IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 9, 2019, accepted November 22, 2019, date of publication November 26, 2019,
date of current version December 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955958

MSPL: Multimodal Self-Paced Learning for Multi-Omics Feature Selection and Data Integration

ZI-YI YANG[1,2], LIANG-YONG XIA[3], HUI ZHANG[1,2], AND YONG LIANG[1,2]

[1]Faculty of Information Technology, Macau University of Science and Technology, Taipa 999078, Macau
[2]State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Taipa 999078, Macau
[3]School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Yong Liang (yliang@must.edu.mo)

This work was partially supported by the Chinese Ministry of Education's Tian Cheng Hui Zhi Innovation and Education Improvement Funds (Grant No. 2018A01014), the Macau Science and Technology Develop Funds (Grant No. 0055/2018/A2) of Macao SAR of China.

ABSTRACT Rapid advances in high-throughput sequencing technology have led to the generation of a large number of multi-omics biological datasets. Integrating data from different omics provides an unprecedented opportunity to gain insight into disease mechanisms from different perspectives. However, integrative analysis and predictive modeling from multi-omics data are facing three major challenges: i) heavy noises; ii) the high dimensions compared to the small samples; iii) data heterogeneity. Current multi-omics data integration approaches have some limitations and are susceptible to heavy noise. In this paper, we present MSPL, a robust supervised multi-omics data integration method that simultaneously identifies significant multi-omics signatures during the integration process and predicts the cancer subtypes. The proposed method not only inherits the generalization performance of self-paced learning but also leverages the properties of multi-omics data containing correlated information to interactively recommend high-confidence samples for model training. We demonstrate the capabilities of MSPL using simulated data and five multi-omics biological datasets, integrating up three omics to identify potential biological signatures, and evaluating the performance compared to state-of-the-art methods in binary and multi-class classification problems. Our proposed model makes multi-omics data integration more systematic and expands its range of applications.

INDEX TERMS Multi-omics data integration, self-paced learning, multimodal data analysis, feature selection, classification.

I. INTRODUCTION

Driven by the development of new high-throughput sequencing techniques, various types of biological data with different formats, sizes, and structures have been increasing at an unprecedented rate. Gene expression, miRNA expression, proteins, DNA methylation and metabolites are some examples of biological data produced by using high-throughput techniques such as microarray [1] and mass spectrometry [2]. Generally, each of these distinct biological data types provides different, partially independent and complementary information of the entire genome [3]. Therefore, deciphering complex human genomes and gene functions may require more complete and complementary information than those are provided by single type of data. The integration of multi-omics data (e.g. genomics, transcriptomics, proteomics and metabolomics, etc.) provides an unprecedented opportunity to gain insight into complex disease mechanisms from different views and levels, predict the subtype of the target disease, and discover potential multi-omics biological signatures [4]–[6].

Effective methods to integrative analysis and predictive modeling from multi-omics data have to overcome at least three computational challenges. i) ***High levels of noise and collection bias present in each type of biological data.*** Random noise and system/collection bias exist in distinct biological data types not only impact the cost and effectiveness of scientific research, but also disrupt precise prediction of disease subtypes that may ultimately impact patients [7]. Moreover, different noise and bias across distinct data types may result in reduced classifier performance and finding unreliable potential biological signatures [8]. ii) ***The high dimensions compared to the small samples.*** The biological data generally contains a large number of features p and small size of samples n, which is called large p and small n problem [9]. From the biological perspective, only a small fraction

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

of features that are highly correlated to the target disease and most of the features are irrelevant. From the machine learning perspective, numerous irrelevant features may prone to overfitting issue and negatively impact the performance of the classifier [10], [11]. iii) ***Data heterogeneity.*** Distinct types of biological data generated from different omics platforms possess heterogeneous information, such as following different statistical distributions, suffering from different levels of imprecisions and containing different kinds of uncertainties [12]. Unfortunately, current multi-omics data integration approaches have yet to address all of these computational challenges together [5]. Therefore, there is an urgent need for a robust method for integrative analysis multi-omics data.

The problem of learning predictive models from multi-omics data can be naturally considered a multimodal learning problem [13], [14]. Commonly, data from multiple modalities contain more complete and complementary information of the object than that is provided by the single modality only. Multi-omics data provides multiple modalities with distinct feature sets in the same set of samples. Current supervised multimodal data integration approaches for predicting cancer subtypes and identifying significant multi-omics signatures can be classified as concatenation-based, ensemble-based, and knowledge-driven approaches [15].

The concatenation-based approach simply combines all features from different types of data into a single large dataset before. And after that, prediction and feature selection are based on a single statistical model [5], [16]. The ensemble-based approach constructs a prediction model on each omics dataset separately and utilizes an average/majority voting scheme to combine the results of prediction [17]. These approaches can be biased towards certain omics data types, and do not consider interactions between omic layers [18], [19]. Recently, classification methods such as Generalized Elastic Net (EN) [20], [21], adaptive Group-Regularized ridge regression [22], and sparse Partial Least Squares Discriminant Analysis (sPLSDA) [23] have incorporated curated biological data such as genetic pathway data, methylation data, and gene expression data. These methods are still limited to single omics data such that, either the concatenation-based or ensemble-based strategy needs to be applied to incorporate additional omics data types. However, neither of these two types of data integration approaches considers the interaction between multiple data types, which limits the understanding of the relationship between different levels of biological function.

Knowledge-driven multimodal data integration considers the relationships between different modalities based on prior knowledge. Very recently, Singh *et al.* [15] proposed Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO), which is dedicated to maximizing the correlated information between multiple omics data. DIABLO actually extends the sparse generalized canonical correlation analysis (SGCCA) [24] to the supervised classification model. It is a multivariate dimension reduction approach that maximizes the covariance between linear combinations of variables from multiple omics according to the given design matrix and combines all potential components for prediction. However, the assumption of a linear relationship between selected significant omics features may not be applicable to some biological research areas. In addition, DIABLO is susceptible to heavy noise, resulting in poor generalization performance.

In this paper, we present ***MSPL***, a robust supervised multimodal method that simultaneously identifies significant multi-omics signatures during the integration process and predicts the cancer subtypes. MSPL (*Multimodal Self-Paced Learning*) adopts a sample reweighting strategy to improve the robustness of the learning process in heavy noise situations. The core idea of MSPL is to interactively recommend high-confidence samples with smaller loss values between multiple omic data types, and automatically select samples from easy to complex to train the model for each modality in a purely self-paced way. Our method is actually established on the self-paced learning (SPL) regime [25], and is a variant of it. Furthermore, to overcome the overfitting issue caused by large p and small n problem, MSPL embeds a regularization method to perform feature selection during the learning process. A series of regularization methods for feature selection have been proposed [26]–[30]. Here, MSPL is performed via L_1 regularization [26]. In the proposed method, MSPL strives to address the three above-mentioned computational challenges faced by integrative analysis and predictive modeling from multi-omics data.

We demonstrate the capability of MSPL and compare its prediction and feature selection performance with other state-of-the-art methods using simulated data and five publicly available multi-omics datasets, including four benchmark cancer datasets and one breast cancer multi-omics dataset. In particular, breast cancer multi-omics dataset has approximately 1000 samples, including four breast cancer subtypes. In these experiments, we integrate up to three omics datasets and evaluate the performance of all competing methods in binary and multi-class classification problems. The results show that MSPL presents competitive performance with existing methods, especially robust in the presence of heavy noises.

The rest of this paper is organized as follows. Section II introduces the related work of self-paced learning, while Section III presents the proposed MSPL algorithm. Experimental results of several competing methods and brief biological analysis are shown in Section IV. A conclusion is given in Section V. Finally, the linkage of this paper code is provided in Section V.

II. RELATED WORK

This section introduces the fundamental concepts of curriculum learning and self-paced learning.

A. CURRICULUM LEARNING

The fundamental definition of Curriculum Learning (CL) was first proposed by [31]. Inspired by human and animal

learning mechanism, learning is better when the samples are organized in a meaningful order, that is start with easier concepts to progressively more complex ones. This learning mechanism gradually included samples from easy to complex correspond to courses that are studied at different stages of the human or animals. CL can accelerate convergence to the global minimum and has been proven by empirical evaluation to help alleviate local optimal problems in non-convex optimization [32], [33]. The main challenge for CL is to identify the easy and complex samples during the learning process. However, providing the ranking of samples may be conceptually difficult for human in many real-world applications. Moreover, what is intuitively ''easy'' for a human may not be in accordance with what is easy for the algorithm in the feature and hypothesis space applied in the given application [25].

B. SELF-PACED LEARNING

To alleviate the deficiency of CL, Kumar *et al.* [25] first proposed Self-Paced Learning (SPL). SPL embeds CL (from easy to progressively more complex samples) as a regularization term into the model learning process. Formally, suppose given a dataset $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the i-th input sample with p features and y_i is the i-th sample with the value 0 or 1 in the classification model. Let $L(y_i, f(\boldsymbol{x}_i, \boldsymbol{\beta}))$ denotes the loss function, which calculates the loss between the real label y_i and the estimated value $f(\boldsymbol{x}_i, \boldsymbol{\beta})$. The $\boldsymbol{\beta}$ represents the model parameter inside the decision function $f(\boldsymbol{x}_i, \boldsymbol{\beta})$. The purpose of the SPL is to jointly learn the model parameter $\boldsymbol{\beta}$ and the latent weight variable $\boldsymbol{v} = [v_1, v_2, \ldots, v_n]$ by minimizing:

$$\min_{\boldsymbol{\beta}, \boldsymbol{v} \in [0,1]^n} E(\boldsymbol{\beta}, \boldsymbol{v}; \gamma) = \sum_{i=1}^{n} v_i L(y_i, f(\boldsymbol{x}_i, \boldsymbol{\beta})) + \lambda ||\boldsymbol{\beta}||_1 + g(v, \gamma) \quad (1)$$

where γ is an age parameter for controlling the learning pace, λ is the L_1 regularizer parameter and $g(v, \gamma)$ represents the self-paced regularizer (SP-regularizer). The traditional objective of SPL is to simultaneously minimize the weighted loss function and the negative L_1-norm regularizer ($g(v_i, \gamma) = -\gamma \sum_{i=1}^{n} v_i, v_i \geq 0$).

