i^\top\mathbf{d}_i = k\lambda_{\min}(\mathbf{F})$, where $\lambda_{\min}(\mathbf{F})$ is the minimum eigenvalue of $\mathbf{F}$ and $\mathbf{I}_k$ denotes the identity matrix, is a constant. Substracting this constant from the objective function will not change the optimal solution to (10). Essentially, the following problem in (12) is equivalent to (10):

$$\mathbf{D}^* = \arg\min_{\mathbf{D}}\{-2\text{trace}(\mathbf{E}\mathbf{D}^\top) + \text{trace}\left(\mathbf{D}(\mathbf{F} - \lambda_{\min}(\mathbf{F})\mathbf{I}_k)\mathbf{D}^\top\right)\} \qquad (12)$$

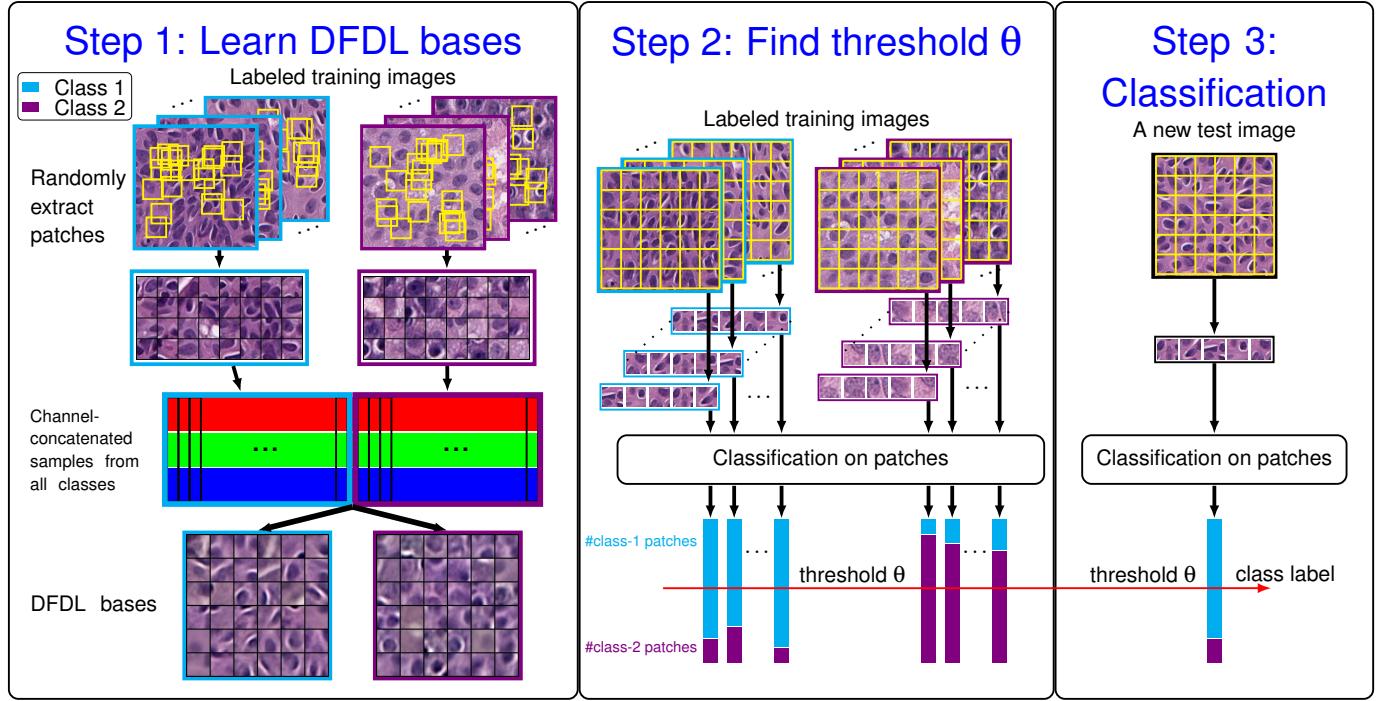$$\text{subject to:}\|\mathbf{d}_i\|_2^2 = 1, i = 1, 2, \ldots, k.$$

Figure 5: IBL/ADL classification procedure

The matrix $\hat{\mathbf{F}} = \mathbf{F} - \lambda_{\min}(\mathbf{F})\mathbf{I}_k$ is guaranteed to be positive semidefinite since all of its eigenvalues now are nonnegative, and hence the objective function in (12) is convex. Now, this optimization problem is very similar to the dictionary update problem in [34]. Then, $\mathbf{D}^*$ could be updated by the following iterations until convergence:

$$\mathbf{u}_j \quad \leftarrow \quad \frac{1}{\hat{\mathbf{F}}_{j,j}}(\mathbf{e}_j - \mathbf{D}\hat{\mathbf{f}}_j) + \mathbf{d}_j. \tag{13}$$

$$\mathbf{d}_j \quad \leftarrow \quad \frac{\mathbf{u}_j}{\|\mathbf{u}_j\|_2}. \tag{14}$$

where $\hat{\mathbf{F}}_{j,j}$ is the value of $\hat{\mathbf{F}}$ at coordinate (j,j) and $\hat{\mathbf{f}}_j$ denotes the j-th column of $\hat{\mathbf{F}}$.

Our DFDL algorithm is summarized in Algorithm 1.

D. Overall classification procedures for three datasets

In this section, we propose a DFDL-based procedure for classifying images in three datasets.

1) IBL and ADL datasets

The key idea in this procedure is that a healthy tissue image largely consists of healthy patches which cover a dominant portion of the tissue. This procedure is shown in Fig. 5 and consists of the following three steps:

Step 1: *Training DFDL bases for each class*. From labeled training images, training patches are randomly extracted (they might be overlapping). The size of these patches is picked based on pathologist input and/or chosen by cross validation [38]. After we have a set of *healthy* patches and a set of *diseased* patches for training, class-specific DFDL dictionaries and the associated classifier are trained by using Algorithm 1.

Step 2: *Learning a threshold θ for proportion of healthy patches in one healthy image*. Labeled training images are now divided into non-overlapping patches. Each of these patches is then classified using the DFDL classifier as described in Eq. (6) and (7). The main purpose of this step is to find the threshold θ such that healthy images have proportion of *healthy* patches greater or equal to θ and diseased ones have proportion of *diseased* patches less than θ. We can consider the proportion of healthy patches in one training image as its one-dimension feature. This feature is then put into a simple SVM to learn the threshold θ.

Step 3: *Classifying test images*. For an unseen test image, we calculate the proportion τ of *healthy* patches in the same way described in Step 2. Now, the identity of the image is determined by comparing the proportion τ to θ. It is categorized as healthy (diseased) if $\tau \geq (<)\theta$. The procedure readily generalizes to multi-class problems.

