Fast Multi-Modal Multi-Instance Support Vector Machine for Fine-grained Chest X-ray Recognition

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

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

2023 IEEE International Conference on Data Mining (ICDM) | 979-8-3503-0788-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDM58522.2023.00164

***Abstract*—Chest X-ray (CXR) analysis plays an important role in patient treatment. As such, a multitude of machine learning models have been applied to CXR datasets attempting automated analysis. However, each patient has a differing number of images per angle, and multi-modal learning should deal with the missing data for specific angles and times. Furthermore, the large dimensionality of multi-modal imaging data with the shapes inconsistent across the dataset introduces the challenges in training. In light of these issues, we propose the Fast Multi-Modal Support Vector Machine (FMMSVM) which incorporates modality-specific factorization to deal with missing CXRs in the specific angle. Our model is able to adjust the fine-grained details in feature extraction and we provide an efficient optimization algorithm scalable to a large number of features. In our experiments, FMMSVM shows clearly improved classification performance.**

***Index Terms*—Scalability, Multi-Instance, Multi-Modal, Support Vector Machine**

I. Introduction

Chest X-ray (CRX) is a vital tool for quick patient triage and as such, there have been great efforts to make computer analysis of X-ray images possible. As imaging technology advances, the number of images per patient continues to grow. Reliable and fast automated analysis can alleviate the workloads for practitioners by offloading some of the work to a computer. However, there are three key challenges in automating the analysis of CRX images.

First, medical images come in various modalities including computed tomography (CT) and traditional X-ray. CXRs may be captured at different angles determined by the patient's status and the physician. Consequently, the analysis model needs to detect patterns varying across these different modalities and learn the relationships between them. The multi-image nature of the data can result in a large number of features that requires significant computational resources to process. Second, multiple images can be captured across different points in time. At the same time, some images at specific times are captured while others are not. For example, X-rays are more accessible and cost-effective than CT which can lead to some angles having X-rays but no CT. Finally, the images can be collected from different devices and hospitals, resulting in discrepancies in the image format and resolution. Existing machine learning models assume a fixed-size image, so they rely on rescaling methods which may incur a loss of information (from down-scaling) or undesirable bias (from up-scaling). Lost information resulting from rescaling negatively impacts a model's performance.

The previous research has framed image analysis as a multi-instance learning problem (MIL) for two reasons: the number of images per patient differ across the dataset and individual images may not be labeled. MIL [1], [2], [3] is a weakly-supervised learning model, which is ideal for this application as each patient is in the form of a labeled "bag". Labels are associated with the bag, not the individual images, so the clinician does not need to label each image individually. There have been extensive studies into machine learning algorithms for MIL including support vector machines (SVMs) and deep learning models. Some examples of SVMs are Multi-Instance Support Vector Machine (MISVM) [4], sparse Multi-Instance Learning (sMIL), sparse balanced MIL (sbMIL) [5], Normalized Set Kernel (NSK), and Statistics Kernel (STK) [6]. These methods have successfully labeled the bags in the testing dataset as either malignant or benign. The multi-instance deep learning models include mi-Net and MI-Net [7], and more recently the attention mechanism-based models such as Attention-based deep Multiple Instance Learning (AMIL) [8] and Loss-based Attention Multiple Instance Learning (LAMIL) [9] are gaining popularity.

Although these models already exceed human performance in some applications, the success of those models depends on extensive training time and computational resources. Considering these difficulties, based on our earlier works [10], [11], [12] in this paper we propose a Fast Multi-modal Multi-instance Support Vector Machine (FMMSVM) method to improve the performance and effectiveness of CXR analysis. Our contributions can be summarized into the following:

- The proposed model simultaneously imputes the missing modality and predicts the clinical outcomes in spite of missing data. This joint imputation is designed to estimate the values of missing entries most helpful for predicting diagnoses.
- We derive an efficient solution algorithm for the proposed FMMSVM which linearly scales to the number of features of input data reducing the need for training time and computing resources.

979-8-3503-0788-7/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDM58522.2023.00164

II. The Method

Throughout the remainder of this paper, we denote matrices with bold upper-case letters (*e.g.*, $\mathbf{M}$), vectors as bold lower-case letters (*e.g.*, $\mathbf{m}$), and scalars as lower-case letters (*e.g.*, m). The i-th row and j-th column of matrix $\mathbf{M}$ are written as $\mathbf{m}^i$ or $[\mathbf{M}]^i$ and $\mathbf{m}_j$ or $[\mathbf{M}]_j$. The scalar value indexed by the i-th row and j-th column of $\mathbf{M}$ are written as m_j^i or $[\mathbf{M}]_j^i$. Each i-th 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 $y_i \in \{1, \cdots, m, \cdots, K\}$. We denote the trace norm of a matrix as $\mathrm{tr}[\cdot]$.

