Scalable Multi-Instance Multi-Shape Support Vector Machine for Whole Slide Breast Histopathology

Hoon Seo
Department of Computer Science
Colorado School of Mines
Golden, Colorado, U.S.A.
seohoon@mines.edu

Lodewijk Brand
Department of Computer Science
Colorado School of Mines
Golden, Colorado, USA
lbrand@mines.edu

Lucia Saldana Barco
Department of Computer Science
Colorado School of Mines
Golden, Colorado, USA
lsaldanabarco@mines.edu

Hua Wang
Department of Computer Science
Colorado School of Mines
Golden, Colorado, USA
huawangcs@gmail.com

2022 IEEE International Conference on Knowledge Graph (ICKG) | 978-1-6654-5101-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICKG55886.2022.00036

***Abstract*—Histopathological image analysis is critical in cancer diagnosis and treatment. Due to the huge size of histopathological images, most existing works analyze the whole slide pathological image (WSI) as a bag and its patches are considered as instances. However, these approaches are limited to analyzing the patches in a fixed shape, while the malignant lesions can form varied shapes. To address this challenge, we propose the Multi-Instance Multi-Shape Support Vector Machine (MIMSSVM) to analyze the multiple images (instances) jointly where each instance consists of multiple patches in varied shapes. In our approach, we can identify the varied morphologic abnormalities of nuclei shapes from the multiple images. In addition to the multi-instance multi-shape learning capability, we provide an efficient algorithm to optimize the proposed model which scales well to a large number of features. Our experimental results show the proposed MIMSSVM method outperforms the existing SVM and recent deep learning models in histopathological classification. The proposed model also identifies the tissue segments in an image exhibiting an indication of an abnormality which provides utility in the early detection of malignant tumors.**

***Index Terms*—scalability, multi-instance, multi-modal, support vector machine, histopathology**

I. Introduction

According to the National Breast Cancer Foundation, one in eight women in the United States develops the breast cancer [1]. Early detection of breast cancer is the key for maximizing the patient's chance of survival, and while strides have been made in both medical and technological fields, this problem remains at the pinnacle of breast cancer research. In the past, traditional diagnostic procedures have leaned heavily on the need for specialized training, especially in pathology and radiology. With the rise of new technological solutions, pathologists have begun digitizing the tissue sections and creating histopathological images or whole-slide images (WSIs). This allows certain features of the image, such as inconsistencies in cell architecture and the presence or vacancy of specific biological qualities, to be detected and analyzed. If such features can be spotted, they can be used as indicators for diseases. For example, cells that are abnormally structured and divided uncontrollably are carcinomas and immediately point to the presence of cancer. Since technologically based diagnoses are founded on features exhibited by tissue samples on a cellular level, being able to analyze the images digitally increases efficiency of the process. Furthermore, current medical procedures have led to an increase in the number of biopsies taken, causing a rise in the number of histopathological images that require unrealistic workloads for pathologists [2].

In light of this issue, recent advancements in artificial intelligence have yielded promise by displaying the ability to analyze large volumes of histopathological images. The Breast Cancer Histopathological Image Classification (BreaKHis) dataset [3] is publicly available and is composed of 7,909 histopathological images of benign and malignant breast cancer tumors. The set of images is organized into groups of the most common types of carcinomas, making it a very valuable dataset for cancer research. Each patient's histopathological images have been classified as malignant or benign and have been assigned to a particular group depending on the type of tumor that they present. Each patient is represented by their collection of histopathological images, which pathologists have correctly classified during the creation of this dataset. To deal with such an extensive dataset, many machine learning algorithms have been employed to locate the abnormal tissue sections and correctly determine whether there is cancer or not. In particular, several multi-instance learning methods (MIL) have achieved satisfactory results in the past when performing similar tasks on the BreaKHis dataset. Some examples are Multi-Instance Support Vector Machine (MISVM) [4], sparse Multi-Instance Learning (sMIL) and sparse balanced MIL (sbMIL) [5], and Normalized Set Kernel (NSK) and Statistics Kernel (STK) [6]. These are all methods that have been deemed successful at correctly labeling the bags in the testing dataset as either malignant or benign.

978-1-6654-5101-7/22/$31.00 ©2022 IEEE
DOI 10.1109/ICKG55886.2022.00036

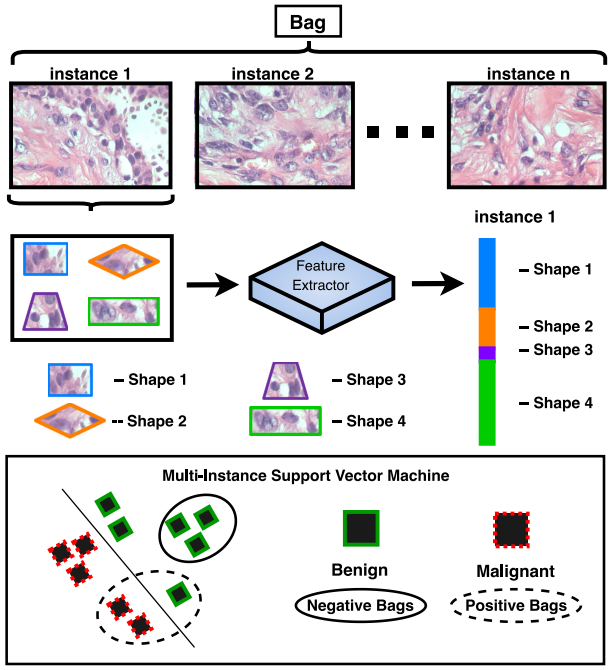

Fig. 1. An illustration of our processing pipeline for our multi-instance multi-shape learning algorithm applied to the BreaKHis dataset. The input is the concatenation of patches in different shapes, and we employ the shape specific regularization to identify the important shape for classification.

Our research also utilizes a multi-instance learning method to determine whether an image shows any indications of carcinoma and, if so, where the abnormality is located within the histopathological image. As mentioned before, multi-instance learning is commonly used for disease detection applications, and these types of algorithms have also been evaluated on the BreaKHis dataset previously. Multi-instance learning [7]–[10] is a common weakly supervised area of machine learning known for organizing the training and testing data into sets of instances called bags. Due to the nature of this approach, the data can be labeled at the bag level instead of at the instance level, which means that clinicians do not need to spend a lot of resources into characterizing each image in the training dataset obtained from a biopsy. Doctors only need to label/diagnose the bag or patient as malignant or benign, and the rest of the instances or images follow suit.

To provide a well-rounded and complete picture of each instance, we extract the multiple patches in the different shapes from each image. Our approach can be particularly effective in the cancer diagnosis, since the malignancy often exhibits the abnormal morphology and varying shape of nuclei in the practice of surgical pathology [11]. We extract the feature vector for each patch, and the feature vectors of all the patches in one instance are concatenated to create a more information-dense vector. Because the features in a same shape are associated each other, we design our model to learn the group structured sparsity respect to the features grouped by their shapes. The patches in different shapes can be regarded as the different views (modalities) on the image, and learning group-wise sparsity has shown performance improvement on the multi-modal data [12], [13]. As illustrated in Fig. 1, each of these shapes can capture an important part of the image, such as a cell nucleus or a possible abnormality in the cell structure.

