Transfer Learning via Optimal Transportation for Integrative Cancer Patient Stratification

Ziyu Liu[1*], **Wei Shao**[2], **Jie Zhang**[3], **Min Zhang**[1] and **Kun Huang**[2,4]
[1]Department of Statistics, Purdue University
[2]Biostatistics and Health Data Science, Indiana University School of Medicine
[3]Department of Medical and Molecular Genetics, Indiana University School of Medicine
[4]Regenstrief Institute, Indianapolis
{liu2301, minzhang}@purdue.edu, {shaowei, jizhan, kunhuang}@iu.edu

Abstract

The Stratification of early-stage cancer patients for the prediction of clinical outcome is a challenging task since cancer is associated with various molecular aberrations. A single biomarker often cannot provide sufficient information to stratify early-stage patients effectively. Understanding the complex mechanism behind cancer development calls for exploiting biomarkers from multiple modalities of data such as histopathology images and genomic data. The integrative analysis of these biomarkers sheds light on cancer diagnosis, subtyping, and prognosis. Another difficulty is that labels for early-stage cancer patients are scarce and not reliable enough for predicting survival times. Given the fact that different cancer types share some commonalities, we explore if the knowledge learned from one cancer type can be utilized to improve prognosis accuracy for another cancer type. We propose a novel unsupervised multi-view transfer learning algorithm to simultaneously analyze multiple biomarkers in different cancer types. We integrate multiple views using non-negative matrix factorization and formulate the transfer learning model based on the Optimal Transport theory to align features of different cancer types. We evaluate the stratification performance on three early-stage cancers from the Cancer Genome Atlas (TCGA) project. Comparing with other benchmark methods, our framework achieves superior accuracy for patient outcome prediction.

1 Introduction

Cancer is one of the major public health problems worldwide. The occurrence of cancer is increasing annually due to the aging of the population, as well as the prevalence of risk factors such as pollution. During the past decade, precision medicine approaches have been developed to reduce mortality rate and increase patient survival time. For precision medicine, the stratification for early-stage cancer patient into correct risk groups is critical for clinical decision-making [Bashiri *et al.*, 2017]. This process helps to estimate prognosis especially survival duration as well as response to treatment. To achieve this goal, advanced statistical and machine learning (ML) algorithms are often needed [Kourou *et al.*, 2015].

Despite the wide applications of ML models in biomedical problems, there still exists large challenges in predicting clinical outcome for cancer patients. Firstly, most of the ML methods only used one specific modality of biomarkers (e.g., image or genomic data) for the prognosis of human cancers. However, such single data modality cannot fully capture the overall mechanism and development of human cancers and thus lacks sufficient predictive power. Accordingly, many researchers have developed integrative analysis pipelines to combine multi-modal data such as imaging (e.g., histopathological and CT images) and genomic data (e.g., gene mutation and gene expression) for cancer outcome prediction [Shao *et al.*, 2019]. Another challenge is that the collection of multi-modal data for cancer patients is costly and difficult while the sample size is often too small to train prediction models on a specific cancer type. Accordingly, it is of great interest to examine if the knowledge learned from other cancer types can be transferred to improve the prediction results on the target cancer type [Hanahan and Weinberg, 2011]. As one widely used ML paradigm, transfer Learning (TL) enables us to leverage knowledge from multiple cancer types and increase the prediction accuracy on the target cancer type. In recent years, Optimal Transport (OT) has been introduced in solving TL problems. In comparison with other domain adaption strategies in TL, OT defines a metric (distance) between probability distributions which can better exploit the geometry of the underlying feature spaces [Flamary and Courty, 2021].

Based on the above consideration, we propose an OT-based multi-view TL algorithm for the prognosis of early-stage cancers by integrating histopathological images with gene expression profiles from multiple types of cancers. Specifically, with the help of the OT framework , our proposed TL approach conducts integrative analysis using non-negative matrix factorization (NMF) techniques and seeking common latent features between two cancer types (e.g., breast invasive carcinoma and kidney renal papillary cell carcinoma), followed by iCluster [Shen *et al.*, 2012], an integrative genomic

*Contact Author

data clustering tool to improve patient stratification performance for different cancer types. To the best of knowledge, this is the first study that utilize the OT framework to solve the multi-view TL algorithm. The experimental results show that the proposed method outperforms other benchmark algorithms on both synthetic data and real datasets from The Cancer Genome Atlas (TCGA). These results not only confirm the superiority of our method for clinical outcome prediction, but also suggest that transferring knowledge between different cancer types has great potential for improving cancer precision medicine. Furthermore, this work demonstrates the advantage of incorporating OT in multi-view TL problems.