A. Our Objective

We start our formulation with the K-class multi-instance SVM [4]:

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

Here N denotes the total number of bags, representing patients, the hyperplane $\mathbf{w}_y$ and bias b_y are associated with the positive class label for i-th bag $\mathbf{X}_i$. However, the conventional multi-instance SVM in Eq. (1) is not able to reach ideal performance with the missing data. To overcome this limitation, we are motivated to develop *FMMSVM* to jointly perform the clinical outcome prediction and imputation as:

$$\begin{aligned}
&\mathcal{L}_{joint} = \mathcal{L}_{imputation} + \mathcal{L}_{classification} \\
&= \min_{\mathbf{W},\mathbf{b},\mathbf{Z},\mathbf{H},\mathbf{F}} \\
&\tau_0\sum_{i=0}^{N}\sum_{g=1}^{G}(\alpha_g\|\mathbf{F}_i^g - \mathbf{H}^g\mathbf{Z}_i\|_F^2) + \tau_1\sum_{i=1}^{N}\sum_{j,k=1}^{n_i}\frac{1}{d_i^{j,k}}\|\mathbf{z}_i^j - \mathbf{z}_i^k\|_2^2 \\
&+ \tau_2\|\mathbf{F}\|_* + \frac{1}{2}\sum_{m=1}^{K}\|\mathbf{w}_m\|_2^2 + \tau_3\sum_{i=1}^{N}\sum_{m=1}^{K} \\
&(1 - [\max_i(\mathbf{w}_m^T\mathbf{Z}_i + \mathbf{1}\mathbf{b}_m) - \max_i(\mathbf{w}_y^T\mathbf{Z}_i + \mathbf{1}\mathbf{b}_y)]y_i^m)_+ \\
&\textit{s.t.}\quad \mathbf{F}_i \odot \mathbf{M}_i = \mathbf{X}_i \odot \mathbf{M}_i,\ \mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_N], \\
&\mathbf{Z}_i\mathbf{Z}_i^T = \mathbf{I},\ \mathbf{H} = [\mathbf{H}^1, \mathbf{H}^2, \cdots, \mathbf{H}^G],
\end{aligned} \tag{2}$$

where $d_i^{j,k}$ is time interval (i.e., temporal distance) between j-th and k-th instance. The imputation loss $\mathcal{L}_{imputation}$ contains factorization [13], [14], locality preserving projection [15], and trace norm [16] terms. The mask $\mathbf{M}_i \in \Re^{d \times n_i}$ contains the binary missingness information of data $\mathbf{X}_i$, where 1 indicates the known entry and 0 indicates the unknown/missing entry. Additionally, $\odot$ is Hadamard product. $\mathbf{F} \in \Re^{d \times n_i}$ is the imputed matrix of $\mathbf{X}_i$ which keeps the known entries in $\mathbf{X}_i$. The trace norm is defined as $\|\mathbf{F}\|_* = \sum_{j=1}^{min\{d,n_i\}}\sigma_j = \mathrm{tr}[(\mathbf{F}^T\mathbf{F})^{\frac{1}{2}}]$, which improves the smoothness between the imputed and known entries.

Both factorization and trace norm terms discover low rank structure of input data $\mathbf{X}_i$ and unknown entries are extrapolated using with linear combination of known entries. This is accomplished by minimizing factorization term $\tau_0\sum_{i=0}^{N}(\|\mathbf{F}_i - \mathbf{H}\mathbf{Z}_i\|_{2,1})$, where the known entry in $\mathbf{X}_i$ is expressed by the product between row of $\mathbf{H} \in \Re^{d \times r}$ and column of $\mathbf{Z}_i \in \Re^{r \times n_i}$, and unknown entries are imputed by $\mathbf{H}$ and $\mathbf{Z}_i$ learned from the known entries. The dimensionality r of $\mathbf{Z}_i$ is typically much less than d of $\mathbf{X}_i$. Therefore, $\mathbf{Z}_i$ represents the enriched version of $\mathbf{X}_i$ which removes the redundant information in $\mathbf{X}_i$, and we replace $\mathbf{X}_i$ in Eq. (1) with our learned representation $\mathbf{Z}_i$. As a result, the decision function is given as such $\tilde{y}_i = \arg\max_{m'}(\max(\mathbf{W}^T\mathbf{H}^+\mathbf{X}_i + \mathbf{b}\mathbf{1}_i)^{m'})$, where $\mathbf{H}^+$ is Moore-Penrose pseudo-inverse of $\mathbf{H}$.

From the imputation integrated SVM in Eq. (2), we should consider the two important aspects in the images learning. First, for each patient the multiple images are captured across the different time points. Therefore, disease patterns in the two consecutive images captured at the similar time points tends to be associated each other. Second, the medical images are provided in the multiple modalities (*e.g.,* X-ray or CT) and some modalities can be more predictive than the others. To account for these two factors, we introduce the graph learning to preserve the temporal locality where the inverse of temporal distances each pair of instances (nodes) are weights (similarities). We also factorize $\mathbf{F}$ modality by modality to learn $\mathbf{H}$ shared across all the bags and $\mathbf{Z}_i$ shared across all the modalities. The hyperparameters α_g adjust the importance of each modality.