2) MVP detection problem in TCGA dataset

As described earlier, MicroVascular Proliferation (MVP) is the presence of blood vessels in a tissue and it is an important indicator of a high-grade tumor in brain glioma. Essentially presence of one such region in the tissue image indicates the high-grade tumor. Detection of such regions in TCGA dataset is an inherently hard problem and unlike classifying images in IBL and ADL datasets which are distinguishable by researching small regions, it requires more effort and investigation on larger connected regions. This is due to the fact that an MVP region may significantly vary in size and is usually surrounded by tumor cells which are actually benign or low grade. In addition, an MVP region is characterized by the presence of enlarged vessels in the tissue with different color shading and thick layers of cell rings inside the vessel (see Fig. 3). We define a patch as *MVP* if it lies entirely within
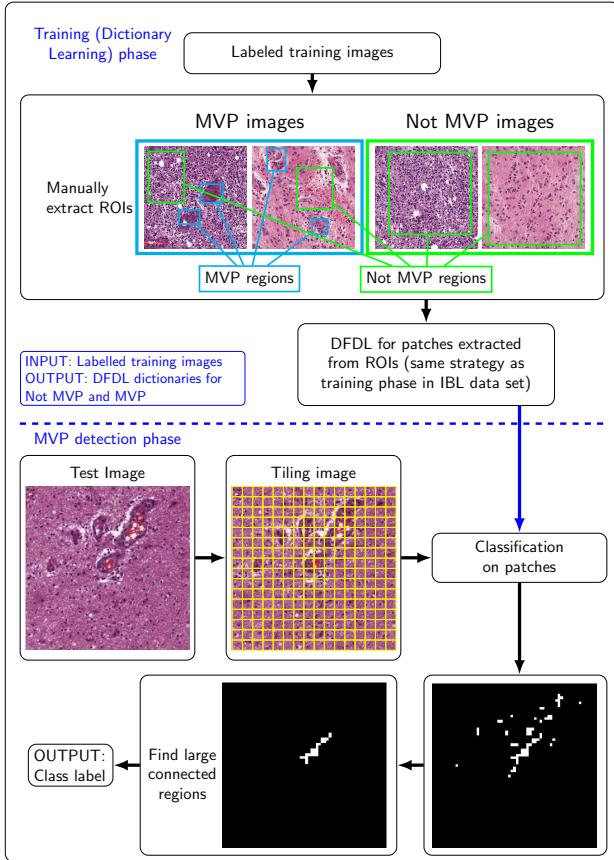
Figure 6: MVP detection procedure

an MVP region and as *Not MVP* otherwise. We also define a region as Not MVP if it does not contain any MVP patch. The procedure consists of two steps:

Step 1: *Training phase*. From training data, MVP regions and Not MVP regions are manually extracted. Note that while MVP regions come from MVP images only, Not MVP regions might appear in all images. From these extracted regions, DFDL dictionaries are obtained in the same way as in step 1 of IBL/ADL classification procedure described in section II-D1 and Fig. 5.

Step 2: *MVP detection phase:* A new unknown image is decomposed into non-overlapping patches. These patches are then classified using DFDL model learned before. After this step, we have a collection of patches classified as MVP. A region with large number of connected classified-as-MVP patches could be considered as an MVP region. If the final image does not contain any MVP region, we categorize the image as a Not MVP; otherwise, it is classified as MVP. The definition of connected regions contains a parameter m, which is the number of connected patches. Depending on m, positive patches might or might not appear in the final step. Specifically, if m is small, false positives tend to be determined as MVP patches; if m is large, true positives are highly likely eliminated. To determine m, we vary it from 1 to 20 and compute its ROC curve for training images and then simply pick the point which is closest to the origin and find the *optimal m*. This procedure is visualized in Fig. 6.

III. VALIDATION AND EXPERIMENTAL RESULTS

In this section, we present the experimental results of applying DFDL to three diverse histopathological image datasets and compare our results with different competing methods:

• WND-CHARM [10], [11] in conjunction with SVM: this method combines state-of-the-art feature extraction and classification methods. We use the collection of features from WND-CHARM, which is known to be a powerful toolkit of features for medical images. While the original paper used weighted nearest neighbor as a classifier, we use a more powerful classifier (SVM [39]) to further enhance classification accuracy. We pick the most relevant features for histopathology [1], including but not limited to (color channel-wise) histogram information, image statistics, morphological features and wavelet coefficients from each color channel. The source code for WND-CHARM is made available by the National Institute of Health online at http://ome.grc.nia.nih.gov/.

• SRC [24]: We apply SRC on the vectorization of the luminance channel of the histopathological images, as proposed initially for face recognition and applied widely thereafter.

• SHIRC [3]: Srinivas *et al.* [2], [3] presented a simultaneous sparsity model for multi-channel histopathology image representation and classification which extends the standard SRC [24] approach by designing three color dictionaries corresponding to the RGB channels. The MATLAB code for the algorithms is posted online at: http://signal.ee.psu.edu/histimg.html.

• LC-KSVD [29] and FDDL [31]: These are two well-known dictionary learning methods which were applied to object recognition such as face, digit, gender, vehicle, animal, etc, but to our knowledge, have not been applied to histopathological image classification. To obtain a fair comparison, dictionaries are learned on the same training patches. Classification is then carried out using the learned dictionaries on non-overlapping patches in the same way described in Section II-D.

• Nayak's: In recent relevant work, Nayak *et al.* [4] proposed a patch-based method to solve the problem of classification of tumor histopathology via sparse feature learning. The feature vectors are then fed into SVM to find the class label of each patch.

A. Experimental Set-Up: Image Datasets

IBL dataset: Each image contains a number of regions of interest (RoIs), and we have chosen a total of 120 images (RoIs), consisting of a randomly selected set of 20 images for training and the remaining 100 RoIs for test. Images are downsampled for computational purposes such that size of a cell is around 20-by-20 (pixels). Examples of images from this dataset are shown in Fig. 2. Experiments in section III-B below are conducted with 10 training images per class, 10000 patches of size 20-by-20 for training per class, $k = 500$ bases for each dictionary, $\lambda = 0.1$ and $\rho = 0.001$. These parameters are chosen using cross-validation [38].

ADL dataset: This dataset contains bovine histopathology images from three sub-datasets of kidney, lung and spleen. Each sub-dataset consists of images of size 4000×3000 pixels
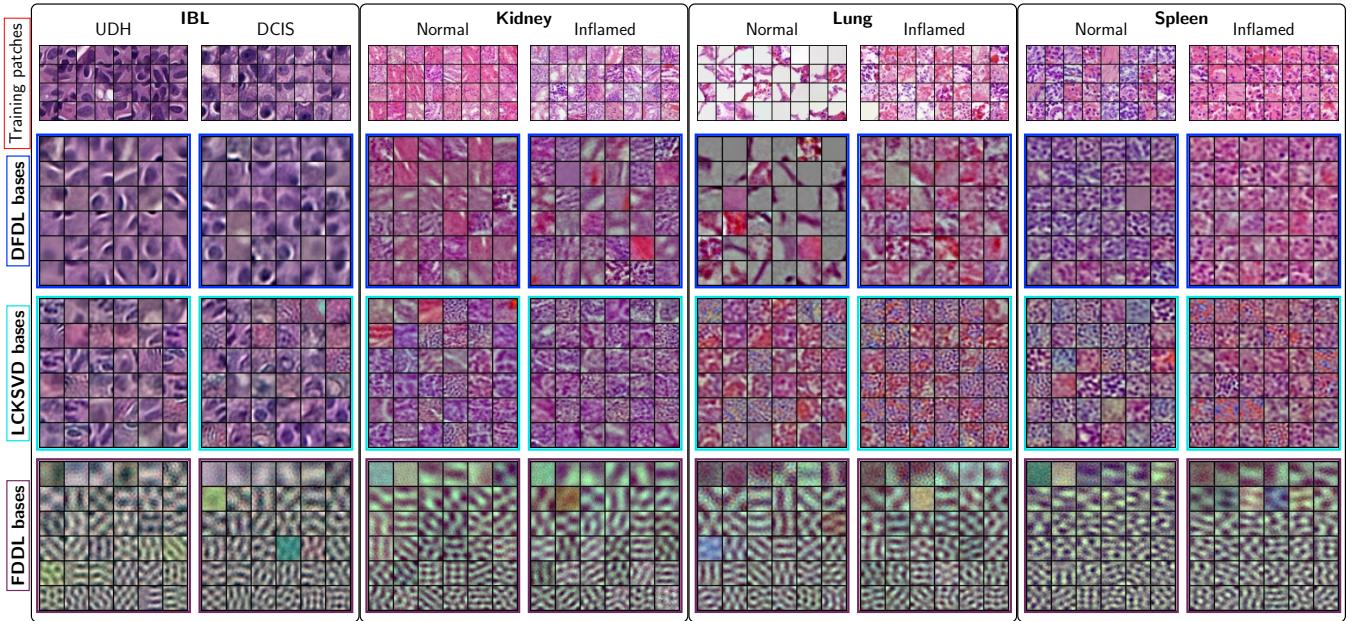
Figure 7: Example bases learned from different methods on different datasets. DFDL, LC-KSVD [29], FDDL [31] in IBL and ADL datasets.

from two classes: healthy and inflammatory. Each class has around 150 images from which 40 images are chosen for training, the remaining ones are used for testing. Number of training patches, bases, λ and ρ are the same as in the IBL dataset. The classification procedure for IBL and ADL datasets is described in Section II-D1.