2 Related Works

A large number of biomarkers have been discovered for cancer prognosis, such as histopathological images, genetic alterations, epigenomic changes, gene, and protein expression signatures [Liu *et al.*, 2018]. With the recent availability of large-scale computing resources and the fast development of digital image technology, we can extract meaningful features from the whole slide histopathological images that are highly associated with the development of cancers. They have shown great promise for the diagnosis and prognosis of different types of cancers such as breast cancer [Lu *et al.*, 2020]. In addition, some researchers also integrated pathological images with genomic profiles for the survival analysis of human cancer, since different biomarkers may provide complementary information that can improve the prognosis performance [Cheng *et al.*, 2017]. However, most of the existing studies treat the prediction task on different cancer types independently and they overlook the fact that clinical knowledge derived from one cancer type can also help the prediction task on other cancer types. Recently, the TL approach has received a lot of attention. Generally, existing TL algorithms can be categorized into the following two strategies: instance weighting strategy and feature transformation strategy. The instance weighting strategy adopts an instance and domain weighting strategy to adapt the source distribution from the source to the target domain [Dai *et al.*, 2007], while the feature transformation strategy aims at minimizing the marginal and the conditional distribution differences while preserving the underlying data structure [Pan *et al.*, 2010; Long *et al.*, 2013]. Our proposed method belongs to the second category, which uses the OT framework to align the distributions of multi-view features between different cancer types. In comparison with previous related works, our proposed Multi-view TL via OT (MVTOT) method has three main contributions: 1) we transfer multiple data modalities from the source to the target domain at the same time without utilizing source domain labels; 2) we design an update rule which ensures the objective function to decrease monotonically at each iteration and eventually converge to the Karush–Kuhn–Tucker (KKT) point; and 3) we not only verify the effectiveness of the proposed MVTOT method on synthetic data but also demonstrate its superiority on multiple TCGA datasets for stratifying early-stage cancer patients.

3 Proposed Method

3.1 Notations

We first introduce the notations for the proposed unsupervised multi-view transfer learning algorithm. Given a set of unlabeled source domain instances with two views, $\mathcal{D}_s = \{\mathbf{x}_i^{s,v}\}_{v=1,2,i=1,\cdots,N_s}$, where $\mathbf{x}_i^{s,1}$ and $\mathbf{x}_i^{s,2}$represent the extracted genomic and pathological image features for the i-th patient in the source domain, respectively. Similar to the source domain instance, we also define the examples in the target domain as $\mathcal{D}_t = \{\mathbf{x}_j^{t,v}\}_{v=1,2,j=1,\cdots,N_t}$. In this study, we have applied the same pre-processing pipelines to extract features from pathological images and gene expression data and their corresponding dimensionalities are identical across different cancer types, i.e. $\mathbf{x}_i^{s,v}$, $\mathbf{x}_j^{t,v} \in \mathbb{R}^{d_v}$, $\forall v = 1, 2, i = 1, \cdots, N_s, j = 1, \cdots, N_t$. For the simplicity of calculation, we rewrite the data matrix of the v-th view of the source and the target domain in a concise form: $\mathbf{X}^{l,v} = [\mathbf{x}_1^{l,v}, \cdots, \mathbf{x}_{N_l}^{l,v}] \in \mathbb{R}^{d_v \times N_l}, v = 1, 2, l = s, t$.

3.2 Non-negative Matrix Factorization for Multi-view Learning

We firstly define the traditional Non-negative Matrix Factorization problem as follows:

Definition 1 (Non-Negative Matrix Factorization (NMF))**.** *Given a non-negative input data matrix* $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}_+^{d \times N}$*, we can factorize* $\mathbf{X}$ *into a basis matrix* $\mathbf{W}$ *and a low-rank coefficient matrix* $\mathbf{H}$ *as* $\mathbf{X} \approx \mathbf{WH}$*, where* $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_K] \in \mathbb{R}_+^{d \times K}$*,* $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_N] \in \mathbb{R}_+^{K \times N}$*, and* $K \ll \min(d, N)$ *(*K *is the number of basis vectors). To preserve as much information as possible, we need to minimize the discrepancy defined by the squared Frobenius norm,*

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_F^2 \quad s.t.\ \mathbf{W} \in \mathbb{R}_+^{d \times K}, \mathbf{H} \in \mathbb{R}_+^{K \times N}. \quad (1)$$

For the better fusion of multi-view data, our goal is to seek one common representation across all modalities. Based on definition 1, we formulate the Non-negative Matrix Factorization problem for multi-view learning as follows:

$$\sum_{v=1}^{2} \|\mathbf{X}^v - \mathbf{W}^v\mathbf{H}^v\|_F^2 + \beta \sum_{v=1}^{2} \|\mathbf{H}^v - \mathbf{H}^*\|_F^2, \quad (2)$$

where the first term aims at factorizing the data of each view $\mathbf{X}^v, v = 1, 2$ into a dictionary matrix $\mathbf{W}^v$ and a low-rank coefficient matrix $\mathbf{H}^v$. The second term is used to seek a common representation $\mathbf{H}^*$ and enforce the disagreement between the representation of each view i.e., $\mathbf{H}^v$ as small as possible.

3.3 Optimal Transportation

By considering the fact that different cancer types share some commonalities, in this paper, we focus on transferring the knowledge from the source cancer cohort to improve the prediction performance on target cancer type.

Traditionally, researchers apply metrics such as Maximum Mean Discrepancy (MMD), Cross Entropy (CE), and Kullback–Leibler (KL) divergence to directly measure the discrepancy of feature distributions on different domains. The

main drawback of CE and KL divergence is that they fail to utilize the geometric information in the space. MMD summaries the distributional differences using the distance between weighted feature means from two domains. It can hardly represent the data distribution unless we have adequate labels. Compared with these methods, the OT framework can capture the intrinsic geometry structure of feature spaces in the setting that no label is available in both domains. In this section, we will briefly introduce the OT framework, and the regularized OT problem can be defined as follows,