B. Primal-dual Support Vector Machine with Smoothness

Although the factorization and regularizations we have introduced in Eq. (2) are highly motivated, it adds many terms and complexity to our objective. Following our previous studies [11], we split the primal variables in Eq. (2) via Alternating Direction Method of Multipliers (ADMM) [17] approach. Another difficulty in the derivation is that the regularization terms in Eq. (2) are non-smooth and the gradients may not exist at some points. To improve the stability of the optimization, we use the optimization framework of the earlier work [18] that propose the iterative reweighted method to minimize non-smooth objective in Eq. (2) in which the key step is minimizing the following smoothed objective:

$$\begin{aligned}
&\min_{\substack{\mathbf{W},\mathbf{b},\mathbf{F},\mathbf{H},\\ \mathbf{Z},\mathbf{B},\mathbf{E},\mathbf{Q},\\ \mathbf{R},\mathbf{T},\mathbf{U}}} \tau_0\sum_{i=0}^{N}\mathrm{tr}\left[(\mathbf{F}_i - \mathbf{H}\mathbf{Z}_i)^T\mathbf{D}_0(\mathbf{F}_i - \mathbf{H}\mathbf{B}_i)\right] \\
&+ \tau_1\sum_{i=1}^{N}\mathrm{tr}\left[\mathbf{Z}_i\mathbf{D}_{1,i}\mathbf{B}_i^T\right] + \tau_2\,\mathrm{tr}\left[\mathbf{F}^T\mathbf{D}_2\mathbf{F}\right] \\
&+ \frac{1}{2}\sum_{m=1}^{K}\|\mathbf{w}_m\|_2^2 + \tau_3\sum_{i=1}^{N}\sum_{m=1}^{K}(y_i^m e_i^m)_+ \\
&\textit{s.t.}\quad \mathbf{F}_i \odot \mathbf{M}_i = \mathbf{X}_i \odot \mathbf{M}_i,\ \mathbf{B}_i\mathbf{Z}_i^T = \mathbf{I},\ \mathbf{B}_i = \mathbf{Z}_i, \\
&e_i^m = y_i^m - q_i^m + r_i^m, \\
&r_i^m = \max(\mathbf{u}_i^m),\ q_i^m = \max(\mathbf{t}_i^m), \\
&\mathbf{t}_i^m = \mathbf{w}_m^T\mathbf{X}_i + \mathbf{1}b_m, \mathbf{u}_i^m = \mathbf{w}_y^T\mathbf{X}_i + \mathbf{1}b_y,
\end{aligned} \tag{3}$$

where $\mathbf{D}_0 \in \Re^{d \times d}$ is a diagonal matrix whose j-th diagonal element is α_g when j-th feature of instance is in g-th features

group (modality). $\mathbf{D}_{1,i} = \tilde{\mathbf{S}}_i - \mathbf{S}_i \in \Re^{r\times r}$ where $[\mathbf{S}_i]_k^j = \frac{1}{2d_i^{j,k}}(\|\mathbf{z}_i^j - \mathbf{z}_i^k\|_2^2 + \delta)^{-\frac{1}{2}}$ and $\tilde{\mathbf{S}}$ is a diagonal matrix where each diagonal element is the row (or column) sum of $\mathbf{S}_i$ such that $[\tilde{\mathbf{S}}_i]_j^j = \sum_k [\mathbf{S}]_j^k$. $\mathbf{D}_2 = \frac{1}{2}(\mathbf{F}\mathbf{F}^T + \delta\mathbf{I})^{-\frac{1}{2}} \in \Re^{D\times D}$ and δ is a small constant value for smoothness.

From Eq. (3) we derive the following equation using the augmented Lagrangian method:

$$
\begin{aligned}
&\min_{\substack{\mathbf{W},\mathbf{b},\mathbf{F},\mathbf{H},\\ \mathbf{Z},\mathbf{B},\mathbf{E},\mathbf{Q},\\ \mathbf{R},\mathbf{T},\mathbf{U}}} \tau_0 \sum_{i=0}^{N} \operatorname{tr}\left[(\mathbf{F}_i - \mathbf{H}\mathbf{Z}_i)^T \mathbf{D}_0 (\mathbf{F}_i - \mathbf{H}\mathbf{B}_i)\right] \\
&+ \tau_1 \sum_{i=1}^{N} \operatorname{tr}\left[\mathbf{Z}_i \mathbf{D}_{1,i} \mathbf{B}_i^T\right] + \tau_2 \operatorname{tr}\left[\mathbf{F}^T \mathbf{D}_2 \mathbf{F}\right] + \frac{1}{2}\sum_{m=1}^{K} \|\mathbf{w}_m\|_2^2 \\
&+ \tau_3 \sum_{i=1}^{N}\sum_{m=1}^{K} (y_i^m e_i^m)_+ + \frac{\mu}{2}\sum_{i=1}^{N} \|(\mathbf{F}_i - \mathbf{X}_i - \frac{\mathbf{\Lambda}_{4,i}}{\mu}) \odot \mathbf{M}_i\|_{2,2}^2 \\
&+ \frac{\mu}{2}\sum_{i=1}^{N} \|\mathbf{B}_i\mathbf{Z}_i^T - \mathbf{I} - \frac{\mathbf{\Lambda}_{1,i}}{\mu}\|_{2,2}^2 + \frac{\mu}{2}\sum_{i=1}^{N} \|\mathbf{Z}_i - \mathbf{B}_i - \frac{\mathbf{\Lambda}_{3,i}}{\mu}\|_{2,}^2{}^2 \\
&+ \frac{\mu}{2}\sum_{i=1}^{N}\sum_{m=1}^{K} \Bigg[\left(e_i^m - (y_i^m - q_i^m + r_i^m - \lambda_{2,i}^m/\mu)\right)^2 \\
&+ (q_i^m - \max(\mathbf{t}_i^m) + \sigma_i^m/\mu)^2 + (r_i^m - \max(\mathbf{u}_i^m) + \omega_i^m/\mu)^2 \\
&+ \operatorname{tr}[(\mathbf{t}_i^m - \left(\mathbf{w}_m^T\mathbf{Z}_i + \mathbf{1}b_m\right) + \boldsymbol{\theta}_i^m/\mu)^T \\
&(\mathbf{t}_i^m - \left(\mathbf{w}_m^T\mathbf{B}_i + \mathbf{1}b_m\right) + \boldsymbol{\theta}_i^m/\mu)] \\
&+ \operatorname{tr}[(\mathbf{u}_i^m - \left(\mathbf{w}_y^T\mathbf{Z}_i + \mathbf{1}b_y\right) + \boldsymbol{\xi}_i^m/\mu)^T \\
&(\mathbf{u}_i^m - \left(\mathbf{w}_y^T\mathbf{B}_i + \mathbf{1}b_y\right) + \boldsymbol{\xi}_i^m/\mu)]\Bigg],
\end{aligned}
\tag{4}
$$

