


RESEARCH ARTICLE

Open Access



Stratifying patients using fast multiple kernel learning framework: case studies of Alzheimer's disease and cancers

Thanh-Trung Giang^{1,2}, Thanh-Phuong Nguyen^{3,4*}  and Dang-Hung Tran⁵

Abstract

Background: Predictive patient stratification is greatly emerging, because it allows us to prospectively identify which patients will benefit from what interventions before their condition worsens. In the biomedical research, a number of stratification methods have been successfully applied and have assisted treatment process. Because of heterogeneity and complexity of medical data, it is very challenging to integrate them and make use of them in practical clinic. There are two major challenges of data integration. Firstly, since the biomedical data has a high number of dimensions, combining multiple data leads to the hard problem of vast dimensional space handling. The computation is enormously complex and time-consuming. Secondly, the disparity of different data types causes another critical problem in machine learning for biomedical data. It has a great need to develop an efficient machine learning framework to handle the challenges.

Methods: In this paper, we propose a fast-multiple kernel learning framework, referred to as fMKL-DR, that optimise equations to calculate matrix chain multiplication and reduce dimensions in data space. We applied our framework to two case studies, Alzheimer's disease (AD) patient stratification and cancer patient stratification. We performed several comparative evaluations on various biomedical datasets.

Results: In the case study of AD patients, we enhanced significantly the multiple-ROIs approach based on MRI image data. The method could successfully classify not only AD patients and non-AD patients but also different phases of AD patients with AUC close to 1. In the case study of cancer patients, the framework was applied to six types of cancers, i.e., glioblastoma multiforme cancer, ovarian cancer, lung cancer, breast cancer, kidney cancer, and liver cancer. We efficiently integrated gene expression, miRNA expression, and DNA methylation. The results showed that the classification model basing on integrated datasets was much more accurate than classification model basing on the single data type.

Conclusions: The results demonstrated that the fMKL-DR remarkably improves computational cost and accuracy for both AD patient and cancer patient stratification. We optimised the data integration, dimension reduction, and kernel fusion. Our framework has great potential for mining large-scale cohort data and aiding personalised prevention.

Keywords: Patient stratification, Alzheimer's diseases, Cancers, Multiple kernel learning, High dimensional data space, Dimension reduction

* Correspondence: nguyentp.dr@gmail.com; phuong.nguyen@ext.uni.lu

³Life Sciences Research Unit, Belval, University of Luxembourg, Luxembourg City, Luxembourg

⁴Megeno S.A., Belval, Esch-sur-Alzette, Luxembourg

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Patient stratification has widespread biomedical and clinical applications, including diagnosis, prognosis, and treatment response prediction. A clinically useful prediction algorithm should be accurate, generalizable, be able to integrate diverse data types, and handle sparse data [1–5]. To achieve effective personalised medicine, patient stratification models are essentially required for all of diseases. Amongst the most emerging diseases, cancer and Alzheimer's disease (AD) have been attracted a lot of research due to the severity, the complication and the high prevalence.

Alzheimer's disease is a neurological disorder in which the death of brain cells causes memory loss and cognitive decline. Aging is the primary cause of AD, however, there are several other reasons related to lifestyle, such as physical inactivity, obesity, unhealthy diets, alcohol abuse, etc. [6]. Among AD's phases, Mild Cognitive Impairment (MCI) are a critical phase because patients with MCI have higher risks for late stage of AD or other dementias. Stratifying AD patients in the early stage is crucial, so that we identify cases whose MCI signs may potentially be converted to the last severe stage of AD [7].

Magnetic Resonance Imaging (MRI) data has been popularly used in AD patient stratification due to MRI's high-quality three-dimensional images of brain. Based on MRI data, regions of interest (ROI) which affect disease development could be revealed, contributing significantly to AD diagnosis and treatment. There are two main approaches based on ROIs, the single-ROI based approach [8] and the multiple-ROI based approach [9–11]. Chupin et al. [8] used probabilistic and anatomical priors for hippocampus segmentation to determine AD, NC, MCI. Ahmed et al. [11] proposed an automatic classification framework for AD, normal controls (NC), MCI, considering visual features from the most involved regions in AD. Several multivariate approaches, such as partial least squares and principal component analysis, were developed to build a discrimination model [12]. Liu et al. [13] constructed an individual network based on ROIs and used it as input of a classification model. Other previous work on multiple ROIs showed that they increased the performance of AD diagnosis [14]. In [15, 16], Liu et al. demonstrated not only ROIs, but also the correlations between ROIs were closely related to AD diagnosis results. Even though the previous methods have achieved remarkable results, they have not completely solved the problem of high dimensional data, in terms of accuracy and computational cost.

