Submitted to the Annals of Applied Statistics

BAYESIAN MULTIPLE INSTANCE CLASSIFICATION BASED ON HIERARCHICAL PROBIT REGRESSION

BY DANYI XIONG[1], SEONGOH PARK[2], JOHAN LIM[3], TAO WANG[4], AND XINLEI WANG[1,*]

[1]*Department of Statistical Science, Southern Methodist University, dxiong@smu.edu; *swang@smu.edu*

[2]*Department of Statistics, Sungshin Women's University, spark6@sungshin.ac.kr*

[3]*Department of Statistics, Seoul National University, johanlim@stats.snu.ac.kr*

[4]*Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Tao.Wang@utsouthwestern.edu*

In multiple instance learning (MIL), the response variable is predicted by features (or covariates) of one or more instances, which are collectively denoted as a bag. Learning the relationship between bags and instances is challenging because of the unknown and possibly complicated data generating mechanism regarding how instances contribute to the bag label. MIL has been applied to solve a variety of real-world problems, which have been mostly focused on supervised tasks, such as molecule activity prediction, protein binding affinities prediction, object detection, and computer-aided diagnosis. However, to date, the majority of the off-the-shelf MIL methods are developed in the computer science domain, and they focus on improving the prediction performance while spending little effort on explainability of the algorithm. In this article, a Bayesian multiple instance learning model based on probit regression (MICProB) is proposed, which contributes a significant portion to the suite of statistical methodologies for MIL. MICProB is composed of two nested probit regression models, where the inner model is estimated for predicting primary instances, which are considered as the "important" ones that determine the bag label, and the outer model is for predicting bag-level responses based on the primary instances estimated by the inner model. The posterior distribution of MICProB can be conveniently approximated using a Gibbs sampler, and the prediction for new bags can be performed in a fully integrated Bayesian way. We evaluate the performance of MICProB against 15 benchmark methods and demonstrate its competitiveness in simulation and real data examples. In addition to its capability of identifying primary instances, as compared to existing optimization-based approaches, MICProB also enjoys great advantages in providing a transparent model structure, straightforward statistical inference of quantities related to model parameters, and favorable interpretability of covariate effects on the bag-level response.

1. Introduction. In contrast to conventional machine learning where the observed response is associated with only one feature (or covariate) vector, multiple instance learning (MIL) assumes that input data are organized as a collection of bags, each containing one or more instances and each instance described by a set of features (Dietterich, Lathrop and Lozano-Pérez (1997)). In supervised problems including multiple instance classification (MIC) and multiple instance regression (MIR), a response variable, or often referred to as a label in literature, is observed at the bag level, but not individually at the instance level. The primary objective of MIL is to predict the bag label based on all its instances by learning

Keywords and phrases: Binary classification, Bayesian inference, Gibbs sampling, primary instance, weakly supervised learning.

the underlying relationship between bags and instances. Besides the voluminous data introduced by multiple instances per bag, the instance labels are not observed directly or even not clearly defined, bringing additional challenge to the learning process. Nevertheless, MIL gains great popularity as it provides an approach to solve many real-life tasks that naturally consist of multiple instance (MI) data. For example, in drug activity prediction, a molecule of different conformations (instances) is treated as a bag (Dietterich, Lathrop and Lozano-Pérez (1997)). In weakly supervised object detection, an image is viewed as a bag with multiple non-overlapping regions (instances) (Carbonneau et al. (2018)).

In the past decades, development of MIL methods for a variety of MI problems has been quite active, especially in the field of computer vision. Several works have reviewed and/or compared existing MIL methods and applications (Foulds and Frank (2010); Amores (2013); Carbonneau et al. (2018)). In Amores (2013), MIL methods for binary classification are categorized into three different paradigms, namely, instance-space (IS), bag-space (BS), and embedded-space (ES), depending on the means that a method takes to learn bag labels from instances in the bags. For IS methods, the learning process occurs at the instance level, where an instance-level classifier is trained to predict scores for instances. Bag labels are then obtained by aggregating instance prediction based on a suitable MI assumption. In this sense, IS methods focus on the characteristics of individual instances and overlook global characteristics of the entire bag. By contrast, both BS and ES methods treat each bag as a whole entity, and train a bag-level classifier utilizing the global, bag-level information. Specifically, BS methods attempt to measure the distance or similarity between each pair of bags and predict the bag labels directly using distance- or kernel-based classifiers such as k-Nearest Neighbors (kNN) and Support Vector Machine (SVM), while ES methods employ a mapping function to embed multiple instances of a bag into a single "meta" instance defined in a new feature space, and then make direct bag-level prediction using standard classifiers. In the first application of MIL, Dietterich, Lathrop and Lozano-Pérez (1997) proposed an IS method named Axis-parallel Rectangles to predict drug activity. Later, more IS methods are developed based on the standard MI assumption that a positive bag has at least one positive instance and a negative bag only has negative instances (Maron and Lozano-Pérez (1998); Zhang and Goldman (2002)). To solve applications in computer vision and text categorization which have more complex data structures, BS and ES methods are developed subsequently and become quite popular (Andrews, Tsochantaridis and Hofmann (2003); Chen, Bi and Wang (2006); Zhang et al. (2007); Cheplygina, Tax and Loog (2015); Ilse, Tomczak and Welling (2018); Wang et al. (2018a)). While MIC problems receive a lot of attention from computer scientists, statisticians are less aware of this type of problems and statistical methodologies in this field are under-developed, except for Chen et al. (2017) that proposed a MI logistic regression model (MILR) with an optional Lasso penalty term and developed an R package. Since MILR associates the bag probability to the predicted instance probabilities of being positive via the standard MI assumption, it belongs to the IS category.

In a recent MIL application on cancer detection using T-cell receptor (TCR) sequences, Xiong et al. (2021) formally discussed two types of data generation mechanisms and evaluated the performance of 16 existing MIC methods under each mechanism. The first relies on witness rate (WR), defined as the proportion of positive instances in positive bags (Carbonneau et al. (2018)). Under this so-called WR framework, only the number or proportion of the positive instances is responsible for labeling a bag as positive, which implies that more positive instances typically indicating a higher confidence of a positive bag. The second formulation is based on primary instances (PIs), a concept first introduced by Ray and Page (2001) in MIR, yet rarely mentioned in the MIC literature. It is assumed that a bag label, whether it is positive or negative, is determined by a (small) number of instances in this bag (called primary instances), while all other instances are irrelevant. Until this work, the PI

framework has not been investigated in the MIC literature. Based on numerical results from both simulation and analyses of sequencing data for multiple cancer types, the authors recommended EMD-SVM (Zhang et al. (2007)) and NSK-SVM (Gärtner et al. (2002)) for their overall better performance over other compared methods. Possibly due to the presence of abundant by-standing TCRs (corresponding to non-primary instances) naive to any antigenic process, the authors also pointed out that results from real data were much more consistent with those from simulated data under the PI framework, hence calling for new MIC methodology to be developed based on the PI framework.

In fact, other applications can also be formulated under the PI framework. For example, Wang et al. (2008) predicted aerosol optical depth from satellite measurements where they treated instances as noisy versions of the primary instance. Recently, Park et al. (2020) developed a Bayesian multiple instance regression model (BMIR) to study the relationship between tumor immune response and immunogenic neoantigens, assuming that each bag contains a single primary instance.

Motivated by Xiong et al. (2021), we develop a Bayesian MIC method based on a two-tier probit regression model (MICProB) under the PI framework. This novel method does not belong to any of the three paradigms (IS, BS, and ES) defined in Amores (2013), and adds to the suite of methodologies to address MIC problems where primary instances need to be identified. MICProB provides a fully integrated Bayesian solution that not only performs training and prediction simultaneously, but also allows for statistical inference and offers great explainability. By contrast, most existing MIC methods are algorithm-driven based on the WR framework, and so cannot offer insights about the mechanism behind data. The two statistical methods, MILR and BMIR, are both model-based, and so are similar to MICProB in terms of explainability; however, MILR is based on the standard MI assumption under the WR framework. While BMIR is based on the PI framework, it requires continuous outcomes and the assumption of a single primary instance per bag, which is rather restrictive in many applications. Also, it relies on an auxiliary random forest model to identify the primary instances and to predict the labels for new bags after the training process is done. Such a sequential approach ignores estimation uncertainty from the training step.

The remainder of this paper is organized as follows. In Section 2, we describe the proposed Bayesian model, computation, inference, and posterior-based prediction. In Section 3, we conduct simulation studies to assess the performance of our method and compare it with 15 benchmark methods, under various design configurations. In Section 4, we illustrate the usage of proposed method with a real application of cancer detection using TCR sequences and also demonstrate its high explainability in the application of modeling immunogenic neoantigens. We summarize our findings and discuss potential extensions of our method in Section 5.

2. Methods. Let B_i denote bag i containing m_i instances, and y_i denote the observed binary bag label (or outcome) for $i = 1, ..., n$, where n is the total number of observed bags (or sample size). Suppose there are d features (or covariates) that characterize each instance j. We use $X_i = (x_{ij})_{j=1}^{m_i}$ to denote the $m_i \times (d+1)$ feature matrix of B_i by stacking x_{ij}'s row-wisely, where $x_{ij} = (1, x_{ij1}, ..., x_{ijd})$ is a row vector of length $d+1$. In many practical situations, not all the instances are necessarily relevant, and there might be instances inside one bag that do not convey any information about its label. Furthermore, its own feature vector x_{ij} may help predict whether an instance is relevant or not. For each bag i, we refer to those relevant instances as its primary instances, collectively denoted by $\widetilde{\mathrm{B}}_i$. Let δ_{ij} be a latent indicator variable, with $\delta_{ij} = 1$ indicating that instance j is a primary instance of B_i and 0 otherwise.