The Alternative Optimization Strategy (AOS) algorithm can effectively solve the SPL problem. It is a biconvex optimization iterative algorithm that divides features used for optimization into two disjoint blocks. The basic procedure of AOS algorithm can be described as: in each iteration, to optimize the target block of features while keeping the other block fixed. For the traditional SPL problem, when latent weight variable $\boldsymbol{v}$ is fixed, the optimal model parameter $\boldsymbol{\beta}$ can be obtained by the state-of-the-art supervised learning approaches. When $\boldsymbol{\beta}$ is fixed, the optimal weight variable $\boldsymbol{v}^* = [v_1^*, v_2^*, \ldots, v_n^*]$ can be calculated by [25]:

$$v_i^* = \begin{cases} 1 & , \ L(y_i, f(\boldsymbol{x}_i, \boldsymbol{\beta})) < \gamma \\ 0 & , \ otherwise \end{cases} \quad (2)$$

This alternative search strategy implies an intuitive explanation: (1) When updating $\boldsymbol{v}$ with a fixed $\boldsymbol{\beta}$, if the loss value of the sample is smaller than the age parameter γ, then the sample is selected as an easy sample ($v_i = 1$) for the classifier training, otherwise, do not select ($v_i = 0$). (2) When updating $\boldsymbol{\beta}$ with a fixed $\boldsymbol{v}$, the classifier is trained only on the selected easy samples ($v_i = 1$). (3) Before starting the next iteration, increase the age parameter γ to control the learning pace, which allows more samples to be used for model training. When γ is small, only easy sample with smaller loss will be selected. With the increase of age parameter γ, more samples with larger loss will be gradually selected to train a more ''mature'' model.

By jointly learning the model parameters $\boldsymbol{\beta}$ and the latent weight variable $\boldsymbol{v} = [v_1, v_2, \ldots, v_n]$, gradually increasing the age parameter γ, SPL can automatically include more samples (from easy to progressively more complex) in the training process with a purely self-paced way. Various machine learning applications provide empirical validation that SPL can be performed robustly in the presence of heavy noises [34]–[36]. Moreover, SPL is also widely used in softmax regression [37], multi-view learning [38], multi-task learning [39], etc. In addition, [40] proved the intrinsic working mechanism of SPL, which naturally explains the effectiveness of SPL, especially its robustness in heavy noises.

III. PROPOSED METHOD

This section presents the proposed *Multimodal Self-paced Learning* (MSPL). The objective of MSPL is first formally defined, and then we present an efficient algorithm to solve the model.

A. THE MSPL MODEL

Multi-omics data naturally has multimodal properties. Multimodal data typically contains more complete description and complementary information than those of single modality. An intuitive way to achieve this is to select samples through the interrelationship between multiple modalities. We assume that the different modalities share common knowledge of sample confidence. In a word, samples with high quality in one omics may be consistent with other omics.

The objective of MSPL can be mathematically described as follows. Suppose given a multimodal dataset $D = \{(\boldsymbol{x}_1^{(j)}, y_1), (\boldsymbol{x}_2^{(j)}, y_2), \ldots, (\boldsymbol{x}_n^{(j)}, y_n)\}$, where $\boldsymbol{x}_i^{(j)} = (x_{i1}^{(j)}, x_{i2}^{(j)}, \ldots, x_{ip^{(j)}}^{(j)})$ is the i-th input sample with $p^{(j)}$ features under the j-th modality. $p^{(j)}$ indicates the number of features in the j-th modality. y_i is the common label of the i-th sample for every modality in the classification model (e.g. $y_i = \{0, 1\}$ in the binary classification problem). Let $L(y_i, f(\boldsymbol{x}_i^{(j)}, \boldsymbol{\beta}^{(j)}))$ denotes the loss function, which calculates the loss between the real label y_i and the estimated value $f(\boldsymbol{x}_i^{(j)}, \boldsymbol{\beta}^{(j)})$ in the j-th modality. The $\boldsymbol{\beta}^{(j)}$ represents the model parameter inside the decision function $f(\boldsymbol{x}_i^{(j)}, \boldsymbol{\beta}^{(j)})$. The objective function of

MSPL can be expressed as:

$$\min_{\boldsymbol{\beta}^{(j)},\boldsymbol{v}^{(j)}\in[0,1]^n,\ j=1,2,\ldots,m} E(\boldsymbol{\beta}^{(j)},\boldsymbol{v}^{(j)};\lambda^{(j)},\gamma^{(j)},\delta)$$
$$=\sum_{j=1}^{m}\sum_{i=1}^{n}v_i^{(j)}L\left(y_i,f^{(j)}\left(\boldsymbol{x}_i^{(j)},\boldsymbol{\beta}^{(j)}\right)\right)+\sum_{j=1}^{m}\lambda^{(j)}||\boldsymbol{\beta}^{(j)}||_1$$
$$-\sum_{j=1}^{m}\sum_{i=1}^{n}\gamma^{(j)}v_i^{(j)}-\delta\sum_{\substack{1\le k,j\le m,\\ k\ne j}}\left(\boldsymbol{v}^{(k)}\right)^T\boldsymbol{v}^{(j)}, \tag{3}$$

where m denotes the total number of modalities. $\boldsymbol{x}_i^{(j)}$ is the i-th input sample ($i=1,2,\ldots,n$) under the j-th modality, and y_i is the corresponding label of $\boldsymbol{x}_i^{(j)}$ for every j. $v_i^{(j)}$ denotes the weight of $\boldsymbol{x}_i^{(j)}$. λ is a tuning parameter, it controls the complexity of the model. $\gamma^{(j)}$ indicates the age parameter, it controls the learning pace in each iteration in the j-th modality. δ is the parameter that controls influence from other modalities when one modality tends to select more training samples.

The proposed MSPL model actually corresponds to the sum of the SPL model under multiple modalities plus a regularization term $\sum_{\substack{1\le k,j\le m\\ k\ne j}}\left(\boldsymbol{v}^{(k)}\right)^T\boldsymbol{v}^{(j)}$. This inner product encodes the relationship between multiple modalities. The new regularizer actually establishes attribute links between multiple modalities by using multimodal data of a sample. It takes advantage of the information content of multimodal data and selects high confident samples through the interrelationship between multiple modalities. Therefore, this new regularizer enforces the weight penalizing the loss of one modality similar to that of other modalities (e.g. a sample with high confidence in one modality is likely the same in other modalities). In addition, each modality of data uses high confident samples to identify potential significant features, which can improve the feature selection performance of the model.

B. THE MSPL ALGORITHM

The proposed MSPL model can be solved by the alternative optimization strategy (AOS) algorithm, as listed in Algorithm 1.

Initialization: Initialize weight parameter $v_i^{(j)}$, age parameter $\gamma^{(j)}$ and δ. $v^{(1)},v^{(2)},\ldots,v^{(m)}$ are zero vectors in R^m. $\gamma^{(1)},\gamma^{(2)},\ldots,\gamma^{(m)}$ are initialized with small values to include few samples in the first iteration of the training process. Set δ to a specific value through the whole learning process. The initial loss of all samples in each modality is obtained by simultaneously training multiple classifiers on all samples of different modalities.

Update $v_i^{(k)}$ ($k=1,2,\ldots,m;\ k\ne j$): This step can obtain the current optimal weight of samples under the k-th modality. Due to the multimodal data intrinsically contains complementary information. Therefore, the physical meaning of this step is to prepare the confident samples ($v_i^{(k)}>0$) for training on the j-th modality. That is, a high confident sample can be selected by the interrelationship between multiple modalities. Based on Equation (3), the first order derivative at $v_i^{(k)}$ can be estimate as:

$$\frac{\partial E}{\partial v_i^{(k)}}=L_i\left(y_i,f^{(k)}\left(\boldsymbol{x}_i^{(k)},\boldsymbol{\beta}^{(k)}\right)\right)-\gamma^{(k)}-\delta\sum_{\substack{1\le j\le m,\\ j\ne k}}v_i^{(j)}. \tag{4}$$

According to Equation (4), the current optimal weight of the i-th sample under the k-th modality can be expressed as:

$$v_i^{(k)}=\begin{cases}1, & L_i\left(y_i,f^{(k)}\left(\boldsymbol{x}_i^{(k)},\boldsymbol{\beta}^{(k)}\right)\right)<\gamma^{(k)}+\delta\sum_{\substack{1\le j\le m,\\ j\ne k}}v_i^{(j)},\\ 0, & otherwise.\end{cases} \tag{5}$$

Update $v_i^{(j)}$: The purpose of this step is to formally define which samples will be confirmed into the training process of the j-th modality. The optimal weight of the i-th sample under the j-th modality can be calculated in the same way as the previous step. Different from the previous step is that the samples selected in this step will be directly employed for training in the j-th modality. According to Equation (5), it is easy to observe that the samples with high confidence from other modalities possess higher chance of being selected for the training of the j-th modality than other samples.

Update $\boldsymbol{\beta}^{(j)}$: The purpose of this step is to train a classifier using the selected samples in the j-th modality. In this work, we select the sparse logistic regression classifier to train the model. In this step, Equation (3) degenerates into the standard sparse logistic regression optimization problem as:

$$\min_{\boldsymbol{\beta}^{(j)}}\sum_{i=1}^{n}L_i\left(y_i,f^{(j)}\left(\boldsymbol{x}_i^{(j)},\boldsymbol{\beta}^{(j)}\right)\right)+\lambda^{(j)}||\boldsymbol{\beta}^{(j)}||_1. \tag{6}$$

where $1<n^{(j)}\le n$. $n^{(j)}$ represents the current number of samples used to train a classifier under the j-th modality. This problem can be readily solved by R package glmnet [41].

