Bioinformatics, 38, 2022, i92–i100
https://doi.org/10.1093/bioinformatics/btac267
ISCB/ISMB 2022

Scaling multi-instance support vector machine to breast cancer detection on the BreaKHis dataset

Hoon Seo, Lodewijk Brand, Lucia Saldana Barco and Hua Wang*

Department of Computer Science, Colorado School of Mines, Golden, CO 80401, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Breast cancer is a type of cancer that develops in breast tissues, and, after skin cancer, it is the most commonly diagnosed cancer in women in the United States. Given that an early diagnosis is imperative to prevent breast cancer progression, many machine learning models have been developed in recent years to automate the histopathological classification of the different types of carcinomas. However, many of them are not scalable to large-scale datasets.

Results: In this study, we propose the novel Primal-Dual Multi-Instance Support Vector Machine to determine which tissue segments in an image exhibit an indication of an abnormality. We derive an efficient optimization algorithm for the proposed objective by bypassing the quadratic programming and least-squares problems, which are commonly employed to optimize Support Vector Machine models. The proposed method is computationally efficient, thereby it is scalable to large-scale datasets. We applied our method to the public BreaKHis dataset and achieved promising prediction performance and scalability for histopathological classification.

Availability and implementation: Software is publicly available at: https://1drv.ms/u/s!AiFpD21bgf2wgRLbQq08ixD0SgRD?e=OpqEmY.

Contact: huawangcs@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Every year, approximately 250 000 women in the United States are diagnosed with breast cancer (CDC, 2020). Differentiating between the different types of carcinomas (ductal, lobular, mucinous and papillary) is essential for making an accurate diagnosis. Histopathology allows for this close examination that leads to patients receiving a personalized treatment and can increase their likelihood of survival. Histopathology is the examination of tissue sections with a microscope to aid in the diagnosis of illnesses such as cancer and inflammatory diseases and increase the likelihood of survival. These tissue sections can also be called whole-slide images (WSIs) or histopathological images when they are digitized. Traditionally, clinical disciplines such as radiology and pathology have relied heavily on specialized training to detect the presence of these diseases in histopathological images. A diagnosis is based on features exhibited by tissue samples on a cellular level. An anomaly in the cell architecture and the presence or absence of certain biological attributes can be strong indicators of a particular disease. For example, abnormal cells that divide uncontrollably, also known as carcinomas, lead to a cancer diagnosis when detected. A pathologist can detect this abnormal growth/tumor from a histopathological image and assess which regimen should be prescribed to halt the progression of the disease. This pattern analysis is an essential component of precision medicine, since it makes a diagnosis based on patient-specific histopathological images.

Modern medical procedures and technologies have increased the number of biopsies performed, and consequently, the number of histopathological images collected has increased far beyond the reasonable workload of pathologists (van der Laak *et al.*, 2021). This poses an impediment to precision medicine, since it requires the analysis of vast amounts of medical data. However, recent advancements in the field of artificial intelligence have shown promise in automating the analysis of histopathological images and improving the accuracy and speed of a diagnosis. Just as a pathologist finds patterns that help detect cellular abnormalities, algorithms can be used to extract features from an image such as pixel intensity (Hamilton *et al.*, 2007), texture (Haralick, 1979) and Zernike moments (Khotanzad and Hong, 1990). The application of computational algorithms to diagnostic fields can aid pathologists in drawing accurate and precise conclusions in an efficient and reproducible manner (Gurcan *et al.*, 2009).

In our research, we focused our analysis efforts on developing a classification model for the public BreaKHis dataset (Spanhol *et al.*, 2015), which is composed of 7909 histopathological images of different types of benign and malignant breast cancer tumors. This dataset has been instrumental in our work, since its structure allows for extensive and precise classification of histopathological images. The dataset is split into benign and malignant categories, and these are further subdivided into different types of carcinomas. The WSIs in each tumor type group are then amplified to four different magnification factors, and they are usually segmented into the patches because of their large size. As a result, the classification problem is naturally formulated as a multi-instance learning (MIL) problem (Brand *et al.*, 2021a,b; Wang *et al.*, 2011) to determine which segments of tissue in an image exhibit an indication of an abnormality. MIL is an area of machine learning in which training and testing

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

data are organized into sets of instances known as bags. MIL is a weakly supervised learning algorithm, which means that the data are frequently provided at the bag level instead of the instance level, therefore 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 and benign, and the rest of the instances or histopathological images follow suit. Despite being a very powerful approach, MIL remains a challenging problem as many standard machine learning approaches rely on fixed-length vector input which are not applicable to the dataset with a varying number of instances per bag. At the same time, MIL models should be translation invariant against the instances of each set input; the prediction of model should not be affected by the order of instances. In our work, breast cancer histopathological images are represented by a bag (set) of patches, as illustrated in Figure 1. The bags, or images, are labeled as either malignant or benign while the instances, or patches, remain unlabeled (Brand *et al.*, 2021a,b; Wang *et al.*, 2011). Taking these facts into account, we propose the Primal-Dual Multi-Instance SVM (*pdMISVM*) method (Brand *et al.*, 2021a), which improves the efficiency of optimization compared to the previously mentioned MIL approaches.

1.1 Related works

To ease the heavy workloads of pathologists, Computer Aided Diagnostics (CAD) has emerged to determine whether an image shows any indication of carcinoma and, if so, where the abnormality is located within the histopathological image. One of the widely used approaches is to use Convolutional Neural Networks (CNNs) trained by the patches extracted from WSIs. CNN is the combination of convolutional layers and consecutive fully connected layers and their concept comes from the working principle of receptive fields and neurons of the human eye and brain. Krizhevsky *et al.* (2012) has shown that the deep structure of the CNN can achieve state-of-the-art performance in image recognition tasks. Their model, AlexNet, has been successfully applied to BreakHis by Titoriya and Sachdeva (2019). However, CAD based on the CNNs still faces obstacles, because training a CNN requires a large amount of training data with the big computational burdens. These requirements make it difficult for predictive models to be combined with the CAD systems. The SVM applications as a practical alternative to deep learning models has also been studied (Zheng *et al.*, 2014). For example, SVM with sparsity inducing regularization (Kahya *et al.*, 2017) can achieve the promising accuracy higher than 90% in image classification. Although SVM models have fewer trainable parameters than deep learning models, and therefore require less time and computational cost to train, their training time and computational complexity increases rapidly as the number of input features increases (Kumar and Rath, 2015; Peng *et al.*, 2016). Another problem of the traditional SVM models is that they are single-instance learning (SIL) models, i.e. they are not able to handle the varying number of input instances, while the WSIs are usually segmented into the multiple patches (instances) because of the large size of WSIs.

In light of the above issue, multiple instance learning would be the better choice for the disease detection applications, and these types of algorithms have also been evaluated on the BreaKHis dataset previously. This is because, in a SIL model, classification becomes difficult when a single patch of insignificant region on the image is given. Meanwhile, MIL models enable the correct classification from some important patches, even if most patches do not include indication of carcinoma. To deal with the multi-instance dataset, several MIL methods have achieved satisfactory results in the past when performing similar tasks especially on the BreaKHis dataset (Sudharshan *et al.*, 2019). For example, Multiple Instance Learning Convolutional Neural Networks (MILCNN) (Sudharshan *et al.*, 2019) take a deep learning approach while others opt for an SVM-based multiple instance learning alternative. Some examples are Multi-Instance Support Vector Machine (MISVM) (Andrews *et al.*, 2002), sparse Multi-Instance Learning (sMIL) and sparse balanced MIL (sbMIL) (Bunescu and Mooney, 2007), and Normalized Set Kernel (NSK) and Statistics Kernel (STK) (Gärtner *et al.*, 2002). These are all methods that have been deemed successful at correctly labeling the bags in the testing dataset as either malignant or benign. However, despite the promising performance of MIL models, we point out that there is a lack of discussions on the scalability of MIL models or they do not scale to the large dataset. In addition, it is difficult to efficiently learn the hypothesis