TCGA dataset: We use a total of 190 images (RoIs) (resolution 3000×3000) from the TCGA, in which 57 images contain MVP regions and 133 ones have no traces of MVP. From each class, 20 images are randomly selected for training. The classification procedure for this dataset is described in Section II-D2.

Each tissue specimen in these datasets is fixed on a scanning bed and digitized using a digitizer at $40\times$ magnification.

B. Validation of Central Idea: Visualization of Discovered Features

This section provides experimental validation of the central hypothesis of this paper: by imposing sparsity constraint on forcing intra-class differences to be small, while simultaneously emphasizing inter-class differences, the class-specific bases obtained are discriminative.

Example bases obtained by different dictionary learning methods are visualized in Fig. 7. By visualizing these bases, we emphasize that our DFDL is able to look for discriminative visual features from which pathologists could understand the reasons behind diseases. In the spleen dataset for example, it is really difficult to realize the differences between two classes by human eyes. However, by looking at DFDL learned bases, we can see that the distribution of cells in two classes are different such that a larger number of cells appears in a normal patch. These differences may provide pathologists one visual cue to classify these images without advanced tools. Moreover, for IBL dataset, UDH bases visualize elongated cells with sharp edges while DCIS bases present more rounded
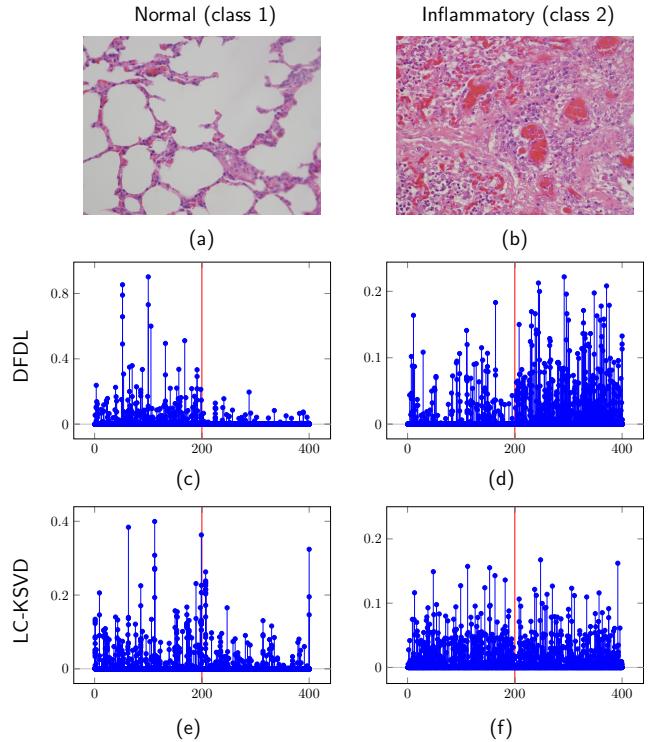

Figure 8: Example of sparse codes using DFDL and LC-KSVD approaches on lung dataset. Left: normal lung (class 1). Right: inflammatory lung (class 2). Row 1: test images. Row 2: Sparse codes visualization using DFDL. Row 3: Sparse codes visualization using LC-KSVD. x axis indicates the dimensions of sparse codes with codes on the left of red lines corresponding to bases of class 1, those on the right are in class 2. y axis demonstrates values of those codes. In one vertical line, different dots represent values of non-zeros coefficients of different patches.

cells with blurry boundaries, which is consistent with their descriptions in [3] and [6]; for ADL-Lung, we observe that a healthy lung is characterized by large clear openings of the
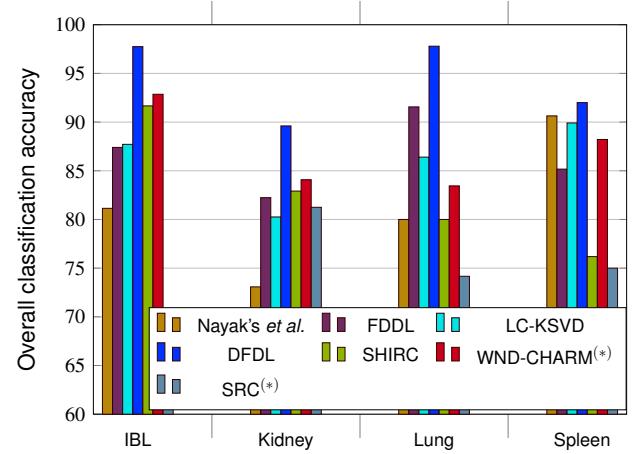
alveoli, while in the inflamed lung, the alveoli are filled with bluish-purple inflammatory cells. This distinction is very clear in the bases learned from DFDL where white regions appear more in normal bases than in inflammatory bases and no such information can be deduced from LC-KSVD or FDDL bases. In comparison, FDDL fails to discover discriminative visual features that are interpretable and LC-KSVD learns bases with the inter-class differences being less significant than DFDL bases. Furthermore, these LC-KSVD bases do not present key properties of each class, especially in lung dataset.

To understand more about the significance of discriminative bases for classification, let us first go back to SRC [24]. For simplicity, let us consider a problem with two classes with corresponding dictionaries $\mathbf{D}_1$ and $\mathbf{D}_2$. The identity of a new patch $\mathbf{y}$, which, for instance, comes from class 1, is determined by equations (6) and (7). In order to obtain good results, we expect most of active coefficients to be present in $\delta_1(\hat{\mathbf{s}})$. For $\delta_2(\hat{\mathbf{s}})$, its non-zeros, if they exists should have small magnitude. Now, suppose that one basis, $\mathbf{d}_1$, in $\mathbf{D}_1$ looks very similar to another basis, $\mathbf{d}_2$, in $\mathbf{D}_2$. When doing sparse coding, if one patch in class 1 uses $\mathbf{d}_1$ for reconstruction, it is highly likely that a similar patch $\mathbf{y}$ in the same class uses $\mathbf{d}_2$ for reconstruction instead. This misusage may lead to the case $\|\mathbf{y} - \mathbf{D}_1\delta_1(\hat{\mathbf{s}})\| > \|\mathbf{y} - \mathbf{D}_2\delta_2(\hat{\mathbf{s}})\|$, resulting in a misclassified patch. For this reason, the more discriminative bases are, the better the performance.

To formally verify this argument, we do one experiment on one normal and one inflammatory image from lung dataset in which the differences of DFDL bases and LCKSVD bases are most significant. From these images, patches are extracted, then their sparse codes are calculated using two dictionaries formed by DFDL bases and LC-KSVD bases. Fig. 8 demonstrates our results. Note that the plots in Figs. 8c) and d) are corresponding to DFDL while those in Figs. 8e) and f) are for LC-KSVD. Most of active coefficients in Fig. 8c) are gathered on the left of the red line, and their values are also greater than values on the right. This means that $\mathbf{D}_1$ contributes more to reconstructing the lung-normal image in Fig. 8a) than $\mathbf{D}_2$ does. Similarly, most of active coefficients in Fig. 8d) locate on the right of the vertical line. This agrees with what we expect since the image in Fig. 8a) belongs to class 1 and the one in Fig. 8b) belongs to class 2. On the contrary, for LC-KSVD, active coefficients in Fig. 8f) are more uniformly distributed on both sides of the red line, which adversely affects classification. In Fig. 8e), although active coefficients are strongly concentrated to the left of the red line, this effect is even more pronounced with DFDL, i.e. in Fig. 8c).