where $\mathbf{W}, \mathbf{b}, \mathbf{F}, \mathbf{H}, \mathbf{Z}, \mathbf{B}, \mathbf{E}, \mathbf{Q}, \mathbf{T}, \mathbf{R}, \mathbf{U}$ are the primal variables,
$\mathbf{\Lambda}_{1,i}, \mathbf{\Lambda}_2, \mathbf{\Lambda}_{3,i}, \mathbf{\Lambda}_{4,i}, \mathbf{\Sigma}, \mathbf{\Theta}, \mathbf{\Omega}, \mathbf{\Xi}, \mathbf{\Gamma}$ are the dual variables.

The algorithm to solve the above objective is provided in Algorithm 1. The detailed derivations of the algorithm is not provided here due to space limit and will be provided in the extended journal version of this paper.

III. Experiments

The chest X-rays are commonly utilized in medical research and clinical practice to detect abnormalities. In our experiments, we use an publicly available dataset of chest X-ray and CT images. Each patient (bag) is labeled by in-hospital mortality and associated with multiple images recorded across the different time points. For each time step we have two CXRs captured from the front and side as well as one CT image. The images are collected from the different public sources, hospitals, and physicians, which results in different shapes between images. The Fig. 1 and 2 shows the widths, heights, and ratios of images in the dataset, and we observe the high variance in the size and ratio distributions. As a result, rescaling the images with the interpolation method can significantly distort the objects in the images. Therefore, instead of rescaling, we divide each image into 3×4 patches. Each patch is then vectorized through Parameter Free Threshold statistics