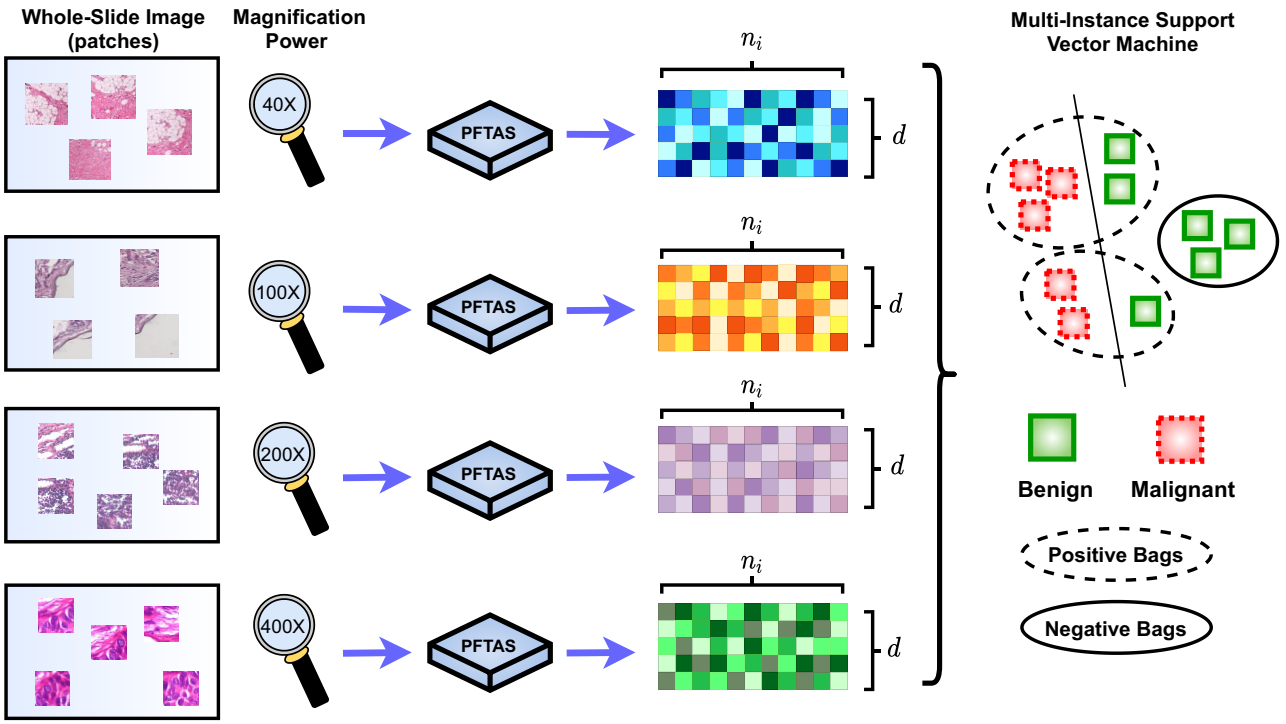

Fig. 1. A visualization of our processing pipeline for our MIL algorithm applied to the BreaKHis dataset. We sample the patches (instances) from the histopathological images (bags) of four different magnification levels, and process the patches with PFTAS method which results in d-features vectors of n_i instances for each image. Finally, our multi-instance SVM classify the bags as malignant or benign

space of MIL models which involves with the multiple instances (Wei *et al.*, 2014). Unlike the previous existing algorithms, our approach scales well to larger datasets, which adds to its value for practical use.

1.2 The paper organization

In the remainder of this manuscript, we present an objective and associated solution of the novel *pdMISVM* that extends to large-scale data. We derive the optimization algorithm based on the multi-block alternating direction method of multipliers (ADMM) (Hong and Luo, 2017) to bypass the quadratic programming problem that comes from the typical SVM and MISVM models. We further improve the ADMM derivation to decrease the complexity with respect to the large number of features. Lastly, we provide an application of the proposed method to classify the bag of patches and identify disease relevant patches, which can reduce the workload of pathologists.

2 Materials and data sources

We perform classifications on the publicly available BreaKHis dataset (https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/). The BreaKHis dataset was built in collaboration with the P&D Laboratory in Parana, Brazil. BreaKHis was first introduced in 'A Dataset for Breast Cancer Histopathological Image Classification' by Spanhol *et al.* (2015). The dataset comprises 7909 microscopic biopsy images of breast tumor tissue images collected in a clinical study from January 2014 to December 2014. The dataset contains 2480 benign and 5429 malignant tissue samples. The images were collected using different magnifying factors (40×, 100×, 200× and 400×), and they were organized into these categories in the dataset. The samples were acquired from 82 patients whose data were anonymized. Samples were generated from breast tissue biopsy slides, stained with hematoxylin and eosin (HE) and collected by surgical open biopsy (SOB). They were labeled in the P&D Laboratory, and the diagnosis of each slide was determined by experienced pathologists (Spanhol *et al.*, 2016).

We segment the histopathological images into patches. In our experiments, each patch contains a 64×64 section of pixels and we extracted a random number (sampled from $\{1, 5, 10\}$) of patches for each of the images. However, in the raw tissue segments, the elements of interest such as nuclei may not be clearly visible. In light of this issue, we extract the feature vector through Parameter Free Threshold Statistics (PFTAS) (Hamilton *et al.*, 2007) for each patch. Based on experimental results of previous study (Spanhol *et al.*, 2015), PFTAS outperforms the other features such as Local Binary Patterns (LBP) (Ojala *et al.*, 2002), Completed LBP (CLBP) (Guo *et al.*, 2010), Local Phase Quantization (LPQ) (Ojansivu and Heikkilä, 2008) and Grey-Level Co-occurrence Matrix (GLCM) (Haralick *et al.*, 1973) in BreaKHis dataset. PFTAS is a method that extracts texture features by counting the number of black pixels in the neighborhood of a pixel. The total count for all the pixels in a given image is stored in a nine-bin histogram (Hamilton *et al.*, 2007). The thresholding is done by Otsu's algorithm (Otsu, 1979) and the extractor returns a 162-dimensional feature vector. The 162 features consist of 3 channels (RGB) × 9 pixels × 3 thresholding ranges concatenated with its bitwise negated version. The Otsu's algorithm iteratively finds the optimal threshold value by maximizing the inter-class intensity variance. As a result, PFTAS features are robust against the varying mean of intensity distribution for each RGB channel across images. To control the number of features, we concatenate several patches, and the final number of features is a multiple of 162. In our experiments, 7909 bags (images) are involved, of which 5429 are malignant and 2480 are benign.

3 Methods

In this section, we develop an objective for the scalable *pdMISVM* algorithm designed to handle multi-instance data. Our formulation for *pdMISVM* provides an efficient solution to avoid dependency on a quadratic programming or least-squares approach.

3.1 Notation

In this article, we denote matrices as $\mathbf{M}$, vectors as $\mathbf{m}$ and scalars as m. The ith row and jth column of $\mathbf{M}$ are $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively. Similarly, m_j^i is the scalar value indexed by the ith row and jth column of $\mathbf{M}$. The matrix $\mathbf{M}_p$ corresponds to the pth column-block of $\mathbf{M}$. Each bag $\mathbf{X}_i = \{\mathbf{x}_i^1, \ldots, \mathbf{x}_i^{n_i}\}$ contains n_i patches and its associated label of mth class is represented by $y_i^m \in \{-1, 1\}$.

3.2 A primal-dual multi-instance support vector machine

The K class multi-instance support vector machine was proposed by Andrews *et al.* (2002), which solve the following objective:

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

where $(\cdot)_+ = \max(\cdot, 0)$ and its decision function is given by:

$$\tilde{y}_i = \underset{m'}{\operatorname{argmax}}(\max(\mathbf{W}^T\mathbf{X}_i + \mathbf{b}1_i)^{m'}), \tag{2}$$

as illustrated in Figure 2.