2.1. *Model and prior specification.* By assuming primary instances of all bags are known, we first consider a probit regression setup to model the relationship between the feature vectors x_{ij}'s of B_i and the outcome y_i. Namely,

$$\begin{aligned} y_i =& \text{sign}(Z_i), \\ Z_i =& \sum_{j=1}^{m_i} \delta_{ij} x_{ij}\beta / C_i + \epsilon_i, \end{aligned}$$

where $\epsilon_i \overset{ind}{\sim} N(0,1)$ for $i = 1, ..., n$, and $\beta = (\beta_r)_{r=0}^d$ is a column vector of intercept and regression coefficients describing the covariate effects on the observed outcome variable. Further, C_i is a normalizing factor that may account for the different number of primary instances in different bags. For example, $C_i = 1$ corresponds to the sum contribution, while $C_i = |\widetilde{\mathrm{B}}_i|$ corresponds to the average contribution of $\widetilde{\mathrm{B}}_i$. In case where $\widetilde{\mathrm{B}}_i = \emptyset$, we let $Z_i = \beta_0 + \epsilon_i$. Without loss of generality, we focus on the sum contribution model. Thus, $\Pr(y_i = 1 | X_i, \beta, \delta_{i1}, ..., \delta_{im_i}) = \Phi\left(\sum_{j=1}^{m_i} \delta_{ij} x_{ij}\beta\right)$ for $C_i = 1$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Next, we model the latent primary indicator of instance j in bag i (i.e., δ_{ij}) through another probit regression model:

$$\begin{aligned} \delta_{ij} =& \text{sign}(U_{ij}), \\ U_{ij} =& x_{ij} b + e_{ij}, \end{aligned}$$

where $e_{ij} \overset{ind}{\sim} N(0,1)$ for $j = 1, ..., m_i$ and $i = 1, ..., n$. Here, $b = (b_r)_{r=0}^d$ is the column vector of intercept and coefficients describing the covariate effects on the instance status (primary vs. non-primary). Similarly, we have $\Pr(\delta_{ij} = 1 | x_{ij}, b) = \Phi(x_{ij} b)$. In each probit model, Z_i's or U_{ij}'s are latent variables which are introduced to make the subsequent Markov Chain Monte Carlo (MCMC) algorithm for posterior sampling become more convenient, due to the data augmentation technique proposed for binary response data in Albert and Chib (1993). The variances of ϵ_i and e_{ij} are both fixed at 1 for model identifiability.

For the above two-tier probit regression model, we employ (conditional) conjugate priors, which are commonly used in the Bayesian regression literature to achieve convenient posterior sampling. The prior for the regression coefficients β is specified as $\beta | \mu_\beta, \Sigma_\beta \sim \text{MVN}(\mu_\beta, \Sigma_\beta)$. It is routine to set $\mu_\beta = (0, 0, ..., 0)$. Similarly, we assign $b | \mu_b, \Sigma_b \sim \text{MVN}(\mu_b, \Sigma_b)$ and set $\mu_b = (0, 0, ..., 0)$. For Σ_β and Σ_b, we adopt the hyperparameter values suggested by Polson, Scott and Windle (2013) and employ a diagonal matrix with $(16, 4, ..., 4)$ on the diagonal entries. Figure 1(a) describes the Bayesian hierarchical model structure.

2.2. *Posterior computation.* Let $y = (y_i)_{i=1}^n$ be a column vector of length n and $X = (X_i)_{i=1}^n$ be the collection of covariate matrices from all bags used for model fitting (i.e., the training cohort). Let Δ and U be a $(\sum_{i=1}^n m_i) \times 1$ column vector of binary indicators δ_{ij}'s and their corresponding latent variables U_{ij}'s, for $i = 1, ..., n$ and $j = 1, ..., m_i$, respectively. Similarly, let $Z = (Z_i)_{i=1}^n$ denote a $n \times 1$ column vector of latent variables Z_i's associated with the bag labels y_i's, for $i = 1, ..., n$. Let $\Theta = (\beta, b, \Delta, Z, U)$ denote the collection of all model parameters and latent variables involved. With hyper-parameters μ_β, Σ_β, μ_b and Σ_b specified, the full probability model is given by

$$p(y, \Theta | X) = p(y|Z) \times p(Z | X, \Delta, \beta) \times p(\Delta | U) \times p(U | X, b)$$

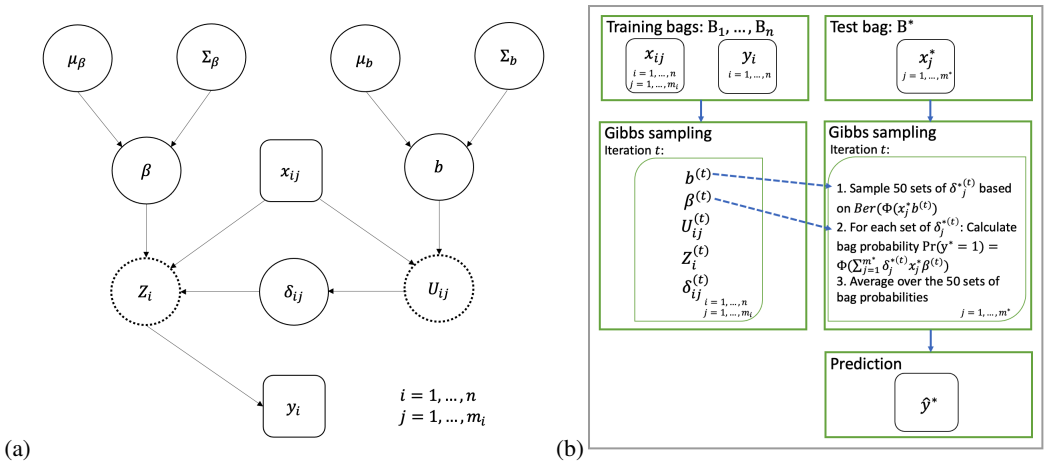

FIG 1. *(a) Bayesian hierarchical model structure of MICProB. Observed data, including instances x_{ij}'s and bag labels y_i's, are showed in square boxes. Latent variables, including Z_i's for response variables and U_{ij}'s for indicators of primary instances, are showed in dashed circles. Hyper-parameters include μ_β, Σ_β, μ_b, and Σ_b. (b) Workflow of MICProB. Left panels explain the model fitting process on training bags and right panels describe prediction steps for a new bag.*

$$
\begin{aligned}
&\times p(\beta|\mu_\beta, \Sigma_\beta) \times p(b|\mu_b, \Sigma_b) \\
&= \prod_{i=1}^{n} \left\{ p(y_i|Z_i) \cdot p(Z_i|x_{\delta_i}, \beta) \cdot \left[\prod_{j=1}^{m_i} p(\delta_{ij}|U_{ij}) \cdot p(U_{ij}|x_{ij}, b) \right] \right\} \\
&\times p(\beta|\mu_\beta, \Sigma_\beta) \times p(b|\mu_b, \Sigma_b).
\end{aligned}
$$

We use MCMC to draw random samples from the joint posterior distribution $p(\Theta|X, y) \propto p(y, \Theta|X)$. One advantage of the proposed modeling is that the conditional posterior distribution of each parameter (or latent variable) given all others, becomes tractable as a known family of distributions, as detailed below.

- $\beta|\cdots \sim \text{MVN}(m_\beta, V_\beta)$, where

$$
\begin{aligned}
m_\beta &= (\Sigma_\beta^{-1} + X_\delta^T X_\delta)^{-1}(\Sigma_\beta^{-1}\mu_\beta + X_\delta^T Z), \\
V_\beta &= (\Sigma_\beta^{-1} + X_\delta^T X_\delta)^{-1}.
\end{aligned}
$$

  Here,

$$
X_\delta = \begin{pmatrix} x_{\delta_1} \\ x_{\delta_2} \\ \vdots \\ x_{\delta_n} \end{pmatrix} = \begin{pmatrix} 1 & \sum_{j=1}^{m_1} \delta_{1j} x_{1j1} & \cdots & \sum_{j=1}^{m_1} \delta_{1j} x_{1jd} \\ 1 & \sum_{j=1}^{m_2} \delta_{2j} x_{2j1} & \cdots & \sum_{j=1}^{m_2} \delta_{2j} x_{2jd} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \sum_{j=1}^{m_n} \delta_{nj} x_{nj1} & \cdots & \sum_{j=1}^{m_n} \delta_{nj} x_{njd} \end{pmatrix}
$$

  is a $n \times (d+1)$ covariate matrix, where the i-th row is formed by the primary instances of bag i, for $i = 1, ..., n$.

