IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 October 2022, accepted 30 October 2022, date of publication 10 November 2022, date of current version 18 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3221089

RESEARCH ARTICLE

MMMF: Multimodal Multitask Matrix Factorization for Classification and Feature Selection

JEONGYOUNG HWANG[1] AND HYUNJU LEE[1,2], for the Alzheimer's Disease Neuroimaging Initiative

[1]AI Graduated School, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea
[2]School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

Corresponding author: Hyunju Lee (hyunjulee@gist.ac.kr)

This work was supported in part by the Bio & Medical Technology Development Program of NRF funded by the Korean Government [Ministry of Science and ICT (MSIT)] under Grant NRF-2018M3C7A1054935; in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant funded by the Korea Government (MSIT) (Development of Intelligent SW Systems for Uncovering Genetic Variation and Developing Personalized Medicine for Cancer Patients With Unknown Molecular Genetic Mechanisms) under Grant 2019-0-00567; in part by the Artificial Intelligence Graduate School Program, Gwangju Institute of Science and Technology, under Grant 2019-0-01842.

ABSTRACT Integration of multiple biological datasets is crucial to understand comprehensive biological mechanisms with the aid of a rapid development of biomedical technology. However, the predictive modeling for such an integrated dataset faces two major challenges, namely, heterogeneity and imbalance in the acquired data. Thus, in this study, we present a method for the integration of multiple biological datasets called multimodal multitask matrix factorization (MMMF) to address these issues. The MMMF uses matrix factorization (MF) to integrate data from multiple heterogeneous biological datasets, and oversampling is applied to resolve the imbalanced data during the training step. Moreover, gradient surgery is used for multitask (MF and classification) learning to increase the quantity of classification information by projecting the gradients of the MF that conflict with the classification gradient onto the normal plane of a classification gradient. We demonstrate that MMMF outperforms other state-of-the-art biomedical classification models in binary and multi-class classification problems using five biological datasets. We also show that MMMF can be used as a feature selection approach for finding biomarkers that help in classification. The source code of the MMMF is available at https://github.com/DMCB-GIST/MMMF.

INDEX TERMS Classification, feature selection, matrix factorization, multimodal, multitask.

I. INTRODUCTION

A rapid development in the field of biomedical technology has facilitated the acquisition of various types of biological data (e.g., gene expression, DNA methylation, and microRNA [miRNA] expression). In general, each type of biological data can only capture a part of the biological complexity and provide independent and complementary information. Therefore, to decipher complex biological mechanisms, it is necessary to find consensus information by integrating several types of biological data. It has been demonstrated, particularly for disease diagnosis, that when multiple biological datasets are integrated, the prediction accuracy is improved compared to that of a single biological dataset [1], [2], [3], [4], [5].

However, the analysis and predictive modeling of such biological datasets can be achieved only after solving the following two challenges. First, the integration of these datasets is difficult owing to their heterogeneous properties. For example, the distributions of various biological data

The associate editor coordinating the review of this manuscript and approving it for publication was Vicente Alarcon-Aquino.

differ, as do the number of features in each of these data; additionally, the types of data differ (e.g., mutations involve category values, but gene expression involves continuous values). Hence, using traditional multimodal fusion methods (e.g., early fusion and late fusion) may lead to an inferior classification performance.

Several models that use new fusion methods to integrate multiple biological datasets have recently been proposed. The canonical correlation analysis (CCA) based data integration analysis for biomarker discovery using latent components (DIABLO) [6] is a supervised framework which has been established by extending the sparse generalized CCA [7]. As DIABLO maximizes the covariance between linear combinations of variables, it is possible to select correlated features from each biological dataset. Autoencoder (AE) based concatenation AE (Concat AE) [8] and cross-modality AE (Cross AE) [8] are proposed for breast cancer survival prediction. Concat AE is a model that predicts by concatenating the hidden layer of AE that obtains complementary information from each biological dataset, and Cross AE maximizes the agreement between multimodal data to achieve modality-invariant representation. The matrix factorization (MF) based collective deep matrix factorization (CDMF) [9] approach exhibits a high performance by building a non-linear hierarchical deep matrix decomposition that decomposes each biological dataset into two matrices according to the classification information. These two matrices include the multiple coefficient matrices that represent specific characteristics of each of the modalities and a common basis matrix with consensus information from multimodal data.

Second, the problem of imbalance in data needs to be resolved. Because most machine learning or deep learning models assume relatively balanced distributions, dealing with imbalanced datasets presents significant challenges [10]. One of the traditional methods for resolving an imbalanced dataset involves matching the number of each class through undersampling or oversampling. However, overfitting occurs when the samples of the insufficient class are increased by oversampling. Hence, a synthetic minority oversampling technique (SMOTE) [11] has been proposed for oversampling through segments joining the minority samples and their ''k'' minority-class nearest neighbors.

In this study, we propose the multimodal multitask matrix factorization (MMMF) approach to solve the two aforementioned challenges using various recently proposed methods. To fuse heterogeneous multimodal data, we divided these data into multiple coefficient matrices and a common basis matrix similar to the CDMF method. During this process, we encountered two problems owing to the simultaneous processing of the two tasks, namely, MF and classification.

First, the goal of MF is to decompose multimodal data into multiple coefficient matrices and a common basis matrix, such that oversampling is not required. However, classification models require oversampling to solve for overfitting and high variance. To solve this problem, we proceed with MF with the raw data itself, and then trained a classifier by oversampling only a common basis matrix, which was used as an input to the classifier. Thus, there are no oversampling effects on MF, and the MMMF can be trained by applying oversampling to the classification model alone.

Second, a common basis matrix is trained by both the MF and classification. Here, MF was guided using relatively little information regarding the classification. To solve the multitask problem, we used a projecting conflicted gradient (PCGrad) [12] that projects each gradient onto the normal plane of the other gradient to minimize gradient interference in case of conflicts between gradients. In this study, PCGrad was used to provide a large amount of classification information to a common basis matrix by projecting the MF gradients that conflicted with the classification gradient onto the normal plane of the classification gradient, which we call gradient surgery.

We demonstrate the classification performance of MMMF compared to other state-of-the-art biomedical classification models in binary and multi-class classification problems using multiple biological datasets, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [13], Religious Orders Study and Memory and Aging Project (ROSMAP), breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), and colon adenocarcinoma (COAD) datasets. In addition, we demonstrate that MMMF can be used for feature selection of important biomarkers for classification.

The remainder of this study is organized as follows. Section II introduces the related work on MF. Section III describes the details of the proposed model. The experimental setup and results are described in Section IV. Finally, the conclusions of the work are proposed in Section V.

II. RELATED WORK

This section briefly reviews MF, collective matrix factorization (CMF) [14], and CDMF related to our model, which can combine heterogeneous multimodal data. Table 1 lists the explanations of the notations used in Eq. 1-6 for clarity.

A. MATRIX FACTORIZATION (MF)

MF is a method of representing a matrix via two low-rank matrices. The original matrix is $X \in \mathbb{R}^{n\times m}$ and the two low-rank matrices are defined by a basis matrix ($U \in \mathbb{R}^{n\times k}$) and coefficient matrix ($V \in \mathbb{R}^{m\times k}$); U, V can be trained using the following objective function:

$$\min_{U,V} \|X - UV^\top\|_F^2, \tag{1}$$

where $\|\cdot\|_F^2$is the frobenius norm, and the rank $k < \min\{n, m\}$. However, MF has two major limitations: (1) it considers linear relationship and (2) it is applicable for a single modality.

B. COLLECTIVE MF (CMF)

The CMF model extends the functionality of MF beyond a single modality. CMF factorizes a common basis matrix

TABLE 1. The notations used in Eq. 1-6.

NOTATION	DIMENSION	DESCRIPTION
n	-	Number of samples.
M	-	Number of modalities.
m_i	-	Dimension of i-th modality.
c	-	Number of label categories.
σ	-	Activation function.
k	-	Dimension of a common basis matrix.
$f(\cdot)$	-	Classifier.
θ	-	Parameters of $f(\cdot)$.
$g_i(\cdot)$	-	Deep neural network of i-th modality with r layers.
$D_{i,j}$	-	Dimension of i-th modality coefficient matrix in j-th layer.
X_i	$n \times m_i$	i-th modality.
Y	$n \times c$	Label.
U	$n \times k$	A common basis matrix.
V_i	$m_i \times k$	i-th modality coefficient matrix.
$g_i(V_i)$	$m_i \times k$	i-th modality coefficient matrix with non-linearity.
$V_{i,j}$	$D_{i,j-1} \times D_{i,j}$	i-th modality coefficient matrix in j-th layer.