Definition 2 (Regularized OT)**.** *Given two discrete probability distributions $\mu_\alpha = \sum_{i=1}^{m} \mathbf{a}_i \delta_{\mathbf{w}_i^s}$ (on $\mathcal{D}_s$) and $\mu_\beta = \sum_{j=1}^{n} \mathbf{b}_j \delta_{\mathbf{w}_j^t}$ (on $\mathcal{D}_t$), where $\delta_{\mathbf{w}}$ is the Dirac function at location $\mathbf{w} \in \mathbb{R}^d$, $\{\mathbf{a}_i\}_{i=1}^m, \{\mathbf{b}_j\}_{j=1}^n$ are probability mass associated with $\{\mathbf{w}_i^s\}_{i=1}^m, \{\mathbf{w}_j^t\}_{j=1}^n$, the relaxed version of earth moving distance (EMD) problem seek to find an optimal coupling matrix $\mathbf{P} \in \mathbb{R}_+^{m\times n}$ to minimize the total cost,*

$$\mathbf{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) =< \mathbf{P}, \mathbf{C} >_F,$$

where $\mathbf{C} \in \mathbb{R}_+^{m\times n}$ is the cost matrix, $\mathbf{C}(i, j)$ represents the cost of moving a probability mass from $\mathbf{w}_i^s$ to $\mathbf{w}_j^t$, $\mathbf{P} \in \mathbf{\Pi}(\mathbf{a}, \mathbf{b}) \coloneqq \{\mathbf{T} \in \mathbb{R}_+^{m\times n} | \mathbf{T}\mathbf{1}_n = \mathbf{a}, \mathbf{T}^T\mathbf{1}_m = \mathbf{b}\}$, $\mathbf{P}(i, j)$ represents the amount of mass we move from location $\mathbf{w}_i^s$ to location $\mathbf{w}_j^t$. Later works [Cuturi, 2013] converted the objective function into computational effective version by adding an entropy regularization term. The regularized OT problem reads,

$$\mathbf{L}_{\mathbf{C}}^{\epsilon}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P}\in\mathbf{\Pi}(\mathbf{a},\mathbf{b})} < \mathbf{P}, \mathbf{C} >_F - \epsilon\mathbf{\Omega}(\mathbf{P}), \tag{3}$$

where $\mathbf{\Omega}(\mathbf{P}) = -\sum_{i,j} \mathbf{P}_{i,j} log(\mathbf{P}_{i,j})$ is the entropy of the coupling matrix.

3.4 Multi-view Transfer Learning via Optimal Transportation (MVTOT).

In the previous studies [Ding *et al.*, 2008], they demonstrated the equivalency of NMF and spectral clustering, matrix factorization and K-means clustering. Specifically, columns of $\mathbf{W}$ represent the clustering centroids while $\mathbf{H}$ contains the cluster membership indicators. Remark that for MMD, the distance between $\mathcal{D}_s$ and $\mathcal{D}_t$ is defined by the distance between weighted sample feature means, which could be interpreted as centroids of data distributions in the feature space. Inspired by MMD, we utilize OT distance to define the distance between empirical distributions of clustering centers,

$$\hat{\mu}^{s,v} = \sum_{i=1}^{K} \mathbf{a}_i \delta_{\mathbf{w}_i^{s,v}}, \quad \hat{\mu}^{t,v} = \sum_{j=1}^{K} \mathbf{b}_j \delta_{\mathbf{w}_j^{t,v}}, \tag{4}$$

where $\mathbf{a}$ and $\mathbf{b}$ are probability mass associated to $\{\mathbf{w}_i^{s,v}\}_{i=1}^K$ and $\{\mathbf{w}_j^{t,v}\}_{j=1}^K$, v is the view indicator ($v \in \{1, 2\}$).

Remark 1. *Although we require a common latent dimension K across different domains, it's easy to extend our method to the condition that the number of basis vectors in $\mathbf{W}^{s,v}$ and $\mathbf{W}^{t,v}$ are different. Another assumption is that weights for each feature is the same, i.e. $\mathbf{a} = \mathbf{b} = \mathbf{1}_K/K \in \mathbb{R}^K$ ($\mathbf{1}_K$ is a K dimensional all one column vector). Nevertheless, it's not an intrinsic constraint.*

Remark 2. *Since we calculate the OT distance between $\hat{\mu}^{s,v}$ and $\hat{\mu}^{t,v}$ for each data modality ($v = 1, 2$) separately, we can parallelly define two cost matrices $\{\mathbf{C}^v\}_{v=1}^2 \in \mathbb{R}^{K\times K}$. In our work, we use the squared Euclidean distance as the cost, i.e. $\mathbf{C}^v(i, j) \coloneqq \left\|\mathbf{w}_i^{s,v} - \mathbf{w}_j^{t,v}\right\|_2^2$. We then re-write the cost matrix in the following matrix function form,*

$$\begin{aligned}\mathbf{C}^v = &diag((\mathbf{W}^{s,v})^T\mathbf{W}^{s,v})\mathbf{1}_K\mathbf{1}_K^T - 2(\mathbf{W}^{s,v})^T\mathbf{W}^{t,v} \\ &+ \mathbf{1}_K\mathbf{1}_K^T diag((\mathbf{W}^{t,v})^T\mathbf{W}^{t,v}),\end{aligned} \tag{5}$$

where $\mathbf{1}_K$ is a K-dimensional all-ones column vector. For a fixed coupling matrix $\mathbf{P}^v$, we can take derivative of the regularized OT distance (3) *with respect to $\mathbf{W}^{s,v}$ and $\mathbf{W}^{t,v}$. It enables us to update the basis matrix based on the gradient information.*