In our work, we also focus on optimizing our model stably and efficiently. Considering the large size of histopathological images, the large number of features of instances can be a hurdle for optimizing MIL models. In our work we take this fact into account and propose the Multi-Instance Multi-Shape SVM (*MIMSSVM*) method, which improves the efficiency of the optimization compared to the previously mentioned MIL approaches. We summarize our contributions as follows:

- We present an novel *MIMSSVM* method which utilizes the multi-shape patches in the images.
- We develop an scalable solution for the proposed method based on optimal line search method [14] to bypass the quadratic programming problem that comes from the typical MISVM models.
- We provide an application of the *MIMSSVM* method to diagnose a patient with breast cancer if their histopathological images show indications of a carcinoma. In regards to the interpretability, our model also identifies the disease relevant patches in the images.

II. Methods

In this section, we will develop the objective of the *MIMSSVM* model for handling multi-instance multi-modal (multi-shape) data and derive its solution algorithm. We provide both exact and inexact solutions, where the inexact solution approximates the exact solution and improves the scalability against the large number of features. In our derivations, we start from standard MISVM objective [4], and add the structured sparsity terms to exploit the multi-modal nature of the instance vectors and employ the iterative reweighted method [15], [16] to optimize the introduced terms in a numerically stable way. Finally, we apply the multi-block alternating direction method of multipliers (ADMM) [17] to minimize the proposed MIMSSVM objective.

A. Notations

In this paper, we write matrices as bold upper-case letters $\mathbf{M}$, vectors as bold lower-case letters $\mathbf{m}$, and scalars as lower-case letters m. The i-th row and j-th column of $\mathbf{M}$ are denoted as $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively. We use m_j^i to represent the scalar value indexed by the i-th row and j-th column of the matrix $\mathbf{M}$. The matrix $\mathbf{M}_p$ and $\mathbf{M}^p$ correspond to the p-th column-block and p-th row-block of $\mathbf{M}$ respectively. Each bag $\mathbf{X}_i = \{\mathbf{x}_{i,1}, \cdots, \mathbf{x}_{i,n_i}\} \in \Re^{d \times n_i}$ contains n_i instances and its associated label is represented by $y_i \in \{1, \cdots, m, \cdots, K\}$, where i is the index of the bag in the dataset. Each instance $\mathbf{x}_{i,j}$ in $\mathbf{X}_i$ is the concatenation of G modalities (shapes) of vectors, such that $\mathbf{x}_{i,j} = [\mathbf{x}_{i,j}^1; \cdots; \mathbf{x}_{i,j}^G]$.

B. The Objective of the Multi-Instance Multi-Shape Support Vector Machine

The K class multi-instance SVM proposed in [4] solves the following objective:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{b}} \ & \frac{1}{2}\sum_{m=1}^{K}\|\mathbf{w}_m\|_2^2 \\ & + C\sum_{i=1}^{N}\sum_{m=1}^{K}(1-[\max(\mathbf{w}_m^T\mathbf{X}_i+\mathbf{1}b_m) \\ & - \max(\mathbf{w}_y^T\mathbf{X}_i+\mathbf{1}b_y)]y_i^m)_+ \ .\end{aligned} \tag{1}$$

Therefore, its decision function is given by:

$$\tilde{y}_i = \arg\max_{m'}(\max(\mathbf{W}^T\mathbf{X}_i+\mathbf{b}\mathbf{1}_i)^{m'}) \ . \tag{2}$$

Motivated by the multi-modal learning using structured-sparsity regularization proposed by [12], [13], we capture the important modalities by applying modality-wise regularization term $\|\mathbf{W}\|_{G_1}$ defined as follows [12], [13]:

$$\|\mathbf{W}\|_{G_1} = \sum_{k=1}^{K}\sum_{g=1}^{G}\|\mathbf{w}_k^g\|_2 \ , \tag{3}$$

where $\mathbf{W} = [\mathbf{W}^1;\mathbf{W}^2;\cdots;\mathbf{W}^G] \in \Re^{d\times K}$ consists of G row-blocks, and each row-block $\mathbf{W}^g \in \Re^{d_g\times K}$ is the weights matrix for g-th modality (shape) of input $\mathbf{X}_i$. To be more specific, for each classification task, the group ℓ_1 norm applies the ℓ_2 norm within each modality and the ℓ_1 norm across different modalities. By minimizing $\|\mathbf{W}\|_{G_1}$, the weights in $\mathbf{w}_k^g$ will approach to zero values, if g-th modality is not discriminative for k-th classification task.

In addition to the group-wise sparsity learning by group ℓ_1 norm, we introduce the additional structured sparsity regularizer term $\|\mathbf{W}\|_*$, which is the trace norm of $\mathbf{W}$ and defined as [18]:

$$\|\mathbf{W}\|_* = \sum_{i=1}^{min\{d,K\}}\sigma_i = \mathrm{tr}[(\mathbf{W}^T\mathbf{W})^{\frac{1}{2}}] \ , \tag{4}$$

where σ_i is the i-th singular value of $\mathbf{W}$. Through the minimization of singular values of $\mathbf{W}$, we can discover the low-rank representation of projection $\mathbf{W}$ and maximize the correlation between G-modalities as its effectiveness has been shown in [19]–[22]. Armed with the structured sparsity regularizer terms $\|\mathbf{W}\|_{G_1}$ and $\|\mathbf{W}\|_*$, we rewrite Eq. (1) as follows:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{b}} \ & \frac{1}{2}\sum_{m=1}^{K}\|\mathbf{w}_m\|_2^2 + \tau_1\|\mathbf{W}\|_{G_1} + \tau_2\|\mathbf{W}\|_* \\ & + C\sum_{i=1}^{N}\sum_{m=1}^{K}(1 \\ & - [\max(\mathbf{w}_m^T\mathbf{X}_i+\mathbf{1}b_m) - \max(\mathbf{w}_y^T\mathbf{X}_i+\mathbf{1}b_y)]y_i^m)_+ \ .\end{aligned} \tag{5}$$

The MIMSSVM objective in Eq. (5) is generally difficult to solve because of the coupled primal variables $\mathbf{w}_k$, b_m with $\max(\cdot)$ operations. Thus we split the primal variables in Eq. (1) via the ADMM approach [17] by converting Eq. (5) into the following constrained optimization problem:

$$\begin{aligned}\min_{\substack{\mathbf{W},\mathbf{b},\mathbf{V},\mathbf{E},\\ \mathbf{Q},\mathbf{R},\mathbf{T},\mathbf{U}}} \ & \frac{1}{2}\sum_{m=1}^{K}\|\mathbf{w}_m\|_2^2 + \tau_1\|\mathbf{V}\|_{G_1} \\ & + \tau_2\|\mathbf{W}\|_* + C\sum_{i=1}^{N}\sum_{m=1}^{K}(y_i^m e_i^m)_+ \\ \textit{s.t.} \ & e_i^m = y_i^m - q_i^m + r_i^m \ , \\ & \mathbf{V} = \mathbf{W}, \ r_i^m = \max(\mathbf{u}_i^m) \ , \\ & q_i^m = \max(\mathbf{t}_i^m) \ , \\ & \mathbf{t}_i^m = \mathbf{w}_m^T\mathbf{X}_i + \mathbf{1}b_m \ , \\ & \mathbf{u}_i^m = \mathbf{w}_y^T\mathbf{X}_i + \mathbf{1}b_y \ .\end{aligned} \tag{6}$$