($U \in \mathbb{R}^{n\times k}$) and multiple coefficient matrices ($V_i \in \mathbb{R}^{m_i\times k}$) that can have different value types. We define M as the number of modalities and $X_i \in \mathbb{R}^{n\times m_i}$ as the i-th modality; U and V_i can be trained using the objective function as follows:

$$\sum_{i=1}^{M} \min_{U,V_i} \|X_i - UV_i^\top\|_F^2, \tag{2}$$

where the rank $k < \min\{n, m_i|\forall i\}$. Although CDF can be applied to multimodal data, its applicability is limited owing to its inconsideration of non-linearity.

C. COLLECTIVE DEEP MATRIX FACTORIZATION

The CDMF approach decomposes the multiple coefficient matrices into a low-rank matrix and adds non-linearity using the activation function (σ). If the dimensions of the i-th modality coefficient matrix are reduced into r factors with $D_{i,0}, D_{i,1}, \ldots, D_{i,r}$, then the dimension of i-th modality coefficient matrix in j-th layer can be obtained as follows:

$$\begin{aligned} &V_{i,j} \in \mathbb{R}^{D_{i,j-1}\times D_{i,j}} \\ &\text{s.t. } D_{i,0} = m_i, D_{i,r} = k, D_{i,j-1} > D_{i,j}, \forall i, j. \end{aligned} \tag{3}$$

When non-linearity is added, the i-th modality coefficient matrix with non-linearity can be represented as follows:

$$g_i(V_i) = \sigma(V_{i,1}(\ldots, \sigma(V_{i,r-1}V_{i,r}))). \tag{4}$$

In Eq. (4), using a non-linear activation function, $g_i(V_i)$ can capture complex biological processes that cannot be achieved through linear matrix multiplication. The $g_i(V_i)$ also serves as a non-linear mapping between the i-th modality (X_i) and a common basis matrix (U). The reconstruction loss resultant expressing the original modality using the two matrices can be computed as follows:

$$\sum_{i=1}^{M} \min_{U,V_i} \|X_i - Ug_i(V_i)^\top\|_F^2. \tag{5}$$

Further, to deliver the classification information to U and V_i, U is used as the input to the classifier. We define $f(\cdot)$ as the classifier, θ as the parameters of the $f(\cdot)$, I_t is the train index, $U[I_t]$ as the rows of U corresponding to the train index, $Y[I_t]$ as the corresponding label, and $L_{\text{clf}}(\cdot)$ as the loss of the classifier; the total objective function of CDMF is expressed as follows:

$$\min_{U[I_t],\theta} L_{\text{clf}}(f(U[I_t]), Y[I_t]) + \min_{U,V_i} \sum_{i=1}^{M} \|X_i - Ug_i(V_i)^\top\|_F^2. \tag{6}$$

From Eq. (6), it is possible to train U, which has consensus information of multimodal data, and V_i, which has complementary information of each modality. Moreover, using $U[I_t]$ as an input to the classifier, it is possible to learn classification information in each matrix as well. Thus, CDMF can account for the non-linear relationship and can be applied to multimodal data in addition to being trained to deliver classification information.

III. PROPOSED METHODOLOGY

In this section, we discuss the details of the proposed model and how to select features that help classification using the proposed model.

A. MULTIMODAL MULTITASK MATRIX FACTORIZATION

Figure 1 shows the forward and backward of the proposed MMMF model. The MMMF model (1) integrates heterogeneous multimodal data, (2) solves the imbalance problem by oversampling to avoid overfitting and reduce variances of the classifier, and (3) solves the multitask problem, implying that a common basis matrix is trained by both MF and classification, using gradient surgery. The process of training MMMF to counter these three issues is described as follows.

The first step involves the integration of the heterogeneous multimodal data, comparable to the CDMF method. The heterogeneous multimodal data $(X_1, X_2, \ldots, X_M)$ matched with the sample are decomposed into a common basis matrix (U) and the multiple coefficient matrices $(V_1, \ldots, V_M)$. Each coefficient matrix is decomposed into r factors, while simultaneously reducing dimensions and adding non-linearity ($g_i(V_i) = \sigma(V_{i,1}(\ldots, \sigma(V_{i,r-1}V_{i,r})))$).

The second step involves solving the imbalance problem through oversampling. To apply oversampling only to the classification task excluding the MF task, $U[I_t]$ is constructed, which is used as the training set of the classifier. Next, the SMOTE oversampling is performed using $U[I_t]$ and the corresponding label $Y[I_t]$. The SMOTE oversampling generates new samples and their labels for relatively insufficient classes in $Y[I_t]$, and they were combined with existing $U[I_t]$ and $Y[I_t]$, resulting in $O(U[I_t])$ and $O(Y[I_t])$,

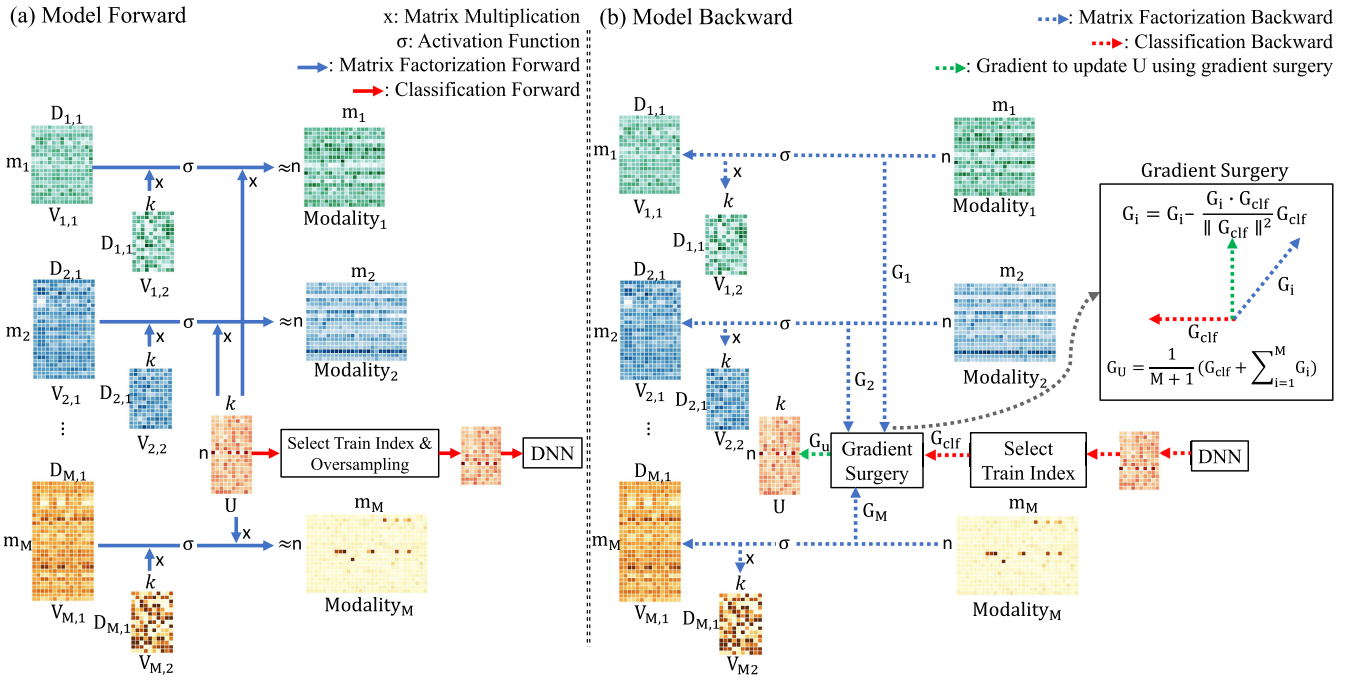

FIGURE 1. Illustration of forward and backward of the MMMF. MMMF can integrate heterogeneous multimodal data by decomposing them into a common basis matrix (U) and multiple coefficient matrices ($V_{1,1}, \ldots, V_{M,2}$); it adds non-linearity using the activation function (σ). (a) Model forward: MMMF performs oversampling by indexing the part of the train index in U that have the consensus information in the heterogeneous multimodal data, and subsequently, supplies it as an input to the classifier to solve the imbalance problem. (b) Model backward: The gradient propagated from the classifier is indexed by the MMMF, except for the gradient that was added due to oversampling in the model forward. After indexing, the gradient for updating U (G_u) is propagated by applying gradient surgery which projects the gradients propagated by MF ($G_{MF} = [G_1, \ldots, G_M]$) that conflicts with the gradient propagated by the classifier (G_{clf}) onto the normal plane of the G_{clf} to solve the multitask problem. By applying gradient surgery, a large amount of classification information can be delivered to U.

respectively. Consequently overfitting and high variance can be reduced because the classifier can be trained using samples with the equal distribution of classes.