To accommodate more general data matrices with mixed signs, we adopt the Semi-nonnegative matrix factorization (Semi-NMF) [Ding *et al.*, 2008], which is an extension of traditional NMF. The objective function is the same as NMF but without the non-negative constraint on $\mathbf{W}$. Then, the objective function of the proposed MVTOT method reads,

$$\begin{aligned}J(\mathbf{W}^{l,v}, \mathbf{H}^{l,v}, \mathbf{H}^{l,*}) = &\sum_{l=s,t}\sum_{v=1}^{2} \left\|\mathbf{X}^{l,v} - \mathbf{W}^{l,v}\mathbf{H}^{l,v}\right\|_F^2 + \\ &\alpha\sum_{v=1}^{2} \mathbf{L}_{\mathbf{C}^v}^{\epsilon}(\frac{\mathbf{1}_K}{K}, \frac{\mathbf{1}_K}{K}) + \beta\sum_{l=s,t}\sum_{v=1}^{2} \left\|\mathbf{H}^{l,v} - \mathbf{H}^{l,*}\right\|_F^2 + \\ &\gamma_1\sum_{l=s,t}\sum_{v=1}^{2} J_{\mathbf{W}}(\mathbf{W}^{l,v}) + \gamma_2\sum_{l=s,t}\sum_{v=1}^{2} J_{\mathbf{H}}(\mathbf{H}^{l,v}),\end{aligned} \tag{6}$$

where $\mathbf{X}^{l,v}, \mathbf{W}^{l,v}, \mathbf{H}^{l,v} (l \in \{s, t\}, v \in \{1, 2\})$ indicate the data, basis and coefficient matrices for the v-th view in the source or target domain, respectively. $\mathbf{H}^{l,*} (l \in \{s, t\})$ denotes the common representation for the source or target domain. The second term is the optimal distance between $\hat{\mu}^{s,v} = \sum_{i=1}^{K} \mathbf{1}_K/K\delta_{\mathbf{w}_i^{s,v}}$ and $\hat{\mu}^{t,v} = \sum_{j=1}^{K} \mathbf{1}_K/K\delta_{\mathbf{w}_j^{t,v}}$. By defining $\mathbf{a} = \mathbf{b} = \mathbf{1}_K/K \in \mathbb{R}^K$, $\mathbf{C}^v$ follows (5), we can represent the regularized OT distance defined by (3) as $\mathbf{L}_{\mathbf{C}^v}^{\epsilon}(\mathbf{1}_K/K, \mathbf{1}_K/K)$. Finally, $J_{\mathbf{W}}$ and $J_{\mathbf{H}}$ are two regularization terms with respect to $\mathbf{W}$ and $\mathbf{H}$. Here, we use the squared Frobenius norm in equation (6) to control the magnitude of $\mathbf{W}$ and $\mathbf{H}$.

4 Description of the Algorithm

Before optimizing the objective function using an alternative manner, we introduce two lemmas about convergence properties. They ensure that the value of the objective function will decrease monotonically after each iteration and the converged $\mathbf{H}$ satisfies the Karush–Kuhn–Tucker (KKT) conditions.

Lemma 1. *If a function $J : \mathbb{R}_+^{n\times K} \to \mathbb{R}$ has the following form:*

$$J(\mathbf{Y}) = tr(-2\mathbf{Y}^T\mathbf{B}) + tr(\mathbf{Y}\mathbf{A}\mathbf{Y}^T),$$

where $\mathbf{B} \in \mathbb{R}^{n\times K}$, $\mathbf{A} \in \mathbb{R}^{K\times K}$, $\mathbf{A}$ is symmetric. By letting $\mathbf{B}_{ij}^+ = (\mathbf{B}_{ij} + |\mathbf{B}_{ij}|)/2$, $\mathbf{B}_{ij}^- = (|\mathbf{B}_{ij}| - \mathbf{B}_{ij})/2$, the update

rule:

$$\mathbf{Y}^{(t+1)} \leftarrow \mathbf{Y}^{(t)} \odot \sqrt{\frac{\mathbf{B}^{+} + (\mathbf{Y}^{(t)}\mathbf{A}^{-})}{\mathbf{B}^{-} + (\mathbf{Y}^{(t)}\mathbf{A}^{+})}}, t = 0, 1, 2, \cdots \quad (7)$$

will lead $J(\cdot)$ to decrease monotonically after each iteration.

Lemma 2. *When the algorithm converge, the limit solution of* (7) *(i.e. $\mathbf{Y}^{(t)} = \mathbf{Y}^{(t+1)}$ for updating time t large enough) satisfies the KKT condition.*

Proof. Both proofs can be found in the supplementary material. □

Based on the above two lemmas, we apply an alternative way to update the coefficient matrix $\mathbf{H}$ and basis matrix $\mathbf{W}$ as follows.

4.1 Update the Coefficient Matrix H

When we update $\mathbf{H}^{l,v}$ for a domain l and view v, we fix all basis matrix $\mathbf{W}^{l',v'}, l' = s, t, v' = 1, 2$, the common expression matrix $\mathbf{H}^{l',*}, l' = s, t$ and $\mathbf{H}^{l'',v''}$ with $l'' \neq l$ or $v'' \neq v$. Then, we can simplify the objective function by dropping superscripts for domains and views,

$$J(\mathbf{H}) = tr(-2\mathbf{H}(\mathbf{X}^T\mathbf{W} + \beta(\mathbf{H}^*)^T)) + tr(\mathbf{H}^T(\mathbf{W}^T\mathbf{W} + (\beta + \gamma_2)\mathbf{I}_K)\mathbf{H}) + const,$$

where $\mathbf{I}_K$ is a K by K indentity matrix. Let $\mathbf{Y} = \mathbf{H}^T$, $\tilde{\mathbf{B}} = \beta(\mathbf{H}^*)^T + \mathbf{X}^T\mathbf{W}$, and $\tilde{\mathbf{A}} = \mathbf{W}^T\mathbf{W} + (\beta + \gamma_2)\mathbf{I}_K$, we can use (7) to update $\mathbf{H}$,

$$\mathbf{H}^T \leftarrow \mathbf{H}^T \odot \sqrt{\frac{\tilde{\mathbf{B}}^{+} + \mathbf{H}^T\tilde{\mathbf{A}}^{-}}{\tilde{\mathbf{B}}^{-} + \mathbf{H}^T\tilde{\mathbf{A}}^{+}}} \quad (8)$$