Before the start of the next iteration, age parameters $\gamma^{(j)}(j=1,2,\ldots,m)$ are increased to allow more samples with larger loss values to enter the next iteration of the training process. We then repeat the above optimization process for different modalities until all samples are used for model training or reach the maximum number of iterations.

From Algorithm 1, we can easily observe that it can obtain the optimal solution under interaction with multiple modalities and the time complexity of it is $O(n^2\times p)$, where $n\ll p$. In the test phase, suppose given a test dataset $D'=\left\{\boldsymbol{x}_1^{(j)},\boldsymbol{x}_2^{(j)},\ldots,\boldsymbol{x}_u^{(j)}\right\}$ with m modalities, where u is the number of test sample. By using the optimal solution of a classifier under each modality $\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)},\ldots,\boldsymbol{\beta}^{(m)}$ to predict the optimal y_k, which can be predicted by the following minimization problem:

$$y_k=\underset{y_k}{argmin}\sum_{j=1}^{m}L_k\left(y_k,f^{(j)}\left(\boldsymbol{x}_k^{(j)},\boldsymbol{\beta}^{(j)}\right)\right) \tag{7}$$

Algorithm 1 Algorithm for Solving MSPL Model

Input: samples $\boldsymbol{x}_1^{(j)}, \boldsymbol{x}_2^{(j)}, \ldots, \boldsymbol{x}_n^{(j)}$ (each sample has m modalities), labels $y_1, \ldots, y_n$, age parameters $\gamma^{(1)}, \ldots, \gamma^{(m)}$, δ and max_iter.
Output: $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(m)}$.
Initialize $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}$, $\gamma^{(1)}, \ldots, \gamma^{(m)}$ and δ
Update $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(m)}$
iter $= 1$
while iter $\leq$ max_iter **do**
  for $j \leftarrow 1$ to m **do**
    for $k \leftarrow 1$ to m **and** $k \neq j$ **do**
      Update $v_i^{(k)}$: Prepare confident samples with non-zeros $v_i^{(k)}$ values for training on the j-th modality
    end for
    Update $v_i^{(j)}$: Confirm which samples will be feeded into the training process of the j-th modality
    Update $\boldsymbol{\beta}^{(j)}$: Train a classifier (e.g. sparse logistic regression model) under the j-th modality
  end for
  Augment $\gamma^{(1)}, \ldots, \gamma^{(m)}$
  iter $\leftarrow$ iter $+ 1$
end while
Return $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(m)}$

C. ALGORITHM ANALYSIS

The proposed MSPL algorithm, as shown in Algorithm 1, mainly differs from current multi-omics data integration approaches in the following four-fold aspects:

1) Instead of "simple and brute" aggregating multi-omics data into a single dataset (e.g. concatenation-based) or ignoring the interrelationship between multiple omic data types (e.g. ensemble-based), the MSPL algorithm uses multimodal interactions to recommend high confident samples to train the model. In MSPL, if the i-th sample in the k-th modality that loss value is smaller than a confidence threshold $\gamma^{(k)} + \delta \sum_{j=1}^{m} v_i^{(j)} (j \neq k)$ is considered to be a high confident sample ($v_i^{(k)} = 1$) and will be selected to train the classifier in the k-th modality. Note that this confidence threshold is related to the age parameter $\gamma^{(k)}$ in the k-th modality and weight of the corresponding samples in other modalities. It implies that we more prefer to select a sample that is high confidence in other modalities than a sample that is not in it. High confident samples are selected between multiple modalities to take full advantages of comprehensive information to characterize an object.
2) When updating samples for training in one modality, in addition to selecting high confident samples that are recommended by other modalities, the MSPL algorithm may picks few high confident samples with very small loss values on the current modality. This strategy preserves some specific characteristics of each modality.
3) For the ensemble-based data integration and DIABLO, both of them apply a majority/average voting scheme during performance evaluation and test dataset prediction. The MSPL algorithm, inspired by [42], predicts the subtype of a sample by solving the minimization problem according to Equation (7). The prediction can be performed more accurately by calculating the sum of the predicted loss values under multiple classifiers.
4) The proposed MSPL model is a variant of the SPL learning regime, which gradually increases the learning pace and automatically select more samples (from smaller losses to progressively larger losses) to train the model and obtain a more "mature" model. Meng *et al.* [40] from the mathematical perspective have proven the effectiveness of the SPL learning regime, especially its robustness in heavy noises situation. Therefore, MSPL achieves a better generalization performance than traditional multimodal data integration methods (see results part).

IV. RESULTS

We evaluate the capability of the proposed MSPL model and compare its performance with other state-of-the-art methods in this section. We applied the logistic regression model/multinomial model with Elastic Net (EN) regularization [27], Random Forest (RF) [43] and Self-paced Learning (SPL) with L_1 penalty [25] in the concatenation and ensemble frameworks. The compared methods include: concatenation-based methods (Concate_EN, Concate_RF, and Concate_SPL), ensemble-based methods (Ensemble_EN, Ensemble_RF, and Ensemble_SPL) and DIABLO [15].

A. SIMULATIONS

We evaluate the robustness and feature selection performance of our proposed MSPL model in simulated experiments with varying noise control parameters and sample sizes.

1) GENERATE SIMULATED DATA

We generate multimodal data with small sample sizes and high dimensionality, and each modality contains a large number of irrelevant features [10]. Beyond that, different modalities with varying dimensionality. We generated the predictor vectors $x_{i1}, x_{i2}, \ldots, x_{ip} (i = 1, \ldots, n)$ independently by the standard normal distribution, where p is the number of features. $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ denotes the i-th sample. The simulated dataset is generated by the logistic regression model and is generated by the follows [11], [44]:

$$\log\left(\frac{y_i}{1 - y_i}\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \sigma \cdot \varepsilon \tag{8}$$

where $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ is the independent random errors generated by $N(0, 4)$, σ is the noise control parameter.

Simulated data are generated by the above procedure. Three simulated datasets (A, B, C) for two classes Y were

TABLE 1. Test accuracy (%) of different methods on simulated data with varying noise parameters and sample sizes. The mean accuracy (±sd) over 30 repetitions of the experiments, and the best results are highlighted in bold.

Sample size/ Noise		Concate_EN	Concate_RF	Concate_SPL	Ensemble_EN	Ensemble_RF	Ensemble_SPL	DIABLO	MSPL
n=100	$\sigma=0$	76.23 ± 5.34	53.23 ± 6.45	78.80 ± 6.42	76.27 ± 6.22	54.47 ± 5.03	80.07 ± 5.39	75.23 ± 4.50	**86.50 ± 3.05**
	$\sigma=0.4$	72.80 ± 7.63	53.00 ± 4.80	74.67 ± 7.42	74.17 ± 5.72	55.83 ± 5.34	78.43 ± 6.93	74.93 ± 5.13	**84.77 ± 4.63**
	$\sigma=0.8$	67.50 ± 6.48	51.30 ± 5.66	69.83 ± 7.85	66.53 ± 5.82	53.67 ± 5.50	70.20 ± 6.12	70.83 ± 6.55	**76.63 ± 2.76**
n=150	$\sigma=0$	86.23 ± 3.97	55.60 ± 3.86	90.10 ± 3.62	88.87 ± 3.85	58.23 ± 4.14	90.80 ± 3.78	83.53 ± 5.55	**92.80 ± 3.19**
	$\sigma=0.4$	86.13 ± 4.51	53.60 ± 5.02	86.63 ± 5.31	84.63 ± 3.94	57.80 ± 5.47	88.27 ± 4.85	81.50 ± 4.57	**91.33 ± 2.37**
	$\sigma=0.8$	81.10 ± 5.96	53.50 ± 4.00	81.40 ± 5.63	77.27 ± 5.44	55.73 ± 5.50	81.17 ± 5.23	79.63 ± 4.85	**85.23 ± 2.67**
n=200	$\sigma=0$	92.93 ± 2.27	56.13 ± 5.54	95.03 ± 3.02	93.03 ± 2.01	57.40 ± 5.30	95.17 ± 2.30	88.00 ± 3.85	**97.23 ± 1.63**
	$\sigma=0.4$	91.30 ± 3.43	56.27 ± 4.82	92.87 ± 2.87	89.40 ± 2.34	56.70 ± 5.31	92.87 ± 3.69	87.97 ± 4.16	**93.86 ± 2.83**
	$\sigma=0.8$	85.20 ± 3.29	55.10 ± 5.90	88.27 ± 4.21	83.03 ± 3.78	54.63 ± 4.34	87.57 ± 4.76	83.57 ± 4.79	**91.78 ± 3.46**

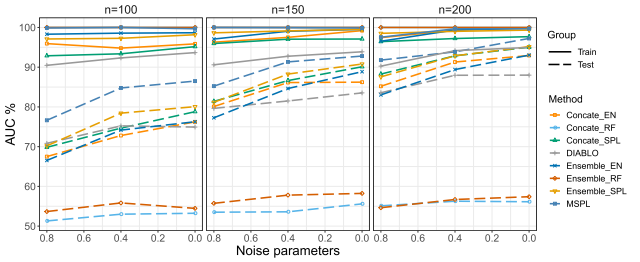


FIGURE 1. Training and test AUC (%) of different methods on simulated data with varying noise parameters and sample sizes.

generated with equal sample sizes by different number of features ($p^{(A)} = 2000, p^{(B)} = 500, p^{(C)} = 1500$). And each simulated dataset (A, B, C) can be treated as one modality of samples. We set the true coefficients $\boldsymbol{\beta}^{(A)}$, $\boldsymbol{\beta}^{(B)}$, and $\boldsymbol{\beta}^{(C)}$ as sparse vectors with $s^{(A)} = 10$, $s^{(B)} = 9$, and $s^{(C)} = 8$ nonzero components, respectively. The locations of each nonzero coefficient are chosen randomly, and the value of each nonzero coefficient is from $\{\pm E\}$ with $E = 2.5$.