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

Data: $\mathbf{X} \in \mathbb{R}^{d\times(n_1+\cdots+n_N)}$ and $\mathbf{Y} \in \{-1,1\}^{K\times N}$.
Hyperparameters: $C > 0$, $\mu > 0$, $\rho > 1$, $tolerance > 0$ and $\tau_0, \tau_1, \tau_2, \tau_3 \geq 0$.
Initialize: primal variables $\mathbf{W}, \mathbf{b}, \mathbf{F}, \mathbf{H}, \mathbf{Z}, \mathbf{B}, \mathbf{E}, \mathbf{Q}, \mathbf{R}, \mathbf{T}, \mathbf{U}$ and dual variables $\mathbf{\Lambda}_1, \mathbf{\Lambda}_2, \mathbf{\Lambda}_3, \mathbf{\Lambda}_4, \mathbf{\Sigma}, \mathbf{\Theta}, \mathbf{\Omega}, \mathbf{\Xi}, \mathbf{\Gamma}$.
while residual $> tolerance$ **do**
  Update $\mathbf{D}_{1,i}$ $(i \in \{1,\cdots,N\})$, $\mathbf{D}_2$ **by** Eq. (3).
  for $m \in K$ **do**
    Update $\mathbf{w}_m \in \mathbf{W}$ **by** Eq. (**??**).
    Update $b_m \in \mathbf{b}$ **by** $b_m = \frac{\sum_{i=1}^{N}[\mathbf{t}_i^m - \mathbf{w}_m^T\mathbf{Z}_i + \boldsymbol{\theta}_i^m/\mu] + \sum_{i'=1}^{N'}\sum_{m=1}^{K}\left[\mathbf{u}_{i'}^m - \mathbf{w}_m^T\mathbf{Z}_{i'} + \boldsymbol{\xi}_{i'}^m/\mu\right]}{N + KN'}$.
  end for
  for $(i,m) \in \{N, K\}$ **do**
    Update $e_i^m \in \mathbf{E}$ **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 $q_i^m \in \mathbf{Q}$ **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 $r_i^m \in \mathbf{R}$ **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 $j \in n_i$ **do**
      Update $t_{i,j}^m \in \mathbf{T}$ **by**
      $t_{i,j}^m = \begin{cases} (1/2)\cdot(q_i^m + \sigma_i^m/\mu) + (1/4)\cdot(\hat{\mathbf{z}}_i^m + \hat{\mathbf{b}}_i^m) \\ \text{when } j = \arg\max(\mathbf{t}_i^m); \\ (1/2)\cdot(\hat{\mathbf{z}}_i^m + \hat{\mathbf{b}}_i^m) \text{ else}; \end{cases}$
      where $\hat{\mathbf{b}}_i^m = \mathbf{w}_m^T\mathbf{B}_i + \mathbf{1}b_m - \boldsymbol{\theta}_i^m/\mu$,
      $\hat{\mathbf{z}}_i^m = \mathbf{w}_m^T\mathbf{Z}_i + \mathbf{1}b_m - \boldsymbol{\theta}_i^m/\mu$.
      Update $u_{i,j}^m \in \mathbf{U}$ **by**
      $u_{i,j}^m = \begin{cases} (1/2)\cdot(r_i^m + \omega_i^m/\mu) + (1/4)\cdot(\bar{\mathbf{z}}_i^m + \bar{\mathbf{b}}_i^m) \\ \text{when } j = \arg\max(\mathbf{u}_i^m); \\ (1/2)\cdot(\bar{\mathbf{z}}_i^m + \bar{\mathbf{b}}_i^m) \text{ else}; \end{cases}$
      where $\hat{\mathbf{b}}_i^m = \mathbf{w}_y^T\mathbf{B}_i + \mathbf{1}b_y - \epsilon_i^m/\mu$,
      $\hat{\mathbf{z}}_i^m = \mathbf{w}_y^T\mathbf{Z}_i + \mathbf{1}b_y - \epsilon_i^m/\mu$.
    end for
    Update $\mathbf{Z}_i$ **by** $[\frac{1}{2}\sum_{m=1}^{K}\mathbf{w}_m(t_i^m - (\mathbf{w}_m^T\mathbf{B}_i + \mathbf{1}b_m) + \frac{\theta_i^m}{\mu})$
    $+\mathbf{w}_y(u_i^m - (\mathbf{w}_y^T\mathbf{B}_i + \mathbf{1}b_y) + \frac{\epsilon_i^m}{\mu}) + 2\mathbf{B}_i + (1/\mu)\mathbf{\Lambda}_{1,i}^T\mathbf{B}_i$
    $+(\tau_0/\mu)\mathbf{H}^T\mathbf{D}_0(\mathbf{F}_i - \mathbf{H}\mathbf{B}_i) + (1/\mu)\mathbf{\Lambda}_{3,i} - (\tau_1/\mu)\mathbf{B}_i\mathbf{D}_{1,i}]$
    $(\mathbf{B}_i^T\mathbf{B}_i + \mathbf{I})^{-1}$.
    Update $\mathbf{B}_i$ **by** $[\frac{1}{2}\sum_{m=1}^{K}\mathbf{w}_m(t_i^m - (\mathbf{w}_m^T\mathbf{Z}_i + \mathbf{1}b_m) + \frac{\theta_i^m}{\mu})$
    $+\mathbf{w}_y(u_i^m - (\mathbf{w}_y^T\mathbf{Z}_i + \mathbf{1}b_y) + \frac{\epsilon_i^m}{\mu}) + 2\mathbf{Z}_i + (1/\mu)\mathbf{\Lambda}_{1,i}^T\mathbf{Z}_i$
    $+(\tau_0/\mu)\mathbf{H}^T\mathbf{D}_0(\mathbf{F}_i - \mathbf{H}\mathbf{Z}_i) + (1/\mu)\mathbf{\Lambda}_{3,i} - (\tau_1/\mu)\mathbf{Z}_i\mathbf{D}_{1,i}]$
    $(\mathbf{Z}_i^T\mathbf{Z}_i + \mathbf{I})^{-1}$.
    Update $\mathbf{F}_i$ **by** Eq. (**??**).
    Update $\Lambda_{1,i}, \Lambda_{3,i}, \Lambda_{4,i}$ **by**
    $\mathbf{\Lambda}_{1,i} = \mathbf{\Lambda}_{1,i} + \mathbf{B}_i\mathbf{Z}_i^T - \mathbf{I}$; $\mathbf{\Lambda}_{3,i} = \mathbf{\Lambda}_{3,i} + \mathbf{Z}_i - \mathbf{B}_i$;
    $\mathbf{\Lambda}_{4,i} = \mathbf{\Lambda}_{4,i} + \mathbf{X}_i - \mathbf{F}_i$.
    Update $\lambda_{2,i}^m, \sigma_i^m, \omega_i^m, \boldsymbol{\theta}_i^m, \boldsymbol{\xi}_i^m$ **by**
    $\lambda_{2,i}^m = \lambda_{2,i}^m + \mu(e_i^m - (y_i^m - q_i^m + r_i^m))$;
    $\sigma_i^m = \sigma_i^m + \mu(q_i^m - \max(\mathbf{t}_i^m))$;
    $\omega_i^m = \omega_i^m + \mu(r_i^m - \max(\mathbf{u}_i^m))$;
    $\boldsymbol{\theta}_i^m = \boldsymbol{\theta}_i^m + \mu(\mathbf{t}_i^m - (\mathbf{w}_m^T\mathbf{X}_i + \mathbf{1}b_m))$;
    $\boldsymbol{\xi}_i^m = \boldsymbol{\xi}_i^m + \mu(\mathbf{u}_i^m - (\mathbf{w}_y^T\mathbf{X}_i + \mathbf{1}b_y))$.
  end for
  Update $\mathbf{H}$ **by** $(\sum_{i=1}^{N}\mathbf{F}_i(\mathbf{Z}_i^T + \mathbf{B}_i^T))$
  $(\sum_{i=1}^{N}(\mathbf{B}_i\mathbf{Z}_i^T + \mathbf{Z}_i\mathbf{B}_i^T))^{-1}$
end while