The MISVM objective in Equation (1) is generally difficult to solve because of the coupled primal variables $\mathbf{w}_k$, b_m by the $\max(\cdot)$ operations. Inspired by Nie *et al.* (2014) and Wang and Zhao (2017), we split the primal variables in Equation (1) via the ADMM approach by introducing the following constraints:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{b},\mathbf{E},\mathbf{Q},\mathbf{R},\mathbf{T},\mathbf{U}} \quad & \frac{1}{2}\sum_{m=1}^{K}||\mathbf{w}_m||_2^2 + C\sum_{i=1}^{N}\sum_{m=1}^{K}(y_i^m e_i^m)_+ \\ \text{subject to} \quad & e_i^m = y_i^m - q_i^m + r_i^m,\ q_i^m = \max(\mathbf{t}_i^m), \\ & r_i^m = \max(\mathbf{u}_i^m), \mathbf{t}_i^m = \mathbf{w}_m^T\mathbf{X}_i + 1b_m, \\ & \mathbf{u}_i^m = \mathbf{w}_y^T\mathbf{X}_i + 1b_y. \end{aligned} \tag{3}$$

From Equation (3) we derive the following augmented Lagrangian function:

$$\begin{aligned} \mathcal{L}_\mu = \ & \frac{1}{2}\sum_{m=1}^{K}||\mathbf{w}_m||_2^2 + \sum_{i=1}^{N}\sum_{m=1}^{K}C(y_i^m e_i^m)_+ \\ & + \frac{\mu}{2}\sum_{i=1}^{N}\sum_{m=1}^{K}\Big[\Big(e_i^m - (y_i^m - q_i^m + r_i^m - \lambda_i^m/\mu)\Big)^2 \\ & + \Big(q_i^m - \max(\mathbf{t}_i^m) + \sigma_i^m/\mu\Big)^2 \\ & + ||\mathbf{t}_i^m - (\mathbf{w}_m^T\mathbf{X}_i + 1b_m) + \boldsymbol{\theta}_i^m/\mu||_2^2 \\ & + \Big(r_i^m - \max(\mathbf{u}_i^m) + \omega_i^m/\mu\Big)^2 \\ & + ||\mathbf{u}_i^m - (\mathbf{w}_y^T\mathbf{X}_i + 1b_y) + \boldsymbol{\xi}_i^m/\mu||_2^2\Big], \end{aligned} \tag{4}$$

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

3.3 The solution algorithm

In this section, we derive the efficient solution algorithm to minimize the proposed objective in Equation (4). In Algorithm 1, we repeat the primal-dual updates until the gap in constraints from the augmented Lagrangian terms in Equation (4) becomes smaller than a predefined tolerance. In order not to distract reading attention and due to space limit, we only provide the derivation details for the class-hyperplane in $\mathbf{w}_m$ and b_m for each mth class in the main article, and leave the derivations for the other variables in the online Supplementary Appendix of this article.

b update. By differentiating Equation (7) element-wise with respect to b_m and setting the result equal to zero, we have the following update:

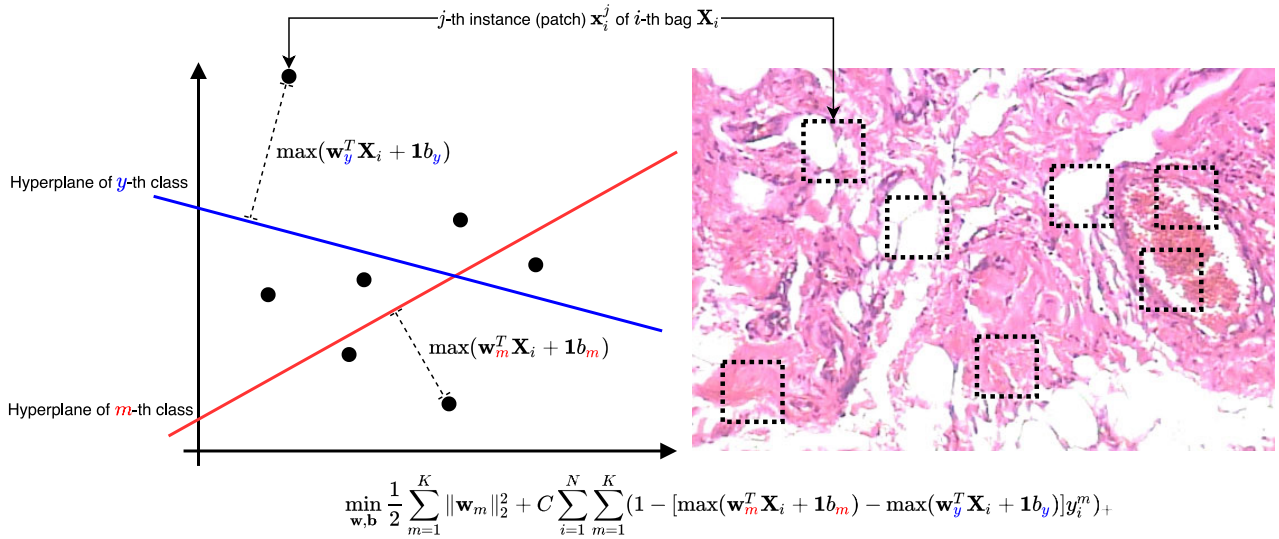

Fig. 2. An illustration for the objective in Equation (1). In our model, each patch corresponds to each instance in a bag. We first calculate the distance from the hyperplane of each mth class to the farthest instance, which is the key instance triggering the bag label. By minimizing Equation (1), we optimize $\mathbf{W}$ and $\mathbf{b}$ of the hyperplanes so that the distance to the hyperplane of the correct class ($m = y$) is greater than the distance to the hyperplane of the incorrect ($m \neq y$) class

$$b_m = \underset{b_m}{\operatorname{argmin}} \sum_{i=1}^{N}\left[||\mathbf{t}_i^m - (\mathbf{w}_m^T\mathbf{X}_i + 1b_m) + \theta_i^m/\mu||_2^2\right] + \sum_{i'=1}^{N'}\sum_{m'=1}^{K}\left[||\mathbf{u}_{i'}^{m'} - (\mathbf{w}_m^T\mathbf{X}_{i'} + 1b_m) + \xi_{i'}^m/\mu||_2^2\right], \tag{5}$$

where i' indicates the column blocks that belong to the mth class are chosen from $\mathbf{X}$. Taking the derivative of Equation (5) with respect to b_m, setting the derivative equal to zero, and solving for b_m gives:

$$b_m = \frac{\sum_{i=1}^{N}[\mathbf{t}_i^m - \mathbf{w}_m^T\mathbf{X}_i + \boldsymbol{\theta}_i^m/\mu] + \sum_{i'=1}^{N'}\sum_{m=1}^{K}[\mathbf{u}_{i'}^m - \mathbf{w}_m^T\mathbf{X}_{i'} + \boldsymbol{\xi}_{i'}^m/\mu]}{N + KN'}, \tag{6}$$

where N' is the total number of patients belonging to the mth class.

W update (without kernel). We discard all terms in Equation (4) which do not include $\mathbf{W}$ and optimize the columns of $\mathbf{W}$ separately by solving the following K problems for $m = 1, \ldots, K$:

$$\mathbf{w}_m^* = \underset{\mathbf{w}_m}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}_m||_2^2 + \frac{\mu}{2}\sum_{i=1}^{N}\left[||\mathbf{t}_i^m - (\mathbf{w}_m^T\mathbf{X}_i + 1b_m) + \boldsymbol{\theta}_i^m/\mu||_2^2\right] + \sum_{i'=1}^{N'}\sum_{m'=1}^{K}\left[\frac{\mu}{2}||\mathbf{u}_{i'}^{m'} - (\mathbf{w}_m^T\mathbf{X}_{i'} + 1b_m) + \boldsymbol{\xi}_{i'}^{m'}/\mu||_2^2\right], \tag{7}$$

where N' is the number of bags which belongs to mth class, and i' denotes the indices of column blocks of $\mathbf{X}$ and the corresponding columns of $\mathbf{U}$ and $\boldsymbol{\Xi}$. Finally $\mathbf{t}_i^m$, $\boldsymbol{\theta}_i^m$, $\mathbf{u}_{i'}^{m'}$ and $\boldsymbol{\xi}_{i'}^{m'}$ are row vectors corresponding to the ith bag and mth class in $\mathbf{T}$, $\boldsymbol{\Theta}$, $\mathbf{U}$ and $\boldsymbol{\Xi}$. By letting the derivative of Equation (7) with respect to $\mathbf{w}_m$ equal zero, we attain the following closed form solution:

$$(\mathbf{w}_m^*)^T = \left(\sum_{i=1}^{N}\left[(\mathbf{t}_i^m - 1b_m + \boldsymbol{\theta}_i^m/\mu)\mathbf{X}_i^T\right] + \sum_{i'=1}^{N'}\sum_{m'=1}^{K}\left[(\mathbf{u}_{i'}^{m'} - 1b_m + \boldsymbol{\xi}_{i'}^{m'}/\mu)\mathbf{X}_{i'}^T\right]\right) * \left(\mathbf{I}/\mu + \sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i^T + K\sum_{i'=1}^{N'}\mathbf{X}_{i'}\mathbf{X}_{i'}^T\right)^{-1}. \tag{8}$$

In the calculation of Equation (8) we can avoid an inverse calculation through a least-squares solver.