We consider the cases with varying training samples size $n = 100, 150, 200$ and varying noise control parameters $\sigma = 0, 0.4, 0.8$, respectively. Each method was evaluated on a test dataset with 100 samples. For the Concate_EN, Concate_SPL, Ensemble_EN, Ensemble_SPL and MSPL methods, we used 10-fold cross-validation to obtain the optimal tuning parameter λ in the sparse logistic regression model. The simulated experiments were repeated 30 times and we report the average measurement.

2) ANALYSIS OF SIMULATION

We demonstrate the average test accuracy of each competing method for each simulated experiment in Table 1. It can be observed that the proposed MSPL method achieves the best performance in all cases compared to other methods. For instance, with sample size $n = 100$, noise parameter $\sigma = 0.4$, the average test accuracy of MSPL is 84.77% obviously superior to 72.80%, 53.00%, 74.67%, 74.17%, 55.83%, 78.43%, and 74.93% obtained by the Concate_EN, Concate_RF, Concate_SPL, Ensemble_EN, Ensemble_RF, Ensemble_SPL and DIABLO, respectively. In addition, when increasing the training sample size n, the test accuracy of all the eight methods are improved. For instance, with $\sigma = 0.4$, the average test accuracy of MSPL are 84.77%, 91.33% and 93.86% with the sample sizes $n = 100$, 150 and 200, respectively. When updating the noise parameter σ, the prediction performance of all methods are deceased. For instance, with sample size $n = 150$, the average test accuracy from MSPL deceased from 92.80% to 85.23%, in which σ increased from 0 to 0.8.

To better illustrate the robustness of the proposed MSPL method towards heavy noises situation, we exhibit the tendency curves of the training and test AUC of different methods on simulated experiments with varying noises parameters and sample sizes in Fig. 1. From this figure, we can easily conclude that the Concate_RF and Concate_RF can readily overfit to the high dimensionality with small sample sizes situation. Beyond that, it can be seen that gaps in predict performance between MSPL and all other methods increase as the training sample sizes deceased with the high noise parameter $\sigma = 0.8$. That is, the robustness of our method outperforms other competing methods in the case of small sample sizes with heavy noise. For instance, with the training sample size $n = 100$, our method achieves more than 6% AUC gain compared with the second best results under noise parameter $\sigma = 0.8$. When the training samples size n increases, the prediction performance of all the methods are improved.

We also evaluate the feature selection performance of each competing method on simulated experiments with varying noise parameters and samples sizes. The β-sensitivity and β-specificity are used to evaluate the feature selection performance, defined as follows [45]:

$$\begin{aligned} &TruePositive(TP)=\left|\beta .* \hat{\beta}\right|_0, TrueNegative(TN)=\left|\bar{\beta} .* \bar{\hat{\beta}}\right|_0 \\ &FalsePositive(FP)=\left|\bar{\beta} .* \hat{\beta}\right|_0, FalseNegative(FN)=\left|\beta .* \bar{\hat{\beta}}\right|_0 \\ &\beta - sensitivity = \frac{TP}{TP+FN}, \beta - specificity = \frac{TN}{TN+FP} \end{aligned} \tag{9}$$

where the $|\cdot|_0$ represents the number of non-zero elements in a vector. The logical not operators of β and $\hat{\beta}$ are $\bar{\beta}$ and $\bar{\hat{\beta}}$, respectively. And $.*$ is the element-wise product. As shown in Table 2, it can be obviously seen that our method gets the best β-sensitivity performance across all cases of simulated experiments. With the training sample size $n = 100$, our method attains more than about 10% β-sensitivity gain compared with the second best results under all noise parameters. It implies that our method is superior to other competing methods in identifying significant features. For the β-specificity, Concate_EN and DIABLO achieve the best and second best results. The proposed MSPL method performs slightly worse than these two methods. Although Concate_EN and Concate_SPL achieved the excellent β-specificity performance in simulated experiments, concatenation-based methods have an imbalance problem for the identified multimodal features (See real dataset experiments part).

TABLE 2. Feature selection performance (%) of different models on simulated data with varying noise parameters and sample sizes. The mean β-sensitivity and β-specificity (±sd) over 30 repetitions of the experiments, and the best results of β-sensitivity and β-specificity are highlighted in bold.

Sample size/ Noise		Concate_EN	Concate_RF	Concate_SPL	Ensemble_EN	Ensemble_RF	Ensemble_SPL	DIABLO	MSPL	Concate_EN	Concate_RF	Concate_SPL	Ensemble_EN	Ensemble_RF	Ensemble_SPL	DIABLO	MSPL
		β-sensiticity								β-specificity							
n=100	$\sigma = 0$	41.48 ± 14.37	23.33 ± 7.42	40.49 ± 7.01	66.17 ± 10.19	53.95 ± 8.18	70.00 ± 9.79	45.43 ± 7.96	**84.69 ± 3.05**	**99.50 ± 0.43**	93.37 ± 0.20	99.23 ± 0.31	97.95 ± 0.96	80.42 ± 0.31	98.00 ± 0.44	99.26 ± 0.32	97.08 ± 0.37
	$\sigma = 0.4$	32.10 ± 12.74	20.62 ± 9.16	33.95 ± 10.88	62.59 ± 11.02	49.26 ± 9.49	67.16 ± 9.26	43.58 ± 8.18	**76.54 ± 8.88**	**99.31 ± 0.49**	93.25 ± 0.17	99.26 ± 0.29	97.97 ± 0.86	80.40 ± 0.32	97.77 ± 0.68	99.24 ± 0.32	96.15 ± 0.66
	$\sigma = 0.8$	29.51 ± 10.85	20.12 ± 6.35	28.02 ± 9.06	46.30 ± 15.77	45.68 ± 9.55	46.54 ± 10.05	34.32 ± 8.96	**54.32 ± 5.92**	99.34 ± 0.62	93.26 ± 0.19	99.22 ± 0.38	97.96 ± 1.16	80.43 ± 0.29	97.55 ± 0.90	**99.39 ± 0.27**	96.53 ± 0.59
n=150	$\sigma = 0$	72.10 ± 9.21	34.81 ± 9.31	75. 68 ± 9.32	88.40 ± 6.51	69.01 ± 8.22	90.62 ± 5.56	61.92 ± 10.90	**93.21 ± 5.33**	99.20 ± 0.45	90.18 ± 0.22	99.14 ± 0.34	97.86 ± 1.10	72.84 ± 0.39	97.04 ± 0.82	**99.38 ± 0.37**	96.41 ± 0.86
	$\sigma = 0.4$	63.70 ± 10.27	32.72 ± 8.26	64.44 ± 9.90	83.83 ± 9.50	66.42 ± 7.84	84.94 ± 7.78	60.62 ± 8.89	**89.51 ± 1.97**	99.25 ± 0.48	90.16 ± 0.27	98.99 ± 0.37	97.76 ± 1.13	72.93 ± 0.45	96.67 ± 0.76	**99.31 ± 0.33**	95.76 ± 0.62
	$\sigma = 0.8$	50.62 ± 14.07	28.77 ± 8.41	52.84 ± 10.29	67.65 ± 11.91	60.74 ± 8.40	71.48 ± 10.62	51.60 ± 8.13	**78.52 ± 3.43**	**99.22 ± 0.61**	90.17 ± 0.27	98.89 ± 0.45	97.65 ± 1.15	72.70 ± 0.41	96.23 ± 0.89	99.14 ± 0.39	94.82 ± 0.53
n=200	$\sigma = 0$	87.90 ± 7.01	47.04 ± 7.73	88.36 ± 9.10	94.67 ± 3.28	78.27 ± 7.70	94.44 ± 4.21	76.30 ± 9.90	**96.78 ± 4.62**	99.09 ± 0.53	87.24 ± 0.25	99.01 ± 0.37	98.17 ± 0.77	66.03 ± 0.39	96.56 ± 0.71	**99.30 ± 0.33**	96.72 ± 0.83
	$\sigma = 0.4$	81.98 ± 8.18	41.98 ± 8.77	82.26 ± 6.85	92.20 ± 5.56	77.16 ± 7.73	92.35 ± 6.22	72.59 ± 9.36	**93.94 ± 4.29**	99.05 ± 0.68	87.01 ± 0.25	98.95 ± 0.41	97.81 ± 1.41	66.05 ± 0.34	96.26 ± 0.87	**99.24 ± 0.35**	95.45 ± 0.93
	$\sigma = 0.8$	69.63 ± 12.20	42.72 ± 8.95	69.66 ± 8.25	82.22 ± 8.50	73.83 ± 6.30	82.22 ± 7.81	64.32 ± 9.40	**90.16 ± 3.34**	98.90 ± 0.94	87.12 ± 0.22	98.55 ± 0.44	97.55 ± 1.29	65.82 ± 0.62	95.38 ± 1.04	**99.23 ± 0.36**	97.23 ± 0.81

TABLE 3. The measurements of sample sizes and the number of features in each omics for four benchmark cancer datasets.

Benchmark datasets	Samples (high/low)	No. of features		
		mRNA	miRNA	methylation
KRCCC	122 (61/61)	17665	329	24960
LSCC	106 (53/53)	12042	352	23074
GBM	213 (105/108)	12042	534	1305
COAD	92 (33/59)	17814	312	23088

B. REAL DATASET EXPERIMENTS

We first compare our proposed method with seven other methods on four benchmark cancer datasets. In addition, we curated approximately 1000 samples from breast cancer multi-omics study, including four cancer subtypes. We use breast cancer multi-omics dataset to evaluate the performance of all competing methods in multi-class classification problem. Besides, we further analyze the significant multi-omics signatures identified by our proposed method in breast cancer data.

1) BENCHMARK CANCER DATASETS

Four benchmark multi-omics cancer datasets (mRNA, miRNA and DNA methylation) were obtained from [5]: Glioblastoma multi-forme (GBM), Kidney renal clear cell carcinoma (KRCCC), Lung squamous cell carcinoma (LSCC), Colon adenocarcinoma (COAD). Survival times were provided for each disease cohort by [5]. By using the median survival time, we dichotomized the samples into two classes in low and high survival times. A brief description of these four benchmark datasets is summarized in Table 3.