4.2 Update the Basis Matrix W

When we update the basis matrix $\mathbf{W}^{s,v}$ for the v-th view of the source domain, since the basis matrix for each view is updated separately, we can drop the superscript for view. We fix coefficient matrices $\mathbf{H}^{l,v}, l = s, t, v = 1, 2$, coupling matrices $\mathbf{P}^v, v = 1, 2$, and the basis matrix of the target domain $\mathbf{W}^{t,v}$. The object function can be re-formulated as:

$$J(\mathbf{W}^s) = \|\mathbf{X}^s - \mathbf{W}^s\mathbf{H}^s\|_F^2 + \gamma_1 \|\mathbf{W}^s\|_F^2$$
$$\alpha < \mathbf{P}, diag((\mathbf{W}^s)^T\mathbf{W}^s)\mathbf{1}_K\mathbf{1}_K^T - 2(\mathbf{W}^s)^T\mathbf{W}^t >_F + const.$$

Since we impose no constraint on $\mathbf{W}^l$, we can calculate the stationary point of $\mathbf{W}^s$ directly and obtain the following update rule,

$$\mathbf{W}^s \leftarrow (\mathbf{X}^s(\mathbf{H}^s)^T + \alpha\mathbf{W}^t\mathbf{P}^T)(\mathbf{H}^s(\mathbf{H}^s)^T + \tilde{\alpha}\mathbf{I}_K)^{-1}, \quad (9)$$

where $\tilde{\alpha} = (\alpha/K + \gamma_1)$. Notice that we use the fact that $\mathbf{P} \in \mathbf{\Pi}(\mathbf{1}_K/K, \mathbf{1}_K/K)$ when we calculate the derivative. The same technique can be applied to calculate the derivative of $\mathbf{W}^t$. More details can be found in the supplementary material.

4.3 Update the Coupling Matrix P

Since we update dictionary matrices for each data modality separately, the superscript for view is dropped. Notice that when we fix $\mathbf{W}^s$ and $\mathbf{W}^t$, the cost matrix (5) is fixed. Then, we can solve the regularized OT problem (3) by directly using the Optimal Transport solver [Flamary and Courty, 2021].

Algorithm 1 Multi-view transfer learning via Optimal Transport

1: **Input**: $\{\mathbf{X}^{l,v}\}_{l=s,t,v=1,2}$ and hyper-parameters K, α, β, γ_1, γ_2, ϵ (regularization weight in (3)).
2: **Initialization**: Use Semi-NMF algorithms [Ding *et al.*, 2008] to factorize $\{\mathbf{X}^{l,v}\}_{l=s,t,v=1,2}$ and get $\{\mathbf{W}^{l,v}\}_{l=s,t,v=1,2}$, $\{\mathbf{H}^{l,v}\}_{l=s,t,v=1,2}$; Initialize $\{\mathbf{C}^v\}_{v=1}^2$, $\{\mathbf{P}^v\}_{v=1}^2$, and $\mathbf{H}^{l,*}$ with regard to (5), (3) and (10) respectively.
3: **repeat**
4: For each data view $v = 1, 2$,
- Update $\mathbf{W}^{s,v}$ with regard to (9); Reconstruct the cost function $\mathbf{C}^v$ with regard to (5); Solve $\mathbf{P}^v$ in (3) using the OT solver [Flamary and Courty, 2021].
- Repeat the same process for $\mathbf{W}^{t,v}$.

5: Update $\{\mathbf{H}^{l,v}\}_{l=s,t,v=1,2}$ with regard to (8); Recompute $\mathbf{H}^{l,*}$ with regard to (10).
6: **until** convergence criterion is satisfied.
7: **Output**: $\mathbf{H}^{t,*}$.

4.4 Update the Common Representation $\mathbf{H}^*$

When updating $\mathbf{H}^{l,*}$, we fix all other matrices. By taking derivative of (6) with regard to $\mathbf{H}^{l,*}$ and setting it to be zero, we obtain,

$$\mathbf{H}^{l,*} \leftarrow \frac{1}{2}\sum_{v=1}^{2}\mathbf{H}^{l,v}. \quad (10)$$

$\mathbf{H}^{l,*}$ is non-negative since it's an average of non-negative matrices.

5 Experiments and Results

In this section, we verify our proposed method by two sets of experiments. The first set of experiments are performed on a synthetic dataset. By comparing classification accuracy, Silhouette score and NMI (normalized mutual information) score for clustering, we demonstrate that our method outperforms other benchmark methods. The second set of experiments are performed on the TCGA datasets. We verify the advantages of our method by comparing P-values of log-rank tests for survival analysis.