As discussed earlier, the structured sparsity regularizers $\|\mathbf{V}\|_{G_1}$ and $\|\mathbf{W}\|_*$ in Eq. (6) are well motivated in the context to analyze multi-shape WSIs. However, they are both not differentiable, therefore the objective in Eq. (6) is non-smooth that is thereby difficult to efficiently solve in general. To ensure the numerical stability of the optimization, we use the optimization framework introduced in our earlier work [15], [16], [23] that proposed the iterative reweighted method to solve non-smooth objectives. Then we can solve Eq. (6) by an iterative procedure in which the key step is minimizing the following smoothed objective:

$$\begin{aligned}\min_{\substack{\mathbf{W},\mathbf{b},\mathbf{V},\mathbf{E},\\ \mathbf{Q},\mathbf{R},\mathbf{T},\mathbf{U}}} \ & \frac{1}{2}\sum_{m=1}^{K}\|\mathbf{w}_m\|_2^2 + C\sum_{i=1}^{N}\sum_{m=1}^{K}(y_i^m e_i^m)_+ \\ & + \tau_1\sum_{g=1}^{G}\mathrm{tr}(\mathbf{V}^g\mathbf{D}_{1,g}(\mathbf{V}^g)^T) \\ & + \tau_2\,\mathrm{tr}(\mathbf{W}^T\mathbf{D}_2\mathbf{W}) \ , \\ \textit{s.t.} \ & e_i^m = y_i^m - q_i^m + r_i^m \ , \\ & \mathbf{V} = \mathbf{W} \ , \\ & r_i^m = \max(\mathbf{u}_i^m) \ , \\ & q_i^m = \max(\mathbf{t}_i^m) \ , \\ & \mathbf{t}_i^m = \mathbf{w}_m^T\mathbf{X}_i + \mathbf{1}b_m \ , \\ & \mathbf{u}_i^m = \mathbf{w}_y^T\mathbf{X}_i + \mathbf{1}b_y \ ,\end{aligned} \tag{7}$$

where $\mathbf{V}^g \in \Re^{d_g\times K}$ is row-block corresponding to the row indices of g-th modality, $\mathbf{D}_{1,g} \in \Re^{K\times K}$ is an diagonal matrix and its j-th diagonal element is computed as:

$$\mathbf{D}_{1,g}(j,j) = \frac{1}{2}(\|\mathbf{v}_j^g\|_2^2+\delta)^{-\frac{1}{2}} \ , \tag{8}$$

and $\mathbf{D}_2$ is computed as:

$$\mathbf{D}_2 = \frac{1}{2}(\mathbf{W}\mathbf{W}^T+\delta\mathbf{I})^{-\frac{1}{2}} \in \Re^{D\times D} \ , \tag{9}$$

where δ is the smoothness term of small positive constant. The hyperparameters τ_1 and τ_2 adjust the impact of corresponding terms.

From Eq. (7) we derive the following augmented Lagrangian:

$$\begin{aligned}\mathcal{L}_\mu = & \frac{1}{2}\sum_{m=1}^{K}\|\mathbf{w}_m\|_2^2 + \sum_{i=1}^{N}\sum_{m=1}^{K} C\,(y_i^m e_i^m)_+ \\ & + \tau_1 \sum_{g=1}^{G} \mathrm{tr}(\mathbf{V}^g \mathbf{D}_{1,g} (\mathbf{V}^g)^T) + \tau_2\,\mathrm{tr}(\mathbf{W}^T\mathbf{D}_2\mathbf{W}) + \frac{\mu}{2}\sum_{i=1}^{N}\sum_{m=1}^{K} \\ & \Big[(e_i^m - (y_i^m - q_i^m + r_i^m - \lambda_i^m/\mu))^2 + (q_i^m - \max(\mathbf{t}_i^m) + \sigma_i^m/\mu)^2 \\ & + \Big\| \mathbf{t}_i^m - \Big(\mathbf{w}_m^T\mathbf{X}_i + \mathbf{1}b_m\Big) + \boldsymbol{\theta}_i^m/\mu \Big\|_2^2 \\ & + (r_i^m - \max(\mathbf{u}_i^m) + \omega_i^m/\mu)^2 \\ & + \Big\| \mathbf{u}_i^m - \Big(\mathbf{w}_y^T\mathbf{X}_i + \mathbf{1}b_y\Big) + \boldsymbol{\xi}_i^m/\mu \Big\|_2^2 \Big] \\ & + \frac{\mu}{2}\|\mathbf{V} - \mathbf{W} + \boldsymbol{\Gamma}/\mu\|_F^2 \ , \end{aligned} \tag{10}$$

where $\mathbf{W}, \mathbf{b}, \mathbf{V}, \mathbf{E}, \mathbf{Q}, \mathbf{T}, \mathbf{R}, \mathbf{U}$ are the primal variables, $\boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \boldsymbol{\Xi}, \boldsymbol{\Gamma}$ are the dual variables, and $\mu > 0$ is a hyperparameter.

C. The Solution Algorithm

In this section, we provide the key derivation details for each class-hyperplane in $\mathbf{W}$ and $\mathbf{b}$ with its associated constraint variable $\mathbf{V}$ as they are most important variables in our model. The full derivation details of the solution algorithm will be provided in our longer journal extension due to space limit.

W & b update. Removing all terms from Eq. (10) that do not include $\mathbf{W}$ and decoupling across columns of $\mathbf{W}$ gives the following K minimization problems:

$$\begin{aligned}\mathbf{w}_m = \arg\min_{\mathbf{w}_m} \ & \frac{1}{2}\|\mathbf{w}_m\|_2^2 + \frac{\mu}{2}\sum_{i=1}^{N}\Big[\Big\|\mathbf{t}_i^m - \big(\mathbf{w}_m^T\mathbf{X}_i + \mathbf{1}b_m\big) + \\ & \boldsymbol{\theta}_i^m/\mu\Big\|_2^2\Big] + \sum_{i'=1}^{N'}\sum_{m'=1}^{K}\Big[\frac{\mu}{2}\Big\|\mathbf{u}_{i'}^{m'} - \big(\mathbf{w}_m^T\mathbf{X}_{i'} + \mathbf{1}b_m\big) \\ & + \boldsymbol{\xi}_{i'}^{m'}/\mu\Big\|_2^2\Big] + \frac{\mu}{2}\Big\|\mathbf{v}_m - \mathbf{w}_m + \boldsymbol{\gamma}_m/\mu\Big\|_2^2 + \tau_2\,\mathrm{tr}(\mathbf{W}^T\mathbf{D}_2\mathbf{W}), \end{aligned} \tag{11}$$

where i' indicates the column blocks in $\mathbf{X}$ (and corresponding columns of $\mathbf{U}$ and $\boldsymbol{\Xi}$) that belong to the m-th class. N' is the total number of bags that belong to the m-th class, and $\mathbf{t}_i^m$ and $\boldsymbol{\theta}_i^m$ are row vectors corresponding to the i-th bag and m-th class in $\mathbf{T}$ and $\boldsymbol{\Theta}$. Finally, $\mathbf{u}_{i'}^{m'}$ and $\boldsymbol{\xi}_{i'}^{m'}$ are row vectors selected from $\mathbf{U}$ and $\boldsymbol{\Xi}$ that correspond to the (i,m) pair belonging to the m-th class.