- $b|\cdots \sim \text{MVN}(m_b, V_b)$, where

$$
\begin{aligned}
m_b &= (\Sigma_b^{-1} + X^T X)^{-1}(\Sigma_b^{-1}\mu_b + X^T U), \\
V_b &= (\Sigma_b^{-1} + X^T X)^{-1}.
\end{aligned}
$$

Let $I(\cdot)$ be an indicator function that equals 1 if the condition inside the parentheses is satisfied (0 otherwise). Latent variables Z_i's and U_{ij}'s can be sampled from truncated normal distributions, respectively:

- $Z_i|\cdots \sim \begin{cases} \mathrm{N}(g(x_{\delta_i},\beta),1)\cdot I(Z_i>0) & \text{if } y_i=1 \\ \mathrm{N}(g(x_{\delta_i},\beta),1)\cdot I(Z_i\le 0) & \text{if } y_i=0 \end{cases}$, where x_{δ_i} corresponds to the i-th row of X_δ and $g(x_{\delta_i},\beta)=\beta_0+\sum_{j=1}^{m_i}\delta_{ij}\sum_{r=1}^{d}x_{ijr}\beta_r$, for $i=1,...,n$.
- $U_{ij}|\cdots \sim \begin{cases} \mathrm{N}(h(x_{ij},b),1)\cdot I(U_{ij}>0) & \text{if } \delta_{ij}=1 \\ \mathrm{N}(h(x_{ij},b),1)\cdot I(U_{ij}\le 0) & \text{if } \delta_{ij}=0 \end{cases}$, where $h(x_{ij},b)=b_0+\sum_{r=1}^{d}x_{ijr}b_r$, for $i=1,...,n$ and $j=1,...,m_i$.

Lastly, the binary indicator for primary instance follows a Bernoulli distribution conditioning on other parameters:

- $\delta_{ij}|\cdots \sim \mathrm{Ber}\left(\frac{A\Phi(h(x_{ij},b))}{A\Phi(h(x_{ij},b))+B[1-\Phi(h(x_{ij},b))]}\right)$, where

$$A=\exp\left\{-\frac{1}{2}\left(Z_i-\beta_0-\sum_{j'\neq j}^{m_i}\delta_{ij'}\sum_{r=1}^{d}x_{ij'r}\beta_r-\sum_{r=1}^{d}x_{ijr}\beta_r\right)^2\right\},$$

$$B=\exp\left\{-\frac{1}{2}\left(Z_i-\beta_0-\sum_{j'\neq j}^{m_i}\delta_{ij'}\sum_{r=1}^{d}x_{ijr'}\beta_r\right)^2\right\},$$

and $h(x_{ij},b)=b_0+\sum_{r=1}^{d}x_{ijr}b_r$, for $i=1,...,n$ and $j=1,...,m_i$.

The above analytical forms allow us to utilize a Gibbs sampler to easily draw samples from $p(\Theta|X,y)$ after proper convergence of the MCMC algorithm.

2.3. *Posterior inference.* Suppose we run the Gibbs sampler for T iterations after the burn-in period. Point estimates of quantities of interest are made based on posterior means (or medians/modes). For example, the covariate effects on the response variable are estimated by $\hat{\beta}=\frac{1}{T}\sum_{t=1}^{T}\beta^{(t)}$, where $\beta^{(t)}$ is the draw of β at iteration t. Estimation of uncertainty is quantified using Bayesian credible intervals (highest posterior density intervals or equal-tailed intervals). This enables us to readily conduct statistical inference about such quantities or their functions and interpret relevant results.

To identify primary instances of a bag in the training cohort, we calculate the posterior inclusion probability:

$$\widehat{\pi}_{ij}=\frac{1}{T}\sum_{t=1}^{T}\delta_{ij}^{(t)},\ j=1,...,m_i,$$

where $\delta_{ij}^{(t)}$ is defined similarly as $\beta^{(t)}$. The instance j in bag i is primary if $\widehat{\pi}_{ij}>\theta$, where θ is a cutoff determined by controlling the Bayesian false discovery rate (FDR) (Newton et al. (2004)) on all instances from the entire dataset. For a given θ, the estimated FDR is

$$\widehat{\mathrm{FDR}}(\theta)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(1-\widehat{\pi}_{ij})\cdot I(\widehat{\pi}_{ij}>\theta)}{\sum_{i=1}^{n}\sum_{j=1}^{m_i}I(\widehat{\pi}_{ij}>\theta)}.$$

In case when the denominator is zero, we define the FDR by 0. We choose the value of θ so that $\widehat{\mathrm{FDR}}(\theta)\le\kappa$, where $\kappa\in(0,1)$ is a pre-specified FDR we aim to control.

The above FDR control method can be used to identify positive bags in the training cohort as well, where the probability that bag i is positive, $\pi_i \equiv \Pr(y_i = 1 \mid X_i)$, can be estimated by

$$\widehat{\pi}_i = \frac{1}{T}\sum_{t=1}^{T} \Phi\left(\sum_{j=1}^{m_i} \delta_{ij}^{(t)} x_{ij}\beta^{(t)}\right).$$

2.4. *Prediction for new bags.* We evaluate the performance of the proposed MICProB on the test bags using posterior-based prediction. Given a new bag B^* with a collection of m^* instances described by the feature matrix $X^* = (x_j^*)_{j=1}^{m^*}$ with $x_j^* \equiv (1, x_{j1}^*, ..., x_{jd}^*)$, we predict the label of B^* (i.e., y^*) based on the probability $\pi^* \equiv \Pr(y^* = 1 \mid X^*, \beta, \Delta^*) = \Phi\left(\sum_{j=1}^{m^*} \delta_j^* x_j^* \beta\right)$, where $\Delta^* = (\delta_j^*)_{j=1}^{m^*}$. This can be computed from the joint distribution of (Δ^*, Θ) given observed data: $p(\Delta^*, \Theta | X^*, X, y) = p(\Delta^* | \Theta, X^*) p(\Theta | X, y)$. To sample from $p(\Delta^*, \Theta | X^*, X, y)$, we sequentially draw (i) Θ from the joint posterior distribution $p(\Theta | X, y)$ and (ii) Δ^* from $p(\Delta^* | \Theta, X^*)$, that is $\delta_j^* \sim \mathrm{Ber}\left(\Phi\left(x_j^* b\right)\right)$ for $j = 1, \ldots, m^*$, where b is obtained in step (i).

Thus, our model takes an integrated approach to produce posterior samples and predict the labels for new bags within the same Gibbs sampling iteration, as described in Figure 1(b). Recall that BMIR (Park et al. (2020)), the only Bayesian method based on the PI framework, has to rely on an auxiliary frequentist model to make prediction for new bags. To reduce the uncertainty in prediction, we sample R replicates of δ^* based on $\beta^{(t)}$ and $b^{(t)}$ within iteration t. In this illustration, $R = 50$. For each replicate $\delta^{*(r,t)}$, we calculate the probability of being positive for each test bag. An averaged probability for a new bag is then computed across all replicates and iterations as

$$\widehat{\pi}^* = \frac{1}{TR}\sum_{t=1}^{T}\sum_{r=1}^{R} \Phi\left(\sum_{j=1}^{m_i} \delta_j^{*(r,t)} x_j^* \beta^{(t)}\right).$$

Similarly, we can compute $\widehat{\pi}_j^* = \frac{1}{TR}\sum_{t=1}^{T}\sum_{r=1}^{R} \delta_j^{*(r,t)}$ to estimate the probability that instance j in B^* is a primary instance. The FDR control method can be used to identify both primary instances and positive bags as in Section 2.3.

3. Simulation. We conduct simulation to illustrate the performance of the proposed MICProB and compare it with 15 existing MIC methods. We consider various simulated scenarios by varying key factors that may affect the performance, including sample size n, bag size m (i.e., the number of instances in a bag), number of features d, and mean proportion of primary instances denoted as $\overline{\mathrm{PPI}}$ (i.e., the total number of primary instances divided by the total number of instances across all bags). We also conduct sensitivity analysis using data generated from the WR framework. For simplicity, we assume different bags in one dataset have a constant number of instances. For each generated dataset, we initialize the parameters of MICProB with random values and run 100,000 iterations of the Gibbs sampler and discard the first half as burn-ins. Standard diagnostic techniques (Gelman et al. (2013)) are used to detect the convergence of our MCMC algorithm.

3.1. *Benchmark methods.* Prior to MICProB, many algorithm-based solutions to MIC have been proposed, which, as mentioned in the introduction, can be categorized as instance-space (IS), bag-space (BS), or embedded-space (ES) methods (Amores (2013)), based on how they extract and exploit information from the MI data. For the purpose of performance

evaluation in various simulated settings, we compare MICProB against 15 benchmark methods, including seven IS methods: EMDD, MI-SVM, mi-SVM, MILR, SI-SVM, SI-kNN, MILBoost (Zhang and Goldman (2002); Andrews, Tsochantaridis and Hofmann (2003); Ray and Craven (2005); Chen et al. (2017); Carbonneau et al. (2018); Babenko et al. (2008)); five BS methods: CkNN, NSK-SVM, EMD-SVM, miGraph, MInD (Wang and Zucker (2000); Gärtner et al. (2002); Zhang et al. (2007); Zhou, Sun and Li (2009); Cheplygina, Tax and Loog (2015)); and three ES methods: MILES, BoW, CCE (Chen, Bi and Wang (2006); Zhou and Zhang (2007); Amores (2013)). We also refer readers to Xiong et al. (2021) for more detail on each selected benchmark method.