The imbalanced problem can be solved by oversampling; however, an additional process needs to be executed in the model backward process. In the model backward process, the gradient propagated from the classifier (G_{clf}) for updating U was also oversampled; thus, it is necessary to exclude the oversampled indexes from G_{clf} before propagating to U. Moreover, as only the train index of U is input to the classifier, the G_{clf} in the validation and test index ($G_{\text{clf}}[\bar{I}_t]$) has no value; thus, $G_{\text{clf}}[\bar{I}_t]$ is declared as 0 to perform gradient surgery.

The third step requires performing gradient surgery to solve the multitask problem. In Eq. (6), it can be seen that both gradients propagated from MF ($G_{\text{MF}} = [G_1, \ldots, G_M]$) and gradient propagated from classifier (G_{clf}) are used to update U, which denotes that gradients from the tasks could be in conflict with each other. Here, while the G_{MF} is propagated M times from the matrix multiplication by $Ug_i(V_i)^\top$, G_{clf} is propagated once. To resolve this problem, we used gradient surgery that can be applied for multitask learning. The gradient surgery projects the G_{MF} that conflicts with the G_{clf} onto the normal plane of the G_{clf} ($G_i - \frac{G_i \cdot G_{\text{clf}}}{||G_{\text{clf}}||^2} G_{\text{clf}}$, $i = 1, \ldots, M$). Thus, U can be updated to be more specific to the classification.

As an additional consideration, U, which is the input of the classifier, continues to change during MMMF training. Moreover, as U is updated by applying gradient surgery, it is significantly changed by G_{clf}. Therefore, to ensure a stable input to the classifier, only G_{MF} is used to update from the beginning to a certain epoch (T_{p}), except for the G_{clf}. Until T_{p}, U is trained by only G_{MF} to have consensus information, and after T_{p}, U is trained by G_{MF} and G_{clf} to have consensus information for classification from the heterogeneous multimodal data. Algorithm 1 provides detailed steps for training the proposed model.

B. FEATURE SELECTION USING MMMF

The MMMF is trained to obtain a common basis matrix (U) with consensus information for classification from the heterogeneous multimodal data. Simultaneously, MMMF is trained to obtain the multiple coefficient matrices ($g_i(V_i)$) with complementary information and non-linearity of each modality. Here, we identify the features that are useful for classification using U and $g_i(V_i)$ after training the MMMF, which is illustrated in Figure 2.

First, modules in U contributing to classification are selected. The criterion for selecting modules contributing to classification is a significant difference (p-value < 0.05) between classes from a t-test for binary class and ANOVA for multi-class. Second, for each module of the coefficient matrix

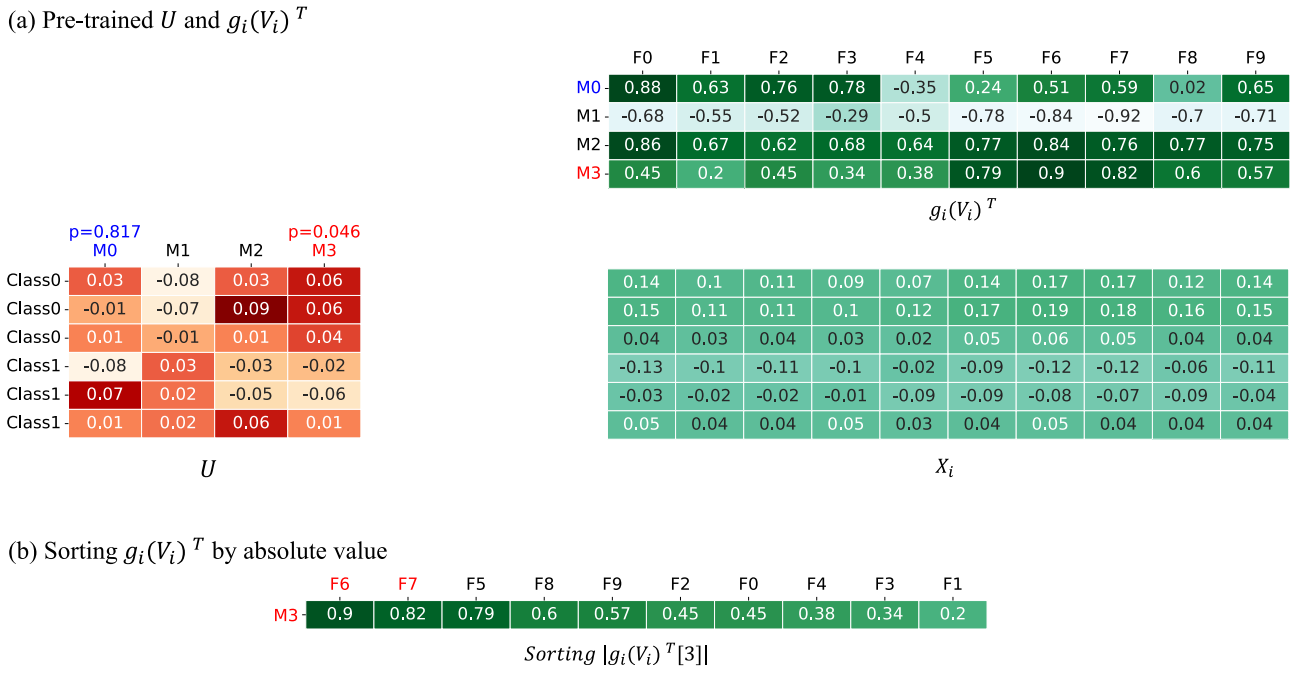

FIGURE 2. Illustration of feature selection using MMMF. (a) Select the modules in U that satisfy the criterion of a significant difference ($p < 0.05$) among the classes. The selected module (M3) has more classification information than the module that does not. Subsequently, select the columns of the coefficient matrix ($g_i(V_i)^T$) corresponding to the selected modules. The selected columns ($g_i(V_i)^T[3]$) have the weights of modality features corresponding to modules with classification information. (b) Select a specific number of features having the largest absolute values among the columns of the coefficient matrix selected in (a). Selected features (F6, F7) are expected to aid in classification.

with non-linearity ($g_i(V_i)$) corresponding to the modules selected in U, a specific number of features having the largest absolute values are selected. For example, in Figure 2(a), the first module of U (M0) does not exhibit a significant difference as $p = 0.817$; whereas, the last module (M3) shows a significant difference as $p = 0.046$. Thus, M3 is selected. Subsequently, F6 and F7 with the largest absolute values in M3 of the coefficient matrix are selected as important features (Figure 2(b)).

IV. RESULTS

A. DATA COLLECTION AND PREPROCESSING

We applied our proposed method to demonstrate its effectiveness at classification and feature selection using five different datasets: ADNI for Alzheimer's Disease (AD) *vs*. normal control (NC) classification, ROSMAP for AD *vs*. NC classification, BRCA for subtypes (luminal A, luminal B, basal-like, HER2-enriched) classification, KIRC for early-stage *vs*. late-stage classification, and COAD for high *vs*. low survival times classification.

In the case of the ADNI dataset, three modalities of structural magnetic resonance imaging (sMRI), positron emission tomography (PET), gene expression (GE), and clinical data were downloaded from the ADNI website.[1] The ADNI database contains data collected from different modalities over a long period. In this study, we conducted experiments by selecting samples based on the GE dataset, which has fewer samples than the neuroimaging dataset. For GE data, 20392 genes were used by averaging the expression values of the same gene symbol. For the neuroimaging dataset, the age and sex are confounding factors that bias the analysis [15], [16], and it is known that the model performance improves when these confounding factors are removed [3], [17]. Therefore, we removed three confounding factors: age, gender, and cohort (ADNI2 or ADNIGO) from the neuroimaging data using generalized linear regression [18].