5.1 Results on Synthetic Data

In the first set of experiment, we generate $N_c = 3$ centroids with dimension $d_1 = 20$ on the source domain $\mathcal{D}_s$, where each centroid represents one class. Then, we generate 100 samples following a Gaussian distribution with standard deviation $\sigma_1 = 2$ around each centroids. In order to generate two different views of each samples, we project each sample onto two higher dimensional spaces ($d_3 = 50, d_4 = 100$) by multiplying two random matrices $\mathbf{V}_1 \in \mathbb{R}^{d_1 \times d_3}$, $\mathbf{V}_2 \in \mathbb{R}^{d_1 \times d_4}$. We use a similar strategy to generate samples (latent dimension $d_2 = 20$) on the target domain, the only difference is that the standard deviation for each centroid is set as $\sigma_2 = 4$. Besides the above synthetic dataset, we generate another dataset

using the same approach but specify $d_1 = 25, d_2 = 18$, $\sigma_1 = 2, \sigma_2 = 3$.

	$d_1 = d_2 = 20$			$d_1 = 25, d_2 = 18$		
Methods	ACC	Silh	NMI	ACC	Silh	NMI
Kmeans	0.70	**0.36**	0.37	0.44	**0.27**	0.16
TCA	0.49	-0.05	0.22	0.34	-0.12	0.01
TAB	0.66	0.31	0.73	0.66	0.18	0.73
mSDA	0.33	/	0	0.33	/	0
OKMSC	0.45	-0.03	0.06	0.95	0.15	0.82
MDT	0.28	0.01	0.02	0.23	0.25	0.04
MVTOT	**0.91**	0.11	**0.71**	**0.97**	0.16	**0.89**

Table 1: The performance on two synthetic datasets. For supervised benchmark methods, we use 80% labeled target samples and all labeled source data as the training set.

We compare our method with other benchmark methods, including: TCA [Pan *et al.*, 2010], TAB [Dai *et al.*, 2007], mSDA [Chen *et al.*, 2012], OKMSC [Zhang *et al.*, 2020], MDT [YANG and Gao, 2013]. From Table 1, we conclude that our method achieves the highest accuracy and NMI score among all methods. This indicates that our MVTOT can more effectively transfer knowledge from the multi-view source data to promote the prediction task on the target domain. The reason is that OT directly compares distributions of features in the source and the target domain while other methods are not as informative as the OT method. In Figure 1 (a), we use OT [Flamary and Courty, 2021] to map samples from the source domain data to the target domain and visualize them using a principle component analysis (PCA) plot. However, this mapping is meaningless since all of source domain samples are mapped to a single point (1st PC ≈ -400, 2nd PC ≈ 0). On the right hand side, by using MVTOT, we extract low-rank representation $\mathbf{H}_s$ and $\mathbf{H}_t$ for the source and the target domain respectively. By using OT to map $\mathbf{H}_s$ on the target feature space, all of source domain samples are mapped to the center of the corresponding class on the target domain. It demonstrates that MVTOT can help us to explore common low-rank structure of datasets from different sources.

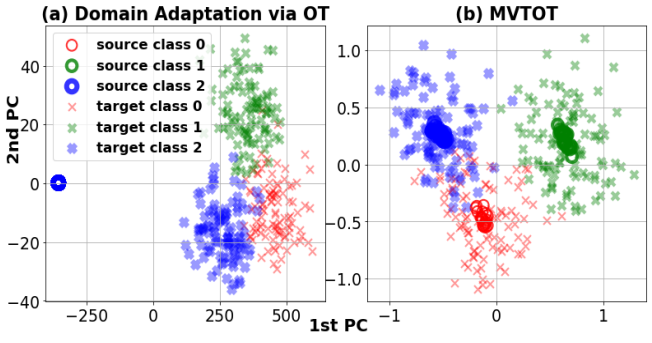

Figure 1: The left panel is the PCA visualization of transferring source samples to the target domain using OT. The right panel represents the result of using our proposed MVTOT to map source samples to the target domain. Marker shape represents domain (source, target) and marker color represents class (0, 1, 2).

5.2 TCGA Patients Stratification

TCGA contains genomic and histopathological image data for 32 cancer types. We focus on three cancer types including breast invasive carcinoma (BRCA), kidney renal papillary cell carcinoma (KIRP), and lung squamous cell carcinoma (LUSC). For each cancer type, we select early-stage (stages I and II) patients with matched gene expression data, histopathological images, and clinical outcomes. We apply the MVTOT method to integrate eigen-genes and tissue morphological features and transfer knowledge from one cancer type to the other. Eigen-genes are obtained by clustering genes into co-expressed modules and summarizing each module into an eigen-vector using singular value decomposition [Cheng *et al.*, 2017]. Finally we get 66 eigen-genes for each sample. For the tissue morphological features derived from the histopathological images, we follow the feature extraction pipeline described in [Cheng *et al.*, 2017]. Using this pipeline to extract and aggregate cell-level morphological features, we obtain a 150-dimensional feature for each patient.