We use the MATLAB "MILSurvey" toolbox, made available by Carbonneau et al. (2018), to implement all benchmark methods except for MILR which is implemented via the R package `milr` (Chen et al. (2017)). For each of the methods implemented, the default setting is used in our evaluation. For example, for SVM-based methods, we use the default kernel function. For model tuning, default ranges of values for hyper-parameters are used. Each selected IS method predicts bag labels from the predicted instance labels based on the standard MI assumption (Dietterich, Lathrop and Lozano-Pérez (1997)). For MILR, it iterates 500 steps for the EM algorithm. Since feature selection is beyond the scope of this paper, we do not impose the LASSO penalty term, which is specified by default in MILR.

3.2. *Settings.* Our proposed MICProB is the only method developed based on the PI framework while all the benchmark methods considered are designed under the WR framework. As pointed out by Xiong et al. (2021), under the PI framework, IS methods are less proper as the non-primary instances introduce irrelevant information to the bag; BS and ES methods may still be suitable for bag classification, as the information of primary instances of each bag could be utilized by the flexible embedding/summary behaviors of these methods. Thus, it would be interesting to compare their performance with that of MICProB using data generated from the PI framework. For instance j in bag i, each covariate x_{ijr} is independently generated from a standard normal distribution, and the primary status indicator δ_{ij} is generated from a Bernoulli distribution $\text{Ber}(p_{ij})$, with $p_{ij} = \Phi\left(b_0 + \sum_{r=1}^{d} x_{ijr} b_r\right)$, where b_0 and b_r for $r = 1, ..., d$ are regression coefficients in the probit regression model for δ_{ij}. Next, we simulate the bag label y_i from $\text{Ber}(\pi_i)$, with $\pi_i = \Phi\left(\beta_0 + \sum_{j=1}^{m} \delta_{ij} \sum_{r=1}^{d} x_{ijr}\beta_r\right)$, where β_0 and β_r for $r = 1, ..., d$ are regression coefficients associated with the probit model for π_i. We adjust the intercepts b_0 to vary $\overline{\text{PPI}}$. We set $b_j = 1$ for $j = 1, ..., d$, $\beta_0 = 0.5$, $\beta_j = 1$ for $j = 1, ..., \lceil d/2 \rceil$ and $\beta_j = -0.5$ for $j = \lceil d/2 \rceil + 1, ..., d$. We vary $n \in \{150, 300, 450, 600\}$; $m \in \{5, 10, 20, 40\}$; $d \in \{2, 15, 30, 45\}$; and $\overline{\text{PPI}} \in \{0.1, 0.4, 0.6, 0.9\}$ and assess their influence on performing multiple instance classification. We employ the vary-one-at-a-time strategy to reduce the work load in this simulation; that is, we vary one and only one factor each time while fixing the others at the basic setting in which $n = 300$, $m = 10$, $d = 30$, and $\overline{\text{PPI}} = 0.4$. We independently generate 50 replication datasets under each of the settings. The performance is measured by evaluating the area under the Receiver Operating Characteristic curve (AUROC) on 300 test bags in each replicate.

To further examine the robustness of MICProB, we generate data from the WR framework as well. Following Xiong et al. (2021), we generate MI datasets by varying WR $\in \{0.05, 0.25, 0.5, 0.75, 1\}$, where WR $= 0.05$ represents the scenario where there is only one positive instance in each bag. For more details, we refer readers to Section 4.1 of Xiong et al. (2021).

3.3. *Results.* Figure 2 compares the performance of MICProB with 15 benchmark methods for bag classification in various simulated scenarios under the PI framework. Each line is

the average AUROC (%) calculated from 50 replications. Across all scenarios, MICProB works best, followed by NSK-SVM, EMD-SVM, and MInD, all from the BS category. MILES from the ES category and MILR from the IS category are middle performers. All the remaining IS methods, miGraph and CkNN from the BS category, do not yield satisfactory performance in most scenarios, with AUROC below 70%, which is slightly better than random guessing. Notably, MILR is the best performing IS method among the selected ones, which shows some promise for statistical model-based approaches in tackling MI problems.

Next, we discuss the impact of each factor on the performance of MICProB and benchmark methods, excluding the bottom performers, which steadily have poor performance. Firstly, increasing the sample size n tends to improve the performance. Secondly, as the bag size m increases, while the performance of MICProB shows a non-monotone pattern, the other methods are not sensitive to the change. In particular, the average AUROC of MICProB is the highest (above 80%) regardless of the bag size. Thirdly, as the number of features d increases, MICProB has improved performance, while all the benchmark methods show an opposite pattern. We note that the signal in data generated from the PI framework becomes stronger in general as d goes up. Among all, only MICProB can capture the stronger signal because it is the only method designed for the PI framework. Lastly, when $\overline{\text{PPI}}$ increases, most methods show higher AUROC by capturing the increased amount of useful information. Individual performance of each method evaluated on 50 replicates at different values of $\overline{\text{PPI}}$ is shown using box plots in Figure 3. We observe that MICProB performs significantly better, with median AUROC greater than 80%, than all the benchmark methods, when there are only 10% primary instances on average in each bag. As $\overline{\text{PPI}}$ increases, the performance of MICProB steadily improves and the spread (i.e., the width of the box) becomes narrower, while the performance for many other methods is more variable across different replications. More detail on individual performance of each method on 50 replicates by varying the sample size, bag size, and the number of features, are shown using box plots in Section S1.1 (Figures S1-S3) of Supplementary Material. In general, MICProB produces consistently better performance with narrow spread.

Figure 4 shows the performance of MICProB for identifying primary instances. Each boxplot is generated by AUROC values calculated from 50 replicates. In many scenarios, the proposed method works quite well, with AUROC greater than 95%. It only dips below 90% in only a few settings. Next, we discuss how each factor affects the performance of MICProB. As in our observations made for bag classification, the performance for instance classification tends to improve with an increased sample size (n) or feature size (d) or mean proportion of primary instances ($\overline{\text{PPI}}$). Among these three factors, it seems that d has a larger impact than n or $\overline{\text{PPI}}$. Secondly, the performance decreases with an increased bag size (m), which could be due to noisy signal induced by more non-primary instances in larger bags.

For robustness checking, Figure 5 compares the performance of MICProB with the 15 benchmark methods for bag classification using data generated from the WR framework. As we expect, MICProB is no longer the best method when WR ≤ 0.25, especially when WR $= 0.05$, where in each bag there is only one positive instance. This is because the presence of a large number of negative instances makes it difficult for MICProB to identify any primary instances assumed under the PI framework. Still, when WR $= 0.25$, it is much better than a few methods that are specially designed under the WR framework including MILBoost, EMDD, SI-kNN, CCE, and BoW. As WR further increases, MICProB has as good performance as other benchmark methods, with AUROC very close to 1. Individual performance of each method on 50 replicates at different values of WR is shown using box plots in Section S1.1 (Figures S4) of Supplementary Material. Figure S4 clearly shows that as WR increases, the performance of MICProB steadily improves and the spread becomes narrower, demonstrating an amazingly high degree of robustness.

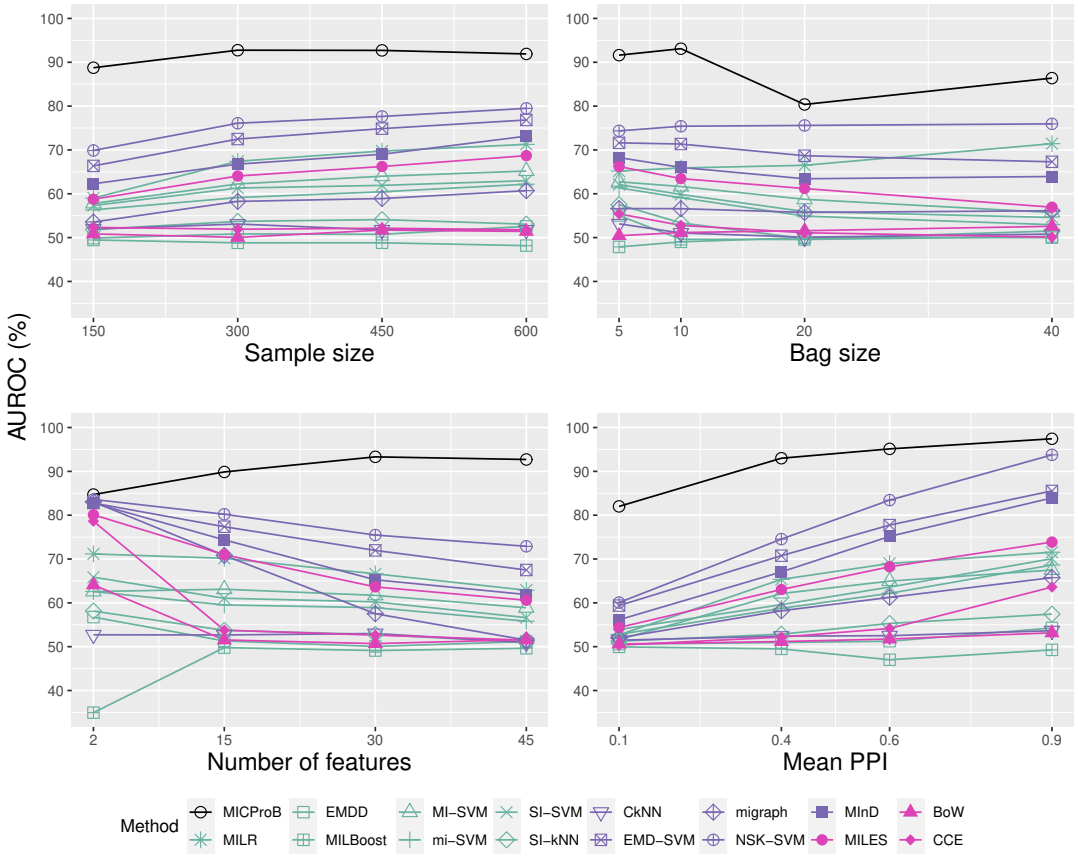

FIG 2. *Simulation evaluation: average AUROC (%) for bag classification using different MIL methods, evaluated on simulation scenarios under the PI framework each with 50 replicates. We vary the sample size, bag size, number of features, and mean PPI, and report the results in the four panels, respectively. Benchmark methods are distinguished by color (green: IS methods; purple: BS methods; magenta: ES methods).*

We provide the runtime information for MICProB and the 15 benchmark methods in Section S1.2 (Figure S5) of Supplementary Material. Due to sequential updates of the MCMC algorithm that cannot be easily parallelized, MICProB is not as computationally efficient as most competing methods. Nevertheless, for data generated from the PI framework, MICProB outperforms all the benchmark methods including MILR, the only other regression-based approach, and enables statistical inference that optimization-based methods do not provide. For data from the WR framework, it appears that MICProB is not overly sensitive and still has strong performance as long as WR is not too low, compared to the top performers.