In the case of ROSMAP dataset, three modalities of GE, DNA methylation (ME), and microRNA expression (MI) and clinical data were obtained from the ROSMAP cohort in the AMP-AD Knowledge Portal.[2] The GE dataset was used by applying log2-transformed to quantile normalized fragments per kilobase of transcript per million mapped read (FPKM) values. The ME dataset was measured β-value using the Infinium HumanMethylation450 BeadChip. The missing β-value of ME was replaced through the k-nearest neighbor algorithm, and probes of CpGs located in the promoter region (TSS200, TSS1500) were assigned to the corresponding gene. All duplicated genes were used by replacing them with the average values. The MI dataset was normalized

[1] http://adni.loni.usc.edu/

[2] https://adknowledgeportal.synapse.org/

Algorithm 1 Details Steps for Training the Proposed MMMF Model

Define:
Number of modalities M, Modalities X_i, Label Y, Classifier $f(\cdot)$ with parameters θ, Train index I_t, Validation index I_v, Cross entropy loss $\mathcal{L}_{\text{CE}}$, Initial patience T_{p}, Max iterations $T_{\max}$, Regularization rates α, β, Learning rates η_1, η_2

while $\mathcal{L}_{val}$ *does not converged or* $t \leq T_{max}$ **do**
  /*Step 1. Integrating heterogeneous multimodal data*/
  for $i \leftarrow 1$ *to* M **do**
    $g_i(V_i) \leftarrow \sigma(V_{i,1}(\ldots, \sigma(V_{i,r-1}V_{i,r})))$
    $\mathcal{L}_{\text{re}} \leftarrow \|X_i - Ug_i(V_i)^\top\|_F^2$
    $V_{i,j} = (1-\alpha)V_{i,j} - \eta_1 \nabla_{V_{i,j}} \mathcal{L}_{\text{re}}$
    $G_i \leftarrow \nabla_U \mathcal{L}_{\text{re}}$
  end
  $G_{\text{MF}} \leftarrow [G_1, G_2, \ldots, G_M]$
  /*For stable input of classifier, G_U **is updated by** G_{MF} **from the beginning to a certain epoch*/**
  if $t \leq T_p$ **then**
    $G_U \leftarrow \text{mean}(G_{\text{MF}})$
  end
  else
    /*Step 2. Solving the imbalance problem as oversampling*/
    $O(U[I_t]), O(Y[I_t]) \leftarrow$ Overampling$(U[I_t], Y[I_t])$
    $\mathcal{L}_{\text{clf}} \leftarrow \mathcal{L}_{\text{CE}}(f(O(U[I_t])), O(Y[I_t]))$
    $\theta = (1-\beta)\theta - \eta_2 \nabla_\theta \mathcal{L}_{\text{clf}}$
    $G_{\text{clf}}[I_t] = (\nabla_{O(U[I_t])} \mathcal{L}_{\text{clf}})[I_t]$
    $G_{\text{clf}}[\bar{I}_t] = 0$
    /*Step 3. Performing gradient surgery to solve the multitask problem*/
    for $i \leftarrow 1$ *to* M **do**
      if $G_i \cdot G_{clf} < 0$ **then**
        $G_i = G_i - \frac{G_i \cdot G_{\text{clf}}}{||G_{\text{clf}}||^2} G_{\text{clf}}$
      end
    end
    $G_U = \text{mean}(G_{\text{clf}}, G_{\text{MF}})$
  end
  /*Updating U **to have consensus information for classification from the heterogeneous multimodal data*/**
  $U = (1-\beta)U - \eta_2 G_U$
  $\mathcal{L}_{\text{val}} \leftarrow \mathcal{L}_{\text{CE}}(f(U[I_v]), Y[I_v])$
  t $\leftarrow$ t $+1$
end

through a variant stabilization normalization method, and the batch effect was corrected using Combat [19].

The original BRCA dataset for three modalities of GE, protein abundance (PROT), and copy number variants (CNV) is publicly available on the Broad GDAC Firehose[3] and the National Cancer Institute GDC Data Portal,[4] where details about data generation can be found. We obtained the BRCA data using the RTCGA R library [20]. For the KIRC dataset, three modalities of GE (Illumina mRNAseq), ME (Illumina HumanMethylation450 BeadArray), and MI (IlluminaHiSeq miRNAseq) were obtained from The Cancer Genome Atlas (TCGA). The ME dataset used in this paper was preprocessed according to Ma et al. [21]. For the COAD dataset, three modalities of GE, ME, and MI were downloaded from http://compbio.cs.toronto.edu/SNF/SNF/Software.html [22], which were originally collected from the TCGA site. For transcripts with the same symbol in the GE data, its expression values were averaged.

We performed a stratified five-fold cross-validation (CV) strategy to evaluate classification and feature selection. The datasets were split into a training dataset (60%), validation dataset (20%), and test dataset (20%) in each CV. Biological data often suffer ''curse of dimensionality'' because the number of features is much greater than the number of samples. To alleviate this problem, we selected the top 1000, 2000, and 3000 features with the highest variances based on the training dataset of each CV. The details of datasets and the number of features used for training are listed in Table 2.

B. EXISTING METHODS FOR PERFORMANCE COMPARISON

We compared the classification performance of the MMMF with the following seven existing methods. We selected (1) support vector machine (SVM) as a machine learning method and (2) deep neural network (DNN) as a deep learning method. Furthermore, as models specialized in integrating multiple biological datasets, the (3) DIABLO, (4) Concat AE, (5) Cross AE, (6) multitask attention learning algorithm for multi-omics data (MOMA) [23], and (7) supervised deep generalized canonical correlation analysis (SDGCCA) [24] were chosen. In addition, all the compared models were trained by applying SMOTE oversampling for fair comparison.

MMMF applied oversampling to solve the imbalance problem, and gradient surgery to the multitask problem, implying that a common basis matrix is trained by MF and classifier. In order to find out whether each solution affects the classification, we did ablation studies on: (8) MMMF without oversampling and gradient surgery (MMMF-Over-GS), (9) MMMF without oversampling (MMMF-Over), and (10) MMMF without gradient surgery (MMMF-GS).

C. EXPERIMENTAL SETTING

Owing imbalance in the data, the balanced accuracy (BA) [25], F1-score (F1), Matthews correlation coefficient (MCC) [26], and area under the receiver operating characteristic curve (AUC) were used for evaluation, which have been also used in previous biomedical studies [27], [28], [29].

[3] https://gdac.broadinstitute.org/
[4] https://portal.gdc.cancer.gov/

TABLE 2. Summary of datasets.

Dataset	Categories	Modality	Number of features in each modality	Number of features after preprocessing in each modality
ADNI	NC: 195, AD: 69	sMRI, PET, GE	311, 111, 20392	311, 111, m_p
ROSMAP	NC: 169, AD: 207	GE, ME, MI	18164, 19353, 309	m_p, m_p, 309
BRCA	Basal-like: 62, HER2-enriched: 37, Luminal A: 126, Luminal B: 80	GE, PROT, CNV	17814, 142, 18050	m_p, 142, m_p
KIRC	Early-stage: 184, Late-stage: 129	GE, ME, MI	16406, 16459, 342	m_p, m_p, 342
COAD	High: 33, Low: 59	GE, ME, MI	17814, 23088, 312	m_p, m_p, 312

m_p: Top 1000, 2000, 3000 features with the highest variances from the training dataset of each cross-validation.
NC: normal control; AD: Alzheimer's Disease; sMRI: structural magnetic resonance imaging; PET: positron emission tomography; GE: gene expression; ME: DNA methylation; MI: microRNA expression; PROT: protein abundance; CNV: copy number variants.

The multi-class BRCA dataset was evaluated as a weighted average for the F1 and One-vs-One AUC. All model hyper-parameters were selected based on the validation AUC of each CV. The MMMF was trained using the following hyper-parameters.

First, the following hyper-parameters were selected through the validation AUC of each CV: Define the dimension of i-th modality multiple coefficient matrices as $D_{i,1}, D_{i,2}$, and define the dimension of the hidden layer of the classifier as h. $D_{i,1}, D_{i,2}$, and h were selected to satisfy $h < D_{i,2} < D_{i,1}$ in [110, 90, 70, 50, 30]. As the minimum value of $D_{i,0}$ denotes 111, which is the dimension of PET from the ADNI dataset, all the conditions of Eq. (3) are satisfied. The initial patience for training U using only the G_{MF} (T_{p}) is 30 or 50.