W update (with kernel). The kernel method (Shawe-Taylor *et al.*, 2004) is widely used in classification tasks to deal with non-linearity of the data. We provide the kernel extension of our method to learn the non-linear relationship between bag and target label. For the arbitrary (possibly non-linear) kernel function ϕ, we map all the columns (instances) of $\mathbf{X}_i \in \Re^{d\times n_i}$ to feature vectors $\phi(\mathbf{X}_i) = \boldsymbol{\Phi}_i \in \Re^{d_z\times n_i}$, and Equation (7) can be rewritten into:

$$\mathbf{w}_m^* = \underset{\mathbf{w}_m}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}_m||_2^2 + \frac{\mu}{2}\sum_{i=1}^{N}\left[||\mathbf{t}_i^m - (\mathbf{w}_m^T\boldsymbol{\Phi}_i + 1b_m) + \boldsymbol{\theta}_i^m/\mu||_2^2\right] + \sum_{i'=1}^{N'}\sum_{m'=1}^{K}\left[\frac{\mu}{2}||\mathbf{u}_{i'}^{m'} - (\mathbf{w}_m^T\boldsymbol{\Phi}_{i'} + 1b_m) + \boldsymbol{\xi}_{i'}^{m'}/\mu||_2^2\right]. \tag{9}$$

We take the derivative with respect to $\mathbf{w}_m$ and set it equal to zero to solve for $\mathbf{w}_m$:

$$(\mathbf{w}_m^*)^T = \left([(\mathbf{t}^m - 1b_m + \boldsymbol{\theta}^m/\mu)\boldsymbol{\Phi}^T] + \sum_{m'=1}^{K}[(\mathbf{u}_{\prime}^{m'} - 1b_m + \boldsymbol{\xi}_{\prime}^{m'}/\mu)\boldsymbol{\Phi}_{\prime}^T]\right) * (\mathbf{I}/\mu + \boldsymbol{\Phi}\boldsymbol{\Phi}^T + K\boldsymbol{\Phi}_{\prime}\boldsymbol{\Phi}_{\prime}^T)^{-1}, \tag{10}$$

where $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_N] \in \Re^{d_z\times N_t}$ and $\boldsymbol{\Phi}_{\prime} = [\boldsymbol{\Phi}_{1'}, \ldots, \boldsymbol{\Phi}_{N'}] \in \Re^{d_z\times N_t'}$. Here $N_t = \sum_{i=1}^{N} n_i$ and $N_t' = \sum_{i'=1}^{N'} n_i$ denote the total number of instances which belongs to all classes and mth class respectively, and $\boldsymbol{\Phi}_{\prime}$ contains the N' column blocks of $\boldsymbol{\Phi}$ corresponding to the mth class.

However, the dimensionality d_z of feature vectors $\boldsymbol{\Phi}$ of kernel function can be very large (possibly infinitely large), thus calculating $(\mathbf{I}/\mu + \boldsymbol{\Phi}\boldsymbol{\Phi}^T + K\boldsymbol{\Phi}_{\prime}\boldsymbol{\Phi}_{\prime}^T)^{-1}$ in Equation (10) may not be computationally feasible. In order to derive the scalable solution against arbitrary kernel function, we rewrite Equation (10) into the following matrix form:

$$(\mathbf{w}_m^*)^T = \mathbf{s}_m\mathbf{D}\hat{\boldsymbol{\Phi}}^T * (\mathbf{I}/\mu + \hat{\boldsymbol{\Phi}}\mathbf{D}\hat{\boldsymbol{\Phi}}^T)^{-1}, \tag{11}$$

where $\mathbf{s}_m = [\mathbf{t}^m - 1b_m + \boldsymbol{\theta}^m/\mu, \quad 1/K\sum_{m'=1}^{K}(\mathbf{u}_{\prime}^{m'} - 1b_m + \boldsymbol{\xi}_{\prime}^{m'}/\mu)]$, $\mathbf{D} = [\mathbf{I}, 0; 0, K\mathbf{I}]$ and $\hat{\boldsymbol{\Phi}} = [\boldsymbol{\Phi}, \boldsymbol{\Phi}_{\prime}]$. Then we can apply the following kernel trick (Welling, 2013) to Equation (11):

$$(\mathbf{P}^{-1} + \mathbf{m}^T\mathbf{R}^{-1}\mathbf{m})^{-1}\mathbf{m}^T\mathbf{R}^{-1} = \mathbf{P}\mathbf{m}^T(\mathbf{m}\mathbf{P}\mathbf{m}^T + \mathbf{R})^{-1},$$

which gives:

$$(\mathbf{w}_m^*)^T = \mathbf{s}_m(\hat{\mathbf{\Phi}}^T\hat{\mathbf{\Phi}} + \mathbf{D}^{-1}/\mu)^{-1}\hat{\mathbf{\Phi}}^T. \tag{12}$$

In Equation (12), we avoid to compute the feature vectors $\mathbf{\Phi}$ in the possibly large dimensionality d_z. Instead we need to compute the inner product of feature vectors $\hat{\mathbf{\Phi}}^T\hat{\mathbf{\Phi}} \in \Re^{(N_t+N_t')\times(N_t+N_t')}$ which is usually more efficient than directly computing $\mathbf{\Phi}\mathbf{\Phi}^T \in \Re^{d_z\times d_z}$.

The algorithm to solve the proposed objective in Equation (4) is summarized in Algorithm 1.

3.4 Avoiding calculations of the least-squares problems

As can be seen in Equation (8), the update for $\mathbf{w}_m$ is reliant on solving a least squares problem in every iteration. However, the least squares solver has complexity $O(Nd^2)$ and will have to be solved every iteration which may not be computationally feasible if the number of features d is very large. To avoid this problem we can instead utilize an optimal line search method (Nie *et al.*, 2014) and update $\mathbf{w}_m$ via gradient descent:

$$\mathbf{w}_m = \mathbf{w}_m - s_m\nabla_{\mathbf{w}_m}, \tag{13}$$

where $\nabla_{\mathbf{w}_m}$ is the analytical gradient of Equation (4) with respect to $\mathbf{w}_m$:

$$\begin{aligned}\nabla_{\mathbf{w}_m} = \quad & \mathbf{w}_m - \mu\mathbf{X}_i\sum_{i=1}^{N}[\mathbf{t}_i^m - \mathbf{w}_m^T\mathbf{X}_i - 1b_m + \boldsymbol{\theta}_i^m/\mu]^T \\ & -\mu\mathbf{X}_{i'}\sum_{i'=1}^{N'}\sum_{m'=1}^{K}[\mathbf{u}_{i'}^{m'} - \mathbf{w}_m^T\mathbf{X}_{i'} - 1b_m + \boldsymbol{\xi}_{i'}^{m'}/\mu]^T,\end{aligned} \tag{14}$$

and it can be used to define a minimization:

$$\begin{aligned}s_m^* = \ & \text{argmin}_{s_m}\frac{1}{2}||\mathbf{w}_m^T - s_m\nabla_{\mathbf{w}_m}^T||_2^2 \\ & +\frac{\mu}{2}\sum_{i=1}^{N}\left[||\mathbf{t}_i^m - (\mathbf{w}_m^T - s_m\nabla_{\mathbf{w}_m}^T)\mathbf{X}_i - 1b_m + \theta_i^m/\mu||_2^2\right] \\ & +\sum_{i'=1}^{N'}\sum_{m=1}^{K}\left[\frac{\mu}{2}||\mathbf{u}_{i'}^m - (\mathbf{w}_m^T - s_m\nabla_{\mathbf{w}_m}^T)\mathbf{X}_{i'} - 1b_m + \xi_{i'}^m/\mu||_2^2\right],\end{aligned} \tag{15}$$