Taking the derivative of Eq. (11) with respect to $\mathbf{w}_m$ and setting the result equal to zero gives the closed form solution

$$\begin{aligned}\mathbf{w}_m^T = & \Big(\textstyle\sum_{i=1}^{N}\big[(\mathbf{t}_i^m - \mathbf{1}b_m + \boldsymbol{\theta}_i^m/\mu)\,\mathbf{X}_i^T\big] + \sum_{i'=1}^{N'}\sum_{m'=1}^{K} \\ & \big[(\mathbf{u}_{i'}^{m'} - \mathbf{1}b_m + \boldsymbol{\xi}_{i'}^{m'}/\mu)\mathbf{X}_{i'}^T\big] + \mathbf{v}_m^T + \boldsymbol{\gamma}_m^T/\mu\Big) * \Big((1/\mu+1)\mathbf{I} \\ & + \frac{2}{\mu}\tau_2\mathbf{D}_2 + \textstyle\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i^T + K\sum_{i'=1}^{N'}\mathbf{X}_{i'}\mathbf{X}_{i'}^T\Big)^{-1} \ , \end{aligned} \tag{12}$$

which can be calculated via a least-squares solver to avoid an inverse calculation. Similarly, differentiating Eq. (10) element-wise with respect to b_m, setting the result equal to zero gives the update:

$$b_m = \frac{\sum_{i=1}^{N}\big[\mathbf{t}_i^m - \mathbf{w}_m^T\mathbf{X}_i + \boldsymbol{\theta}_i^m/\mu\big] + \sum_{i'=1}^{N'}\sum_{m=1}^{K}\big[\mathbf{u}_{i'}^m - \mathbf{w}_m^T\mathbf{X}_{i'} + \boldsymbol{\xi}_{i'}^m/\mu\big]}{N + KN'} \ . \tag{13}$$

V update. By discarding all terms not including $\mathbf{V}$ from Eq. (10) and setting the derivative respect to $\mathbf{V}^g$ to zero matrix, we have:

$$\mathbf{V}^g = (\mu\mathbf{W}^g - \boldsymbol{\Gamma}^g)(2\tau_1\mathbf{D}_{1,g} + \mu\mathbf{I})^{-1} \ . \tag{14}$$

The full steps to solve the objective Eq. (10) is summarized in Algorithm 1.

D. Improving Scalability Against Features

The update for $\mathbf{w}_m$ in Eq. (12) depends on solving a least squares problem in each iteration. Considering the computational complexity $O((N+d)d^2)$ of least squares solver, updating $\mathbf{w}_m$ with Eq. (12) for every iteration may not be computationally feasible in case the number of features d is very large. Due to the multi-modal nature of instances, the dimensionality d is typically large and we are motivated to avoid solving the least squares problem in Eq. (12). To improve the scalability against the number of features d, we employ an optimal line search method [14] and update $\mathbf{w}_m$ via following gradient descent:

$$\mathbf{w}_m = \mathbf{w}_m - s_m\nabla_{\mathbf{w}_m} \ , \tag{15}$$

where $\nabla_{\mathbf{w}_m}$ is the analytical gradient of Eq. (10) with respect to $\mathbf{w}_m$:

$$\begin{aligned}\nabla_{\mathbf{w}_m} = & \ \mathbf{w}_m - \mu\sum_{i=1}^{N}[\mathbf{X}_i(\mathbf{t}_i^m - \mathbf{w}_m^T\mathbf{X}_i - \mathbf{1}b_m + \boldsymbol{\theta}_i^m/\mu)^T] \\ & - \mu\sum_{i'=1}^{N'}\sum_{m'=1}^{K}[\mathbf{X}_{i'}(\mathbf{u}_{i'}^{m'} - \mathbf{w}_m^T\mathbf{X}_{i'} - \mathbf{1}b_m + \boldsymbol{\xi}_{i'}^{m'}/\mu)^T] \\ & - \mu(\mathbf{v}_m - \mathbf{w}_m + \boldsymbol{\gamma}_m/\mu) + 2\tau_2\mathbf{D}_2\mathbf{w}_m \ . \end{aligned} \tag{16}$$

Then we optimize the amount of update s_m along the direction of gradient $\nabla_{\mathbf{w}_m}$:

$$\begin{aligned}s_m = \arg\min_{s_m} \ & \frac{1}{2}\big\|\mathbf{w}_m^T - s_m\nabla_{\mathbf{w}_m}^T\big\|_2^2 \\ & + \tau_2\,\mathrm{tr}((\mathbf{w}_m^T - s_m\nabla_{\mathbf{w}_m}^T)\mathbf{D}_2(\mathbf{w}_m^T - s_m\nabla_{\mathbf{w}_m}^T)^T) \\ & + \frac{\mu}{2}\sum_{i=1}^{N}\Big[\Big\|\mathbf{t}_i^m - (\mathbf{w}_m^T - s_m\nabla_{\mathbf{w}_m}^T)\mathbf{X}_i - \mathbf{1}b_m + \theta_i^m/\mu\Big\|_2^2\Big] \\ & + \sum_{i'=1}^{N'}\sum_{m'=1}^{K}\Big[\frac{\mu}{2}\Big\|\mathbf{u}_{i'}^{m'} - (\mathbf{w}_m^T - s_m\nabla_{\mathbf{w}_m}^T)\mathbf{X}_{i'} - \mathbf{1}b_m \\ & + \boldsymbol{\xi}_{i'}^{m'}/\mu\Big\|_2^2\Big] + \frac{\mu}{2}\|\mathbf{v}_m - (\mathbf{w}_m - s_m\nabla_{\mathbf{w}_m}) + \boldsymbol{\gamma}_m/\mu\|_2^2 \ . \end{aligned} \tag{17}$$

Algorithm 1 The multiblock ADMM updates to optimize Eq. (6).