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

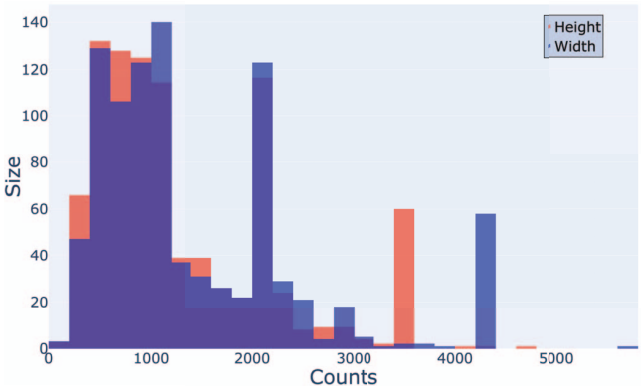

Fig. 1: The histogram of image sizes (width in blue and height in orange).

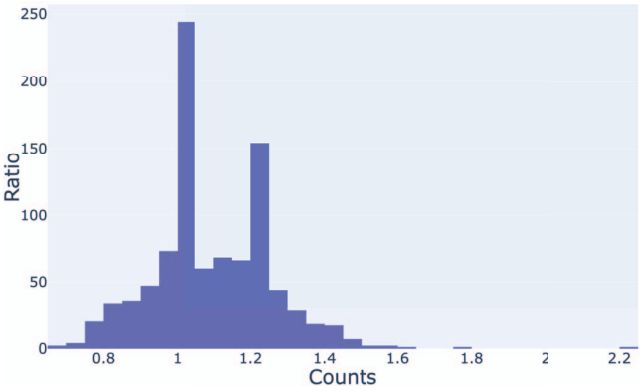

Fig. 2: The ratio (width/height) distribution of the image sizes.

(PFTAS) [19] method. These patches from two CXRs and one CT image at each time point is concatenated to create each information-dense instance. The PFTAS extracts the texture features by counting the number of neighboring black pixels for each pixel. Then the total count for all the pixels in a given patch is stored in a nine-bin histogram [19]. The thresholding is conducted by Otsu's algorithm [20] which generates a 162-dimensional feature vector for each patch. As a result, each instance is a vector of $162 \times 12 \times 3$ features. We list the details of this dataset:

- Cohort size: The number of patients is 323. The number of all patients associated with this dataset is 472, however 149 patients have been dropped because their labels are not provided.
- Label distribution: In our experiments, we have aimed to estimate the severity of patients from their CXR images. The severity is determined by whether the given patient was dead, or needs supplemental oxygen or intubation during admission in ICU (based on clinicians decision). As a result, 262 and 61 patients (bags) are labeled as the severe and non-severe status. Because of the limited number of bags, we have augmented the images in the bags by applying flipping, random rotations, and random translations. As a result of data augmentation, 1,000 bags have been involved in our experiments.
- The number of missing images: Total number of instances is 635. The number of front view CXRs captured is 611 (e.g., 24 front view CXRs are missing). The number of side view CXRs captured is 53. The number of CT images captured is 27.

A. Comparison Methods

We compare the classification performance and scalability of proposed *FMMSVM* to the following models:

- (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.
- The five multi-instance deep learning (DL) models: The (4) mi-Net and (5) MI-Net [7] approach to the MIL problem through instance space and embedded space (learning vectorial representation of bag) paradigm respectively. (6) The Multi-Modal Multi-Instance deep learning model (MMMI-deep) [21] learns the global cross-modal representation. (7) Attention-based deep Multiple Instance Learning (AMIL) [8] calculates the parameterized attention score for each instance to generate the probability distribution of bag labels. (8) Loss-based Attention for deep Multiple Instance Learning (LAMIL) [9] learns the instance scores and predictions jointly.
- (9) An variation of the *FMMSVM* for the purpose of ablation study: We discard the locality preserving (Ours w/o LP in Table. I) and smoothness learning (Ours w/o SL) capability from our model (Ours) to evaluate their effectiveness. We set τ_1 and τ_2 to zero to remove the impact of each term.

B. Hyperparameters

For the classification models used in Table I, we report the following hyperparameters found by grid search on the balanced accuracies of five test sets. For SIL, NSK, and STK the regularization tradeoff is set to 1.0. We set τ_0, τ_1, τ_2, τ_3, initial μ to $1e+2$, $1e-1$, $1e-2$, $1e+2$, $1e-5$ for our exact *FMMSVM* model and $1e+2$, $1e-2$, $1e-2$, $1e+3$, $1e-10$ for our inexact *FMMSVM* model. The tolerance is set to $1e-5$ for both. We set α_g to 5.0, 3.0, and 1.0 for front-view CXR, side-view CXR, and CT image modality. The deep learning models (mi-Net, MI-Net, MMMI-deep, AMIL, and LAMIL) are implemented using the codes provided by their respective papers [7], [21], [8], [9].