2) ANALYSIS ON BENCHMARK DATASETS

We evaluate the prediction and feature selection performance of the eight methods on benchmark cancer datasets using random partition. We randomly divide the datasets that approximate 70% of the datasets as the training samples and the rest as the test samples. We repeated the experiments 30 times, and report the average measurement.

Fig. 2 plots the box plot analysis of training and test accuracy calculated on four benchmark cancer datasets under 30 repetitions. For training accuracy, all methods get desirable results, except the DIABLO. For instance, the average training accuracy of DIABLO is 74.78%, 77.52% and 79.56% in three datasets GBM, KRCCC and LSCC respectively, while other methods have reached more than 90%. For the test accuracy, it can be easily seen that our method performs

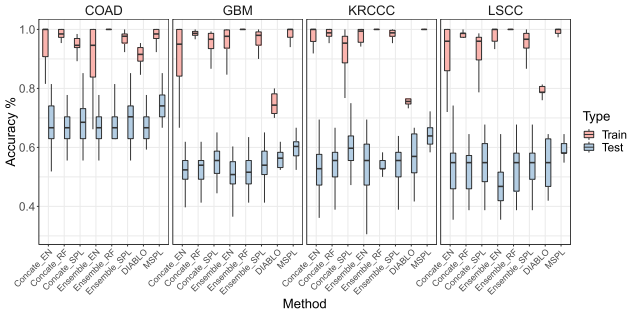


FIGURE 2. Boxplot diagram of training and test accuracy (%) for the methods with 30 repetitions in each benchmark cancer dataset.

best performance across all benchmark datasets. Our method demonstrates the best generalization performance, it attains approximate 5% test accuracy gain compared with other methods in almost all benchmark datasets, except the LSCC dataset. Moreover, methods with self-paced learning have better generalization performance than the corresponding without self-paced learning. For example, the average accuracy of Ensemble_SPL is superior to Ensemble_EN in all benchmark datasets.

Fig. 3 indicates the number of significant multi-omics signatures identified by all methods in the benchmark datasets. From the figure, we can easily find that the concatenation-based methods tend to be biased towards the more predictive signatures (mRNA and methylation). For instance, in KRCCC dataset, the average number of signatures of mRNA, miRNA and methylation selected by the Concate_EN are 14.37, 0.17 and 22.67, respectively. Compared to other omics significant features of mRNA and methylation, the concatenation-based methods almost failed to select the significant miRNA. Our proposed MSPL method, DIABLO and ensemble-based methods are robust in multi-omics feature selection.

3) BREAST CANCER MULTI-OMICS DATASET

We curated breast cancer multi-omics dataset (mRNA, miRNA and methylation) from the Cancer Genome Atlas (TCGA, data version 2015_11_01 for BRCA) [46] in order to achieve a systems characterization of breast cancer subtypes with multiple omics. This dataset contains four subtypes of breast cancer: Luminal A (LumA), Luminal B (LumB), Her2-enriched (Her2) and Basal-like (Basal), which have been reported the most replicated subtypes of human breast cancer [47]. The miRNA dataset was derived from two different

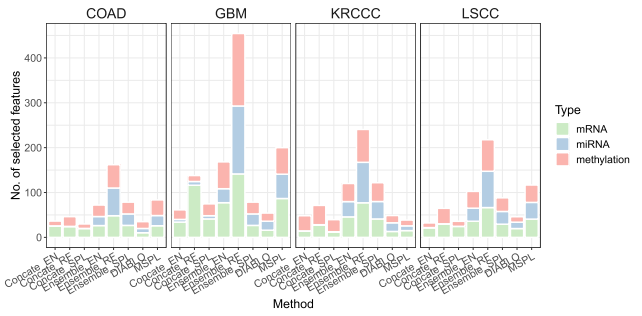


FIGURE 3. The stacked bar chart shows the average number of significant multi-omics signatures identified by all methods with 30 repetitions in each benchmark cancer dataset.

TABLE 4. A brief description of original and pre-processed breast cancer multi-omics data.

Omics types	Original data		Pre-processed data	
	Sample	No. of features	Sample	No. of features
mRNA	1212	20502	989	20502
miRNA (Genome Analyzer)	341	1046	989	1046
miRNA (Hiseq)	849	1046		
methylation (Illumina 27)	347	27578	989	25978
methylation (Illumina 450K)	889	485577		

TABLE 5. The number of each subtype in breast cancer dataset used to the training and test dataset.

Subtype	Basal	Her2	LumA	LumB	No. of Samples
Training dataset	102	40	346	122	610
Test dataset	76	38	188	77	379

Illumina technologies: The Illumina Genome Analyzer and the Illumina Hiseq. The methylation data was derived from two different platforms: the Illumina Methylation 27 and the Illumina 450K.

Normalization and pre-processing were used to clean the original multi-omics breast cancer dataset. Each omics data was normalized to log2-counts per million (logCPM) [48]. After normalization, we removed genes that counted to 0 in 70% of the samples. And the pre-processing of miRNA transcripts and methylation was the same as the mRNA. Besides that, we removed PAM50 labels in mRNA according to the [15]. Original and pre-processed breast cancer multi-omics data are described in Table 4. Table 5 demonstrates the number of each subtype in breast cancer dataset used to the training and test dataset.

4) ANALYSIS ON BREAST CANCER DATASETS

We evaluate the prediction performance of the eight methods in multiple cancer subtypes classification. Classification accuracy and Cohen's kappa (KAPPA) [49] are used as indicators for evaluating all methods. As shown in Table 6, we can conclude that the Concate_RF and the Ensemble_RF methods easily overfit to the training dataset, whose test accuracy and KAPPA are inferior to other methods, except for DIABLO. For the DIABLO method, the accuracy and KAPPA of the training and test dataset are both worst compared with other methods. While our method gets the third best result in the training dataset, it is less prone to overfitting issue, the test KAPPA achieves over 81%, which higher than other methods.

TABLE 6. Training and test prediction performance (%) of different methods on breast cancer dataset. The best results are highlighted in bold.

Method	Train		Test	
	Accuracy	KAPPA	Accuracy	KAPPA
Concate_EN	93.93	89.68	86.28	78.26
Concate_RF	99.18	98.64	79.16	66.70
Concate_SPL	94.10	90.10	86.81	79.38
Ensemble_EN	92.62	87.35	85.75	77.55
Ensemble_RF	**99.84**	**99.73**	79.42	66.87
Ensemble_SPL	92.79	87.67	86.28	78.43
DIABLO	83.11	73.79	74.41	64.04
MSPL	97.7	96.18	**88.13**	**81.66**

TABLE 7. The 20 top-ranked significant mRNA selected by all methods from breast cancer multi-omics dataset.

Concate_EN	Concate_RF	Concate_SPL	Ensemble_EN	Ensemble_RF	Ensemble_SPL	DIABLO	MSPL
SPDEF	SOX10	CCDC170	CCDC170	ANXA8	CCDC170	AGR3	IGBP1
CCDC170	CT62	SPDEF	SPDEF	SGO1	SPDEF	GATA3	CCDC170
AGR3	KIF11	AGR3	STARD3	CA12	STARD3	SLC44A4	ZNF552
GATA3	COL14A1	GATA3	AGR3	FAM107A	AGR3	XBP1	SUOX
CA12	AVPR2	RGMA	TBC1D31	COL17A1	TBC1D31	CA12	PGAP3
REEP6	KCNG2	TBC1D31	REEP6	RAD51AP1	REEP6	PRR15	GATA3
TBC1D31	KNTC1	PGAP3	GATA3	RSPH1	GATA3	TBC1D9	SLC44A4
PGAP3	ATAD2	SLC44A4	ZNF552	SGO2	ZNF552	AGR2	AGR3
RGMA	SLC25A5	STARD3	ARSG	TRPM6	ARSG	CT62	ZP1
SLC44A4	GYPC	CA12	CA12	ENG	CA12	DEGS2	DEFB1
STARD3	SRP54	KNSTRN	SLC44A4	CENPA	SLC44A4	THSD4	FZD9
GREB1	STIL	REEP6	GREB1	PLPP3	GREB1	ERBB4	TSLP
SRSF12	CDCA8	TMEM86A	SRSF12	CTDSP1	SRSF12	CCDC170	TCAM1P
TPX2	GATA3	ROPN1	PGAP3	LRRC47	PGAP3	TFF3	CKS1B
ZNF552	ART3	CDKN3	ENPP3	FOXM1	ENPP3	SPDEF	P3H2
KNSTRN	RMI2	LOC100130148	KNSTRN	SNAPC1	KNSTRN	DNALI1	MYLIP
TMEM86A	CRIM1	TSLP	RGMA	EML2	RGMA	ZNF552	CHEK2
AURKA	DEPDC1B	GREB1	DEFB1	LOC442454	DEFB1	SRARP	KRT6B
ROPN1	FANCB	SPNS2	TSLP	WT1	TSLP	C5AR2	STARD3
OSR1	PFKP	FAM83D	MRPS16	CSGALNACT1	MRPS16	KCNJ11	CA12

Fig. 4 shows the normalized confusion matrix to visualize the test performance of all methods. Since sample sizes of each subtype are an imbalance, we normalized the confusion matrix so that it contains only numbers between 0 and 1. It can be seen more intuitively from the figure that our method performs better than other methods in multi-class classification. For Concate_RF and Ensemble_RF, there is huge confusion of Her2 with other subtypes. Meanwhile, they are difficult to distinguish between LumA and LumB. The Concate_EN and Ensemble_EN methods also make it difficult to distinguish Her2 with other subtypes. Compared to Concate_EN and Ensemble_EN, Concat_SPL and Ensemble _SPL have improved in distinguishing Her2 and LumB, but inferior to DIABLO and our method. Although MSPL with slightly weak separation of Her2 and LumB compared to DIABLO, DIABLO is confused with LumA and LumB, and only 63% of the samples are correctly predicted to be LumA. This is significantly worse than other competing methods.