For the implementation of the proposed MVTOT method, we adapt knowledge from the source domain (e.g., BRCA) to the target domain (e.g., LUSC) and represent target samples by the low-rank matrix $\mathbf{H}^{t,*}$ shown in Eq.(6). Then, We apply iCluster to cluster early-stage cancer patients based on $\mathbf{H}^{t,*}$. Finally, we test if these two groups have distinct survival times using the Kaplan-Meier (KM) estimator and log-rank test. We compare that the prognostic power of different approaches by stratifying cancer patients into two subgroups (i.e., the high and low survival risk groups) as shown in Fig. 2. In Fig. 2, *cancer 1* $\rightarrow$ *cancer 2* means that we transfer knowledge from cancer type 1 to cancer type 2 and perform stratification on cancer type 2. We can observe that most of the KM curves for low-risk and high-risk patients are separable. In Fig. 2 (a), (b), (c), (e), and (f), gaps between survival curves increasing as time goes by, indicating that MVTOT can stratify patients into groups with distinct survival rates (P-value = 0.006, 0.008, 0.005, 0.0004, and 0.006 respectively). The only insignificant (P-value = 0.0758) scenario is transferring KIRP dataset to LUSC dataset (Fig. 2 (d)). We summarize the proposed methods and ablation study results in Table 2. We observe that the P-values for MVTOT is smaller than single view TL via OT (SVTOT) in most conditions (except K $\rightarrow$ L and L $\rightarrow$ K). This means combining these two data

Methods	B→K	B→L	K→B	K→L	L→B	L→ K
Clinical	0.4569	0.0635	0.8651	0.0635	0.8651	0.4569
MVTOT	**0.0062**	**0.0078**	0.0053	0.0758	**0.0004**	0.0061
SVTOT (g)	0.0143	0.0797	0.0456	0.0899	0.0028	**0.0046**
SVTOT (i)	0.0189	0.0971	0.0174	**0.0390**	0.0007	0.0643
MVNMF	0.0106	0.0503	**0.0029**	0.0503	0.0029	0.0106

Table 2: The performance of stratifying cancer patients into high and low survival risk groups by MVTOT and its competitors. The p-values are calculated via log-rank test. $\mathbf{B}$, $\mathbf{L}$, and $\mathbf{K}$ represent breast, lung, and kidney cancers respectively. $\mathbf{B} \rightarrow \mathbf{L}$ represents transferring knowledge from breast cancer to lung cancer. P-values for clinical is obtained by using the clinical stages of cancer patients. SVTOT (g) means only use the eigen-gene data modality (g: eigen-gene, i: image features).

TL Pairs	JDA (g)	JDA (i)	TCA (g)	TCA (i)	DSAN (g)	DSAN (i)	mSDA (g)	mSDA (i)	RGrad (g)	RGrad (i)
B → K	0.212	0.968	0.128	0.986	0.080	0.526	0.064	0.294	0.791	0.533
B → L	0.492	0.032	0.248	0.032	0.774	/	0.178	0.021	0.405	0.527
K → B	0.216	0.318	0.006	0.399	0.373	0.802	0.513	0.638	0.606	0.099
K → L	0.197	0.990	0.366	0.875	0.207	0.782	0.910	0.526	0.155	0.363
L → B	0.572	0.110	0.093	0.110	0.950	0.592	0.681	0.630	0.130	0.186
L → K	0.891	0.656	0.242	0.956	0.047	0.211	0.590	0.316	0.204	0.915

Table 3: The performance of stratifying cancer patients into high and low survival risk groups by different TL benchmarks. The character in the bracket indicate what data modality we use (g: eigen-gene, i: image feature). Slash line means this method fails to produce two distinguish classes for the desired target task.

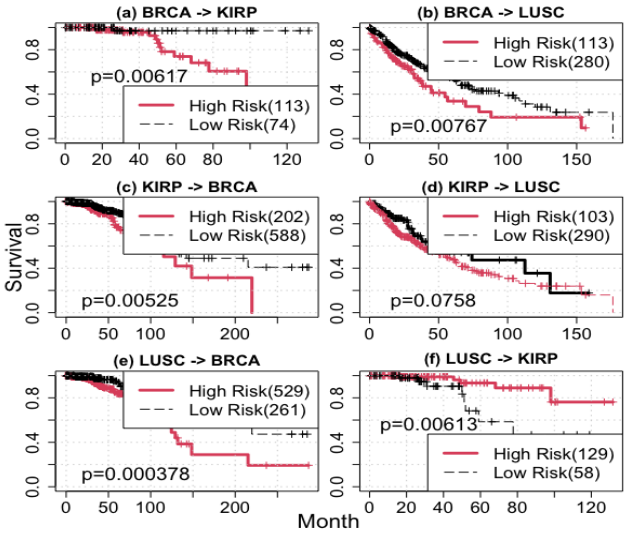

Figure 2: KM curves by applying MVTOT on pair-wise cancer types

modalities with compatible and complementary information can improve the performance of early-stage cancer stratification. We also investigate the condition when TL components in our model is removed. This multi-view NMF (MVNMF) model is inferior or equivalent to MVTOT in most conditions (B → K, B → L, L → K, L→B). It provides superior results than MVTOT in stratification breast and lung patients without borrowing knowledge from kidney cancer to cluster them. Nevertheless, its performance on the K → L task is worse than SVTOT (i). It implies that knowledge from kidney cancer may be ineffective for understanding breast cancer.

Methods	B→K	B→L	K→B	K→L	L→B	L→K
LMSC	0.137	0.728	0.050	0.728	0.050	0.137
OKMSC	0.886	0.160	0.837	0.160	0.837	0.886
DMFMV	/	/	0.190	/	0.190	/
MRAN	0.002	/	/	/	0.800	/
MDT	0.745	0.957	0.673	0.051	0.196	0.130
MDTM	0.428	0.998	0.821	0.132	0.064	0.021

Table 4: The performance of stratifying cancer patients into high and low survival risk groups by different Multi-view Learning or Multi-view TL methods. Slash line means this method fails to produce two distinguish classes on the desired target task.