C. Overall Classification Accuracy

To verify the performance of our idea, for IBL and ADL datasets, we present overall classification accuracies in the form of bar graphs in Fig. 9. It is evident that DFDL outperforms other methods in both datasets. Specifically, in IBL and ADL Lung, the overall classification accuracies of DFDL are over 97.75%, the next best rates come from WND-CHARM (92.85% in IBL) and FDDL (91.56% in ADL-Lung), respectively, and much higher than those reported in [3] and our own previous results in [32]. In addition, for ADL-Kidney

$^{(*)}$ Images are classified in whole image level.

Figure 9: Bar graphs indicating the overall classification accuracies (%) of the competing methods.
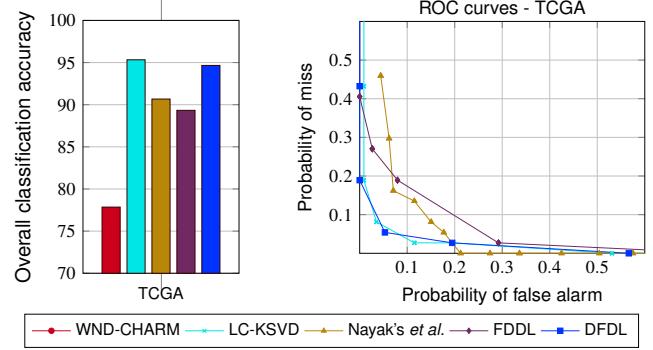

Figure 10: Bar graphs (left) indicating the overall classification accuracies (%) and the receiver operating characteristic (right) of the competing methods for TCGA dataset.

and ADL-Spleen, our DFDL also provides the best result with accuracy rates being nearly 90% and over 92%, respectively.

For the TCGA dataset, overall accuracy of competing methods are shown in Fig. 10, which reveals that DFDL performance is the second best, bettered only by LC-KSVD and by less than 0.67% (i.e. one more misclassified image for DFDL).

D. Complexity analysis

In this section, we compare the computational complexity for the proposed DFDL and competing dictionary learning methods: LC-KSVD [29], FDDL [31], and Nayak's [4]. The complexity for each dictionary learning method is estimated as the (approximate) number of operations required by each method in learning the dictionary (see Appendix for details). From Table II, it is clear that the proposed DFDL is the least expensive computationally. Note further, that the final column of Table II shows *actual run times* of each of the methods. The parameters were as follows: $c = 2$ (classes), $k = 500$ (bases per class), $N = 10,000$ (training patches per class), data dimension $d = 1200$ (3 channels $\times 20 \times 20$), sparsity level $L = 30$. The run time numbers in the final column of Table II are in fact consistent with numbers provided in Table III, which are calculated by plugging the above parameters into the second column of Table II.

Table IV: CONFUSION MATRIX: ADL (%).

Class	Kidney		Lung		Spleen		Method
	Health	inflammatory	Health	inflammatory	Health	inflammatory	
Health	83.27	16.73	83.20	16.80	87.23	12.77	WND-CHARM[*] [11]
	87.50	12.50	72.50	27.50	70.83	29.17	SRC[*] [24]
	82.50	17.50	75.00	25.00	65.00	35.00	SHIRC [3]
	83.26	16.74	*93.15*	6.85	86.94	13.06	FDDL [31]
	86.84	13.16	85.59	15.41	*89.75*	10.25	LC-KSVD [29]
	73.08	26.92	89.55	10.45	86.44	13.56	Nayak's *et al.* [4]
	88.21	11.79	**96.52**	3.48	**92.88**	7.12	DFDL
inflammatory	14.22	85.78	14.31	83.69	10.48	89.52	WND-CHARM[*] [11]
	25.00	75.00	24.17	75.83	20.83	79.17	SRC[*] [24]
	16.67	*83.33*	15.00	85.00	11.67	88.33	SHIRC [3]
	19.88	80.12	10.00	*90.00*	8.57	91.43	FDDL [31]
	19.25	81.75	10.89	89.11	8.57	91.43	LC-KSVD [29]
	26.92	73.08	25.90	74.10	6.05	**93.95**	Nayak's *et al.* [4]
	9.92	**90.02**	2.57	**97.43**	7.89	*92.01*	DFDL

[*] Images are classified in whole image level.

Table I: CONFUSION MATRIX: IBL.

Class	UDH	DCIS	Method
UDH	91.75	8.25	WND-CHARM[*] [11]
	68.00	32.00	SRC[*] [24]
	93.33	6.67	SHIRC [3]
	84.80	15.20	FDDL [31]
	90.29	9.71	LC-KSVD [29]
	85.71	14.29	Nayak's *et al.* [4]
	96.00	4.00	DFDL
DCIS	5.77	*94.23*	WND-CHARM[*] [11]
	44.00	56.00	SRC[*] [24]
	10.00	90.00	SHIRC [3]
	10.00	90.00	FDDL [31]
	14.86	85.14	LC-KSVD [29]
	23.43	76.57	Nayak's *et al.* [4]
	0.50	**99.50**	DFDL

[*] Images are classified in whole image level.

Table II: Complexity analysis for different dictionary learning methods.

Method	Complexity	Running time
DFDL	$c^2 kN(2d + L^2)$	~ 0.5 hours
LC-KSVD [29]	$c^2 kN(2d + 2ck + L^2)$	~ 3 hours
Nayak's *et al.* [4][*]	$c^2 kN(2d + 2qck) + c^2 dk^2$	~ 8 hours
FDDL [31][*]	$c^2 kN(2d + 2qck) + c^3 dk^2$	> 40 hours

[*] q is the number of iterations required for l_1-minimization in sparse coding step.

Table III: Estimated number of operations required in different dictionary learning methods.

Method	$q = 1$	$q = 3$	$q = 10$
DFDL	6.6×10^{10}	6.6×10^{10}	6.6×10^{10}
LC-KSVD [29]	1.06×10^{11}	1.06×10^{11}	1.06×10^{11}
Nayak's *et al.* [4]	8.92×10^{10}	1.692×10^{11}	4.492×10^{11}
FDDL [31]	9.04×10^{10}	1.704×10^{11}	4.504×10^{11}

Table V: CONFUSION MATRIX: TCGA (%).

Class	Not MVP	MVP	Method
Not VMP	76.68	23.32	WND-CHARM [11]
	92.92	7.08	Nayak's *et al.* [4]
	96.46	3.54	LC-KSVD [29]
	92.04	7.96	FDDL [31]
	94.69	5.31	DFDL
MVP	21.62	78.38	WND-CHARM [11]
	16.22	83.78	Nayak's *et al.* [4]
	8.10	*91.90*	LC-KSVD [29]
	18.92	81.08	FDDL [31]
	5.41	**94.59**	DFDL

Fig. 11 and Fig. 10 (right) show the ROC curves for all three datasets. The lowest curve (closest to the origin) has the best overall performance and the optimal operating point minimizes the sum of the miss and false alarm probabilities. It is evident that ROC curves for DFDL perform best in comparison to those of other state-of-the-art methods.

Remark: Note for ROC comparisons, we compare the different flavors of dictionary learning methods (the proposed DFDL, LC-KSVD, FDDL and Nayak's), this is because as Table V shows, they are the most competitive methods. Note for the IBL and ADL datasets, θ, as defined in Fig. 5, is changed from 0 to 1 to acquire the curves; whereas for the TCGA dataset, number of connected classified-as-MVP patches, *m*, is changed from 1 to 20 to obtain the curves. It is worth re-emphasizing that DFDL achieves these results even as its complexity is lower than competing methods.

E. Statistical Results: Confusion Matrices and ROC Curves

Next, we present a more elaborate interpretation of classification performance in the form of confusion matrices and ROC curves. Each row of a confusion matrix refers to the actual class identity of test images and each column indicates the classifier output. Table I, IV and V show the mean confusion matrices for all of three dataset. In continuation of trends from Fig. 9, in Table IV, DFDL offers the best disease detection accuracy in almost all datasets for each organ, while maintaining high classification accuracy for healthy images.