in terms of s_m instead of $\mathbf{w}_m$. Differentiating Equation (15) with respect to s_m, setting the result equal to zero gives:

$$s_m^* = \frac{\left(\mathbf{w}_m^T - \mu\sum_{i=1}^{N}\hat{\mathbf{t}}_i^m\mathbf{X}_i^T - \mu\sum_{i'=1}^{N'}\sum_{m'}^{K}\hat{\mathbf{u}}_{i'}^{m'}\mathbf{X}_{i'}^T\right)\nabla_{\mathbf{w}_m}}{\nabla_{\mathbf{w}_m}^T\left(\mathbf{I} + \mu\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i^T + \mu K\sum_{i'=1}^{N'}\mathbf{X}_{i'}\mathbf{X}_{i'}^T\right)\nabla_{\mathbf{w}_m}} \tag{16}$$

where $\hat{\mathbf{t}}_i^m = \mathbf{t}_i^m - \mathbf{w}_m^T\mathbf{X}_i - 1b_m + \boldsymbol{\theta}_i^m/\mu$ and $\hat{\mathbf{u}}_{i'}^{m'} = \mathbf{u}_{i'}^{m'} - \mathbf{w}_m^T\mathbf{X}_{i'} - 1b_m + \xi_{i'}^{m'}/\mu$. Finally we plug Equations (14) and (16) into Equation (13) to earn an efficient update equation which avoids the least squares problem in Equation (8). The time complexity of the proposed method is $O(Nd\bar{n})$, where $\bar{n}$ is the average number of instances per bag. The number of instances $\bar{n}$ is typically smaller than the number of features d (the multiple of 162 in our experiments), therefore our model with the solution in Equation (13) (inexact *pdMISVM*) is more scalable compared to Equation (8) (exact *pdMISVM*).

Algorithm 1 The multiblock ADMM updates to optimize Equation (4).

Data: $\mathbf{X} \in \mathbb{R}^{D\times(n_1+\cdots+n_N)}$ and $\mathbf{Y} \in \{-1,1\}^{K\times N}$.
Hyperparameters: $C>0$, $\mu>0$, $\rho>1$ and *tolerance* >0.
Initialize: primal variables $\mathbf{W},\mathbf{b},\mathbf{E},\mathbf{Q},\mathbf{R},\mathbf{T},\mathbf{U}$ and dual variables $\mathbf{\Lambda},\mathbf{\Sigma},\mathbf{\Theta},\mathbf{\Omega},\mathbf{\Xi}$.
while residual $>$ *tolerance* **do**
  for $m \in K$ **do**
    Update $\mathbf{w}_m \in \mathbf{W}$ **by** *Eq.* (13).
    Update $b_m \in \mathbf{b}$ **by** $b_m =$

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

  end for
  for $(p,m) \in \{N,K\}$ **do**

$$\text{Update } e_p^m \in \mathbf{E} \text{ by } e_i^m = \begin{cases} n_i^m - \frac{C}{\mu}y_i^m & \text{when} \quad y_i^m n_i^m > \frac{C}{\mu}, \\ 0 & \text{when} \quad 0 \le y_i^m n_i^m \le \frac{C}{\mu}, \\ n_i^m & \text{when} \quad y_i^m n_i^m < 0, \end{cases}$$

    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**

$$q_i^m = \frac{(y_i^m - e_i^m + r_i^m - \lambda_i^m/\mu + \max(\mathbf{t}_i^m) - \sigma_i^m/\mu)}{2}.$$

    Update $r_p^m \in \mathbf{R}$ **by**

$$r_i^m = \frac{(e_i^m - y_i^m + q_i^m + \lambda_i^m/\mu + \max(\mathbf{u}_i^m) - \omega_i^m/\mu)}{2}.$$

    for $j \in n_p$ **do**
      Update $t_p^m \in \mathbf{T}$ **by**

$$t_{i,j}^m = \begin{cases} \frac{\max(\phi_i^m) + q_i^m + \sigma_i^m/\mu}{2} & \text{if } j = \text{argmax}(\phi_i^m), \\ \phi_{i,j}^m & \text{else}, \end{cases}$$

      where $\phi_i^m = \mathbf{w}_m^T\mathbf{X}_i + 1b_m - \boldsymbol{\theta}_i^m/\mu$.
      Update $u_p^m \in \mathbf{U}$ **by**

$$u_{i,j}^m = \begin{cases} \frac{\max(\boldsymbol{\psi}_i^m) + r_i^m + \omega_i^m/\mu}{2} & \text{if } j = \text{argmax}(\boldsymbol{\psi}_i^m), \\ \psi_{i,j}^m & \text{else}, \end{cases}$$

      where $\boldsymbol{\psi}_i^m = \mathbf{w}_y^T\mathbf{X}_i + 1b_y - \boldsymbol{\xi}_i^m/\mu$.
    end for
    Update $\lambda_p^m, \sigma_p^m, \omega_p^m, \boldsymbol{\theta}_p^m, \boldsymbol{\xi}_p^m$ **by**

$$\begin{aligned}\lambda_i^m &= \lambda_i^m + \mu(e_i^m - (y_i^m - q_i^m + r_i^m)); \\ \sigma_i^m &= \sigma_i^m + \mu(q_i^m - \max(\mathbf{t}_i^m)); \\ \omega_i^m &= \omega_i^m + \mu(r_i^m - \max(\mathbf{u}_i^m)); \\ \boldsymbol{\theta}_i^m &= \boldsymbol{\theta}_i^m + \mu(\mathbf{t}_i^m - (\mathbf{w}_m^T\mathbf{X}_i + 1b_m)); \\ \boldsymbol{\xi}_i^m &= \boldsymbol{\xi}_i^m + \mu(\mathbf{u}_i^m - (\mathbf{w}_y^T\mathbf{X}_i + 1b_y)).\end{aligned}$$

  end for
  Update $\mu = \rho\mu$.
end while
return $(\mathbf{w}_m,\ldots,\mathbf{w}_K) \in \mathbf{W}$ and $(b_1,\ldots,b_K) \in \mathbf{b}$.

4 Results

In our experiments, we evaluate the classification performance and scalability of the proposed exact and inexact *pdMISVM* implementations. The scalability of *pdMISVM* is assessed across the increasing number of bags and features. Regarding the interpretability of our model, we also identify the disease relevant patches (instances) of each bag (image).

4.1 Benchmarks and hyperparameters

The classification performance and scalability of *pdMISVM* is compared against the following standard MIL benchmarks:

- A 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 bag-based methods; Normalized Set Kernel (NSK) and Statistics Kernel (STK) (Gärtner *et al.*, 2002), which map the entire bag to a single-instance by a way of kernel function.
- An iterated discrimination Axis-Parallel Rectangles algorithm (APR) (Dietterich *et al.*, 1997): the APR is a MIL model which starts from a single positive instance and grows the APR by expanding it to cover the remaining positive instances.
- The two multi-instance deep learning methods: The mi-Net and MI-Net (Wang *et al.*, 2018) approach to the MIL problem in a way of instance space and embedded space (learning vectorial representation of the bag) paradigm respectively.
- The two attention mechanism-based MIL models: Ilse *et al.* (2018) (AMIL) calculate the parameterized attention (importance) score for each instance to generate the probability distribution of bag labels. Shi *et al.* (2020) (LAMIL) propose to learn the instance scores and predictions jointly by integrating the attention mechanism with the loss function.

For these SVM models, the regularization tradeoff is set to 1.0. For the exact and inexact *pdMISVM*, the regularization tradeoff C is set to $1e-3$ and $1e+4$ respectively, the tolerance is set to $1e-5$ for both, and μ is initialized with $1e-10$ and $1e-8$ respectively. We use the radial basis kernel function for all SVM models (except inexact *pdMISVM* which uses linear kernel). For the deep learning models (mi-Net, MI-Net, AMIL and LAMIL), we use the same hyperparameters as in their articles.