Data: $\mathbf{X} \in \mathbb{R}^{D\times(n_1+\cdots+n_N)}$ and $\mathbf{Y} \in \{-1,1\}^{K\times N}$.
Hyperparameters: $C > 0$, $\mu > 0$, $\rho > 1$, $\tau_1 \geq 0$, $\tau_2 \geq 0$ and $tolerance > 0$.
Initialize: primal variables $\mathbf{W}, \mathbf{b}, \mathbf{V}, \mathbf{E}, \mathbf{Q}, \mathbf{R}, \mathbf{T}, \mathbf{U}$ and dual variables $\mathbf{\Lambda}, \mathbf{\Sigma}, \mathbf{\Theta}, \mathbf{\Omega}, \mathbf{\Xi}, \mathbf{\Gamma}$.
while residual $> tolerance$ **do**
  Update $\mathbf{D}_{1,g}$ $(g \in G)$ **by** Eq. (8).
  Update $\mathbf{D}_2$ **by** Eq. (9).
  for $m \in K$ **do**
    Update $\mathbf{w}_m \in \mathbf{W}$ **by** Eq. (15).
    Update $b_m \in \mathbf{b}$ **by** $b_m = \frac{\sum_{i=1}^{N}\left[\mathbf{t}_i^m - \mathbf{w}_m^T\mathbf{X}_i + \boldsymbol{\theta}_i^m/\mu\right] + \sum_{i'=1}^{N'}\sum_{m=1}^{K}\left[\mathbf{u}_{i'}^m - \mathbf{w}_m^T\mathbf{X}_{i'} + \boldsymbol{\xi}_{i'}^m/\mu\right]}{N+KN'}$
  end for
  for $(p,m) \in \{N,K\}$ **do**
    Update $e_p^m \in \mathbf{E}$ **by**
    $$e_i^m = \begin{cases} n_i^m - \frac{C}{\mu}y_i^m & \text{when} \quad y_i^m n_i^m > \frac{C}{\mu} \\ 0 & \text{when} \quad 0 \leq y_i^m n_i^m \leq \frac{C}{\mu} \\ n_i^m & \text{when} \quad y_i^m n_i^m < 0 \end{cases},$$
    where $n_i^m = y_i^m - q_i^m + r_i^m - \lambda_i^m/\mu$.
    Update $q_p^m \in \mathbf{Q}$ **by**
    $$q_i^m = \frac{\left(y_i^m - e_i^m + r_i^m - \lambda_i^m/\mu + \max\left(\mathbf{t}_i^m\right) - \sigma_i^m/\mu\right)}{2}$$
    Update $r_p^m \in \mathbf{R}$ **by**
    $$r_i^m = \frac{\left(e_i^m - y_i^m + q_i^m + \lambda_i^m/\mu + \max\left(\mathbf{u}_i^m\right) - \omega_i^m/\mu\right)}{2}$$
    for $j \in n_p$ **do**
      Update $t_{p,j}^m \in \mathbf{T}$ **by**
      $$t_{i,j}^m = \begin{cases} \frac{\max(\boldsymbol{\phi}_i^m) + q_i^m + \sigma_i^m/\mu}{2} & \text{if } j = \arg\max(\boldsymbol{\phi}_i^m) \\ \phi_{i,j}^m & \text{else} \end{cases},$$
      where $\boldsymbol{\phi}_i^m = \mathbf{w}_m^T\mathbf{X}_i + \mathbf{1}b_m - \boldsymbol{\theta}_i^m/\mu$.
      Update $u_{p,j}^m \in \mathbf{U}$ **by**
      $$u_{i,j}^m = \begin{cases} \frac{\max(\boldsymbol{\psi}_i^m) + r_i^m + \omega_i^m/\mu}{2} & \text{if } j = \arg\max(\boldsymbol{\psi}_i^m) \\ \psi_{i,j}^m & \text{else} \end{cases},$$
      where $\boldsymbol{\psi}_i^m = \mathbf{w}_y^T\mathbf{X}_i + \mathbf{1}b_y - \boldsymbol{\xi}_i^m/\mu$.
    end for
    Update $\lambda_p^m, \sigma_p^m, \omega_p^m, \boldsymbol{\theta}_p^m, \boldsymbol{\xi}_p^m$ **by**
    $\lambda_i^m = \lambda_i^m + \mu(e_i^m - (y_i^m - q_i^m + r_i^m))$;
    $\sigma_i^m = \sigma_i^m + \mu(q_i^m - \max(\mathbf{t}_i^m))$;
    $\omega_i^m = \omega_i^m + \mu(r_i^m - \max(\mathbf{u}_i^m))$;
    $\boldsymbol{\theta}_i^m = \boldsymbol{\theta}_i^m + \mu(\mathbf{t}_i^m - (\mathbf{w}_m^T\mathbf{X}_i + \mathbf{1}b_m))$;
    $\boldsymbol{\xi}_i^m = \boldsymbol{\xi}_i^m + \mu(\mathbf{u}_i^m - (\mathbf{w}_y^T\mathbf{X}_i + \mathbf{1}b_y))$
  end for
  for $g \in G$ **do**
    Update $\mathbf{V}^g$ **by** $\mathbf{V}^g = (\mu\mathbf{W}^g - \mathbf{\Gamma}^g)(2\tau_1\mathbf{D}_{1,g} + \mu\mathbf{I})^-$
  end for
  Update $\mathbf{\Gamma} = \mathbf{\Gamma} + \mu(\mathbf{V} - \mathbf{W})$
  Update $\mu = \rho\mu$.
end while

return $(\mathbf{w}_m, \ldots, \mathbf{w}_K) \in \mathbf{W}$ and $(b_1, \ldots, b_K) \in \mathbf{b}$.

We differentiate Eq. (17) with respect to s_m, and set the result equal to zero to earn the solution for s_m:

$$s_m = \frac{\left(\mathbf{w}_m^T(\mathbf{I}+2\tau_2\mathbf{D}_2) - \mu\sum_{i=1}^{N}\hat{\mathbf{t}}_i^m\mathbf{X}_i^T - \mu\sum_{i'=1}^{N'}\sum_{m'}^{K}\hat{\mathbf{u}}_{i'}^{m'}\mathbf{X}_{i'}^T - \mu\hat{\mathbf{v}}_m^T\right)\nabla_{\mathbf{w}_m}}{\nabla_{\mathbf{w}_m}^T\left((1+\mu)\mathbf{I} + 2\tau_2\mathbf{D}_2 + \mu\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i^T + \mu K\sum_{i'=1}^{N'}\mathbf{X}_{i'}\mathbf{X}_{i'}^T\right)\nabla_{\mathbf{w}_m}}, \tag{18}$$

where $\hat{\mathbf{t}}_i^m = \mathbf{t}_i^m - \mathbf{w}_m^T\mathbf{X}_i - \mathbf{1}b_m + \boldsymbol{\theta}_i^m/\mu$ and $\hat{\mathbf{u}}_{i'}^{m'} = \mathbf{u}_{i'}^{m'} - \mathbf{w}_m^T\mathbf{X}_{i'} - \mathbf{1}b_m + \boldsymbol{\xi}_{i'}^{m'}/\mu$ and $\hat{\mathbf{v}}_m = \mathbf{v}_m - \mathbf{w}_m + \boldsymbol{\gamma}_m/\mu$. Finally we plug the gradient $\nabla_{\mathbf{w}_m}$ in Eq. (16) and amount s_m in Eq. (18) into Eq. (15).The gradient descent update in Eq. (15) does not contain the matrix inversion which requires the least squares solver and only depend on the matrix by vector multiplication. Note that $\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i^T$ and $K\sum_{i'=1}^{N'}\mathbf{X}_{i'}\mathbf{X}_{i'}^T$ can be pre-computed in the data preparing step. The time complexity of the inexact solution in Eq. (15) is $O(Nd(n_1 + n_2 + \cdots + n_N))$.