Tables 7, 8 and 9 summarize the 20 top-ranked significant features of mRNA, miRNA and methylation identified by all methods in breast cancer dataset, respectively. According to these three tables, we can intuitively find that the biological features selected by the concatenation-based methods are still unbalanced, consistent with the results of previous experiments. To explore the multi-omics features that are selected by MSPL in depth, we examine the interplay between 20 top-ranked selected features by our

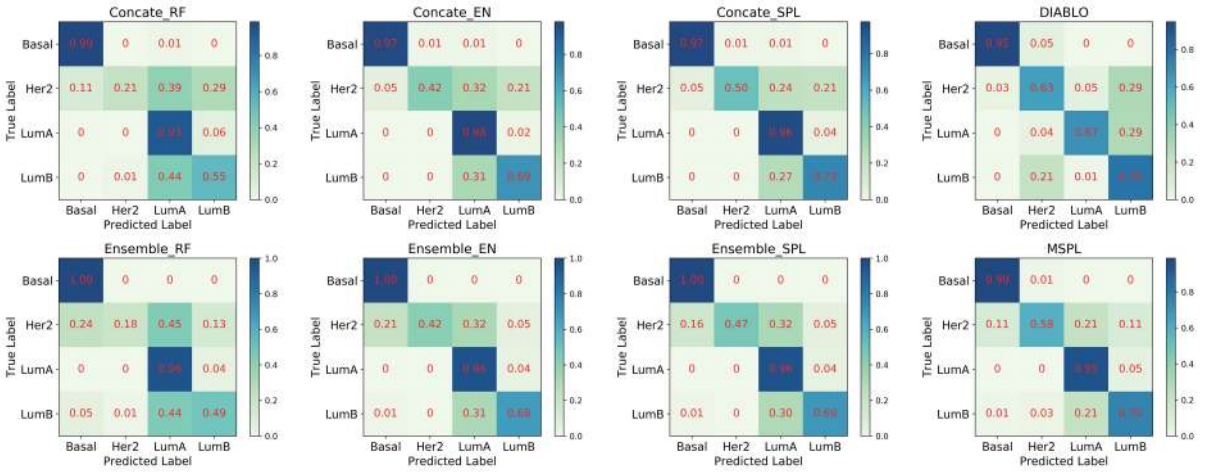


FIGURE 4. Normalized confusion matrix of all methods. The columns of each sub-figure are the truth labels and the rows of each sub-figure are the predicted labels. Therefore, the values of the diagonal elements represent the degree of the correctly predicted classes. The confusion is expressed by the false classified offdiagonal elements, since they are mistakenly confused with another class.

TABLE 8. The 20 top-ranked significant miRNA selected by all methods from breast cancer multi-omics dataset. .

Concate_EN	Concate_RF	Concate_SPL	Ensemble_EN	Ensemble_RF	Ensemble_SPL	DIABLO	MSPL
hsa-mir-1307	hsa-mir-377	hsa-mir-1307	hsa-mir-190b	hsa-let-7a-1	hsa-mir-190b	hsa-mir-190b	hsa-mir-190b
			hsa-mir-18a	hsa-let-7b	hsa-mir-18a	hsa-mir-18a	hsa-mir-455
			hsa-mir-584	hsa-let-7c	hsa-mir-584	hsa-mir-577	hsa-mir-135b
			hsa-mir-135b	hsa-let-7d	hsa-mir-135b	hsa-mir-17	hsa-mir-532
			hsa-mir-17	hsa-let-7i	hsa-mir-17	hsa-mir-135b	hsa-mir-584
			hsa-mir-532	hsa-mir-101-1	hsa-mir-532	hsa-mir-505	hsa-mir-500b
			hsa-mir-130b	hsa-mir-101-2	hsa-mir-130b	hsa-mir-584	hsa-mir-29c
			hsa-let-7c	hsa-mir-103-2	hsa-let-7c	hsa-mir-532	hsa-mir-106b
			hsa-mir-30a	hsa-mir-106b	hsa-mir-30a	hsa-mir-106b	hsa-let-7c
			hsa-mir-577	hsa-mir-1180	hsa-mir-577	hsa-mir-92a-2	hsa-mir-130a
			hsa-mir-1307	hsa-mir-1226	hsa-mir-378	hsa-mir-20a	hsa-mir-130b
			hsa-mir-378	hsa-mir-127	hsa-mir-1307	hsa-mir-19b-1	hsa-mir-342
			hsa-mir-29c	hsa-mir-1270-2	hsa-mir-29c	hsa-mir-500a	hsa-mir-30a
			hsa-mir-342	hsa-mir-128-2	hsa-mir-342	hsa-mir-224	hsa-mir-16-2
			hsa-mir-15b	hsa-mir-1287	hsa-mir-15b	hsa-mir-942	hsa-mir-1266
			hsa-mir-455	hsa-mir-1291	hsa-mir-455	hsa-mir-92a-1	hsa-mir-1468
			hsa-mir-204	hsa-mir-1301	hsa-mir-204	hsa-mir-455	hsa-mir-18a
			hsa-mir-16-2	hsa-mir-1307	hsa-mir-16-2	hsa-mir-502	hsa-mir-877
			hsa-mir-375	hsa-mir-130b	hsa-mir-375	hsa-mir-452	hsa-mir-132
			hsa-mir-224	hsa-mir-132	hsa-mir-224	hsa-mir-197	hsa-mir-185

TABLE 9. The 20 top-ranked significant methylation selected by all methods from breast cancer multi-omics dataset.

Concate_EN	Concate_RF	Concate_SPL	Ensemble_EN	Ensemble_RF	Ensemble_SPL	DIABLO	MSPL
LYPD4	ATP8B1	LYPD4	MIA	ABHD4	MIA	MIA	IL17B
VAT1	BBS9	VAT1	MUC1	ACR	MUC1	KRTDAP	MUC1
KLHDC8B	APOLD1	TNFRSF10A	ELK4	ACVR1	ELK4	RRM2	MIA
CYBRD1	BRCA2	TKTL2	RRM2	AKAP5	RRM2	A2ML1	NOL3
	FAM30A	CYBRD1	ADIPOQ	AKIRIN1	ADIPOQ	SRRM3	ADIPOQ
	KIR3DL3		IL17B	AMBP	IL17B	KRT39	KRTDAP
	MCM3		PREX1	ANXA9	PREX1	IFI35	ELK4
	CST8		FOXH1	AOAH	FOXH1	SPRTN	ORMDL3
	SUN1		BRCA2	AQP2	BRCA2	SCOC	CDKL1
	HSF4		KRT34	ART3	KRT34	BRCA2	VAT1
	RBM46		TNFRSF10A	ATP7B	TNFRSF10A	DOP1B	HMGA2
	NCAPH2		GSDMB	ATP8B1	GSDMB	DNALI1	TMEM71
	PREX1		ABCG8	ATRX	ABCG8	C6ORF15	PREX1
	PCSK1		HMGA2	BNIP2	HMGA2	EPHX1	CLCC1
	AHCY		HERPUD1	BRIX1	HERPUD1		EPHX1
	RBBP8NL		A2ML1	LRRC74A	A2ML1		MDP1
	PAGE5		SRRM3	RAB5IF	SRRM3		EFNA3
			MGC14436	ANXA2R	MGC14436		SLC7A11
			TMEM71	CAMKK1	HSD17B8		PROZ
			TAS2R13	CARD9	TMEM71		LGI4

method. Fig. 5 shows the interactive network of the 20 top-ranked features of mRNA and methylation selected by MSPL. We construct an integrative network of interactions among these features using the cBioPortal [50], [51] by integrating the biological interaction from publicly breast cancer dataset (METABRIC [52], [53]). Fig. 5 shows that mRNA features

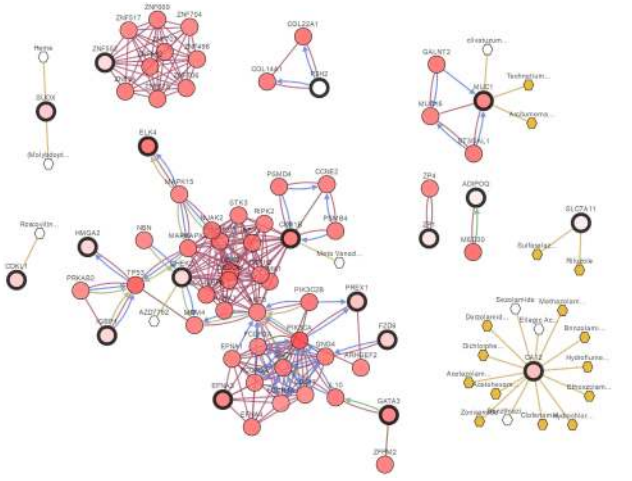


FIGURE 5. Integrative network view of 20 top-ranked features from mRNA and methylation selected by MSPL. The genes with thick border represent the selected features. The rest genes with thin border represent genes that are frequently altered in the public databases. The genes are gradient color-coded according to the alteration frequency based on data derived from METABRIC breast cancer database. The hexagons represent drug targets gene, and with yellow color represents FDA approved drug.