4. Real data examples. We present two data examples: the first is on cancer detection using sequencing data from The Cancer Genome Atlas (TCGA), to evaluate the performance of our proposed MICProB against benchmark methods on detecting various types of cancer; the second is on modeling immunogenic neoantigens, to illustrate the explainability of MICProB.

4.1. *Cancer detection using T-cell receptor sequences.* Early diagnosis, especially for aggressive cancer types, is crucial for patients to receive appropriate treatments for best pos-

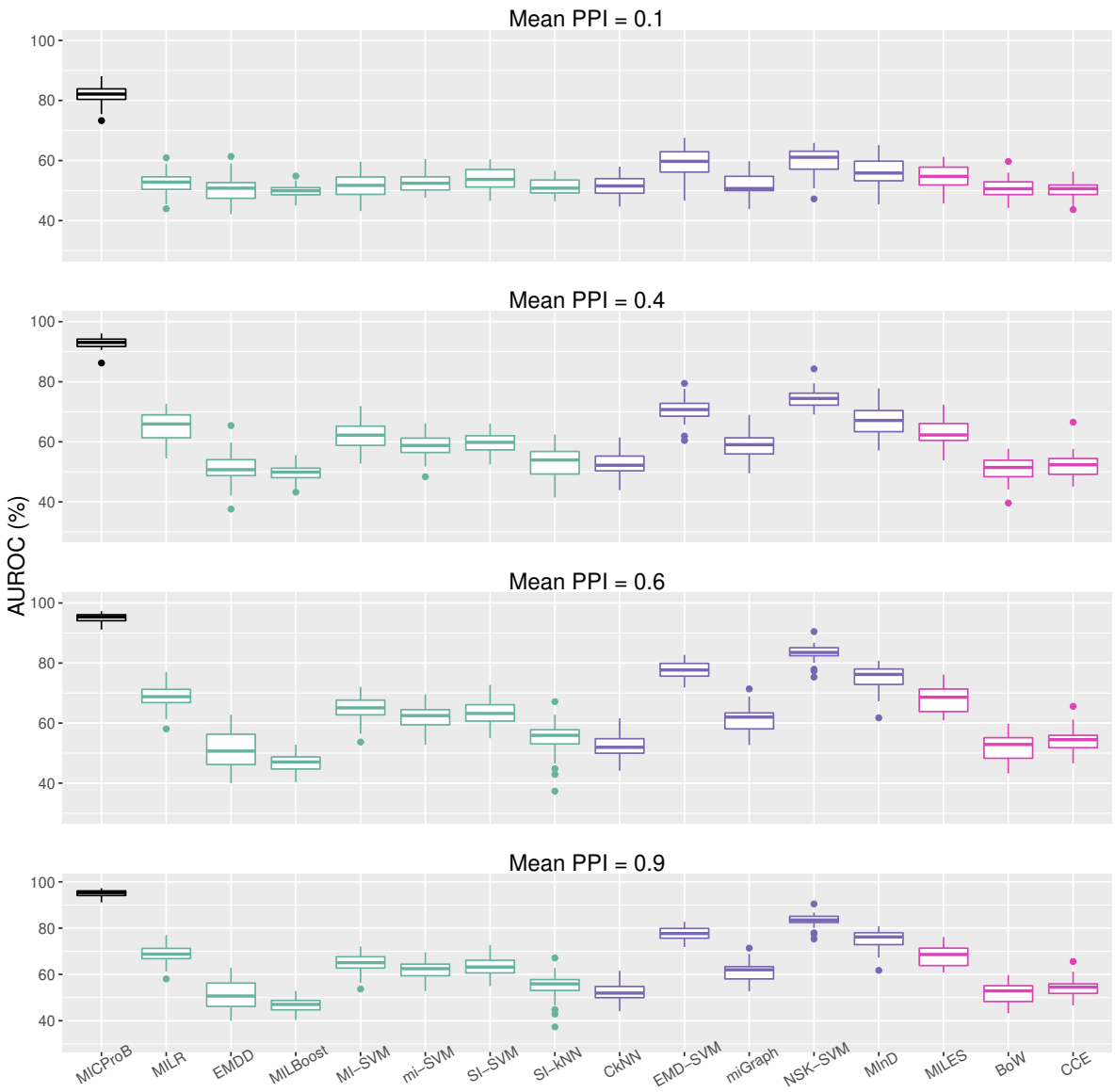

FIG 3. *Simulation evaluation: AUROC (%) for bag prediction using different MIL methods. We vary the mean PPI and report results for each of the methods in each setting under the PI framework using a box plot (based on 50 replicates). Benchmark methods are distinguished by color (green: IS methods; purple: BS methods; magenta: ES methods).*

sible prognosis. Various tools have been developed to facilitate cancer screening, which, however, are less ideal for detecting certain types of cancer (Clarke-Pearson (2009); Byers and Rudin (2015); Singhi et al. (2019)). One possible new approach for cancer detection is to examine the TCR sequences in peripheral blood of patients, as TCRs are used by the T cells to target and initiate the destruction of tumor cells, and may contain critical information regarding tumor progression in the human body. This approach also has the advantage of being non-invasive as it requires blood samples. The problem can be formulated into the MIL framework by treating each patient as a bag with voluminous TCR data (instances).

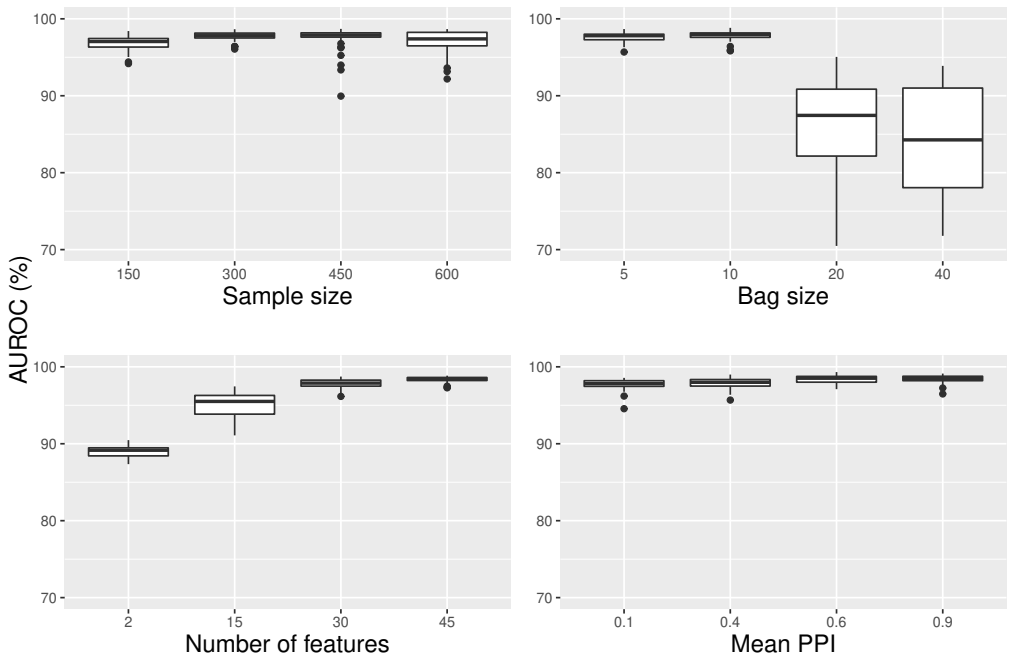

FIG 4. *Simulation evaluation: AUROC (%) for identifying primary instances using MICProB. We vary the sample size, bag size, number of features, and mean PPI, and report results for each of the methods in each setting under the PI framework using a box plot (based on 50 replicates) in the four panels, respectively. Note that all benchmark methods do not offer the functionality of identifying primary instances.*