Typically in medical image classification problems, pathologists desire algorithms that reduce the probability of miss (diseased images are misclassified as healthy ones) while also ensuring that the false alarm rate remains low. However, there is a trade-off between these two quantities, conveniently described using receiver operating characteristic (ROC) curves.

F. Performance vs. size of training set

Real-world histopathological classification tasks must often contend with lack of availability of large training sets. To understand training dependence of the various techniques, we present a comparison of overall classification accuracy as a function of the training set size for the different methods. We
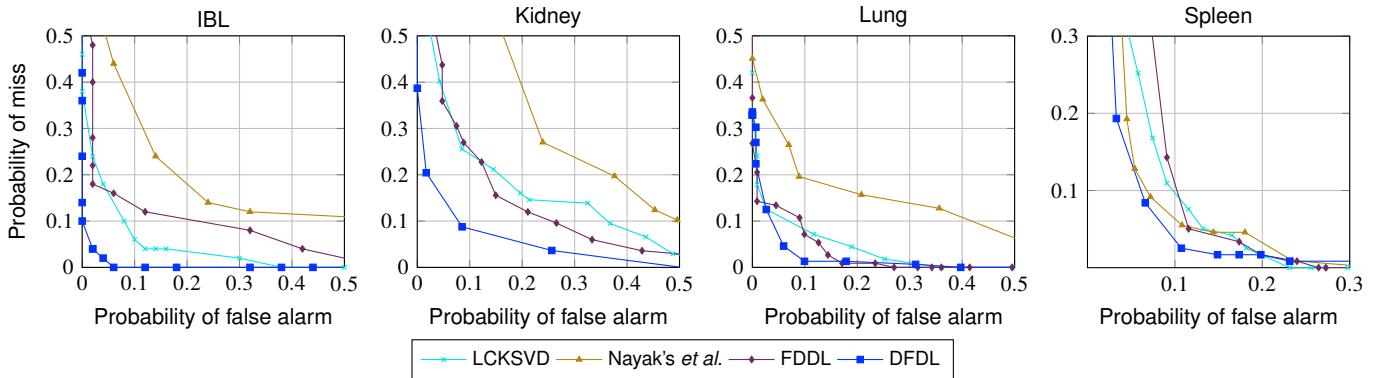
Figure 11: Receiver operating characteristic (ROC) curves for different organs, methods, and datasets (IBL and ADL).
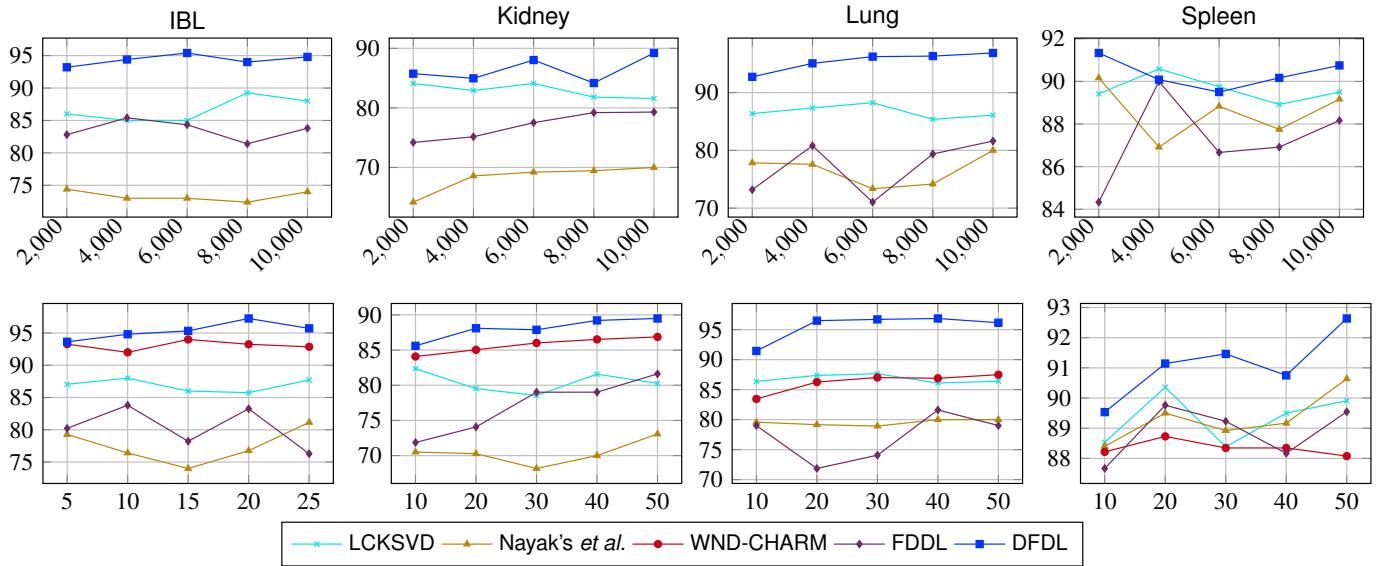

Figure 12: Overall classification accuracy (%) as a function of training set size per class. Top row: number of training patches. Bottom row: number of training images.
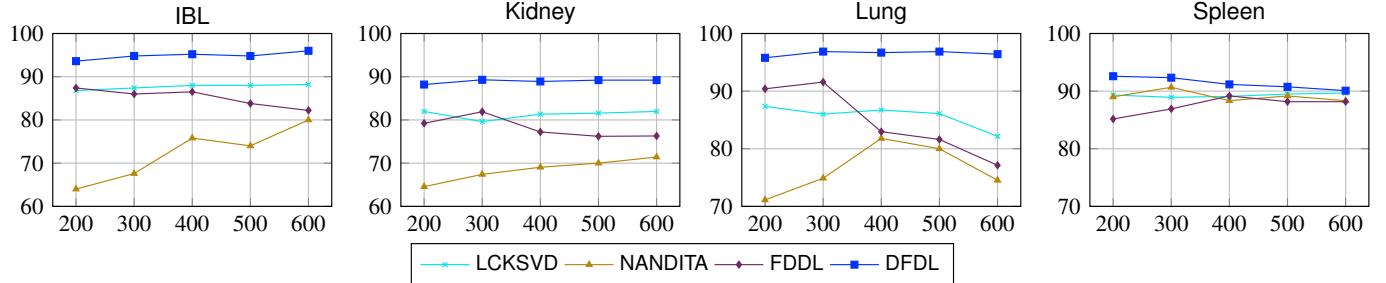

Figure 13: Overall classification accuracy (%) as a function of number of training bases.

also present a comparison of classification rates as a function of the number of training patches for different dictionary learning methods[4]. In Fig. 12, overall classification accuracy is reported for IBL and ADL datasets corresponding to five scenarios. It is readily apparent that DFDL exhibits the most graceful decline as training is reduced.

G. Performance vs. number of training bases

We now compare the behavior of each dictionary learning method as the number of bases in each dictionary varies from 200 to 600 (with patch size being fixed at 20×20 pixels). Results reported in Fig. 13 confirm that DFDL again outperforms other methods. In general, overall accuracies of DFDL on different datasets remain high when we reduce number of training bases. Interpreted another way, these results illustrate that DFDL is fairly robust to changes in parameters, which is a highly desirable trait in practice.

[4]Since WND-CHARM is applied in the whole image level, there is no result for it in comparison of training patches.

IV. DISCUSSION AND CONCLUSION

In this paper, we address the histopathological image classification problem from a feature discovery and dictionary learning standpoint. This is a very important and challenging problem and the main challenge comes from the geometrical richness of tissue images, resulting in the difficulty of obtaining reliable discriminative features for classification. Therefore, developing a framework capable of capturing this structural richness and being able to discriminate between different types is investigated and to this end, we propose the DFDL method which learns discriminative features for histopathology images. Our work aims to produce a more versatile histopathological image classification system through the design of discriminative, class-specific dictionaries which is hence capable of automatic feature discovery using example training image samples.