4.2 Classification performance

In this section, we evaluate the classification models to investigate whether our exact/inexact *pdMISVM* achieves the better or comparable performance to the best performing classical or recent models. In Table 1, we report the performance of our *pdMISVM* compared against the other MIL algorithms in the classification of benign/malignant bags. For each model, we provide the precision, recall, F1-score, accuracy and balanced accuracy (BACC) across the 10 6-fold cross-validation experiments (six repetitions per experiment).

From the results reported in Table 1, the proposed exact/inexact *pdMISVM* show promising performance across the various magnification levels. In particular, our exact *pdMISVM* outperforms the other models based on recall. A high recall rate is critical in the medical domain, as false negatives may result the serious consequences. This result shows the clinical utility of our model as it is crucial not to miss a malignant tumor in the diagnosis. When the SIL model is compared to the other MIL models, SIL performed the worst because it is difficult to accurately classify labels from individual patches. For example, evidence of malignancy may appear only in some patches of the bag. In this case, it is difficult to classify a patch as a malignancy from a patch where no evidence of a malignant tumor appeared. Our experimental results support the assumption that MIL models will classify better than SIL model.

Interestingly, our exact *pdMISVM* performs better than the inexact version at the smaller magnification levels, and while the opposite results are observed in the larger magnification levels. These results show that the classification pattern of *pdMISVM* can vary depending on the choice of optimization approach, just like the impact of the optimization algorithm on the deep learning models (Wang *et al.*, 2019). Although our derivation for inexact *pdMISVM* does not obtain the exact optimal solution of the MISVM objective in Equation (1), our experimental results show that the inexact solution may improve the classification performance when compared to the exact solution. This is well supported by the previous finding (Chang *et al.*, 2008) that some implementations of SVM achieve the highest accuracy before the objective reaches its minimum. Our exact/inexact *pdMISVM* has gained the overall improved accuracy/BACC as well, and this validates their usefulness in the field of MIL and the early detection of a malignant tumor.

4.3 The scalability against bags and features

The main contribution of this study is that the derived Algorithm 1 scales to the large dataset. In this timing experiment, our goal is to verify the analytical complexity calculated in Section 3.4 on the real world dataset. We plot the training time of the classifiers on the BreaKHis dataset to verify this improved scalability against the number of bags in Figure 3 and the number of features in Figure 4. In this timing experiment, we use the linear kernel function for all SVM models. The deep learning models are excluded in this experiment as their training times exceed the reasonable limit (5 h). In Figure 3, the running time of NSK increases rapidly while the other models maintain the linear trend. Our *pdMISVM* outperforms the other models in training a large number of bags. This result validates the superior scalability of the proposed primal-dual approach over the other SVM models which rely on repeatedly solving a quadratic programming problem.

Despite the fact that the initial derivation with Equation (8) scales well with respect to the bags, the update for $\mathbf{w}_k$ in Equation (8) requires solving a least-squares problem that scales quadratically as the number of features d increases. To tackle this difficulty, we adapt an optimal line search method in Equation (13) to achieve the linear complexity against the number of features. In Figure 4, we compare the training time of the exact/inexact versions of our models to the other competing models. Among all models, Ours-inexact and APR spend the smallest training time when trained with the large number of features. Interestingly, inexact variation of *pdMISVM* scales significantly better than the exact *pdMISVM* against the increasing number of features where the number of bags is fixed at 1000. This is well represented by the analytical complexity of the two derivations ($O(Nd^2)$ versus $O(Nd\bar{n})$) as discussed in Section 3.4.

4.4 Patch identification

Along with the improved prediction performance and scalability, our model *pdMISVM* can identify disease-relevant locations. The interpretability is crucial as it can add confidence to the generated predictions and help clinicians use histopathological references to make a diagnosis. We calculate the patch-wise importance $\max(\mathbf{W}^T\mathbf{x}_i^j + \mathbf{b})$ which is the response of jth patch to the decision function in Equation (2). Figures 5 and 6 show the identified patches in the benign and malignant images at 400× magnification level. In Figures 5 and 6, the 10 boxes (patches) represent the 10 instances of each bag (image).

The patches identified by our model are in accordance with the clinical insights. The color, shape and size morphologic abnormalities of the nuclei are regarded as the key characteristics that categorize a digitized biopsy as cancerous or non-cancerous (Rajbongshi *et al.*, 2018). For example, in the third image in Figure 5, our model highlights the regions containing the cell's nuclei. From the identified patches, our model can reveal that the nuclear to cell volume ratio is consistent throughout which is a distinctive feature of non-carcinoma (Jevtić and Levy, 2014). Because of this, our model correctly classifies the bag as benign. A previous study (Kumar *et al.*, 2015) explains that a disorganized arrangement of cells is one of the characteristics of cancerous cells. In the second image in Figure 5, our model identifies a continuous, organized distribution of cells so this is another indication that our model was correct in labeling this image as benign. For the three malignant samples in Figure 6, our model focuses on the variation in the size and shape of nuclei. Based on the literature (Fischer, 2020), the loss of normal morphology and large/varying shape of nuclei are essential for the diagnosis of malignancy in the practice of surgical pathology. The accurately identified regions validate the correctness of our model in the histopathological image classification and add value to its clinical practicability.

5 Discussion

We demonstrated that the MIL SVM can detect the malignancy in the patches. With the development of image acquisition technology, and it has become crucial to train the models with the large amount of images to improve classification performance. Accordingly, scalability has emerged as a major issue, and the improved scalability can increase the performance and decrease the cost in response to

Table 1. The classification performance of our *pdMISVM* and competing models with the different magnification levels