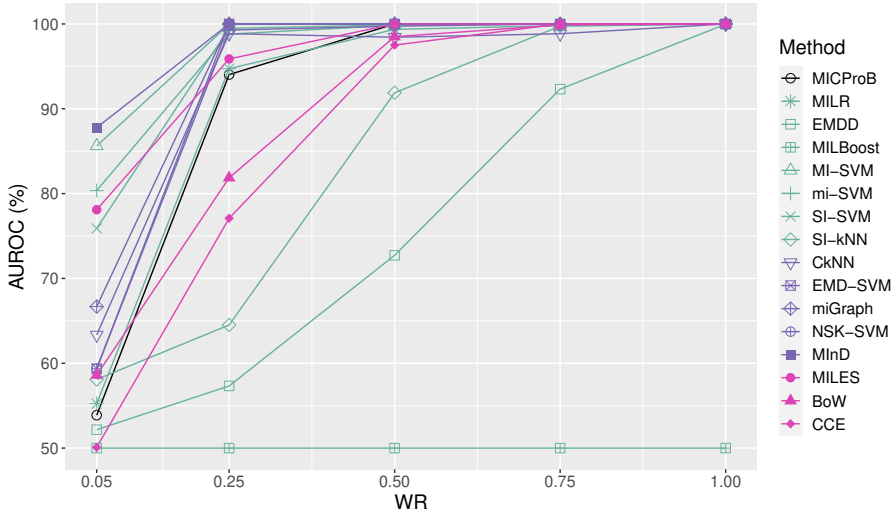

FIG 5. *Simulation evaluation for robustness checking: average AUROC (%) for bag classification using different MIL methods, evaluated on simulation scenarios under the WR framework each with 50 replicates by varying WR. Benchmark methods are distinguished by color (green: IS methods; purple: BS methods; magenta: ES methods).*

Aiming to comprehensively explore genomic changes involved in human cancer, TCGA collected and analyzed tissue samples from patients of over thirty cancer types and obtained genomic data for each sample using next-generation sequencing techniques, such as RNA-sequencing, whole exome-sequencing, etc. We use MiTCR (Bolotin et al. (2013)), a commonly used TCR reconstruction software to reconstruct TCRs from the RNA-sequencing data. MiTCR also records the number (abundance) of each unique TCR in each sample (bag). We exclude TCRs whose abundance is 1, because they are most likely the ones that have not been exposed to any antigens.

Under the MIL framework, each sample is considered as a bag consisting of TCR sequences (instances) represented by text strings of amino acids. In order to make it convenient for MIL methods to utilize the physicochemical properties of TCRs, we embed each TCR sequence into a d-dimensional numeric vector using a deep learning auto-encoder, which has been systematically validated in our previous work (Zhang et al. (2021); Lu et al. (2021)). In this study, each instance is described by 31 features. The first 30 dimensions represent the embedded TCR and the last feature is the log-transformed abundance for each TCR sequence appeared in each bag.

We apply MICProB to tissue samples of ten cancer types in the TCGA database, including skin cutaneous melanoma (SKCM), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), breast invasive carcinoma (BRCA), stomach adenocarcinoma (STAD), ovarian serous cystadenocarcinoma (OV), thymoma (THYM), and esophageal carcinoma (ESCA) (Network (2013); Lambrechts et al. (2018); Liu et al. (2018)), to illustrate its utility on distinguishing cancer patients from healthy individuals via TCRs. These cancer types are selected as they have reasonably large sample sizes (i.e., the number of normal + tumor tissue samples) and bag sizes (i.e., the number of TCRs in each sample).

In the TCGA data, the number of positive bags (tumor samples) is much greater than that of negative bags (normal tissue samples), as TCGA is mainly focused on studying cancer patients. To adjust for the imbalanced TCGA data (more positive bags than negative bags), we randomly sample from positive bags so that the resulting dataset only includes a subset of positive bags for each cancer type. Furthermore, we combine all normal tissue samples available from more than 30 cancer types in the TCGA data to increase the number of negative bags to 405. Mixing negative bags across datasets for different cancer types is reasonable because the characteristics of normal tissue samples should be similar across patients.

We randomly sample about 50% of the 405 negative bags (i.e., 202 normal tissues samples) to reduce the computation time. For each of the selected cancer types, we create a balanced dataset with 50% positive and 50% negative bags, as advised in He and Garcia (2009) that it is often preferred to apply machine learning methods to balanced data. As a result, for DLBC, THYM, and ESCA, due to a small number of positive bags, the sample sizes are 90, 216, and 332, respectively; for each of the remaining cancers, the total sample size is 404. Figure 6 shows the number of instances for selected cancer types after pre-processing, reflecting a more realistic situation in real data that the number of instances varies across bags. We also observe for each of the ten cancer types, the distribution of bag size is severely right-skewed, where most bags have a relatively small number of instances but a few can have many more instances. We standardize input variables so that they all have zero mean and unit standard deviation.

For MICProB model training and validation, we employ a 10-fold cross-validation (CV) procedure. We run the derived Gibbs sampler for $100,000$ iterations and discard the first half as burn-ins. We illustrate convergence diagnostics for MICProB using five independent MCMC chains in Section S2.1 (Figures S6 and S7) of Supplementary Material. Prediction for bags in the held-out fold is performed in an integrated manner, as shown in Figure 1 (b).

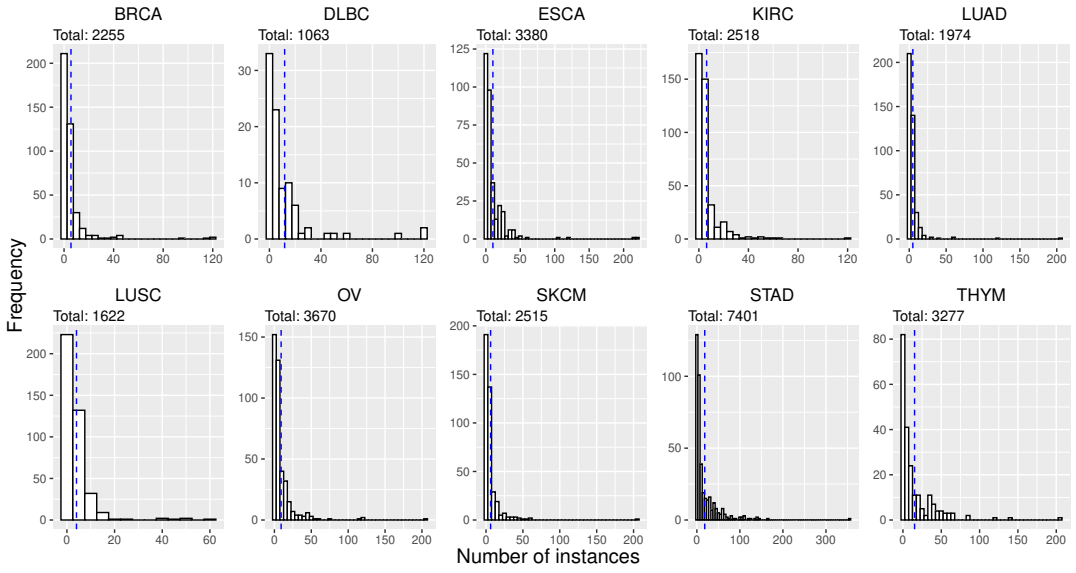

FIG 6. *TCGA data: the number of instances for selected cancer types. Blue dashed line indicates sample mean.*

Further, in this study, we take the average of the primary instances to measure their contribution to the bag. For benchmark methods, a nested cross-validation (CV) procedure (Cawley and Talbot (2010); Carbonneau et al. (2018)) is deployed, in which the model is tuned (i.e., the hyper-parameters are optimized over a range of values) in the inner layer CV and the performance of the fitted model is evaluated in the outer layer CV.In our implementation, both inner and outer layers have ten folds. We calculate the average performance from CV in terms of AUROC of each method.

Table 1 shows the performance of each method by cancer type. We only include seven cancers with average AUROC across all methods greater than 60% and exclude the other three (BRCA, LUAD, and LUSC) for which none of themethods works adequately. For each cancer type, the performance of MICProB is always (much) higher than the average. More importantly, MICProB works best in four (KIRC, SKCM, DLBC, and THYM) out of seven cancers. Notably, KIRC and SKCM are well known immunogenic cancer types with high levels of T-cell infiltration (Wang et al. (2018b)). MICProB achieves higher performance than the benchmark methods in the presence of bystander effects (Iwahori et al. (2015); Whiteside et al. (2018)), that is, the strong T-cell activation in these cancers may have caused infiltration of both abundant tumor-specific and non-specific T cells in the tumor, creating additional difficulty for MIL to distinguish tumor versus normal samples. As a result, with average AUROC at 78.3% across all cancers, MICProB gives the best performance compared to benchmark methods, followed by EMD-SVM, with average AUROC at 78.1%, and NSK-SVM, with average AUROC at 76.3%. For THYM, the performance of MICProB (AUROC of $(87.7 \pm 2.2)\%$) is significantly better than the second best method MInD (AUROC of $(84.7 \pm 0.6)\%$). For STAD and ESCA, the performance of MICProB also stays in the upper range. Lastly, the performance of the methods depends on cancer type. For STAD, except EMDD, MILBoost, and CkNN, all other methods can achieve AUROC at least 75%. For KIRC and SKCM, while MICProB performs the best, the average performance across all methods is below 70%.