We also compare our method with other benchmark methods, including: 1) TL benchmarks: JDA [Long *et al.*, 2013], TCA [Pan *et al.*, 2010], DSAN [Zhu *et al.*, 2020], mSDA [Chen *et al.*, 2012], and RevGrad [Ganin and Lempitsky, 2015]; 2) Multi-view clustering benchmarks: LMSC [Zhang *et al.*, 2018], OKMSC [Zhang *et al.*, 2020], and deepMFMV [Zhao *et al.*, 2017]; 3) Multi-view TL benchmarks: MRAN [Zhu *et al.*, 2019], MDT [YANG and Gao, 2013], and MDTM [He *et al.*, 2019]. Since these TL methods require training data on the source domain to be labeled, we use the clinical stage (stage I and stage II) as their labels. As shown in Table 2, 3, 4, MVTOT achieves superior stratification results (P-value < 0.05) than other methods. Some of these methods (i.e. JDA (i), TCA (i), mSDA (i), TCA, MRAN, and MDTM) yield significant stratification results (P-val < 0.05) on some TL pairs. However, they fail to provide plausible performances on other TL tasks. The main reason is that their labels (clinical stage I and II) are less reliable for early-stage cancer patients stratification (see the first row in Table 2). Furthermore, these benchmarks methods are highly dependent on the choice of training set and hyper-parameters. These results indicate that our method is more reliable for making clinical decisions. More numerical experiments for convergence, stability, and complexity of our methods can be found in the supplementary material.

6 Conclusion

Our proposed method MVTOT uses NMF and OT to integrate multiple data views and transfer knowledge from one data domain to another one. Since OT gives us an informative description of the distance between two empirical distributions, it helps us to compare feature spaces of the source and the target domain. Although MVTOT is an unsupervised method, it outperforms all semi-supervised benchmark methods when we evaluate them on synthetic data classification task and early-stage cancer patient stratification task. Our work demonstrated the feasibility and advantage of transferring knowledge learnt from one cancer type to improve the accuracy of prognosis to another cancer type.

Acknowledgments

This work is partially supported by IU Precision Health Initiative (ZL, JZ), NIH R01EB025018 grant (KH), and NIH R01 GM131491 (MZ).

References

[Bashiri *et al.*, 2017] Azadeh Bashiri, Marjan Ghazisaeedi, Reza Safdari, Leila Shahmoradi, and Hamide Ehtesham. Improving the prediction of survival in cancer patients by using machine learning techniques: experience of gene expression data: a narrative review. *Iranian journal of public health*, 46(2):165, 2017.

[Chen *et al.*, 2012] Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683*, 2012.

[Cheng *et al.*, 2017] Jun Cheng, Jie Zhang, Yatong Han, Xusheng Wang, Xiufen Ye, Yuebo Meng, Anil Parwani, Zhi Han, Qianjin Feng, and Kun Huang. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. *Cancer research*, 77(21):e91–e100, 2017.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[Dai *et al.*, 2007] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200, 2007.

[Ding *et al.*, 2008] Chris HQ Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55, 2008.

[Flamary and Courty, 2021] Rémi Flamary and Nicolas Courty. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[Hanahan and Weinberg, 2011] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[He *et al.*, 2019] Yiwei He, Yingjie Tian, and Dalian Liu. Multi-view transfer learning with privileged learning framework. *Neurocomputing*, 335:131–142, 2019.

[Kourou *et al.*, 2015] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[Liu *et al.*, 2018] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.

[Long *et al.*, 2013] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207, 2013.

[Lu *et al.*, 2020] Zixiao Lu, Siwen Xu, Wei Shao, Yi Wu, Jie Zhang, Zhi Han, Qianjin Feng, and Kun Huang. Deep-learning–based characterization of tumor-infiltrating lymphocytes in breast cancers from histopathology images and multiomics data. *JCO Clinical Cancer Informatics*, 4:480–490, 2020.

[Pan *et al.*, 2010] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

[Shao *et al.*, 2019] Wei Shao, Zhi Han, Jun Cheng, Liang Cheng, Tongxin Wang, Liang Sun, Zixiao Lu, Jie Zhang, Daoqiang Zhang, and Kun Huang. Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis. *IEEE Transactions on Medical Imaging*, 39(1):99–110, 2019.

[Shen *et al.*, 2012] Ronglai Shen, Qianxing Mo, Nikolaus Schultz, Venkatraman E Seshan, Adam B Olshen, Jason Huse, Marc Ladanyi, and Chris Sander. Integrative subtype discovery in glioblastoma using icluster. *PloS one*, 7(4):e35236, 2012.

[YANG and Gao, 2013] Pei Yang YANG and Wei Gao. Multi-view discriminant transfer learning. In *Proceedings of the 23rd Int. Joint Conf. Artif. Intell*, page 1848–1854, 2013.

[Zhang *et al.*, 2018] Changqing Zhang, Huazhu Fu, Qinghua Hu, Xiaochun Cao, Yuan Xie, Dacheng Tao, and Dong Xu. Generalized latent multi-view subspace clustering. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):86–99, 2018.

[Zhang *et al.*, 2020] Guang-Yu Zhang, Yu-Ren Zhou, Xiao-Yu He, Chang-Dong Wang, and Dong Huang. One-step kernel multi-view subspace clustering. *Knowledge-Based Systems*, 189:105126, 2020.

[Zhao *et al.*, 2017] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[Zhu *et al.*, 2019] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Jingwu Chen, Zhiping Shi, Wenjuan Wu, and Qing He. Multi-representation adaptation network for cross-domain image classification. *Neural Networks*, 119:214–221, 2019.

[Zhu *et al.*, 2020] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. Deep subdomain adaptation network for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.