Second, the fixed hyper-parameters are as follows: Learning rate (η_1) was 1e−3 and regularization rate (α) was 1e−4, both of which were used for training multiple coefficient matrices ($V_{i,r}$). Learning rate (η_2) was 1e−3 and regularization rate (β) was 1e−3, both of which were used for training the classifier ($f(\cdot)$) and a common basis matrix (U). Both U and $V_{i,r}$ were initialized based on singular value decomposition (SVD) [30]. However, when using the COAD dataset, and when $D_{i,1}$ was 110, the number of samples (92) was less than $D_{i,1}$. Thus, initialization through SVD was impossible. In this case, He initialization [31] was used.

The MMMF was implemented in the 1.7.0 version of PyTorch and experimented on a single NVIDIA Geforce RTX 3090 GPU. The hyper-parameters of other models are described in Supplementary Section S1 and Supplementary Table S1.

D. PERFORMANCE OF CLASSIFICATION

The classification performance in terms of BA is summarized in Table 3 and F1, MCC, and AUC values are summarized in Supplementary Table S2. In addition, the visualization of receiver operating characteristic (ROC) curve and precision-recall (PR) curve are presented in Supplementary Figure S1. The proposed MMMF method exhibited the best performance in 10 out of 12 cases in the ADNI dataset and 5 out of 12 cases in the BRCA dataset; thereby, proving the superiority of MMMF compared to the previously developed models. However, MMMF is the second best to MOMA on the ROSMAP and KIRC datasets, and also second best to Cross AE on the COAD dataset.

To estimate the statistical significance of the performance of our model compared to other models, we performed the Wilcoxon signed rank test by obtaining the average of the five-fold CV classification results from all datasets (5 datasets × 3 feature sets after preprocessing); the results are summarized in Table 4. We found that MMMF statistically significantly outperformed other competing models in 22 out of 28 cases (p-value < 0.05).

To estimate the effect of oversampling and gradient surgery in classification, we performed a Wilcoxon signed rank test using the averages of five-fold CV classification results of all metrics;, and the results are summarized in Table 5. The performance was observed to improve statistically: oversampling (p-value = 2.9E-12), and gradient surgery (p-value = 3.3E-6). However, when comparing MMMF-Over and MMMF-Over-GS, it can be seen that only 24 out of 60 performances were improved (p-value = 0.99), which adversely affected its performance. Gradient surgery forcibly emphasizes the gradient of a common basis matrix in the direction of the gradient of the classifier, and if the performance is good, it helps in classification; but in the opposite case, it hinders the classification performance. The proposed MMMF model solves the imbalance problem by applying oversampling and exhibited a better performance on applying gradient surgery to solve the multitask problem.

E. FEATURE SELECTION RESULTS

In Section III-A, we mentioned that a large quantity of classification information can be delivered to a common basis matrix (U) by applying gradient surgery. Hence, the method of feature selection using the U and multiple coefficient matrix with complementary information ($g_i(V_i)$) was proposed in Section III-B. Thus, we investigated whether the selected features are helpful in classification.

Feature selection using MMMF trained with training data in each CV was performed in the following three steps: First, only the validation indices were selected from a common basis matrix ($U[I_{\mathrm{v}}]$) to exclude test information. Second, we selected significant modules using ANOVA for the

TABLE 3. Balanced accuracy of classification with the stratified five-fold cross-validation.

m_p[a]	Model	ADNI	ROSMAP	BRCA	KIRC	COAD
1000	SVM	0.752±0.043	0.637±0.067	0.708±0.058	0.677±0.056	0.650±0.141
	DNN	0.810±0.025	0.665±0.036	0.735±0.042	0.680±0.028	0.617±0.081
	DIABLO	0.765±0.108	0.672±0.022	0.545±0.049	0.649±0.032	0.693±0.106
	Concat AE	0.808±0.059	0.687±0.056	0.749±0.036	0.651±0.040	0.643±0.110
	Cross AE	0.793±0.021	0.655±0.039	0.725±0.040	0.676±0.051	0.540±0.066
	MOMA	0.830±0.015	0.686±0.046	**0.754±0.023**	**0.711±0.021**	0.639±0.102
	SDGCCA	0.817±0.043	0.695±0.045	0.752±0.055	0.703±0.035	0.647±0.108
	MMMF-Over-GS	0.729±0.060	0.700±0.032	0.697±0.053	0.690±0.056	0.658±0.038
	MMMF-Over	0.729±0.060	0.694±0.041	0.667±0.058	0.690±0.054	0.706±0.059
	MMMF-GS	0.855±0.033	0.709±0.012	0.737±0.045	0.692±0.061	0.704±0.064
	MMMF	**0.870±0.023**	**0.711±0.011**	0.752±0.046	0.700±0.043	**0.720±0.064**
2000	SVM	0.775±0.058	0.688±0.051	0.691±0.061	0.671±0.052	0.650±0.123
	DNN	0.784±0.052	0.703±0.029	0.734±0.039	0.690±0.035	0.687±0.087
	DIABLO	0.756±0.051	0.669±0.013	0.578±0.079	0.627±0.038	0.645±0.147
	Concat AE	0.794±0.043	0.654±0.046	0.742±0.037	0.691±0.061	0.547±0.113
	Cross AE	0.839±0.036	0.700±0.040	0.763±0.074	0.676±0.048	**0.712±0.055**
	MOMA	0.852±0.029	0.708±0.056	**0.760±0.054**	0.705±0.033	0.677±0.093
	SDGCCA	0.805±0.064	**0.714±0.033**	0.730±0.034	0.650±0.017	0.634±0.144
	MMMF-Over-GS	0.712±0.034	0.692±0.024	0.697±0.036	0.660±0.018	0.601±0.082
	MMMF-Over	0.712±0.034	0.682±0.031	0.688±0.012	0.668±0.005	0.615±0.078
	MMMF-GS	0.865±0.046	0.687±0.022	0.744±0.022	0.693±0.024	0.647±0.096
	MMMF	**0.870±0.038**	0.685±0.031	0.759±0.021	**0.708±0.033**	0.682±0.134
3000	SVM	0.739±0.070	0.689±0.088	0.703±0.051	0.669±0.065	0.598±0.058
	DNN	0.762±0.045	0.708±0.044	0.770±0.053	0.667±0.055	0.700±0.072
	DIABLO	0.729±0.074	0.664±0.015	0.580±0.058	0.647±0.059	0.672±0.100
	Concat AE	0.756±0.077	0.640±0.026	0.755±0.050	0.690±0.032	0.678±0.091
	Cross AE	0.823±0.044	0.695±0.024	0.757±0.074	0.668±0.028	**0.704±0.115**
	MOMA	0.840±0.045	**0.721±0.058**	0.766±0.064	0.709±0.039	0.658±0.112
	SDGCCA	0.824±0.040	0.718±0.035	0.721±0.061	0.700±0.031	0.580±0.103
	MMMF-Over-GS	0.772±0.090	0.686±0.044	0.715±0.024	0.708±0.050	0.603±0.115
	MMMF-Over	0.708±0.077	0.677±0.040	0.720±0.027	0.708±0.050	0.607±0.082
	MMMF-GS	0.862±0.029	0.715±0.025	0.781±0.015	0.700±0.022	0.657±0.145
	MMMF	**0.865±0.034**	0.716±0.030	**0.786±0.026**	**0.717±0.040**	0.672±0.159

[a] The number of features after preprocessing.
The best performances are marked in bold.

multi-class BRCA data and t-test for all binary class datasets (p-value < 0.05). If there is no case which satisfies the criterion of p-value < 0.05 (i.e., third CV with COAD dataset and the number of features after preprocessing is 1000), one module with the lowest p-value was selected. Finally, the top 5, 10, 20, and 30 features with the largest absolute values among coefficients corresponding to the significant modules (top n) were selected. The number of selected features for each CV is shown in Supplementary Table S3 that is calculated as the number of significant modules multiplied by the top n after excluding duplicated features.

SVM was used as a classifier to evaluate the performance of the selected features. For comparison, the same number of randomly selected features and all features were used for classification. For the randomly selected features, MMMF was performed 100 times in each CV; subsequently, the mean and standard deviation were measured. After the performance measurement, the p-value was measured through the t-test to check whether the selected features using MMMF and random features showed a significant difference (p-value < 0.05).