4.2. *Modeling immunogenic neoantigens.* We present another data example of modeling immunogenic neoantigens to demonstrate the high explainability of MICProB. Neoantigens

		KIRC	SKCM	DLBC	ESCA	OV	THYM	STAD
	Overall	63.2	62.0	66.5	71.3	72.1	75.2	77.8
MICProB	**78.3**	**68.1** (2.5)	**69.6** (2.2)	**82.2** (5.1)	78.2 (3.7)	76.7 (3.0)	**87.7**(2.2)	85.6 (1.5)
MILR	74.7	66.1 (4.0)	66.0 (2.1)	64.4 (7.5)	80.5 (3.5)	78.7 (1.6)	81.0 (2.4)	86.0 (2.9)
EMDD	60.1	54.4 (4.1)	50.1 (3.6)	55.3 (4.3)	67.5 (4.2)	57.8 (2.3)	65.0 (5.7)	70.3 (6.5)
MILBoost	51.3	56.5 (2.4)	48.9 (1.6)	47.0 (4.6)	49.9 (1.3)	55.1 (4.1)	49.7 (3.0)	51.9 (0.6)
MI-SVM	72.3	66.2 (1.5)	64.6 (1.4)	69.6 (3.5)	72.9 (1.3)	72.6 (1.4)	80.2 (1.3)	79.9 (1.2)
mi-SVM	71.2	66.1 (1.6)	66.5 (1.7)	66.0 (1.8)	69.4 (2.1)	73.2 (1.3)	77.0 (1.4)	80.3 (1.3)
SI-SVM	71.3	66.3 (1.0)	66.5 (1.1)	66.6 (2.3)	72.2 (2.2)	73.3 (0.8)	75.1 (2.0)	78.8 (1.5)
SI-kNN	71.0	65.1 (1.1)	65.1 (1.5)	65.0 (2.7)	72.6 (0.7)	74.7 (1.0)	74.7 (1.8)	79.5 (0.9)
CkNN	50.2	52.4 (1.0)	54.7 (2.2)	55.2 (4.2)	46.4 (1.3)	60.9 (1.3)	36.5 (1.8)	45.6 (2.0)
EMD-SVM	78.1	66.4 (1.2)	65.3 (1.5)	78.1 (2.5)	**83.1** (0.5)	**82.7** (0.6)	84.0 (1.0)	**87.3** (0.5)
miGraph	67.8	61.6 (1.6)	60.4 (0.7)	65.9 (5.0)	60.7 (1.4)	68.7 (1.0)	78.9 (0.7)	78.7 (0.8)
NSK-SVM	76.3	67.5 (0.7)	65.4 (1.0)	68.6 (2.6)	80.8 (0.7)	81.7 (0.8)	83.8 (1.2)	86.2 (1.0)
MInD	73.2	58.3 (1.3)	58.5 (2.6)	78.1 (3.7)	76.9 (1.6)	69.5 (1.2)	84.7 (0.6)	86.4 (0.6)
MILES	75.0	**68.1** (0.9)	63.7 (1.7)	67.7 (3.1)	77.8 (1.1)	79.6 (0.5)	83.1 (0.8)	84.9 (0.7)
BoW	74.0	65.2 (0.9)	64.1 (1.8)	65.4 (4.6)	78.6 (0.6)	74.7 (1.2)	84.2 (1.5)	85.9 (0.3)
CCE	75.6	65.4 (1.1)	66.1 (0.6)	66.9 (2.5)	82.0 (0.7)	79.8 (0.6)	83.2 (1.0)	85.8 (0.5)

TABLE 1

TCGA data: average AUROC (%) with standard error given in parentheses for predicting bag labels for each method across seven cancers. The highest AUROC is highlighted in bold. The average performance across all methods is shown below each cancer type.

are short peptides presented by the major histocompatibility complex (MHC) proteins on the surface of tumor cells, which serve as recognition markers for cytotoxic T cells via T-cell receptors. As one of the most fundamental and unsolved questions in tumor immunology, the relationship between tumor immune responses and tumor neoantigens is the key to understanding the inefficiency of immunotherapy observed in many cancer patients. However, it is often a challenging task to quantify this relationship as the the properties of neoantigens that can elicit immune responses remain unclear. This biological problem is investigated in the MIR context by Park et al. (2020), who modeled multiple instances (neoantigens) within each bag (patient specimen) with the continuous response (T-cell infiltration). Each instance is characterized by covariates of neoantigen qualities.

We use neoantigen data from several existing studies (Network (2013); Sato et al. (2013); Miao et al. (2018); Wang et al. (2018b)). To apply MICProB to the neoantigen data, we code samples with T-cell infiltration greater than or equal to 5 as one (high-level infiltration) and smaller than 5 as zero (low-level infiltration), resulting in 36% positive bags out of 728 bags. The distribution of numbers of neoantigens (instances) from different patients (bags) is shown in Figure S8(a) of Supplementary Material, with mean being 113, median being 30, and maximum being 664. Figure S8(b) of Supplementary Material shows the distribution for each of the six covariates that describe neoantigens along the x-axis. These covariates include hydrophobicity (hydro), similarity to pathogenic epitopes (blast), rank of binding affinities to major histocompatibility complex molecules (perc_rank), an immunogenicity score previously established for class I neoantigens only (neoantigens could bind to both class I and class II human leukocyte antigen (HLA) molecules) (immune), transport efficiency (TAP), and the mutation type (mut_type, 1 for missense mutations, and 0 for insertions/deletions, stoploss mutations).We standardize all continuous variables to have a zero mean and unit standard deviation.

We randomly split the dataset ten times and in each split, we use 75% data to train MICProB and the remaining 25% data to evaluate the performance. We run our sampler for 100,000 iterations and discard the first half as burn-ins. The average AUROC across ten random splits is 99%. The posterior mean estimates for the intercept, hydro, blast, perc_rank,

immune, TAP, and mut_type are $-1.162, 0.087, 0.092, 6.473, -0.129, -0.111, 0.420$, respectively. Thus, according to the signs of these estimates, higher hydrophobicity, higher similarity to pathogenic epitopes, and higher rank of binding affinities to major histocompatibility complex moleculesare, and missense mutations tend to increase the likelihood of high infiltration; meanwhile, the estimates of class I immunogenicity scores and TAP activity, are negative, and thus tend to decrease the likelihood of high infiltration. We also provide interval estimates of covariate effects from MICProB and estimates with standard errors from MILR using the default setting in Table S2 of Supplementary Material. We find that the two regression-based methods agree on the directions of effects of blast, perc_rank, TAP, and mut_type. They also agree that the intercept and the effect of perc_rank is statistically signifiant while the others are not. Lastly, we point out that other benchmark methods do not offer such interpretability via regression coefficients reflecting covariate effects.

5. Discussion. In MIL literature, methods for classification are exclusively based on the WR framework, under which the functionality of identifying primary instances is not offered. Further, these methods are mainly optimization-based and hence suffer from poor explainability. Under the PI framework (Xiong et al. (2021)) that is much less explored in the MIC literature, we develop a novel Bayesian hierarchical model, MICProB, to learn from multiple instance data with a binary response and identify both primary instances and bag labels. Specifically, MICProB is composed of two nested probit regression models, where the inner model is estimated for predicting primary instances (i.e., predicting δ_{ij} from x_{ij}), and the outer model is estimated for predicting bag-level responses based on the primary instances. Thanks to its fully Bayesian formulation, prediction for new bags can be performed in an integrated manner via posterior predictive sampling. Furthermore, MICProB enables convenient statistical inference for quantities related to model parameters with posterior samples drawn. Regression coefficients that reflect covariate effects on the bag-level response are explicitly estimated, hence offering high explainability to the model, as demonstrated in the application to the neoantigen data.

Due to its special design for the PI data generation mechanism, we recognize that MICProB does not belong to any of the existing categories of MIC methods of IS, BS, or ES paradigm. MICProB is not an IS method as the prediction step does not occur at the instance level (i.e., predicting whether an instance is positive or negative). Further, MICProB is not a BS method as there is no distance computed between each pairs of bags. Lastly, it might be tempting to view MICProB as an ES method. But as opposed to mainstream ES methods that are vocabulary-based, which use instances from training data to build a dictionary for feature embedding, MICProB does not have the embedding step that maps the original feature space to a new one. Thus, the proposed method does not fall under the ES category. Given above, MICProB provides a fresh perspective to the development of new MIL methods that are model-based and tailored for MI data with the concept of primary instances.

MICProB yields significantly better performance in various simulated scenarios than the 15 benchmark methods. Based on the assumption that the bag label is determined by primary instances in each bag, our model is capable of identifying these instances with high accuracy, even though they are not known in advance. Given that there is no such method that works universally well in real data application of cancer diagnosis, the proposed method performs the best in four out of seven cancer types. Across the seven cancer types, MICProB also has the highest performance on average, suggesting that the PI framework is more likely to represent the underlying data generation mechanism for this particular application.

We make our code available at `https://github.com/danyixiong/MICProB`. A user of MICProB has the flexibility of choosing between the "sum" or "average" contribution of primary instances, according to the user's perception of the underlying data generation

process or results from cross validation. MICProB can also be implemented using different priors, to incorporate various forms of prior knowledge. For example, in the real data application, we experiment with marginally non-informative prior distributions for covariance matrices Σ_β and Σ_b, as suggested by Huang and Wand (2013), to reflect our vague information on the regression coefficients. The resulting MCMC algorithm can still be conveniently implemented via Gibbs sampling with data augmentation technique (see Section S2.2 of Supplementary Material for technical detail). However, this prior elicitation is not preferred as its resulting performance (Table S1 of Supplementary Material) is worse than that of simpler prior specifications for the original MICProB. Finally, it is also possible to consider a logistic model for the binary response variable. Under this formulation, however, the Bayesian inference would become harder due to the analytically inconvenient form of the model's likelihood function. A potential direction would be to employ a data augmentation strategy using Pólya-Gamma latent variables (Polson, Scott and Windle (2013)). With these flexible options of design at different parts of the model, MICProB thereby provides a generic framework for developing future model-based statistical methods that are dedicated to addressing MI problems where primary instances need to be identified.