III. Experiments

In this section, we empirically evaluation our proposed method by two parts of experiments. We compare the proposed exact/inexact *MIMSSVM* model to the various SVM and deep learning models (1) in classification performance and (2) in computation time across the increasing number of features.

A. Dataset

We test the classification models on BreaKHis[1] dataset [3]. The BreaKHis dataset was built in collaboration with the P&D Laboratory in Parana, Brazil. The dataset contains 7,909 microscopic biopsy images of breast tumor tissue taken between January and December 2014 in a clinical study. The dataset contains 2,480 samples of benign tissue and 5,429 samples of malignant tissue, therefore the classification on this imbalanced dataset has been considered as challenging task. The images were collected based on the various magnifying factors (40X, 100X, 200X, and 400X), and they were categorized into benign or malignant class in the dataset. The images are sampled from the anonymized 82 patients and in a format of 3-channel RGB PNG and 700 $\times$ 460 pixels. The images were generated from breast tissue biopsy slides, stained with hematoxylin and eosin (HE), and collected by surgical open biopsy (SOB). The diagnosis of each slide was labeled by experienced pathologists in the P&D Laboratory [24].

We design the multi-instance learning problem by binding the multiple histopathological images in a same WSI as a bag, where each image in a bag represents an instance. We segment each image into the patches in the various sizes and extract the features from each patch which will be vectorized and concatenated into an instance. This multi-shape instance enables the classification model to detect the various shapes and sizes of tumors and cells.

[1]The dataset is publicly available and can be accessed at https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

B. Feature Extraction

We vectorize each patch or image by extracting the feature vector through Parameter Free Threshold statistics (PFTAS) [25]. PFTAS extracts the texture features by counting the number of black pixels in the neighborhood of each pixel. Then the total count for all the pixels in a given patch is stored in a nine-bin histogram [25]. The thresholding is conducted by Otsu's algorithm [26] which returns a 162-dimensional feature vector for each patch. In our evaluation of classification performance, each bag contains the random number (in {3, 5, 10}) of images and each image is segmented into non-overlapping two 64 × 64 patches, one 64 × 32 patch, one 32 × 32 patch at the random positions. As a result, each instance is a vector of 162 × 4 features.

C. Comparison Methods

We compare the classification performance and scalability of proposed exact/inexact *MIMSSVM* to the following three SVM models (SIL, NSK, STK) and five deep learning models (mi-Net, MI-Net, MMMI-deep, AMIL, and LAMIL):

- (1) A single-instance learning (SIL) method that assigns the bags' labels to all instances during training and produces the maximum response for each bag/class pair at testing time for the training bag's instances.
- The two multi-instance SVM methods: (2) Normalized Set Kernel (NSK) and (3) Statistics Kernel (STK) [6] map the entire bag to a single-instance via kernel function.
- The five multi-instance deep learning (DL) models: The (4) mi-Net and (5) MI-Net [27] approach to the MIL problem in a way of instance space and embedded space (learning vectorial representation of bag) paradigm respectively. (6) The Multi-Modal Multi-Instance deep learning model (MMMI-deep) [28] learns the global cross-modal representation from the features of each modality. Finally, the two attention mechanism based MIL models are introduced. (7) Attention-based deep Multiple Instance Learning (AMIL) [29] calculates the parameterized attention (importance) score for each instance to generate the probability distribution of bag labels. (8) Loss-based Attention for deep Multiple Instance Learning (LAMIL) [30] proposes to learn the instance scores and predictions jointly by integrating the attention mechanism with the loss function.
- (9) An variation of inexact *MIMSSVM* for the ablation study: We discard the multi-modal learning capability from our model (Ours (MISM) in Table. I) to evaluate the effectiveness of structured sparsity terms introduced to utilize the multi-modality of instances. We set τ_1 and τ_2 to zero to remove the impact of group norm $\|\mathbf{W}\|_{G_1}$ and trace norm $\|\mathbf{W}\|_*$.

For these classification models, we use the following hyper-parameters found by grid search. For SIL, NSK, and STK the regularization tradeoff is set to 1.0. For our exact and inexact *MIMSSVM* models, we set the regularization tradeoff C to $1e+3$ and $1e+4$ respectively. The tolerance is set to $1e-5$ for both, and μ is initialized with $1e-5$ and $1e-10$ respectively. For exact *MIMSSVM* model, τ_1 and τ_2 are set to $1e{-}4$ and $1e{+}3$, and for inexact *MIMSSVM* models, τ_1 and τ_2 are set to 50 and $1e-5$. We use the linear kernel function for SIL, NSK, STK. The deep learning models (mi-Net, MI-Net, MMMI-deep, AMIL, and LAMIL) are implemented using the codes provided as a companion th their papers [27]–[30] and we use the implementations of SVM models (SIL, NSK, and STK) provided by Doran *et al* [31]. Because MMMI-deep, AMIL, and LAMIL take the image as an input, we provide the raw patches to them without extracting features.

D. Classification Performance

In table I, we report the precision, recall, F1-score, accuracy, and balanced accuracy (BACC) in the classification of benign/malignant bags with the PFTAS features input. We employ balanced accuracy considering the imbalanced dataset of many malignant bags. We split the bags into 80% for training and 20% for test set and follow the five-fold cross validation scheme to report the averaged scores and their standard deviations of five test sets.

The comparison between the classification models in table I show that the proposed inexact/exact *MIMSSVM* model clearly surpasses the other existing multi-instance models. We interpret the reason as our model has the better capability in detecting the various shapes and sizes of nucleus and cells. When Ours are compared to the ablation model Ours (MISM), we observe that the introduced structured sparsity terms improve the prediction. This shows the good generalization capability of our multi-modal (multi-shape) learning method for the different cancer types. Another interesting observation from the results at lower magnification levels (100X, 40X) is that our inexact *MIMSSVM* achieves the encouraging performance compared to exact *MIMSSVM*. These results demonstrate that classification pattern for *MIMSSVM* can vary based on the optimization strategy used, much as the optimization algorithm's impact on deep learning models can vary [32]. While our derivation of inexact *MIMSSVM* may not yield the exact optimal solution for the objective function in Eq. (10), our results indicate that the inexact solution in Eq. (15) can improve the prediction compared to the exact solution in Eq. (12). It is supported by the previous finding [33] which has shown some implementations of SVM obtain the highest accuracy before the objective reaches its minimum.