IGBP1, GATA3, FZD9, CKS1B, CHEK2 and methylation features ELK4, HMGA2, PREX1, EFNA3 in the maximum interactive network, which are connected to other frequently altered genes. In particular, GATA binding protein 3 (GATA3) is frequently mutated in breast cancer [54] and it is a critical transcription factor in mammary gland development and differentiation [55]. Checkpoint kinase 2 (CHEK2) is a tumor suppressor gene, which is a key component of the DNA damage-signaling pathway [56]. CHEK2 pathogenic variants are associated with breast cancer and colorectal families, and the risk of developing breast cancer is higher in carries of CHEK2 mutations [57]. Expression of high mobility group AT-hook 2(HMGA2) in cancer is associated with poor prognosis for patients. In the latest research, [58] suggest that HMGA2 is an attractive therapeutic target for com-

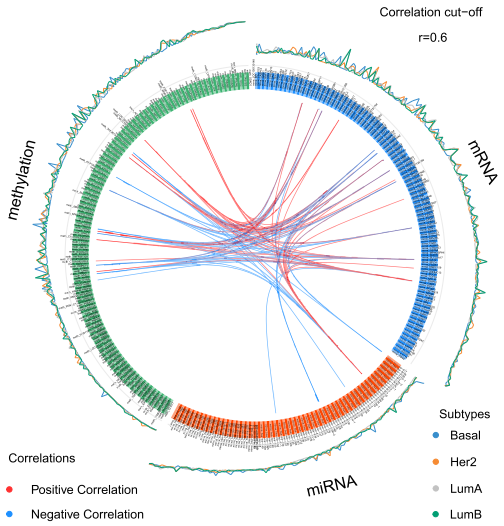


FIGURE 6. The circos plot shows the multi-omics significant signatures identified by the MSPL method. Each link indicates a Pearson correlation coefficient. The selected features are represented on the side of the circos plot, the side color indicates each omics type, and the optional line represents the expression level in each cancer subtype.

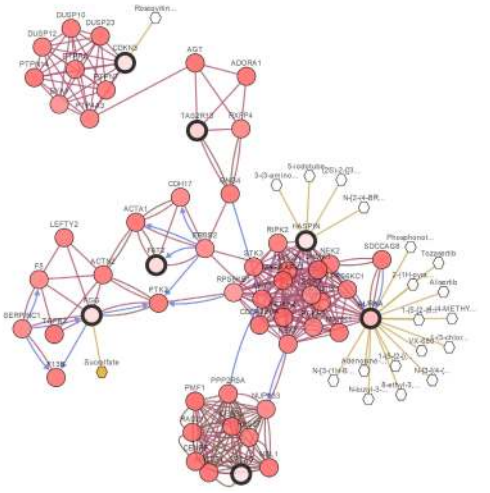


FIGURE 7. Integrative network view of 20 top-ranked correlation features from mRNA and methylation.

bination therapies using DNA damaging drugs. Moreover, CHEK2 and CKS1B are targeted by several cancer drugs. In other small networks of Fig. 5, there are several genes are connected to other frequently altered genes associated with breast cancer. For instance, MUC1 has been used in clinical practice as a serum tumor marker (CA15-3) for monitoring recurrence and response to the treatment of breast cancers [59]. Besides, we can easily find that SUOX, CDKL1, MUC1, SLC7A11 and CA12 are targeted by several cancer drugs, the yellow hexagon represents FDA approved drug targets gene.

We also use circos plot to present correlations between features identified by MSPL in Fig. 6. We use a Pearson correlation coefficient to calculate the association between features. The association between features is shown as a color link inside the figure to indicate a positive or negative correlation. Fig. 6 shows the balance of multi-omics significant features selected in our method. And, we construct an interactive network of the high correlation features from mRNA and methylation according to the 20 top-ranked Pearson correlation values. Fig. 7 demonstrates the maximum interactive network between these features, mRNA features AURKA, CDKN3, FAT2, HASPIN, SPC25 and methylation features FGG and TAS2R13 in the same interactive network and linked to other frequently altered genes. AURKA is a molecular barrier to the efficacy of PI3K-pathway inhibitors in breast cancer [60], [61]. And [62] discovered a novel AURKA-MEK1 interaction in breast cancer cells as a potential therapeutic target. Reference [63] found that SPC25 expression is quite high in basal-like subtype compared with other subtypes.

These above mentioned multi-omics features demonstrate that our proposed MSPL method can efficiently and robustly identify significant multi-omics signatures associated with breast cancer. MSPL not only efficiently selects signatures with high correlations between multi-omics, but also successfully identifies significant biological signatures that are associated with other frequently altered breast cancer genes.

V. CONCLUSION

Driven by technological advances, large-scale molecular omics datasets are in strong need of integrative machine learning methods for better utilize the multiple sources data to gain insight into complex biological systems from different levels and the development of predictive models. However, heavy noises, large p and small n problem, and data heterogeneity of omics data present significant computational challenges in applying the state-of-the-art machine learning methods for integrative analysis and predictive modeling from multi-omics data. In this paper, we propose a novel multi-omics data integration method MSPL that simultaneously identifies significant multi-omics signatures during the integration process and predicts the cancer subtypes. Compared with current state-of-the-art methods, our method performs robust in the presence of heavy noises and possesses excellent generalization performance. In addition, our method achieves the best performance in both binary and multi-class classification problems. Moreover, the proposed method also can work well on other classifiers (e.g. Support Vector Machine (SVM)). At last, the significant multi-omics signatures selected by our method in breast cancer multi-omics dataset are introduced in detail, which verifies the effectiveness and robustness of our method in feature selection. This work in progress is aimed at further developing effective machine learning method for integrative analysis and predictive modeling from multi-omics data, and discover potential biological signatures. This learning mechanism is hopeful to be extended to other multimodal problems and expands its range of applications.

DATA AVAILABILITY

The code is available at https://github.com/must-bio-team/MSPL. All computation is done in R.

REFERENCES

[1] M. Schena, D. Shalon, R. Davis, and P. Brown, "Quantitative monitoring of gene expression patterns with a complementary dna microarray," *Science*, vol. 270, no. 5235, pp. 467–470, Oct. 1995.

[2] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[3] J. S. Hamid, P. Hu, N. M. Roslin, V. Ling, C. M. Greenwood, and J. Beyene, "Data integration in genetics and genomics: Methods and challenges," *Hum. Genomics Proteomics*, vol. 2009, Jan. 2009, Art. no. 869093, doi: 10.4061/2009/869093.

[4] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen, "Pattern discovery and cancer gene identification in integrated cancer genomic data," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 11, pp. 4245–4250, 2013.

[5] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, p. 333, 2014.

[6] R. Burk *et al.*, "Integrated genomic and molecular characterization of cervical cancer," *Nature*, vol. 543, no. 7645, p. 378, 2017.

[7] R. R. Kitchen, V. S. Sabine, A. A. Simen, J. M. Dixon, J. M. S. Bartlett, and A. H. Sims, "Relative impact of key sources of systematic noise in affymetrix and illumina gene-expression microarray experiments," *BMC Genomics*, vol. 12, no. 1, 2011, Art. no. 589.

[8] G. Tini, L. Marchetti, C. Priami, and M.-P. Scott-Boyer, "Multi-omics integration—A comparison of unsupervised clustering methodologies," *Briefings Bioinf.*, vol. 20, no. 4, pp. 1269–1279, 2017.

[9] Y. Wang, D. J. Miller, and R. Clarke, "Approaches to working in high-dimensional data spaces: Gene expression microarrays," *Brit. J. Cancer*, vol. 98, no. 6, pp. 1023–1028, 2008.

[10] Y. Liang, C. Liu, X.-Z. Luan, K.-S. Leung, T.-M. Chan, Z.-B. Xu, and H. Zhang, "Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification," *BMC Bioinf.*, vol. 14, no. 1, 2013, Art. no. 198.

[11] Z.-Y. Yang, Y. Liang, H. Zhang, H. Chai, B. Zhang, and C. Peng, "Robust sparse logistic regression with the $L_q(0 < q < 1)$ regularization for feature selection using gene expression data," *IEEE Access*, vol. 6, pp. 68586–68595, 2018.

[12] Y. Li, F. X. Wu, and A. Ngom, "A review on machine learning principles for multi-view biological data integration," *Briefings Bioinform.*, vol. 19, no. 2, pp. 325–340, 2018.

[13] B. Ray, M. Henaff, S. Ma, E. Efstathiadis, E. R. Peskin, M. Picone, T. Poli, C. F. Aliferis, and A. Statnikov, "Information content and analysis methods for multi-modal high-throughput biomedical data," *Sci. Rep.*, vol. 4, Mar. 2014, Art. no. 4411.

[14] A. Mandal and P. Maji, "FaRoC: Fast and robust supervised canonical correlation analysis for multimodal omics data," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1229–1241, Apr. 2018.

[15] A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K.-A. Lê Cao, "DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays," *Bioinformatics*, vol. 35, no. 17, pp. 3055–3062, 2019.

[16] Y. Liu, V. Devescovi, S. Chen, and C. Nardini, "Multilevel omic data integration in cancer cell lines: Advanced annotation and emergent properties," *BMC Syst. Biol.*, vol. 7, no. 1, 2013, Art. no. 14.

[17] O. P. Günther, V. Chen, G. C. Freue, R. F. Balshaw, S. J. Tebbutt, Z. Hollander, M. Takhar, W. R. McMaster, B. M. McManus, P. A. Keown, and R. T. Ng, "A computational pipeline for the development of multi-marker bio-signature panels and ensemble classifiers," *BMC Bioinf.*, vol. 13, no. 1, 2012, Art. no. 326.

[18] N. Aben, D. J. Vis, M. Michaut, and L. F. A. Wessels, "TANDEM: A two-stage approach to maximize interpretability of drug response models based on multiple molecular data types," *Bioinformatics*, vol. 32, no. 17, pp. i413–i420, 2016.

[19] S. Ma, J. Ren, and D. Fenyö, "Breast cancer prognostics using multi-omics data," *AMIA Joint Summits Transl. Sci.*, vol. 2016, pp. 52–59, Jul. 2016.

[20] J. J. Hughey and A. J. Butte, "Robust meta-analysis of gene expression using the elastic net," *Nucleic Acids Res.*, vol. 43, no. 12, p. e79, 2015.

[21] A. Sokolov, D. E. Carlin, E. O. Paull, R. Baertsch, and J. M. Stuart, "Pathway-based genomics prediction using generalized elastic net," *PLoS Comput. Biol.*, vol. 12, no. 3, 2016, Art. no. e1004790.

[22] M. A. van de Wiel, T. G. Lien, W. Verlaat, W. N. van Wieringen, and S. M. Wilting, "Better prediction by use of co-data: Adaptive group-regularized ridge regression," *Statist. Med.*, vol. 35, no. 3, pp. 368–381, 2016.