Our DFDL algorithm learns these dictionaries by leveraging the idea of sparse representation of in-class and out-of-class samples. This idea leads to an optimization problem which encourages intra-class similarities and emphasizes the inter-class differences. Ultimately, the optimization in (10) is done by solving the proposed equivalent optimization problem using a convexifying trick. Similar to other dictionary learning (machine learning approaches in general), DFDL also requires a set of regularization parameters. Our DFDL requires only one parameter, ρ, in its training process which is chosen by cross validation [38] – plugging different sets of parameters into the problem and selecting one which gives the best performance on the validation set. In the context of application of DFDL to real-world histopathological image slides, there are quite a few other settings should be carefully chosen, such as patch size, tiling method, number of connected components in the MVP detection etc. Of more importance is the patch size to be picked for each dataset which is mostly determined by consultation with the medical expert in the specific problem under investigation and the type of features that we should be looking for. For simplicity we employ regular tiling; however, using prior domain knowledge this may be improved. For instance in the context of MVP detection, informed selection of patch locations using existing disease detection and localization methods such as [5] can be used to further improve the detection of disease.

Experiments are carried out on three diverse histopathological datasets to show the broad applicability of the proposed DFDL method. It is illustrated our method is competitive with or outperforms state of the art alternatives, particularly in the regime of realistic or limited training set size. It is also shown that with minimal parameter tuning and algorithmic changes, DFDL method can be easily applied on different problems with different natures which makes it a good candidate for automated medical diagnosis instead of using customized and problem specific frameworks for every single diagnosis task. We also make a software toolbox available to help deploy DFDL widely as a diagnostic tool in existing histopathological image analysis systems. Particular problems such as grading and detecting specific regions in histopathology may be investigated using our proposed techniques.

APPENDIX
COMPLEXITY ANALYSIS

In this section, we compare the computational complexity for the proposed DFDL and competing dictionary learning methods: LC-KSVD [29], FDDL [31], and Nayak's [4]. The complexity for each dictionary learning method is estimated as the (approximate) number of operations required by each method in learning the dictionary. For simplicity, we assume that number of training samples, number of dictionary bases in each class are the same, which means: $N_i = N_j = N, k_i = k_j = k, \forall i, j = 1, 2, \ldots, c$, and also $L_i = L_j = L, \forall i, j = 1, 2, \ldots, c$. For the consitence, we have changed notations in those methods by denoting $\mathbf{Y}$ as training samples and $\mathbf{S}$ as the sparse code.

In most of dictionary learning methods, the complexity of sparse coding step, which is often a l_0 or l_1 minimization problem, dominates that of dictionary update step, which is typically solved by either block coordinate descent [34] or singular value decomposition [33]. Then, in order to compare the complexity of different dictionary learning methods, we focus on comparing the complexity of sparse coding steps in each iteration.

A. Complexity of the DFDL

The most expensive computation in DFDL is solving an Orthogonal Matching Pursuit (OMP [36]) problem. Given a set of samples $\mathbf{Y} \in \mathbb{R}^{d \times N}$, a dictionary $\mathbf{D} \in \mathbb{R}^{d \times k}$ and sparsity level L, the OMP problem is:

$$\mathbf{S}^* = \arg \min_{\|\mathbf{S}\|_0 \leq L} \|\mathbf{Y} - \mathbf{D}\mathbf{S}\|_F^2.$$

R. Rubinstein *et al.* [40] reported the complexity of Batch-OMP when the dictionary is stored in memory in its entirety as: $T_{\text{b-omp}} = N(2dk + L^2k + 3Lk + L^3) + dk^2$. Assuming an asymptotic behavior of $L \ll k \approx d \ll N$, the above expression can be simplified to:

$$T_{\text{b-omp}} \approx N(2dk + L^2k) = kN(2d + L^2). \quad (15)$$

This result will also be utilized in analyzing complexity of LC-KSVD.

The sparse coding step in our DFDL consists of solving c sparse coding problems: $\hat{\mathbf{S}} = \arg\min_{\|\mathbf{S}\|_0 \leq L} \|\hat{\mathbf{Y}} - \mathbf{D}_i \hat{\mathbf{S}}_i\|_F^2$. With $\hat{\mathbf{Y}} \in \mathbb{R}^{d \times cN}, \mathbf{D}_i \in \mathbb{R}^{d \times k}$, each problem has complexity of $k(cN)(2d + L^2)$. Then the total complexity of these c problems is: $T_{\text{DFDL}} \approx c^2 kN(2d + L^2)$.

B. Complexity of LC-KSVD

We consider LC-KSVD1 only (LC-KSVD2 has a higher complexity) whose optimization problem is written as [29]:

$$(\mathbf{D}, \mathbf{A}, \mathbf{S}) = \arg \min_{\mathbf{D}, \mathbf{A}, \mathbf{S}} \|\mathbf{Y} - \mathbf{D}\mathbf{S}\|_F^2 + \alpha \|\mathbf{Q} - \mathbf{A}\mathbf{S}\|_F^2 \text{ s.t. } \|\mathbf{s}_i\|_0 \leq L.$$

and it is rewritten in the K-SVD form:

$$(\mathbf{D}, \mathbf{A}, \mathbf{S}) = \arg \min_{\mathbf{D}, \mathbf{A}, \mathbf{S}} \left\| \begin{bmatrix} \mathbf{Y} \\ \sqrt{\alpha}\mathbf{Q} \end{bmatrix} - \begin{bmatrix} \mathbf{D} \\ \sqrt{\alpha}\mathbf{A} \end{bmatrix} \mathbf{S} \right\|_F^2 \text{ s.t. } \|\mathbf{s}_i\|_0 \leq L. \quad (16)$$

Since $\mathbf{Q} \in \mathbb{R}^{ck \times cN}$ and $\mathbf{A} \in \mathbb{R}^{ck \times ck}$, $\tilde{\mathbf{Y}} = \begin{bmatrix} \mathbf{Y} \\ \sqrt{\alpha}\mathbf{Q} \end{bmatrix} \in \mathbb{R}^{(d+ck) \times cN}$ and $\tilde{\mathbf{D}} = \begin{bmatrix} \mathbf{D} \\ \sqrt{\alpha}\mathbf{A} \end{bmatrix} \in \mathbb{R}^{(d+ck) \times ck}$. Neglecting the computation of

scalar multiplications, the complexity of (16) is:

$$T_{\text{LC-KSVD}} \approx (ck)(cN)(2(d+ck)+L^2) = c^2 kN(2d+2ck+L^2).$$

C. Complexity of Nayak's

The optimization problem in Nayak's [4] is:

$$(\mathbf{D},\mathbf{S},\mathbf{W}) = \arg\min_{\mathbf{D},\mathbf{S},\mathbf{W}} \|\mathbf{Y} - \mathbf{DS}\|_F^2 + \lambda \|\mathbf{S}\|_1 + \|\mathbf{S} - \mathbf{WY}\|_F^2.$$