C. Classification Performance

In table I, we report precision, recall, F1-score, accuracy, and balanced accuracy (BACC) in classification of survival/death bags. We split the bags into 80% for training and

TABLE I: The classification performance of our *FMMSVM* and competing models on in-hospital mortality are below. We highlight the best scores in **bold**

Model	Precision	Recall	F1Score	Accuracy	BACC
SIL	0.862±0.013	0.796±0.013	0.823±0.036	0.781±0.043	0.804±0.024
NSK	0.891±0.024	0.901±0.025	0.881±0.019	0.847±0.031	0.860±0.022
STK	0.879±0.030	0.880±0.021	0.861±0.032	0.844±0.024	0.847±0.026
mi-Net	0.899±0.024	0.871±0.019	0.881±0.019	0.867±0.015	0.877±0.026
MI-Net	0.900±0.021	0.899±0.028	0.898±0.019	0.899±0.021	0.896±0.027
MMMI-deep	0.901±0.019	0.904±0.021	0.902±0.027	0.905±0.024	0.884±0.023
AMIL	0.881±0.032	0.886±0.026	0.890±0.019	0.849±0.047	0.846±0.020
LAMIL	0.893±0.028	0.895±0.041	0.894±0.031	0.867±0.026	0.879±0.026
Ours	0.919±0.021	0.904±0.029	0.898±0.021	0.917±0.029	0.911±0.023
Ours w/o LP	0.869±0.034	0.885±0.031	0.877±0.054	0.904±0.065	0.894±0.027
Ours w/o SL	0.908±0.024	0.903±0.031	0.906±0.035	0.908±0.024	0.916±0.012

20% for test set, then we train the classifier with training set of 3 folds, and tune the hyperparameters based on the accuracy on the validation set of 1 fold. Finally, the performance is measured on the test (held out) set of 1 fold and this is repeated 5 times and scores are averaged across 5 results following 5-fold cross validation scheme. The comparison between the classification models in Table I shows that the proposed *FMMSVM* models outperform the other existing multi-instance models. We interpret that our model has the better capability in joint imputation-prediction and learning temporal relationships between the instances. When Ours is compared to Ours w/o LP and Ours w/o SL, we observe that the introduced locality preserving and smoothness loss terms improve the prediction. These results demonstrate that classification pattern for *FMMSVM* can vary based on the optimization strategy.

D. Experiments with Multi-modal Brain Imaging Dataset

Although the results from CXR dataset are promising, the cohort size (323 patients) is limited and the classification score gaps between our model and best performing baseline model (MMMI-deep) in Table I is within 3σ. Therefore we have extended our experiments to the additional dataset, which can provide the better measurement on the clinical applicability. We have conducted experiments on Alzheimer's Disease Neuroimaging Initiative (ADNI) database collected from 818 participants. In this dataset, each participant (bag) has 1 to 5 brain MRIs (1 to 5 instances), and we perform FreeSurfer and voxel-based morphometry [22] to extract the gray matter measures for 90 regions of interest for each brain MRI. To reproduce the missing modalities, we discard each MRI randomly with 50% probability. We classify each patient's Alzheimer's Diesase progression (AD) in {Alzheimer's Disease (337 AD patients), Mild Cognitive Impairment (251 MCI patients), Healthy Control (230 HC patients)}. Considering this ternary classification is usually more difficult than binary classification, the results are promising and they show the significantly improved performance of our model as the score gaps are larger than 3σ.

Besides the improved performance, our model has identified the AD risk factor as shown in Fig. 3. We have analyzed the learned weights of our model on each ROI feature. Our model identifies hippocampus [23], caudate nucleus [23], lateral ventricle [24], and amygdala [25] regions. The identified regions have been shown in the medical literatures to be related to AD. These results additionally validate the correctness of disease progression prediction from our model.

IV. Conclusion

Information in medical image datasets usually represent a large number of features and are delivered in a variety of modalities. As data mining technologies develop, multi-modal methods are attracting more attention in the field of machine learning researches. In this study, we present a novel multi-instance learning model that is scalable to a large number of features. We employ the factorization based joint imputation-prediction approach to handle the missing data in the modalities and PFTAS methods to control the fine-grained details of imaging information. In our experiments, we have observed 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.

Acknowledgment

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

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] 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.

[2] H. Wang, F. Nie, and H. Huang, "Learning instance specific distance for multi-instance classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, 2011, pp. 507–512.

[3] K. Liu, H. Wang, F. Nie, and H. Zhang, "Learning multi-instance enriched image representations via non-greedy ratio maximization of the l1-norm distances," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7727–7735.

[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.

TABLE II: The classification performance on AD progression

Model	AD-Precision	MCI-Precision	HC-Precision	AD-Recall	MCI-Recall	HC-Recall	Accuracy
SIL	0.45±0.02	0.53±0.02	0.51±0.02	0.44±0.02	0.49±0.02	0.48±0.03	0.54±0.02
NSK	0.42±0.03	0.50±0.02	0.54±0.03	0.48±0.03	0.53±0.02	0.54±0.02	0.53±0.02
STK	0.50±0.01	0.54±0.04	0.52±0.04	0.58±0.02	0.48±0.03	0.55±0.02	0.55±0.03
mi-Net	0.57±0.06	0.52±0.03	0.45±0.02	0.57±0.02	0.46±0.02	0.49±0.03	0.58±0.02
MI-Net	0.55±0.01	0.54±0.02	0.57±0.02	0.49±0.04	0.52±0.03	0.46±0.02	0.57±0.01
MMMI-deep	0.68±0.02	0.62±0.00	0.59±0.02	0.57±0.02	0.52±0.05	0.54±0.02	0.54±0.03
AMIL	0.62±0.02	0.66±0.01	0.60±0.02	0.54±0.02	0.57±0.02	0.52±0.03	0.56±0.01
LAMIL	0.65±0.02	0.61±0.02	0.62±0.01	0.57±0.03	0.58±0.01	0.57±0.02	0.59±0.02
Ours	0.76±0.02	0.69±0.01	0.75±0.03	0.65±0.02	0.69±0.01	0.77±0.03	0.72±0.02
Ours w/o LP	0.69±0.00	0.73±0.02	0.68±0.01	0.62±0.03	0.59±0.02	0.67±0.02	0.69±0.02
Ours w/o SL	0.70±0.01	0.70±0.04	0.73±0.02	0.68±0.02	0.62±0.01	0.69±0.04	0.67±0.02