In the era of big data, we envision an increasing need for MIL to handle increasingly complex structures of real-world data. Advances in the biomedical research domain, in particular, propel the development of novel MIL techniques, especially Bayesian methodologies that can naturally incorporate prior beliefs into observed data, because the complicated nature of biological and medical applications necessitates consideration of prior knowledge available from domain experts or past studies, to narrow the search space of the MIL model for the observed new data. By developing MICProB, we provide a successful example of how statistical learning tackles an MIL problem, and we believe there is a broad space for new Bayesian MIL methods with diverse capacities to emerge to capture various characteristics of real-world data.

Acknowledgments. In loving memory of Ze Zhang, who was a great friend and colleague, we would like to express our deepest gratitude for her kindness and support.

Funding. This work was supported by NIH grants R01CA258584 (PIs: T. Wang and X. Wang), R15GM131390 (PI: X. Wang), Cancer Prevention and Research Institute of Texas (CPRIT) grant RP190208 (PI: T. Wang), and National Research Foundation of Korea (NRF) grant 2021R1G1A1005641 funded by the Korean Ministry of Science and ICT (PI: S. Park).

SUPPLEMENTARY MATERIAL

Supplement to "Bayesian Multiple Instance Classification Based on Hierarchical Probit Regression"
Supplementary Material A (MICProB_Supp_v5.pdf): tables, figures, and Bayesian full conditionals mentioned in the paper. In the first section of this Supplementary Material, we provide additional figures for simulation results. In the second section of this Supplementary Material, we provide the full posterior conditionals for the Gibbs sampler with marginally non-informative priors for covariance matrices and corresponding results on the TCGA data. The remaining section of this Supplementary Material provides additional results on the neoantigen data.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88** 669–679.

AMORES, J. (2013). Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence* **201** 81–105.

ANDREWS, S., TSOCHANTARIDIS, I. and HOFMANN, T. (2003). Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems* 577–584.

BABENKO, B., DOLLÁR, P., TU, Z. and BELONGIE, S. (2008). Simultaneous learning and alignment: Multi-instance and multi-pose learning. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*.

BOLOTIN, D. A., SHUGAY, M., MAMEDOV, I. Z., PUTINTSEVA, E. V., TURCHANINOVA, M. A., ZVYAGIN, I. V., BRITANOVA, O. V. and CHUDAKOV, D. M. (2013). MiTCR: Software for T-cell receptor sequencing data analysis. *Nature Methods* **10** 813.

BYERS, L. A. and RUDIN, C. M. (2015). Small cell lung cancer: Where do we go from here? *Cancer* **121** 664–672.

CARBONNEAU, M.-A., CHEPLYGINA, V., GRANGER, E. and GAGNON, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition* **77** 329–353.

CAWLEY, G. C. and TALBOT, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research* **11** 2079–2107.

CHEN, Y., BI, J. and WANG, J. Z. (2006). MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28** 1931–1947.

CHEN, P.-Y., CHEN, C.-C., YANG, C.-H., CHANG, S.-M. and LEE, K.-J. (2017). milr: Multiple-Instance Logistic Regression with Lasso Penalty. *The R Journal* **9** 446.

CHEPLYGINA, V., TAX, D. M. and LOOG, M. (2015). Multiple instance learning with bag dissimilarities. *Pattern Recognition* **48** 264–275.

CLARKE-PEARSON, D. L. (2009). Screening for ovarian cancer. *New England Journal of Medicine* **361** 170–177.

DIETTERICH, T. G., LATHROP, R. H. and LOZANO-PÉREZ, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* **89** 31–71.

FOULDS, J. and FRANK, E. (2010). A review of multi-instance learning assumptions. *The Knowledge Engineering Review* **25** 1–25.

GÄRTNER, T., FLACH, P. A., KOWALCZYK, A. and SMOLA, A. J. (2002). Multi-instance kernels. In *Proceedings of the 19th International Conference on Machine Learning* 179–186.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*. Chapman Hall, London.

HE, H. and GARCIA, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21** 1263–1284.

HUANG, A. and WAND, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis* **8** 439–452.

ILSE, M., TOMCZAK, J. and WELLING, M. (2018). Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning* 2127–2136.

IWAHORI, K., KAKARLA, S., VELASQUEZ, M. P., YU, F., YI, Z., GERKEN, C., SONG, X.-T. and GOTTSCHALK, S. (2015). Engager T cells: a new class of antigen-specific T cells that redirect bystander T cells. *Molecular Therapy* **23** 171–178.

LAMBRECHTS, D., WAUTERS, E., BOECKX, B., AIBAR, S., NITTNER, D., BURTON, O., BASSEZ, A., DECALUWÉ, H., PIRCHER, A., VAN DEN EYNDE, K. et al. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature Medicine* **24** 1277–1289.

LIU, J., LICHTENBERG, T., HOADLEY, K. A., POISSON, L. M., LAZAR, A. J., CHERNIACK, A. D., KOVATICH, A. J., BENZ, C. C., LEVINE, D. A., LEE, A. V. et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173** 400–416.

LU, T., ZHANG, Z., ZHU, J., WANG, Y., JIANG, P., XIAO, X., BERNATCHEZ, C., HEYMACH, J. V., GIBBONS, D. L., WANG, J. et al. (2021). Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nature Machine Intelligence* 1–12.

MARON, O. and LOZANO-PÉREZ, T. (1998). A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems* 570–576.

MIAO, D., MARGOLIS, C. A., GAO, W., VOSS, M. H., LI, W., MARTINI, D. J., NORTON, C., BOSSÉ, D., WANKOWICZ, S. M., CULLEN, D. et al. (2018). Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* **359** 801–806.

NETWORK, C. G. A. R. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499** 43–49.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

PARK, S., WANG, X., LIM, J., XIAO, G., LU, T. and WANG, T. (2020). Bayesian multiple instance regression for modeling immunogenic neoantigens. *Statistical Methods in Medical Research* **29** 3032–3047.

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association* **108** 1339–1349.

RAY, S. and CRAVEN, M. (2005). Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the 22nd International Conference on Machine Learning* 697–704.

RAY, S. and PAGE, D. (2001). Multiple instance regression. In *Proceedings of the 18th Internatinal Conference on Machine Learning* 425–432.

SATO, Y., YOSHIZATO, T., SHIRAISHI, Y., MAEKAWA, S., OKUNO, Y., KAMURA, T., SHIMAMURA, T., SATO-OTSUBO, A., NAGAE, G., SUZUKI, H. et al. (2013). Integrated molecular analysis of clear-cell renal cell carcinoma. *Nature Genetics* **45** 860–867.

SINGHI, A. D., KOAY, E. J., CHARI, S. T. and MAITRA, A. (2019). Early detection of pancreatic cancer: opportunities and challenges. *Gastroenterology* **156** 2024–2040.

WANG, J. and ZUCKER, J.-D. (2000). Solving multiple-instance problem: A lazy learning approach. In *Proceedings of the 17th International Conference on Machine Learning* 1119–1126.

WANG, Z., RADOSAVLJEVIC, V., HAN, B., OBRADOVIC, Z. and VUCETIC, S. (2008). Aerosol optical depth prediction from satellite observations by multiple instance regression. In *Proceedings of the 2008 SIAM International Conference on Data Mining* 165–176.

WANG, X., YAN, Y., TANG, P., BAI, X. and LIU, W. (2018a). Revisiting multiple instance neural networks. *Pattern Recognition* **74** 15–24.

WANG, T., LU, R., KAPUR, P., JAISWAL, B. S., HANNAN, R., ZHANG, Z., PEDROSA, I., LUKE, J. J., ZHANG, H., GOLDSTEIN, L. D. et al. (2018b). An empirical approach leveraging tumorgrafts to dissect the tumor microenvironment in renal cell carcinoma identifies missing link to prognostic inflammatory factors. *Cancer Discovery* **8** 1142–1155.

WHITESIDE, S. K., SNOOK, J. P., WILLIAMS, M. A. and WEIS, J. J. (2018). Bystander T cells: a balancing act of friends and foes. *Trends in Immunology* **39** 1021–1035.

XIONG, D., ZHANG, Z., WANG, T. and WANG, X. (2021). A comparative study of multiple instance learning methods for cancer detection using T-cell receptor sequences. *Computational and Structural Biotechnology Journal* **19** 3255.

ZHANG, Q. and GOLDMAN, S. A. (2002). EM-DD: An improved multiple-instance learning technique. In *Advances in Neural Information Processing Systems* 1073–1080.

ZHANG, J., MARSZAŁEK, M., LAZEBNIK, S. and SCHMID, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision* **73** 213–238.

ZHANG, Z., XIONG, D., WANG, X., LIU, H. and WANG, T. (2021). Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nature Methods* **18** 92–99.

ZHOU, Z.-H., SUN, Y.-Y. and LI, Y.-F. (2009). Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th International Conference on Machine Learning* 1249–1256.

ZHOU, Z.-H. and ZHANG, M.-L. (2007). Solving multi-instance problems with classifier ensemble based on constructive clustering. *Knowledge and Information Systems* **11** 155–170.