Model	Magnification	Precision	Recall	F1Score	Accuracy	BACC
SIL	40×	0.874 ± 0.018	0.752 ± 0.043	0.829 ± 0.030	0.787 ± 0.036	0.808 ± 0.032
NSK	40×	0.906 ± 0.016	0.902 ± 0.025	0.904 ± 0.014	0.868 ± 0.019	0.848 ± 0.022
STK	40×	**0.911 ± 0.031**	0.907 ± 0.036	0.908 ± 0.017	0.875 ± 0.021	0.857 ± 0.024
APR	40×	0.881 ± 0.019	0.813 ± 0.034	0.856 ± 0.036	0.793 ± 0.024	0.817 ± 0.035
mi-Net	40×	0.883 ± 0.016	0.872 ± 0.037	0.883 ± 0.025	0.836 ± 0.032	0.850 ± 0.027
Mi-Net	40×	0.891 ± 0.070	0.884 ± 0.045	0.887 ± 0.023	0.852 ± 0.029	0.842 ± 0.030
AMIL	40×	0.901 ± 0.025	0.897 ± 0.019	0.899 ± 0.031	0.870 ± 0.031	0.861 ± 0.024
LAMIL	40×	0.896 ± 0.041	0.900 ± 0.023	0.894 ± 0.047	0.863 ± 0.024	0.859 ± 0.027
Ours	40×	0.894 ± 0.018	**0.924 ± 0.036**	0.903 ± 0.023	0.853 ± 0.028	0.842 ± 0.034
Ours (inexact)	40×	0.902 ± 0.014	0.916 ± 0.024	**0.916 ± 0.009**	**0.879 ± 0.009**	**0.863 ± 0.023**
SIL	100×	0.908 ± 0.014	0.797 ± 0.034	0.848 ± 0.023	0.804 ± 0.027	0.808 ± 0.025
NSK	100×	0.918 ± 0.019	0.926 ± 0.008	0.922 ± 0.011	**0.892 ± 0.013**	0.872 ± 0.017
STK	100×	0.895 ± 0.024	0.929 ± 0.023	0.911 ± 0.010	0.876 ± 0.012	0.844 ± 0.012
APR	100×	0.896 ± 0.025	0.854 ± 0.039	0.879 ± 0.032	0.818 ± 0.031	0.812 ± 0.034
mi-Net	100×	0.860 ± 0.019	0.917 ± 0.025	0.889 ± 0.019	0.879 ± 0.032	0.862 ± 0.023
Mi-Net	100×	0.876 ± 0.026	0.928 ± 0.029	0.891 ± 0.025	0.869 ± 0.027	0.870 ± 0.019
AMIL	100×	0.889 ± 0.027	0.935 ± 0.038	0.914 ± 0.039	0.867 ± 0.029	0.869 ± 0.029
LAMIL	100×	0.898 ± 0.046	0.924 ± 0.035	0.910 ± 0.041	0.870 ± 0.034	0.859 ± 0.032
Ours	100×	0.873 ± 0.016	**0.944 ± 0.019**	0.919 ± 0.009	0.883 ± 0.017	0.864 ± 0.026
Ours (inexact)	100×	**0.923 ± 0.025**	0.942 ± 0.022	**0.925 ± 0.020**	0.891 ± 0.024	**0.876 ± 0.041**
SIL	200×	0.903 ± 0.015	0.812 ± 0.018	0.863 ± 0.007	0.821 ± 0.013	0.826 ± 0.016
NSK	200×	0.902 ± 0.020	0.935 ± 0.022	0.923 ± 0.016	0.893 ± 0.022	0.867 ± 0.025
STK	200×	0.898 ± 0.025	0.927 ± 0.020	0.922 ± 0.011	0.892 ± 0.017	0.871 ± 0.025
APR	200×	0.889 ± 0.021	0.897 ± 0.024	0.895 ± 0.013	0.857 ± 0.018	0.864 ± 0.027
mi-Net	200×	0.879 ± 0.021	0.909 ± 0.032	0.891 ± 0.028	0.876 ± 0.021	0.846 ± 0.023
Mi-Net	200×	0.885 ± 0.020	0.918 ± 0.029	0.896 ± 0.025	0.885 ± 0.019	0.851 ± 0.024
AMIL	200×	**0.905 ± 0.024**	0.918 ± 0.031	0.900 ± 0.027	0.881 ± 0.024	0.849 ± 0.021
LAMIL	200×	0.891 ± 0.031	0.914 ± 0.037	0.907 ± 0.030	0.875 ± 0.024	0.853 ± 0.028
Ours	200×	0.903 ± 0.017	**0.936 ± 0.023**	**0.924 ± 0.012**	**0.898 ± 0.017**	**0.872 ± 0.026**
Ours (inexact)	200×	0.890 ± 0.019	0.931 ± 0.017	0.918 ± 0.012	0.889 ± 0.021	0.859 ± 0.020
SIL	400×	0.860 ± 0.021	0.744 ± 0.042	0.819 ± 0.027	0.778 ± 0.032	0.797 ± 0.029
NSK	400×	0.890 ± 0.022	0.910 ± 0.013	0.886 ± 0.009	0.863 ± 0.014	0.836 ± 0.025
STK	400×	0.889 ± 0.024	0.898 ± 0.029	0.893 ± 0.014	0.854 ± 0.020	0.832 ± 0.023
APR	400×	0.891 ± 0.025	0.803 ± 0.038	0.871 ± 0.033	0.816 ± 0.028	0.819 ± 0.036
mi-Net	400×	0.837 ± 0.021	0.901 ± 0.025	0.864 ± 0.029	0.831 ± 0.031	0.822 ± 0.024
Mi-Net	400×	0.849 ± 0.020	0.895 ± 0.026	0.871 ± 0.024	0.841 ± 0.028	0.820 ± 0.025
AMIL	400×	0.852 ± 0.019	0.897 ± 0.022	0.880 ± 0.023	0.846 ± 0.024	0.818 ± 0.027
LAMIL	400×	0.867 ± 0.025	0.889 ± 0.029	0.891 ± 0.031	0.857 ± 0.023	0.819 ± 0.029
Ours	400×	**0.909 ± 0.012**	**0.932 ± 0.016**	**0.899 ± 0.012**	**0.868 ± 0.016**	**0.838 ± 0.018**
Ours (inexact)	400×	0.875 ± 0.023	0.923 ± 0.019	0.898 ± 0.015	0.858 ± 0.019	0.823 ± 0.020

Note: The reported metrics and their standard deviations are calculated across 10 6-fold cross-validation experiments. The best scores are highlighted in bold font.

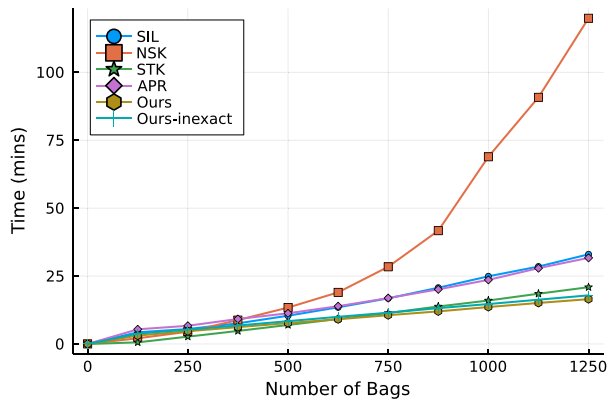

Fig. 3. Computation time over the increasing number of bags. The number of features is fixed at 162

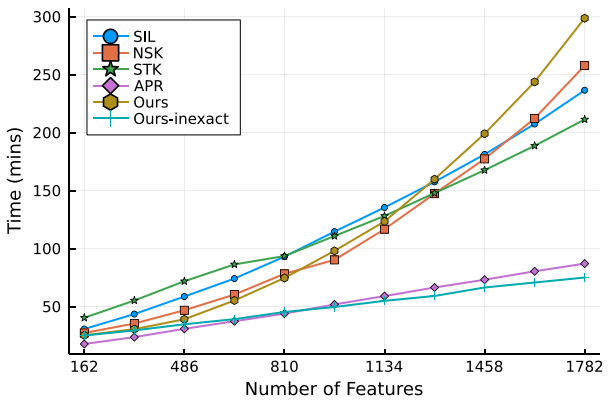

Fig. 4. Computation time over the increasing number of features. The number of features are controlled by concatenating the multiple patches of image

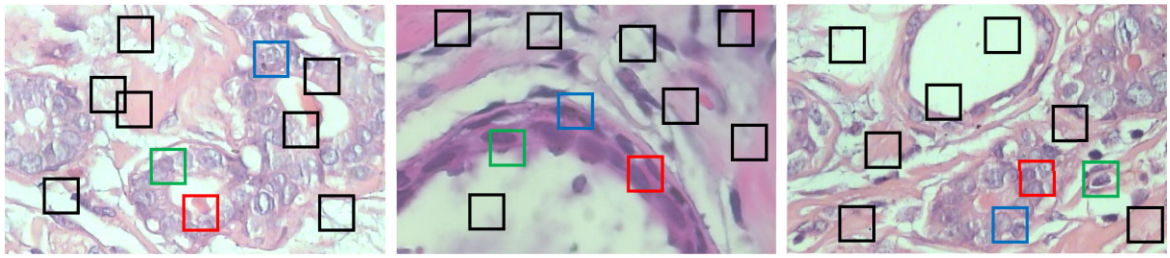

Fig. 5. The identified patches in the *adenosis (benign)* images, where the red, green and blue boxes denote the first, second and third most important patches

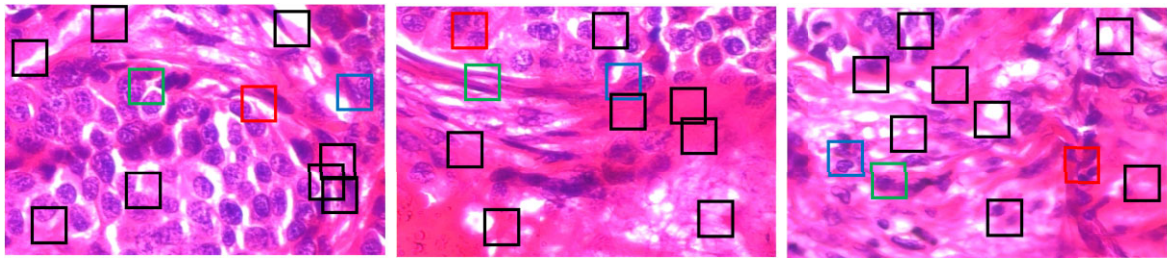

Fig. 6. The identified patches in the *ductal carcinoma (malignant)* images, where the red, green and blue boxes denote the first, second and third most important patches

the system processing demands of CAD. Therefore, this study proposes a new optimization method for SVM with improved scalability. The proposed method reduces the computational complexity against the large number of features of instances by approximating the optimal point of SVM, but nevertheless, the experimental results show that the classification performance of SVM is not sacrificed, and rather improved in certain cases (in the lower magnification levels). In addition, the permutation invariant property is satisfied in Equation (1), which is desirable in the MIL. The proposed optimization method can be applied regardless of whether the kernel function is used, however we plan to deal with the improved scalability of kernelized SVM in the future study. In this study, we have sampled the patches of WSIs at the random locations, and we plan to integrate the attention mechanism to automatically sample the patches important for malignant tumor detection. In this study, we propose a general framework for MIL and the other models stemming from our approach can be flexibly applied to solve the various MIL problems.