TABLE 4. Pairwise comparison of classification performance between MMMF against other models based on the Wilcoxon signed rank test (5 datasets × 3 feature sets after preprocessing).

Model	BA	F1	AUC	MCC
SVM	**1.2E-4**	**4.3E-3**	0.095	**4.3E-4**
DNN	**0.01**	**0.03**	**0.03**	**0.01**
DIABLO	**1.2E-4**	**1.2E-4**	**6.1E-5**	**1.2E-4**
Concat AE	**1.8E-4**	**2.6E-3**	**1.2E-3**	**3.1E-4**
Cross AE	**0.02**	0.15	**3.4E-3**	**0.03**
MOMA	0.06	0.39	0.89	0.07
SDGCCA	**4.3E-3**	**2.0E-3**	**4.3E-4**	**3.4E-4**

The p-values < 0.05 are marked in bold.

Figure 3 shows the performance of models with selected features and the visualization of ROC curves and RP curves are depicted in Supplementary Figure S2. For ADNI (Figure 3 (a)), MMMF exhibited a greater performance compared with using all the features and significantly outperformed randomly selected features in 37 out of 48 cases. It was also observed that the performances of randomly

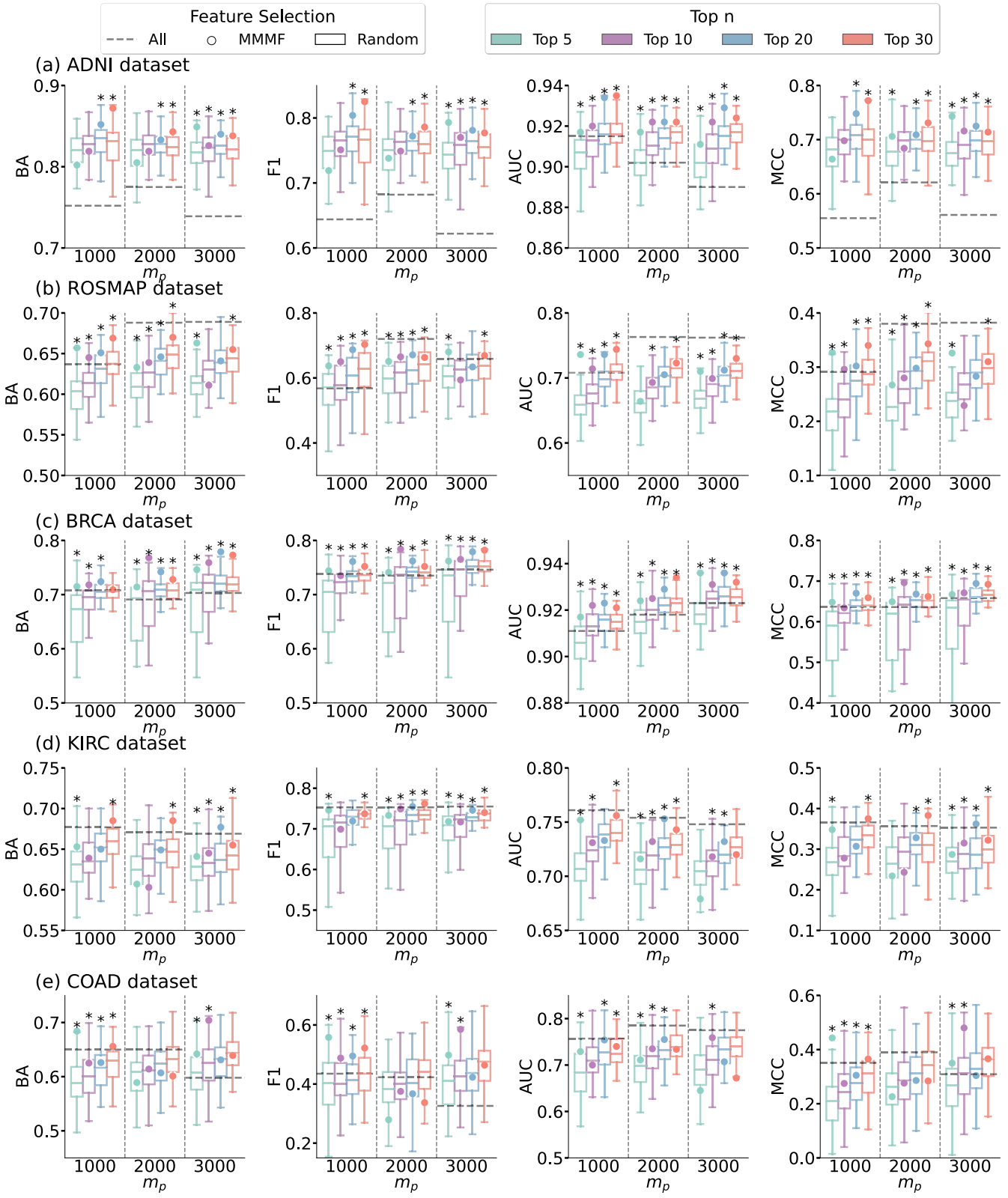

FIGURE 3. Classification performances with selected features. Results of the (a) ADNI dataset, (b) ROSMAP dataset, (c) BRCA dataset, (d) KIRC dataset, (e) COAD dataset. m_p denotes the number of features after preprocessing. The top n is the number of features with the largest absolute value among the coefficients corresponding to the significant modules selected using MMMF. "MMMF", "Random", and "All" indicate the results of using features selected by MMMF, the same number of randomly selected features as MMMF, and all features (m_p), respectively. * indicates that the performance difference between "MMMF" and "Random" measured by the t-test was significant with a p-value < 0.05.

TABLE 5. Pairwise comparison of classification performance for the estimation of effects of oversampling and gradient surgery based on the Wilcoxon signed rank test (5 datasets × 3 feature sets after preprocessing × 4 metrics).

	Effect	p-value	# of improvements
(MMMF + MMMF-GS) $vs.$ (MMMF-Over + MMMF-Over-GS)	Oversampling	**2.9E-12**	89/120
(MMMF + MMMF-Over) $vs.$ (MMMF-GS + MMMF-Over-GS)	Gradient surgery	**3.3E-6**	77/120
(MMMF-Over) $vs.$ (MMMF-Over-GS)	Gradient surgery	0.99	24/60
(MMMF) $vs.$ (MMMF-GS)	Gradient surgery	**4.0E-10**	53/60

The p-values < 0.05 are marked in bold.

TABLE 6. Pairwise comparison of classification performance between MMMF against other models based on the Wilcoxon signed rank test (5 datasets × 3 feature sets after preprocessing).

Top n[a]	BA	F1	AUC	MCC
5	**4.3E-3**	**0.01**	**0.04**	**1.5E-3**
10	0.07	**0.01**	**1.5E-3**	0.11
20	**3.4E-3**	**8.4E-3**	**6.7E-3**	**0.02**
30	**5.4E-3**	**8.4E-3**	**0.03**	**4.3E-3**

[a] The number of features with the largest absolute value among the coefficients corresponding to the significant modules.
The p-values < 0.05 are marked in bold.

selected features were better than those using all the features. This can be attributed to the interference of the features with each other for classification. For the ROSMAP and KIRC datasets (Figure 3(b), Figure 3(d)), the classification performances using all the features were the highest for most cases. However, MMMF significantly outperformed randomly selected features for 40 out of 48 cases for ROSMAP and 34 out of 48 cases for KIRC. For BRCA (Figure 3(c)), MMMF outperformed randomly selected features and using all features for 47 and 46 out of 48 cases, respectively. For COAD (Figure 3(e)), comparative performances of MMMF, randomly selected features, and all features differ depending on cases and none of the approaches significantly outperform other approaches. This could be attributed to the insufficiency in the number of samples in COAD compared to other datasets.

To estimate the statistical significance of the performance of MMMF compared to the case of using randomly selected features, we performed the Wilcoxon signed rank test using the average of five-fold CV feature selection results from all datasets;, the results are summarized in Table 6. The features selected using MMMF statistically outperformed randomly selected features in 14 out of 16 cases. Among the 12 average of five-fold CV test AUCs (3 feature sets after preprocessing$\times$ 4 top n) for each dataset, the best performance feature sets were described in Supplementary Tables S4-8.