E. Scalability Evaluation

One of the primary contributions of our study is that the derived Algorithm 1 scales to the large number of features. In Fig. 2, we plot the training time to the convergence of loss function of our exact/inexact methods and the other SVM models on PFTAS features input. The DL models (mi-Net, MI-Net, MMMI-deep, AMIL, and LAMIL) are excluded in this timing experiment since their training requires more than three hours. To control the number of features of instances, we concatenate the multiple number of 64 × 64 patches which will be processed into 162 feature vectors. From the

Model	Magnification	Precision	Recall	F1Score	Accuracy	BACC
SIL	**40X**	0.881±0.019	0.806±0.029	0.846±0.030	0.809±0.031	0.831±0.026
NSK	**40X**	0.915±0.014	0.913±0.021	0.911±0.016	0.871±0.020	0.880±0.019
STK	**40X**	0.901±0.028	0.896±0.031	0.889±0.018	0.863±0.019	0.864±0.021
mi-Net	**40X**	0.908±0.019	0.883±0.016	0.892±0.016	0.887±0.010	0.876±0.020
MI-Net	**40X**	0.913±0.018	0.918±0.026	0.914±0.013	0.919±0.016	0.907±0.019
MMMI-deep	**40X**	0.911±0.027	0.919±0.018	0.915±0.019	0.918±0.038	0.902±0.017
AMIL	**40X**	0.893±0.042	0.904±0.026	0.900±0.031	0.868±0.031	0.869±0.020
LAMIL	**40X**	0.907±0.036	0.916±0.053	0.909±0.017	0.883±0.020	0.890±0.024
Ours	**40X**	0.912±0.016	0.918±0.034	0.913±0.023	0.921±0.021	0.926±0.021
Ours (inexact)	**40X**	**0.935±0.017**	0.920±0.021	**0.921±0.010**	**0.945±0.012**	**0.931±0.021**
Ours (MISM)	**40X**	0.893±0.016	0.901±0.014	0.904±0.015	0.916±0.014	0.908±0.023
SIL	**100X**	0.854±0.016	0.837±0.028	0.847±0.021	0.822±0.031	0.829±0.022
NSK	**100X**	0.869±0.021	0.901±0.012	0.892±0.019	0.889±0.016	0.903±0.019
STK	**100X**	0.853±0.031	0.912±0.020	0.880±0.015	0.893±0.018	0.881±0.014
mi-Net	**100X**	0.876±0.011	0.842±0.014	0.856±0.008	0.891±0.011	0.909±0.019
MI-Net	**100X**	0.893±0.020	0.876±0.023	0.893±0.019	0.901±0.018	0.889±0.020
MMMI-deep	**100X**	**0.932±0.035**	0.923±0.015	0.919±0.024	0.915±0.017	0.921±0.021
AMIL	**100X**	0.908±0.028	0.897±0.034	0.906±0.031	0.928±0.019	0.931±0.024
LAMIL	**100X**	0.916±0.030	0.892±0.027	0.909±0.029	0.919±0.023	0.921±0.029
Ours	**100X**	0.903±0.014	**0.936±0.020**	0.921±0.014	0.916±0.015	0.923±0.024
Ours (inexact)	**100X**	0.921±0.014	0.931±0.021	**0.926±0.019**	**0.942±0.021**	**0.938±0.026**
Ours (MISM)	**100X**	0.881±0.018	0.896±0.017	0.890±0.023	0.919±0.026	0.909±0.009
SIL	**200X**	0.879±0.013	0.840±0.019	0.857±0.017	0.851±0.019	0.832±0.023
NSK	**200X**	0.900±0.016	0.882±0.019	0.893±0.019	0.911±0.017	0.913±0.026
STK	**200X**	0.883±0.021	0.876±0.035	0.884±0.016	0.876±0.027	0.863±0.023
mi-Net	**200X**	0.893±0.015	0.847±0.021	0.868±0.029	0.907±0.016	0.893±0.009
MI-Net	**200X**	0.914±0.016	0.901±0.040	0.908±0.021	0.917±0.020	0.916±0.017
MMMI-deep	**200X**	0.913±0.017	**0.926±0.028**	0.918±0.022	0.918±0.017	0.912±0.015
AMIL	**200X**	0.879±0.032	0.919±0.040	0.889±0.025	0.884±0.034	0.898±0.040
LAMIL	**200X**	0.895±0.019	0.871±0.025	0.889±0.041	0.899±0.029	0.876±0.026
Ours	**200X**	0.916±0.013	0.922±0.008	**0.920±0.012**	**0.921±0.019**	0.918±0.026
Ours (inexact)	**200X**	**0.917±0.024**	0.908±0.015	0.911±0.016	0.918±0.025	0.916±0.012
Ours (MISM)	**200X**	0.907±0.020	0.913±0.024	0.898±0.012	0.912±0.021	0.912±0.024
SIL	**400X**	0.905±0.024	0.810±0.031	0.852±0.021	0.817±0.017	0.802±0.021
NSK	**400X**	0.881±0.020	0.905±0.017	0.881±0.011	0.846±0.016	0.881±0.021
STK	**400X**	0.892±0.023	0.901±0.027	0.899±0.016	0.861±0.017	0.853±0.018
mi-Net	**400X**	0.893±0.015	0.847±0.021	0.868±0.029	0.907±0.016	0.893±0.009
MI-Net	**400X**	0.901±0.021	0.887±0.022	0.897±0.019	0.929±0.031	0.917±0.031
MMMI-deep	**400X**	0.895±0.024	0.851±0.028	0.883±0.024	0.915±0.016	0.904±0.023
AMIL	**400X**	0.906±0.030	0.874±0.031	0.908±0.032	0.890±0.027	0.910±0.029
LAMIL	**400X**	0.911±0.027	0.890±0.028	0.901±0.061	0.897±0.031	0.904±0.028
Ours	**400X**	0.929±0.026	0.919±0.019	0.923±0.016	**0.950±0.018**	0.942±0.020
Ours (inexact)	**400X**	0.925±0.023	0.860±0.020	0.891±0.016	0.931±0.019	0.913±0.025
Ours (MISM)	**400X**	0.912±0.014	0.918±0.017	0.915±0.015	0.906±0.019	0.901±0.027

TABLE I

THE CLASSIFICATION PERFORMANCE OF OUR *MIMSSVM* AND COMPETING MODELS OVER THE VARIOUS MAGNIFICATION LEVELS. THE PFTAS FEATURES ARE PROVIDED AS INPUT EXCEPT FOR MMMI-DEEP, AMIL, AND LAMIL WHICH RECEIVE RAW PATCHES AS INPUT. THE BEST SCORES ARE HIGHLIGHTED IN BOLD.

results in Fig. 2, the inexact variation of *MIMSSVM* scales significantly better than the exact *MIMSSVM* and other SVM models against the increasing number of features. This is well represented by the analytical complexity of two derivations ($O(Nd(n_1+n_2+\cdots+n_N))$ *v.s.* $O((N+d)d^2)$) as discussed in Section II-D. This result validates the superior scalability of the proposed *MIMSSVM* over the other SVM models which rely on repeatedly solving a quadratic programming problem.