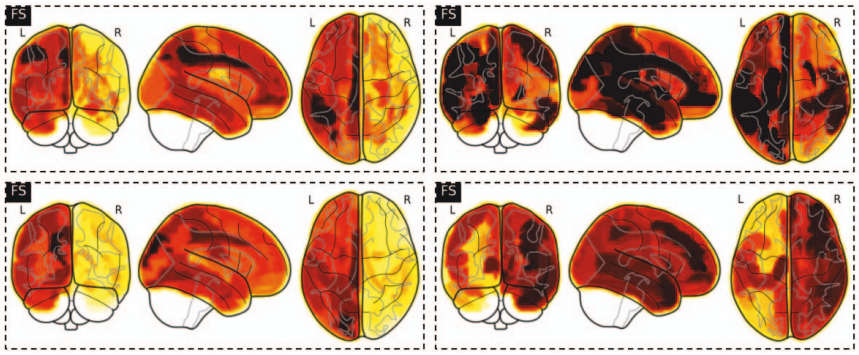

Fig. 3: Identification results on brain ROIs. The darker color indicates that the corresponding region is identified as the important region in predicting AD progression by our model. The following regions have been identified in **top-left**: hippocampus and caudate nucleus, **top-right**: lateral ventricle, **bottom-left**: amygdala, **bottom-right**: caudate.

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

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

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

[9] 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.

[10] H. Seo, L. Brand, L. S. Barco, and H. Wang, "Scaling multi-instance support vector machine to breast cancer detection on the breakhis dataset," *Bioinformatics*, vol. 38, no. Supplement_1, pp. i92–i100, 2022.

[11] L. Brand, H. Seo, L. Z. Baker, C. Ellefsen, J. Sargent, and H. Wang, "A linear primal–dual multi-instance svm for big data classifications," *Knowledge and Information Systems*, pp. 1–32, 2023.

[12] H. Seo, L. Brand, L. S. Barco, and H. Wang, "Scalable multi-instance multi-shape support vector machine for whole slide breast histopathology," in *2022 IEEE International Conference on Knowledge Graph (ICKG)*. IEEE, 2022, pp. 225–232.

[13] M. Ranjbar, P. Moradi, M. Azami, and M. Jalili, "An imputation-based matrix factorization method for improving accuracy of collaborative filtering systems," *Engineering Applications of Artificial Intelligence*, vol. 46, pp. 58–66, 2015.

[14] L. Brand, L. Z. Baker, and H. Wang, "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*, 2021, pp. 1–6.

[15] X. He and P. Niyogi, "Locality preserving projections," *Advances in neural information processing systems*, vol. 16, 2003.

[16] F. Nie, H. Wang, H. Huang, and C. Ding, "Joint schatten p-norm and l p-norm robust matrix completion for missing value recovery," *Knowledge and Information Systems*, vol. 42, no. 3, pp. 525–544, 2015.

[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] 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.

[19] 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.

[20] 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.

[21] 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.

[22] S. L. Risacher, L. Shen, J. D. West, S. Kim, B. C. McDonald, L. A. Beckett, D. J. Harvey, C. R. Jack Jr, M. W. Weiner, A. J. Saykin *et al.*, "Longitudinal mri atrophy biomarkers: relationship to conversion in the adni cohort," *Neurobiology of aging*, vol. 31, no. 8, pp. 1401–1418, 2010.

[23] L. G. Apostolova, M. Beyer, A. E. Green, K. S. Hwang, J. H. Morra, Y.-Y. Chou, C. Avedissian, D. Aarsland, C. C. Janvin, J. P. Larsen *et al.*, "Hippocampal, caudate, and ventricular changes in parkinson's disease with and without dementia," *Movement Disorders*, vol. 25, no. 6, pp. 687–695, 2010.

[24] O. T. Carmichael, L. H. Kuller, O. L. Lopez, P. M. Thompson, R. A. Dutton, A. Lu, S. E. Lee, J. Y. Lee, H. J. Aizenstein, C. C. Meltzer *et al.*, "Ventricular volume and dementia progression in the cardiovascular health study," *Neurobiology of aging*, vol. 28, no. 3, pp. 389–397, 2007.

[25] S. P. Poulin, R. Dautoff, J. C. Morris, L. F. Barrett, B. C. Dickerson, A. D. N. Initiative *et al.*, "Amygdala atrophy is prominent in early alzheimer's disease and relates to symptom severity," *Psychiatry Research: Neuroimaging*, vol. 194, no. 1, pp. 7–13, 2011.