$\mathbf{S}$ is estimated via the gradient descent method that is an iterative method whose main computational task in each iteration is to calculate the gradient of $Q(\mathbf{S}) = \|\mathbf{Y} - \mathbf{DS}\|_F^2 + \|\mathbf{S} - \mathbf{WY}\|_F^2$ with respect to $\mathbf{S}$. We have:

$$\frac{\partial Q(\mathbf{S})}{\partial \mathbf{S}} = 2\Big((\mathbf{D}^\top\mathbf{D}+\mathbf{I})\mathbf{S} - (\mathbf{D}^\top - \mathbf{W})\mathbf{Y}\Big).$$

where $\mathbf{D}^\top\mathbf{D}+\mathbf{I}$, and $(\mathbf{D}^\top - \mathbf{W})\mathbf{Y}$ could be precomputed and at each step, only $(\mathbf{D}^\top\mathbf{D}+\mathbf{I})\mathbf{S}$ need to be recalculated after $\mathbf{S}$ is updated. With $\mathbf{D} \in \mathbb{R}^{d\times ck}, \mathbf{S} \in \mathbb{R}^{ck\times cN}, \mathbf{Y} \in \mathbb{R}^{d\times cN}, \mathbf{W} \in \mathbb{R}^{ck\times d}$, the complexity of the sparse coding step can be estimated as:

$$
\begin{aligned}
T_{\text{Nayak's}} &\approx (ck)d(ck)+2(ck)d(cN)+2q(ck)^2cN, &(17)\\
&= c^2 kN(2d+2qck)+c^2 dk^2. &(18)
\end{aligned}
$$

with q being the average number of iterations needed for convergence. Here we have ignored matrix subtractions, additions and scalar multiplications and focused on matrix multiplications only. We have also used the approximation that complexity of $\mathbf{AB}$ is $2mnp$ where $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times p}$. The first term in (17) is of $\mathbf{D}^\top\mathbf{D}+\mathbf{I}$ (note that this matrix is symmetric, then it needs only half of regular operations), the second term is of $(\mathbf{D}^\top - \mathbf{W})\mathbf{Y}$ and the last one comes from q times complexity of calculating $(\mathbf{D}^\top\mathbf{D}+\mathbf{I})\mathbf{S}$.

D. Complexity of FDDL

The sparse coding step in FDDL [31] requires solving c class-specific problems:

$$
\mathbf{S}_i = \arg\min_{\mathbf{S}_i} \Big\{ \|\mathbf{Y}_i - \mathbf{DS}_i\|_F^2 + \|\mathbf{Y}_i - \mathbf{D}_i\mathbf{S}_i^i\|_F^2 + \sum_{j=1, j\neq i}^c \|\mathbf{D}_j\mathbf{S}_i^j\|_F^2
$$
$$
+\lambda_2\big\{ \|\mathbf{S}_i - \mathbf{M}_i\|_F^2 - \sum_{k=1}^c \|\mathbf{M}_k - \mathbf{M}\|_F^2 + \eta\|\mathbf{S}_i\|_F^2 \big\} + \lambda_1 \|\mathbf{S}_i\|_1 \Big\},
$$

with $\mathbf{D} = [\mathbf{D}_1,\dots,\mathbf{D}_c], \mathbf{S}_i^\top = [(\mathbf{S}_i^1)^\top,\dots,(\mathbf{S}_i^c)^\top]$, and $\mathbf{M}_k = [\mathbf{m}_k,\dots,\mathbf{m}_k] \in \mathbb{R}^{ck\times N}, \mathbf{M} = [\mathbf{m},\dots,\mathbf{m}] \in \mathbb{R}^{ck\times N}$ where $\mathbf{m}_k$ and $\mathbf{m}$ are the mean vector of $\mathbf{S}_i$ and $\mathbf{S} = [\mathbf{S}_1,\dots,\mathbf{S}_c]$ respectively. The algorithm for solving this problem uses Iterative Projective Method [41] whose complexity depends on computing gradient of six Frobineous-involved terms in the above optimization problem at each iteration.

For the first three terms, the gradient could be computed as:

$$
2(\mathbf{D}^\top\mathbf{D})\mathbf{S}_i - 2\mathbf{D}^\top\mathbf{Y}_i + \begin{bmatrix} 2(\mathbf{D}_1^\top\mathbf{D}_1)\mathbf{S}_i^1 \\ \vdots \\ 2(\mathbf{D}_i^\top\mathbf{D}_i)\mathbf{S}_i^i - \mathbf{D}_i^\top\mathbf{Y}_i \\ \vdots \\ 2(\mathbf{D}_c^\top\mathbf{D}_c)\mathbf{S}_i^c \end{bmatrix}, \quad (19)
$$

where $\mathbf{D}^\top\mathbf{D}$, and $\mathbf{D}^\top\mathbf{Y}_i$ could be precomputed with the total cost of $(ck)d(ck)+2(ck)dN = cdk(2N+ck)$; $\mathbf{D}_i^\top\mathbf{D}_i$, and $\mathbf{D}_i^T\mathbf{Y}_i$ could be extracted from $\mathbf{D}^\top\mathbf{D}$, and $\mathbf{D}^\top\mathbf{Y}_i$ at no cost; at each iteration, cost of computing $(\mathbf{D}^\top\mathbf{D})\mathbf{S}_i$ is $2(ck)^2N$, each of $(\mathbf{D}_j^\top\mathbf{D}_j)\mathbf{S}_i^j$ could be attained in the intermediate step of computing $(\mathbf{D}^\top\mathbf{D})\mathbf{S}_i$. Therefore, with q iterations, the computational cost of (19) is:

$$cdk(2N+ck)+2qc^2k^2N. \quad (20)$$

For the last three terms, we will prove that:

$$
\begin{aligned}
\frac{\partial}{\partial\mathbf{S}_i}\|\mathbf{S}_i - \mathbf{M}_i\|_F^2 &= 2(\mathbf{S}_i - \mathbf{M}_i), &(21)\\
\frac{\partial}{\partial\mathbf{S}_i}\sum_{k=1}^c \|\mathbf{M}_k - \mathbf{M}\|_F^2 &= 2(\mathbf{M}_i - \mathbf{M}), &(22)\\
\frac{\partial}{\partial\mathbf{S}_i}\eta\|\mathbf{S}_i\|_F^2 &= 2\eta\mathbf{S}_i. &(23)
\end{aligned}
$$

Indeed, let $\mathbf{E}_{m,n}$ be a all-one matrix in $\mathbb{R}^{m\times n}$, one could easily verify that:

$$\mathbf{M}_k = \frac{1}{N}\mathbf{S}_k\mathbf{E}_{N,N}; \quad \mathbf{M} = \frac{1}{cN}\mathbf{S}\mathbf{E}_{cN,N} = \frac{1}{cN}\sum_{i=1}^c \mathbf{S}_i\mathbf{E}_{N,N};$$

$$\mathbf{E}_{m,n}\mathbf{E}_{n,p} = n\mathbf{E}_{m,p}; \quad (\mathbf{I} - \frac{1}{N}\mathbf{E}_{N,N})(\mathbf{I} - \frac{1}{N}\mathbf{E}_{N,N})^\top = (\mathbf{I} - \frac{1}{N}\mathbf{E}_{N,N}).$$

Thus, (21) can be obtained by:

$$
\begin{aligned}
\frac{\partial}{\partial\mathbf{S}_i}\|\mathbf{S}_i - \mathbf{M}_i\|_F^2 &= \frac{\partial}{\partial\mathbf{S}_i}\|\mathbf{S}_i - \frac{1}{N}\mathbf{S}_i\mathbf{E}_{N,N}\|_F^2 \\
&= \frac{\partial}{\partial\mathbf{S}_i}\|\mathbf{S}_i(\mathbf{I} - \frac{1}{N}\mathbf{E}_{N,N})\|_F^2 = 2\mathbf{S}_i(\mathbf{I} - \frac{1}{N}\mathbf{E}_{N,N})(\mathbf{I} - \frac{1}{N}\mathbf{E}_{N,N})^\top \\
&= 2\mathbf{S}_i(\mathbf{I} - \frac{1}{N}\mathbf{E}_{N,N}) = 2(\mathbf{S}_i - \mathbf{M}_i).
\end{aligned}
$$

For (22), with simple algebra, we can prove that:

$$\frac{\partial}{\partial\mathbf{S}_i}\|\mathbf{M}_i - \mathbf{M}\|_F^2 = \frac{2(c-1)}{cN}(\mathbf{M}_i - \mathbf{M})\mathbf{E}_{N,N} = \frac{2(c-1)}{c}(\mathbf{M}_i - \mathbf{M}).$$

$$\frac{\partial}{\partial\mathbf{S}_i}\|\mathbf{M}_k - \mathbf{M}\|_F^2 = \frac{2}{cN}(\mathbf{M} - \mathbf{M}_k)\mathbf{E}_{N,N} = \frac{2}{c}(\mathbf{M} - \mathbf{M}_k), (k\neq i).$$

Compared to (19), calculating (21), (22) and (23) require much less computation. As a result, the total cost of solving $\mathbf{S}_i$ approximately equals to (20); and the total estimated cost of sparse coding step of FDDL is estimated as c times cost of each class-specific problem and approximately equals to:

$$T_{\text{FDDL}} \approx c^2 dk(2N+ck)+2qc^3k^2N = c^2 kN(2d+2qck)+c^3 dk^2.$$

Final analyzed results of four different dictionary learning methods are reported in Table II.