[23] K.-A. Lê Cao, S. Boitard, and P. Besse, "Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems," *BMC Bioinf.*, vol. 12, no. 1, 2011, Art. no. 253.

[24] A. Tenenhaus, C. Philippe, V. Guillemot, K.-A. Lê Cao, J. Grill, and V. Frouin, "Variable selection for generalized canonical correlation analysis," *Biostatistics*, vol. 15, no. 3, pp. 569–583, 2014.

[25] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1189–1197.

[26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[27] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B (Stat. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.

[28] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, nos. 5–6, pp. 877–905, 2008.

[29] Z. Xu, H. Zhang, Y. Wang, X. Chang, and Y. Liang, "$L_{1/2}$ regularization," *Sci. China Inf. Sci.*, vol. 53, no. 6, pp. 1159–1169, 2010.

[30] X.-Y. Liu, S. Wang, H. Zhang, H. Zhang, Z.-Y. Yang, and Y. Liang, "Novel regularization method for biomarker selection and cancer classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

[31] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[32] E. A. Ni and C. X. Ling "Supervised learning with minimal effort," in *Advances in Knowledge Discovery and Data Mining (PAKDD)* (Lecture Notes in Computer Science), vol 6119, M. J. Zaki, J. X. Yu, B. Ravindran, and V. Pudi, (Eds). Berlin, Germany: Springer, 2010.

[33] S. Basu and J. Christensen, "Teaching classification boundaries to humans," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 109–115.

[34] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 547–556.

[35] H. Chai, Z.-N. Li, D.-Y. Meng, L.-Y. Xia, and Y. Liang, "A new semi-supervised learning model combined with Cox and SP-AFT models in cancer survival analysis," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 13053.

[36] L.-Y. Xia and Q.-Y. Wang, "QSAR classification modeling for bioactivity of molecular structure via SPL-logsum," 2018, *arXiv:1804.08615*. [Online]. Available: https://arxiv.org/abs/1804.08615

[37] Y. Ren, P. Zhao, Y. Sheng, D. Yao, and Z. Xu, "Robust softmax regression for multi-class classification with self-paced learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2641–2647.

[38] C. Xu, D. Tao, and C. Xu, "Multi-view self-paced learning for clustering," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 3974–3980.

[39] Y. Ren, X. Que, D. Tao, and Z. Xu, "Self-paced multi-task clustering," *Neurocomputing*, vol. 350, no. 1, pp. 212–220, Jul. 2019.

[40] D. Meng, Q. Zhao, and L. Jiang, "A theoretical understanding of self-paced learning," *Inf. Sci.*, vol. 414, pp. 319–328, Nov. 2017.

[41] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, p. 1, 2010.

[42] F. Ma, D. Meng, Q. Xie, Z. Li, and X. Dong, "Self-paced co-training," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 2275–2284.

[43] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[44] L.-Y. Xia, Y.-W. Wang, D.-Y. Meng, X.-J. Yao, H. Chai, and Y. Liang, "Descriptor selection via log-sum regularization for the biological activities of chemical structure," *Int. J. Mol. Sci.*, vol. 19, no. 1, p. 30, 2017.

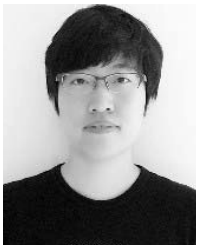
[45] W. Zhang, Y.-W. Wan, G. I. Allen, K. Pang, M. L. Anderson, and Z. Liu, "Molecular pathway identification using biological network-regularized logistic models," *BMC Genomics*, vol. 14, no. 8, 2013, Art. no. S7.

[46] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, pp. 1113–1120, 2013.

[47] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lønning, P. O. Brown, A.-L. Børresen-Dale, and D. Botstein, "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 14, pp. 8418–8423, 2003.

[48] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts," *Genome Biol.*, vol. 15, no. 2, 2014, Art. no. R29.

[49] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The kappa statistic," *Family Med.*, vol. 37, no. 5, pp. 360–363, 2005.
[50] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci. Signaling*, vol. 6, no. 269, p. pl1, 2013.
[51] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schul, "The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data," *Cancer Res.*, vol. 2, no. 5, pp. 401–404, 2012.
[52] C. Curtis *et al.*, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, 2012.
[53] B. Pereira *et al.*, "The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes," *Nature Commun.*, vol. 7, May 2016, Art. no. 11479.
[54] M. Takaku, S. A. Grimm, J. D. Roberts, K. Chrysovergis, B. D. Bennett, P. Myers, L. Perera, C. J. Tucker, C. M. Perou, and P. A. Wade, "GATA3 zinc finger 2 mutations reprogram the breast cancer transcriptional network," *Nature Commun.*, vol. 9, no. 1, 2018, Art. no. 1059.
[55] V. Larsen *et al.*, "Germline genetic variants in GATA3 and breast cancer treatment outcomes in SWOG S8897 trial and the pathways study," *Clin. Breast Cancer*, vol. 19, no. 4, pp. 225.e2–235.e2, 2019.
[56] R. Kumar, X. Pei, P. Selenica, H. Y. Wen, S. Powell, M. Robson, N. Riaz, J. S. Reis-Filho, B. Weigelt, and D. Mandelker, "Abstract P4-04-01: The landscape of somatic genetic alterations in breast cancers from CHEK2 germline mutation carriers," Tech. Rep., 2019.
[57] P. Apostolou and I. Papasotiriou, "Current perspectives on CHEK2 mutations in breast cancer," *Breast Cancer, Targets Therapy*, vol. 9, pp. 331–335, May 2017.
[58] S. Hombach-Klonisch, F. Kalantari, M. R. Medapati, S. Natarajan, S. N. Krishnan, A. Kumar-Kanojia, T. Thanasupawat, F. Begum, F. Y. Xu, G. M. Hatch, M. Los, and T. Klonisch, "HMGA2 as a functional antagonist of PARP1 inhibitors in tumor cells," *Mol. Oncol.*, vol. 13, no. 2, pp. 153–170, 2019.
[59] A. Shimizu, K. Hatanaka, Y. Hatanaka, K. Naruchi, M. Sato, H. Kase, T. Mitsuhashi, H. Yamashita, and Y. Matsuno, "Cancer-associated muc1 epitope-recognizing antibody as a novel immunohistochemical marker for breast carcinoma," *Cancer Res.*, vol. 78, no. 13, p. 4616, 2018.
[60] H. J. Donnella, J. T. Webber, R. S. Levin, R. Camarda, O. Momcilovic, N. Bayani, K. N. Shah, J. Korkola, K. M. Shokat, A. Goga, J. Gordan, and S. Bandyopadhyay, "Kinome rewiring reveals AURKA limits PI3K-pathway inhibitor efficacy in breast cancer," *Nature Chem. Biol.*, vol. 14, pp. 768–777, 2018. doi: 10.1038/s41589-018-0081-9.
[61] H. J. Donnella, J. T. Webber, R. S. Levin, R. Camarda, O. Momcilovic, N. Bayani, K. N. Shah, J. E. Korkola, K. M. Shokat, A. Goga, J. D. Gordan, and S. Bandyopadhyay, "Kinome rewiring reveals AURKA limits pi3k-pathway inhibitor efficacy in breast cancer," *Nature Chem. Biol.*, vol. 14, no. 8, pp. 768–777, 2018.
[62] S. Gandhi, M. Gil, T. Khoury, K. Takabe, I. Puzanov, I. Gelman, A. D'Assoro, and M. Opyrchal, "A novel interaction of AURKA with MAPK pathway in breast cancer cells as a potential therapeutic target [abstract]," in *Proc. San Antonio Breast Cancer Symp.*, San Antonio, TX, USA, Dec. 2018, p. P2-06.
[63] R. Pathania, S. Ramachandran, G. Mariappan, P. Thakur, H. Shi, J.-H. Choi, S. Manicassamy, R. Kolhe, P. D. Prasad, S. Sharma, B. L. Lokeshwar, V. Ganapathy, and M. Thangaraju, "Combined inhibition of DNMT and HDAC blocks the tumorigenicity of cancer stem-like cells and attenuates mammary tumor growth," *Cancer Res.*, vol. 76, no. 11, pp. 3224–3235, 2016.



ZI-YI YANG received the B.Sc. degree from the Tongji University Zhejiang College, China, in 2013, and the M.Sc. degree from the Macau University of Science and Technology, Macau, in 2015, where she is currently pursuing the Ph.D. degree. From 2015 to 2017, she was a Bioinformatics Analysis Engineer with BGI and China National GeneBank, Shenzhen, China. Her current research interests include machine learning, feature selection, data integration, bioinformatics, self-paced learning, and few-shot learning.

LIANG-YONG XIA received the B.Sc., M.Sc., and Ph.D. degrees from the Macau University of Science and Technology, Macau, in 2014, 2016, and 2019, respectively. He is currently a Postdoctoral Researcher with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests are in computational intelligent, including machine learning, big data analysis, feature selection, and systems biology.

HUI ZHANG received the B.Sc. degree from Northwest University, Xi'an, China, in 2013, and the M.Sc. degree from Northwest University, Xi'an, China, in 2016. She is currently pursuing the Ph.D. degree with the Macau University of Science and Technology. From 2016 to 2017, she was an Algorithm Engineer with Merit Data, Xi'an, China. Her current research interests include machine learning, feature selection, non-convex regularization, high-dimensional statistics, and bioinformatics.

YONG LIANG received the B.Sc. and M.Sc. degrees in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 1996 and 1999, respectively, and the Ph.D. degree in computer science from The Chinese University of Hong Kong, Hong Kong, in 2003. From 2004 to 2006, he was a Postdoctoral Researcher with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, and later joined Shantou University, China. He also joined the Faculty of Information Technology, Macau University of Science and Technology, in 2007, where he is currently a Professor. He has authored and coauthored more than 60 articles and the articles have cited more than 1 300. His research interests are in computational intelligent, including machine learning, big data analysis, evolutionary computation, and bioinformatics.