F. OVERSAMPLING AND GRADIENT SURGERY EFFECTS IN FEATURE SELECTION

To examine the effect of oversampling and gradient surgery on feature selection, we compared performances of selected features when MMMF-Over-GS, MMMF-Over, MMMF-GS, and MMMF were used. For each approach, the top 50 features corresponding to the module with the lowest p-value were selected. The SVM was evaluated as a classifier. Randomly selected features were also compared, where the same numbers of features as MMMF were randomly selected 100 times and the average and standard deviations were calculated in each CV. Performances of other methods were measured by obtaining the average of five-fold CVs.

Supplementary Figure S3 shows the performance comparison of selected features using these methods. To estimate the effect of oversampling and gradient surgery in feature selection, we performed a Wilcoxon signed rank test using the averages of five-fold CV feature selection results of all metrics, and the results are summarized in Table 7. When oversampling was applied, 78 out of 120 performances were significantly improved statistically (p-value = 2.9E-4). However, when gradient surgery was applied, only 62 out of 120 cases exhibited an improved performance; therefore, the improvement was deduced to be statistically insignificant (p-value =0.7). When gradient surgery was used without oversampling, only 18 out of 60 performances were improved, exhibiting a statically significantly reduced performance (p-value=3.0E-5). Nonetheless, similar to the results of Section IV-D, the gradient surgery improved the performance in 44 out of 60 cases and showed a significant p-value of 3.6E-3 on the application of oversampling.

The results of Section IV-E and those of this section not only show that MMMF is not a model wherein the performance is largely determined by a classifier, but also one in which the classification information is well transmitted to a common basis matrix (U). In addition, it can be seen that oversampling has a great influence on the process of transmitting classification information, which can be emphasized by applying gradient surgery in situations where classification performance is improved through oversampling.

G. BIOMARKERS IDENTIFIED BY THE MMMF

We further examined whether selected features by MMMF are previously known disease related genes for ADNI and BRCA datasets, in which MMMF had the best performance compared to other 11 models in the classification experiment. For each dataset, one module with the best AUC among the 15 experiments (3 feature sets after preprocessing x 5 CVs) in Section IV-F was used. From the module, five genes

TABLE 7. Pairwise comparison of feature selection performance for the estimation of effects of oversampling and gradient surgery based on the Wilcoxon signed rank test (5 datasets × 3 feature sets after preprocessing × 4 metrics).

	Effect	p-value	# of improvements
(MMMF + MMMF-GS) $vs.$ (MMMF-Over + MMMF-Over-GS)	Oversampling	**2.9E-4**	78/120
(MMMF + MMMF-Over) $vs.$ (MMMF-GS + MMMF-Over-GS)	Gradient surgery	0.7	62/120
(MMMF-Over) $vs.$ (MMMF-Over-GS)	Gradient surgery	**3.0E-5**	18/60
(MMMF) $vs.$ (MMMF-GS)	Gradient surgery	**3.6E-3**	44/60

The p-values < 0.05 are marked in bold.

TABLE 8. Average number of parameters, GPU memory usage, and computational time required to train models.

	DNN	Concat AE	Cross AE	MOMA	SDGCCA	MMMF-Over-GS	MMMF-Over	MMMF-GS	MMMF
# of Parameters (K)	166.71	356.01	341.77	543.81	186.97	191.34	191.34	191.34	191.34
GPU Memory usage (KB)	4851.62	13460.67	15886.13	34388.99	17715.18	27588.57	27621.36	27621.4	27621.4
Computational time (s)	12.677	18.835	45.027	19.336	10.251	14.906	16.472	19.106	20.628

with the largest absolute values were selected. The following genes were selected through this process: DSC1, CARD16, METTL7B, RGMB, and CHEK1 for ADNI and NLRP3, PRAM1, ZNF804A, SFRS9, and GNG10 for BRCA.

In AD pathogenesis, DSC1 is known to be a potential involvement of the adrenergic signaling pathways [32]. The overexpression of METTL7B is known to reduce Aβ generation [33], [34]. CSF levels of RGMB were decreased in AD and Aβ negative mild cognitive impairment (MCI) and plasma levels of RGMB were decreased in AD and Aβ positive MCI. [35]. CHEK1 is related to cognitive function in the APP/PS1 mouse model and mediates tau/APP hyperphosphorylation in primary neurons [36]. Although the association of CARD16 to AD is not known, it was differentially expressed in both Parkinson's disease and bipolar disorder [37].

In breast cancer, NLRP3 is a well known inflammasome involving in inflammatory and immunity [38], [39]. SFRS9 is an mRNA splicing factor protein and phosphorylation of SFRS9 was found in ErbB2-overexpressing breast and ovarian cancer cell lines. Moreover, the depletion of SFRS9 reduced the migration rate of ErbB2-overexpressing ovarian cancer cells [40]. GNG10, PRAM1, and ZNF804A were not well studied in breast cancer, but they were known to be related to other cancer types. In a recent study, GNG10 is known to be involved in the prognosis of lung adenocarcinoma (LUAD) and efficacy of chemotherapy in LUAD [41], and its overexpression is related to the progression of colorectal cancer [42]. PRAM1 was suggested as one of the 10-gene signature for tumor immune microenvironment related to prognosis of non-small cell lung cancer [43]. Finally, ZNF804A was found to be correlated with the survival of pancreatic cancer patients [44].

H. COMPARISON OF MEMORY AND COMPUTATIONAL TIME

We compared MMMF and other DNN-based models to check the GPU memory and computational time during the classification experiment. Moreover, MMMF-Over-GS, MMMF-Over, and MMMF-GS were compared to check the additional increase in GPU memory and computational time when oversampling and gradient surgery were used. Supplementary Table S9 summarized the number of parameters of the model, GPU memory, and computational time for each model for all the datasets. Table 8 shows the average of all datasets. The other multi-omics models, Concat AE and Cross AE, employ the decoder to ensure that the number of parameters is twice that of DNN, and MOMA using attention has the largest number of parameters. However, for MMMF, the number of parameters is comparable to that of DNN, although it requires larger GPU memory. In addition, it can be seen that when oversampling and gradient surgery are applied, there is no increase in GPU memory; however, the computational time increases.

V. CONCLUSION

In this study, we defined two challenges when modeling using multiple biological datasets: data heterogeneity and imbalanced classes of samples. To address these challenges, we proposed the MMMF method that performs oversampling and gradient surgery for multimodal biological datasets. MMMF significantly improved classification performance compared to other biomedical classification models when it was applied to five biomedical tasks. Furthermore, it exhibited a significantly enhanced feature selection performance compared to using random or all features. Thus, MMMF is a prominent model that can be used to extract important features for classification tasks in the field of biomedicine.

ACKNOWLEDGMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_ to_ apply/ADNI_Acknowledgement_ List.pdf (accessed on 29 March 2022). ADNI is funded by the National Institute on Aging, the National Institute of

Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd., and its affiliated company Genentech Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Company Inc.;Meso Scale Diagnostics LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The authors thank Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, for making their data available. Data collection was supported through funding by NIA (Grants P30AG10161, RO1AG15819, RO1AG17917, RO1AG30146, RO1AG36042, and RO1AG36836); the Illinois Department of Public Health; and the Translational Genomics Research Institute.

REFERENCES

[1] M. Chierici, N. Bussola, A. Marcolini, M. Francescatto, A. Zandonà, L. Trastulla, C. Agostinelli, G. Jurman, and C. Furlanello, "Integrative network fusion: A multi-omics approach in molecular profiling," *Frontiers Oncol.*, vol. 10, p. 1065, Jun. 2020.

[2] T. Zhou, M. Liu, K.-H. Thung, and D. Shen, "Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2411–2422, Oct. 2019.

[3] T. Zhou, K.-H. Thung, X. Zhu, and D. Shen, "Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis," *Hum. Brain Mapping*, vol. 40, no. 3, pp. 1001–1016, 2019.

[4] Z. Huang, X. Zhan, S. Xiang, T. S. Johnson, and B. Helm, "SALMON: Survival analysis learning with multi-omics neural networks on breast cancer," *Frontiers Genet.*, vol. 10, p. 166, Mar. 2019.

[5] T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, and K. Huang, "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nature Commun.*, vol. 12, no. 1, pp. 1–13, Dec. 2021.