REFERENCES

[1] M. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: a review," *IEEE Rev. Biomed. Eng.*, vol. 2, 2009.

[2] U. Srinivas, H. S. Mousavi, C. Jeon, V. Monga, A. Hattel, and B. Jayarao, "SHIRC: A simultaneous sparsity model for histopathological image representation and classification," *Proc. IEEE Int. Symp. Biomed. Imag.*, pp. 1118–1121, Apr. 2013.

[3] U. Srinivas, H. S. Mousavi, V. Monga, A. Hattel, and B. Jayarao, "Simultaneous sparsity model for histopathological image representation and classification," *IEEE Trans. on Medical Imaging*, vol. 33, no. 5, pp. 1163–1179, May 2014.

[4] N. Nayak, H. Chang, A. Borowsky, P. Spellman, and B. Parvin, "Classification of tumor histopathology via sparse feature learning," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2013, pp. 1348–1351.

[5] H. S. Mousavi, V. Monga, A. U. Rao, and G. Rao, "Automated discrimination of lower and higher grade gliomas based on histopathological image analysis," *Journal of Pathology Informatics*, 2015.

[6] M. M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, and M. N. Gurcan, "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Trans. on Biomed. Engineering*, vol. 58, no. 7, pp. 1977–1984, 2011.

[7] A. B. Tosun and C. Gunduz-Demir, "Graph run-length matrices for histopathological image segmentation," *transmi*, vol. 30, no. 3, pp. 721–732, 2011.

[8] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *Proc. IEEE Int. Symp. Biomed. Imag.* IEEE, 2008, pp. 496–499.

[9] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Efficient multiple instance convolutional neural networks for gigapixel resolution image classification," *arXiv preprint arXiv:1504.07947*, 2015.

[10] N. Orlov, L. Shamir, T. Macuraand, J. Johnston, D. Eckley, and I. Goldberg, "WND-CHARM: Multi-purpose image classification using compound image transforms," *Pattern Recogn. Lett.*, vol. 29, no. 11, pp. 1684–1693, 2008.

[11] L. Shamir, N. Orlov, D. Eckley, T. Macura, J. Johnston, and I. Goldberg, "Wndchrm–an open source utility for biological image analysis," *Source Code Biol. Med.*, vol. 3, no. 13, 2008.

[12] T. Gultekin, C. Koyuncu, C. Sokmensuer, and C. Gunduz-Demir, "Two-tier tissue decomposition for histopathological image representation and classification," *IEEE Trans. on Medical Imaging*, vol. 34, pp. 275–283.

[13] J. Shi, Y. Li, J. Zhu, H. Sun, and Y. Cai, "Joint sparse coding based spatial pyramid matching for classification of color medical image," *Computerized Medical Imaging and Graphics*, 2014.

[14] S. Minaee, Y. Wang, and Y. W. Lui, "Prediction of longterm outcome of neuropsychological tests of mtbi patients using imaging features," in *Signal Proc. in Med. and Bio. Symp. IEEE*, 2013.

[15] F. Zana and J.-C. Klein, "Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation," *IEEE Trans. on Image Processing*, vol. 10, no. 7, pp. 1010–1019, 2001.

[16] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1055–1064, 1999.

[17] M. Unser, A. Aldroubi, and A. Laine, "Guest editorial: wavelets in medical imaging," *IEEE Trans. on Medical Imaging*, vol. 22, no. LIB-ARTICLE-2003-004, pp. 285–288, 2003.

[18] E. Ozdemir and C. Gunduz-Demir, "A hybrid classification model for digital pathology using structural and statistical pattern recognition," *IEEE Trans. on Medical Imaging*, vol. 32, no. 2, pp. 474–483, 2013.

[19] I. Kopriva, M. P. Hadžija, M. Hadžija, and G. Aralica, "Offset-sparsity decomposition for automated enhancement of color microscopic image of stained specimen in histopathology," *Journal of biomedical optics*, vol. 20, no. 7, 2015.

[20] Y. Yu, J. Huang, S. Zhang, C. Restif, X. Huang, and D. Metaxas, "Group sparsity based classification for cervigram segmentation," *Proc. IEEE Int. Symp. Biomed. Imag.*, pp. 1425–1429, 2011.

[21] Y. Song, W. Cai, H. Huang, Y. Zhou, Y. Wang, and D. Feng, "Locality-constrained subcluster representation ensemble for lung image classification," *Medical image analysis*, vol. 22, no. 1, pp. 102–113, 2015.

[22] Y. Song, W. Cai, H. Huang, Y. Zhou, D. Feng, Y. Wang, M. Fulham, and M. Chen, "Large margin local estimate with applications to medical image classification," vol. 34, no. 6, pp. 1362–1377, 2015.

[23] H. Chang, N. Nayak, P. T. Spellman, and B. Parvin, "Characterization of tissue histopathology via predictive sparse decomposition and spatial pyramid matching," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013.* Springer, 2013, pp. 91–98.

[24] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on Pattern Analysis and Machine Int.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[25] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Advances in Neural Information Processing Systems*, 2006, pp. 609–616.

[26] S. Bahrampour, A. Ray, N. Nasrabadi, and K. Jenkins, "Quality-based multimodal classification using tree-structured sparsity," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2014, pp. 4114–4121.

[27] H. S. Mousavi, U. Srinivas, V. Monga, Y. Suo, M. Dao, and T. Tran, "Multi-task image classification via collaborative, hierarchical spike-and-slab priors," in *Proc. IEEE Conf. on Image Processing*, 2014, pp. 4236–4240.

[28] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Proc. IEEE Conf. Computer Vision Pattern Recognition.* IEEE, 2010, pp. 2691–2698.

[29] Z. Jiang, Z. Lin, and L. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. on Pattern Analysis and Machine Int.*, vol. 35, no. 11, pp. 2651–2664, 2013.

[30] Y. Suo, M. Dao, T. Tran, H. Mousavi, U. Srinivas, and V. Monga, "Group structured dirty dictionary learning for classification," in *Proc. IEEE Conf. on Image Processing*, 2014, pp. 150–154.

[31] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Conf. on Computer Vision*, Nov. 2011, pp. 543–550.

[32] T. H. Vu, H. S. Mousavi, V. Monga, U. Rao, and G. Rao, "Dfdl: Discriminative feature-oriented dictionary learning for histopathological image classification," *Proc. IEEE Int. Symp. Biomed. Imag.*, pp. 990–994, 2015.

[33] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[34] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[35] N. I. of Health, "The Cancer Genome Atlas (TCGA) database," http://cancergenome.nih.gov, accessed: 2014-11-09.

[36] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. on Info. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[37] "SPArse Modeling Software," http://spams-devel.gforge.inria.fr/, accessed: 2014-11-05.

[38] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection." Morgan Kaufmann, 1995, pp. 1137–1143.

[39] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[40] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit," *CS Technion*, vol. 40, no. 8, pp. 1–15, 2008.

[41] L. Rosasco, A. Verri, M. Santoro, S. Mosci, and S. Villa, "Iterative projection methods for structured sparsity regularization," *MIT Technical Reports,MIT-CSAIL-TR-2009-050, CBCL-282*, 2009.