IV. CONCLUSION

Information in the medical dataset is usually delivered in a variety of modalities as the development of data mining technologies, and multi-modal study is attracting more attention in the machine learning researches. In the image analysis, the patches in the different shapes sampled from the image can be regarded as the different views on the image. In this study we present a novel *Multi-Instance Multi-Shape SVM* method which scales to the large number of features. Our model employs structured sparsity regularizations to achieve the modality-wise sparsities and extract the predictive shapes of patches in the prediction. The multi-modal learning capability of our model is not limited in the different shapes of patches and our model can be applied to the various multi-modal data analysis. We have conducted extensive experiments on the BreaKHis dataset and observed the promising performance and scalability of the proposed method when compared to the existing SVM and deep learning models. In addition to the improved performance and scalability, our model identifies the disease relevant regions in the images supported by the multiple histopathological studies.

ACKNOWLEDGMENT

Corresponding author: Hua Wang (huawangcs@gmail.com).

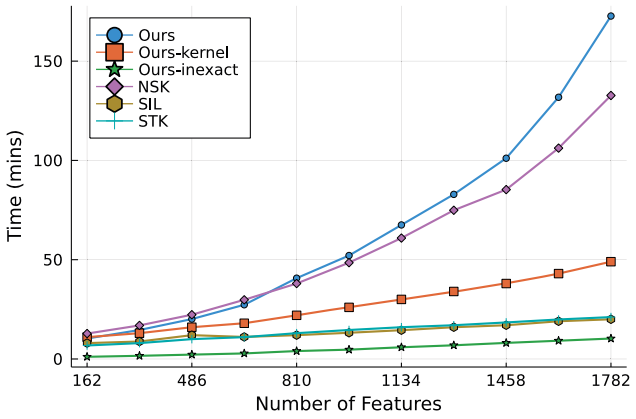

Fig. 2. Training time over the increasing number of features.

This work was supported in part by the National Science Foundation (NSF) under the grants of IIS 1652943, IIS 1849359, CNS 1932482 and CCF 2029543.

REFERENCES

[1] CDC, "The Basics on Hereditary Breast and Ovarian Cancer — CDC," https://www.cdc.gov/genomics/disease/breast_ovarian_cancer/basics_hboc.htm, 2020, [Online; accessed 1-August-2021].

[2] J. van der Laak, G. Litjens, and F. Ciompi, "Deep learning in histopathology: the path to the clinic," *Nature medicine*, vol. 27, no. 5, pp. 775–784, 2021.

[3] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *Ieee transactions on biomedical engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[4] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning." in *NIPS*, vol. 2. Citeseer, 2002, pp. 561–568.

[5] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 105–112.

[6] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *ICML*, vol. 2, 2002, p. 7.

[7] H. Wang, F. Nie, and H. Huang, "Learning instance specific distance for multi-instance classification," in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[8] H. Wang, H. Huang, F. Kamangar, F. Nie, and C. Ding, "Maximum margin multi-instance learning," *Advances in neural information processing systems*, vol. 24, 2011.

[9] K. Liu, H. Wang, F. Nie, and H. Zhang, "Learning multi-instance enriched image representations via non-greedy ratio maximization of the ℓ_1-norm distances," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7727–7735.

[10] L. Brand, L. Z. Baker, C. Ellefsen, J. Sargent, and H. Wang, "A linear primal-dual multi-instance svm for big data classifications," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 21–30.

[11] E. G. Fischer, "Nuclear morphology and the biology of cancer cells," *Acta cytologica*, vol. 64, no. 6, pp. 511–519, 2020.

[12] H. Wang, F. Nie, and H. Huang, "Multi-view clustering and feature learning via structured sparsity," in *International conference on machine learning*. PMLR, 2013, pp. 352–360.

[13] H. Wang, F. Nie, H. Huang, and C. Ding, "Heterogeneous visual features fusion via sparse multimodal machine," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3097–3102.

[14] F. Nie, Y. Huang, X. Wang, and H. Huang, "New primal svm solver with linear computational cost for big data classifications," in *Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32*, 2014, pp. II–505.

[15] Y. Liu, Y. Guo, H. Wang, F. Nie, and H. Huang, "Semi-supervised classifications via elastic and robust embedding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[16] H. Yang, K. Liu, H. Wang, and F. Nie, "Learning strictly orthogonal p-order nonnegative laplacian embedding via smoothed iterative reweighted method," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 4040–4046.

[17] M. Hong and Z.-Q. Luo, "On the linear convergence of the alternating direction method of multipliers," *Mathematical Programming*, vol. 162, no. 1-2, pp. 165–199, 2017.

[18] F. Nie, H. Wang, X. Cai, H. Huang, and C. Ding, "Robust matrix completion via joint schatten p-norm and ℓ_p-norm minimization," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 566–574.

[19] H. Wang, F. Nie, H. Huang, S. L. Risacher, A. J. Saykin, L. Shen, and A. D. N. Initiative, "Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning," *Bioinformatics*, vol. 28, no. 12, pp. i127–i136, 2012.

[20] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, and L. Shen, "High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction," *Advances in neural information processing systems*, vol. 25, pp. 1277–1285, 2012.

[21] L. Brand, H. Wang, H. Huang, S. Risacher, A. Saykin, L. Shen *et al.*, "Joint high-order multi-task feature learning to predict the progression of alzheimer's disease," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 555–562.

[22] L. Brand, K. Nichols, H. Wang, L. Shen, and H. Huang, "Joint multi-modal longitudinal regression and classification for alzheimer's disease prediction," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1845–1855, 2019.

[23] L. Lu, S. Elbeleidy, L. Baker, H. Wang, L. Shen, and H. Heng, "Improved prediction of cognitive outcomes via globally aligned imaging biomarker enrichments over progressions," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 11, pp. 3336–3346, 2021.

[24] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *2016 international joint conference on neural networks (IJCNN)*. IEEE, 2016, pp. 2560–2567.

[25] N. A. Hamilton, R. S. Pantelic, K. Hanson, and R. D. Teasdale, "Fast automated cell phenotype image classification," *BMC bioinformatics*, vol. 8, no. 1, pp. 1–8, 2007.

[26] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[27] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[28] H. Li, F. Yang, X. Xing, Y. Zhao, J. Zhang, Y. Liu, M. Han, J. Huang, L. Wang, and J. Yao, "Multi-modal multi-instance learning using weakly correlated histopathological images and tabular clinical information," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 529–539.

[29] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[30] X. Shi, F. Xing, Y. Xie, Z. Zhang, L. Cui, and L. Yang, "Loss-based attention for deep multiple instance learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5742–5749.

[31] G. Doran and S. Ray, "A theoretical and empirical analysis of support vector machine methods for multiple-instance classification," *Machine learning*, vol. 97, no. 1-2, pp. 79–102, 2014.

[32] Y. Wang, J. Liu, J. Mišić, V. B. Mišić, S. Lv, and X. Chang, "Assessing optimizer impact on dnn model sensitivity to adversarial examples," *IEEE Access*, vol. 7, pp. 152 766–152 776, 2019.

[33] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, "Coordinate descent method for large-scale l2-loss linear support vector machines." *Journal of Machine Learning Research*, vol. 9, no. 7, 2008.