6 Conclusion

The improvement of the scalability of methods is attracting more attention from machine learning studies as the amount of available data is increasing due to the development of data mining technologies. In this work, we present a novel *Primal-Dual Multi-Instance SVM* method and the associated derivations, which scale to a large number of bags and features. We have conducted extensive experiments on the BreaKHis dataset to show the promising performance and scalability of the proposed method when compared to the traditional SVM-based MIL techniques. In addition to the improved classification performance and scalability, the key patches for the classification identified by our model are well supported by previous medical studies. The experimental results illustrate the clinical utility of our approach on the detection of cancerous abnormalities in a large dataset to prevent the progression of breast cancer in a patient.

Funding

This work was supported in part by the National Science Foundation (NSF) under the grants of Information and Intelligent Systems (IIS) [1652943, 1849359], Computer and Network Systems (CNS) [1932482].

Conflict of Interest: none declared.

References

Andrews,S. *et al.* (2002) Support vector machines for multiple-instance learning. In: *Advances in neural information processing systems (NIPS)*, **Vol. 2**. Citeseer, pp. 561–568.

Brand,L. *et al.* (2021a) A linear primal-dual multi-instance svm for big data classifications. In: *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, pp. 21–30.

Brand,L. *et al.* (2021b) A multi-instance support vector machine with incomplete data for clinical outcome prediction of covid-19. In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–6.

Bunescu,R.C. and Mooney,R.J. (2007) Multiple instance learning for sparse positive bags. In *Proceedings of the 24th international conference on Machine learning*, pp. 105–112.

CDC. (2020) *The Basics on Hereditary Breast and Ovarian Cancer—CDC*. https://www.cdc.gov/genomics/disease/breast_ovarian_cancer/basics_hboc.htm (1 August 2021, date last accessed).

Chang,K.-W. *et al.* (2008) Coordinate descent method for large-scale l2-loss linear support vector machines. *J. Mach. Learn. Res.*, **9**, 1369–1398.

Dietterich,T.G. *et al.* (1997) Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, **89**, 31–71.

Fischer,E.G. (2020) Nuclear morphology and the biology of cancer cells. *Acta Cytol.*, **64**, 511–519.

Gärtner,T. *et al.* (2002) Multi-instance kernels. In: *International Conference on Machine Learning (ICML)*, Vol. 2, p. 7.

Guo,Z. *et al.* (2010) A completed modeling of local binary pattern operator for texture classification. *IEEE Trans. Image Process.*, **19**, 1657–1663.

Gurcan,M.N. *et al.* (2009) Histopathological image analysis: a review. *IEEE Rev. Biomed. Eng.*, **2**, 147–171.

Hamilton,N.A. *et al.* (2007) Fast automated cell phenotype image classification. *BMC Bioinformatics*, **8**, 110–118.

Haralick,R.M. (1979) Statistical and structural approaches to texture. *Proc. IEEE*, **67**, 786–804.

Haralick,R.M. *et al.* (1973) Textural features for image classification. *IEEE Trans. Syst, Man, Cybern.*, **SMC-3**, 610–621.

Hong,M. and Luo,Z.-Q. (2017) On the linear convergence of the alternating direction method of multipliers. *Math. Program.*, **162**, 165–199.

Ilse,M. *et al.* (2018) Attention-based deep multiple instance learning. In: *International Conference on Machine Learning*. PMLR, pp. 2127–2136.

Jevtić,P. and Levy,D.L. (2014) Mechanisms of nuclear size regulation in model systems and cancer. *Cancer Biol. Nuclear Envelope*, **773**, 537–569.

Kahya,M.A. *et al.* (2017) Classification of breast cancer histopathology images based on adaptive sparse support vector machine. *J. Appl. Math. Bioinf.*, **7**, 49.

Khotanzad,A. and Hong,Y.H. (1990) Invariant image recognition by Zernike moments. *IEEE Trans. Pattern Anal. Machine Intell.*, **12**, 489–497.

Krizhevsky,A. *et al.* (2012) Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, **25**, 1097–1105.

Kumar,M. and Rath,S.K. (2015) Classification of microarray using mapreduce based proximal support vector machine classifier. *Knowledge Based Syst.*, **89**, 584–602.

Kumar,R. *et al.* (2015) Detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features. *J. Med. Eng.*, **2015**, 457906.

Nie,F. *et al.* (2014) 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*, Vol. 32, pp. II–505.

Ojala,T. *et al.* (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Machine Intell.*, **24**, 971–987.

Ojansivu,V. and Heikkilä,J. (2008) Blur insensitive texture classification using local phase quantization. In: *International Conference on Image and Signal Processing*. Springer, pp. 236–243.

Otsu,N. (1979) A threshold selection method from gray-level histograms. *IEEE Trans. Syst, Man Cybern.*, **9**, 62–66.

Peng,X. *et al.* (2016) L1-norm loss based twin support vector machine for data recognition. *Inf. Sci.*, **340-341**, 86–103.

Rajbongshi,N. *et al.* (2018) Analysis of morphological features of benign and malignant breast cell extracted from FNAC microscopic image using the Pearsonian system of curves. *J. Cytol.*, **35**, 99–104.

Shawe-Taylor,J. *et al.* (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY.

Shi,X. *et al.* (2020) Loss-based attention for deep multiple instance learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 5742–5749.

Spanhol,F.A. *et al.* (2015) A dataset for breast cancer histopathological image classification. *IEEE Trans. Biomed. Eng.*, **63**, 1455–1462.

Spanhol,F.A. *et al.* (2016) Breast cancer histopathological image classification using convolutional neural networks. In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 2560–2567.

Sudharshan,P. *et al.* (2019) Multiple instance learning for histopathological breast cancer image classification. *Expert Syst. Appl.*, **117**, 103–111.

Titoriya,A. and Sachdeva,S. (2019) Breast cancer histopathology image classification using Alexnet. In: *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*. IEEE, pp. 708–712.

van der Laak,J. *et al.* (2021) Deep learning in histopathology: the path to the clinic. *Nat. Med.*, **27**, 775–784.

Wang,J. and Zhao,L. (2021) Nonconvex generalization of Alternating Direction Method of Multipliers for nonlinear equality constrained problems. Results in Con. Optim., 2, 100009. https://doi.org/10.1016/j.rico.2021.100009.

Wang,H. *et al.* (2011) Learning instance specific distance for multi-instance classification. In: *Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, CA, USA*. Association for the Advancement of Artificial Intelligence (AAAI), Palo Alto, California, USA. pp. 507–512.

Wang,X. *et al.* (2018) Revisiting multiple instance neural networks. *Pattern Recognit.*, **74**, 15–24.

Wang,Y. *et al.* (2019) Assessing optimizer impact on DNN model sensitivity to adversarial examples. *IEEE Access*, **7**, 152766–152776.

Wei,X.-S. *et al.* (2014) Scalable multi-instance learning. In: *2014 IEEE International Conference on Data Mining*. IEEE, pp. 1037–1042.

Welling,M. (2013) Kernel ridge regression. *Max Welling's Classnotes Mach. Learn.*, 1–3.

Zheng,B. *et al.* (2014) Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Exp. Syst. Appl.*, **41**, 1476–1482.