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

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

[8] L. Tong, J. Mitchel, K. Chatlin, and M. D. Wang, "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–12, Dec. 2020.

[9] Q. Wang, M. Sun, L. Zhan, P. Thompson, S. Ji, and J. Zhou, "Multi-modality disease modeling via collective deep matrix factorization," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1155–1164.

[10] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[12] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," 2020, *arXiv:2001.06782*.

[13] C. R. Jack, M. A. Bernstein, and N. C. Fox, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Reson. Imag., Off. J. Int. Soc. Magn. Reson. Med.*, vol. 27, no. 4, pp. 685–691, 2008.

[14] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2008, pp. 650–658.

[15] J. Dukart, K. Mueller, A. Villringer, F. Kherif, B. Draganski, R. Frackowiak, and M. L. Schroeter, "Relationship between imaging biomarkers, age, progression and symptom severity in Alzheimer's disease," *NeuroImage: Clin.*, vol. 3, pp. 84–94, 2013.

[16] S. L. Tyas, J. Manfreda, L. A. Strain, and P. R. Montgomery, "Risk factors for Alzheimer's disease: A population-based, longitudinal study in Manitoba, Canada," *Int. J. Epidemiol.*, vol. 30, no. 3, pp. 590–597, Jun. 2001.

[17] F. Falahati, D. Ferreira, H. Soininen, P. Mecocci, and B. Vellas, "The effect of age correction on multivariate classification in Alzheimer's disease, with a focus on the characteristics of incorrectly and correctly classified subjects," *Brain Topography*, vol. 29, no. 2, pp. 296–307, 2016.

[18] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *J. Roy. Stat. Soc., A (Gen.)*, vol. 135, no. 3, pp. 370–384, 1972.

[19] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007.

[20] M. Kosinski and P. Biecek. (2019). *RTCGA: The Cancer Genome Atlas Data Integration*. R Package Version 1.16.0. [Online]. Available: https://rtcga.github.io/RTCGA/

[21] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, and F. Song, "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data," *Comput. Biol. Med.*, vol. 121, Jun. 2020, Art. no. 103761.

[22] 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, pp. 333–337, 2014.

[23] S. Moon and H. Lee, "MOMA: A multi-task attention learning algorithm for multi-omics data interpretation and classification," *Bioinformatics*, vol. 38, no. 8, pp. 2287–2296, Apr. 2022.

[24] J. Hwang, S. Moon, and H. Lee, "SDGCCA: Supervised deep generalized canonical correlation analysis for multi-omics integration," *J. Comput. Biol.*, vol. 29, no. 8, pp. 892–907, 2022.

[25] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3121–3124.

[26] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[27] N. Q. K. Le and Q.-T. Ho, "Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes," *Methods*, vol. 204, pp. 199–206, Aug. 2022.

[28] B. P. Nguyen, "Prediction of FMN binding sites in electron transport chains based on 2-D CNN and PSSM profiles," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2189–2197, Nov./Dec. 2021.

[29] R. S. Olson, W. L. Cava, Z. Mustahsan, A. Varik, and J. H. Moore, "Data-driven advice for applying machine learning to bioinformatics problems," in *Proc. Pacific Symp. Biocomput., Pacific Symp.* Singapore: World Scientific, 2018, pp. 192–203.

[30] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognit.*, vol. 41, no. 4, pp. 1350–1362, Apr. 2008.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[32] M. C. Ramos, R. Tenorio, A. Martínez-García, I. Sastre, E. Vilella-Cuadrada, A. Frank, M. Rosich-Estragó, F. Valdivieso, and M. J. Bullido, "Association of DSC1, a gene modulated by adrenergic stimulation, with Alzheimer's disease," *Neurosci. Lett.*, vol. 408, no. 3, pp. 203–208, Nov. 2006.
[33] D. Franjic, J. Choi, M. Skarica, C. Xu, Q. Li, and S. Ma, "Molecular diversity among adult human hippocampal and entorhinal cells," *BioRxiv*, doi: 10.1101/2019.12.31.889139.
[34] A. Kobro-Flatmoen, M. J. Lagartos-Donate, Y. Aman, P. Edison, M. P. Witter, and E. F. Fang, "Re-emphasizing early Alzheimer's disease pathology starting in select entorhinal neurons, with a special focus on mitophagy," *Ageing Res. Rev.*, vol. 67, May 2021, Art. no. 101307.
[35] C. D. Whelan, N. Mattsson, M. W. Nagle, S. Vijayaraghavan, C. Hyde, and S. Janelidze, "Multiplex proteomics identifies novel CSF and plasma biomarkers of early Alzheimer's disease," *Acta Neuropathologica Commun.*, vol. 7, no. 1, pp. 1–14, 2019.
[36] W. Hu, Z. Wang, H. Zhang, Y. A. R. Mahaman, and F. Huang, "Chk1 inhibition ameliorates Alzheimer's disease pathogenesis and cognitive dysfunction through CIP2A/PP2A signaling," *Neurotherapeutics*, vol. 19, pp. 570–591, Mar. 2022.
[37] M. B. Hossain, M. K. Islam, A. Adhikary, A. Rahaman, and M. Z. Islam, "Bioinformatics approach to identify significant biomarkers, drug targets shared between Parkinson's disease and bipolar disorder: A pilot study," *Bioinf. Biol. Insights*, vol. 16, Feb. 2022, Art. no. 11779322221079232.
[38] M. Sonnessa, A. Cioffi, O. Brunetti, N. Silvestris, F. A. Zito, C. Saponaro, and A. Mangia, "NLRP3 inflammasome from bench to bedside: New perspectives for triple negative breast cancer," *Frontiers Oncol.*, vol. 10, p. 1587, Sep. 2020.
[39] S. S. Faria, S. Costantini, V. C. C. de Lima, V. P. de Andrade, M. Rialland, R. Cedric, A. Budillon, and K. G. Magalhães, "NLRP3 inflammasome-mediated cytokine production and pyroptosis cell death in breast cancer," *J. Biomed. Sci.*, vol. 28, no. 1, pp. 1–15, Dec. 2021.
[40] M. Mukherji, L. M. Brill, S. B. Ficarro, G. M. Hampton, and P. G. Schultz, "A phosphoproteomic analysis of the ErbB2 receptor tyrosine kinase signaling pathways," *Biochemistry*, vol. 45, no. 51, pp. 15529–15540, Dec. 2006.
[41] Y. Song, Z. Qu, H. Feng, L. Xu, Y. Xiao, Z. Zhao, D. Wu, C. Sun, X. Fan, and D. Zhou, "Genomic and immunological characterization of pyroptosis in lung adenocarcinoma," *J. Oncol.*, vol. 2022, pp. 1–20, Jul. 2022.
[42] N. Wang, J. Li, J. He, Y.-G. Jing, W.-D. Zhao, W.-J. Yu, and J. Wang, "Knockdown of lncRNA CCAT1 inhibits the progression of colorectal cancer via hsa-miR-4679 mediating the downregulation of GNG10," *J. Immunol. Res.*, vol. 2021, pp. 1–14, Dec. 2021.
[43] J. Li, X. Li, C. Zhang, C. Zhang, and H. Wang, "A signature of tumor immune microenvironment genes associated with the prognosis of non-small cell lung cancer," *Oncol. Rep.*, vol. 43, no. 3, pp. 795–806, 2020.
[44] H. Sun, R. Xin, C. Zheng, and G. Huang, "Aberrantly DNA methylated-differentially expressed genes in pancreatic cancer through an integrated bioinformatics approach," *Frontiers Genet.*, vol. 12, Mar. 2021, Art. no. 583568.

JEONGYOUNG HWANG received the B.S. degree in information communications engineering from the Hankuk University of Foreign Studies, South Korea, in 2019. He is currently pursuing the M.S. degree with the AI Graduated School, Gwangju Institute of Science and Technology (GIST). His research interests include bioinfomatics, multimodal classification, and feature selection.

HYUNJU LEE received the B.S. degree in computer science from the Korea Institute of Science and Technology, South Korea, in 1997, the M.S. degree in computer engineering from Seoul National University, South Korea, in 1999, and the Ph.D. degree in computer science from the University of Southern California, USA, in 2006. From 2006 to 2007, she was a Postdoctoral Research Fellow with Brigham and Women's Hospital, Harvard Medical School. Since 2007, she has been with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea. Her research interests include machine learning, natural language processing, and bioinformatics.