Cancer is not only threatening but also very diverse. Cancer patient stratification has been one of most challenging topics in biomedical informatics. Previous work have either focused on a specific data type of interest, such as gene expression, DNA methylation [17–20] or

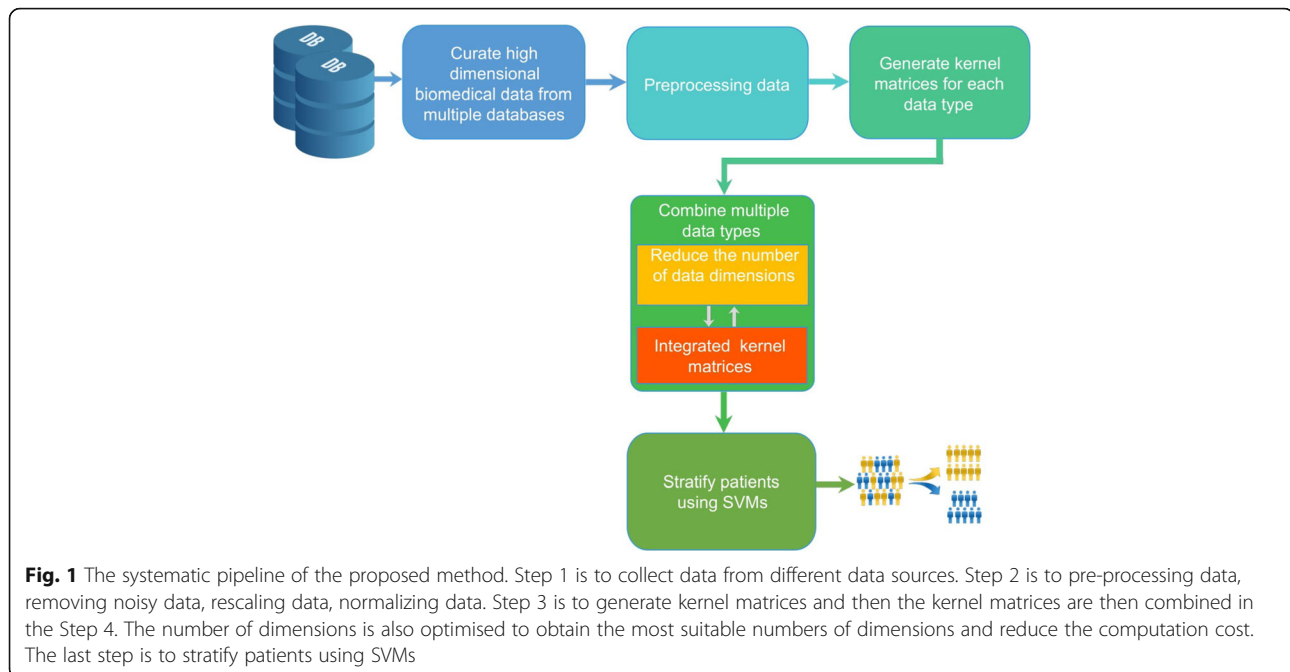
combined multiple data [21, 22]. Nowadays patient data has immensely been available with diverse information, such as gene expression, DNA methylation, miRNA expression, protein expression, exon expression, etc. [23]. It has been shown that there are multiple factors contributing to cancer pathology. Developing novel methods to combine a wide range of data is emerging and challenging [24–28].

The computation for learning high dimensional space is extremely complex and time-consuming. Since data types are of great difference (for example, categorical data, numerical data, imaging data), it is therefore essential to unify data measurement before integrating them. Lin et al. in [28] have proposed an effective method to combine data from different sources. They used multiple kernel learning to solve the second challenge and reduced data dimension basing on graph embedding. The method was applied to the image processing problem for ten data types in the Caltech-101 dataset [29]. In a recent study, Spacher et al. employed this method to cluster cancer patients and attained good results [30]. However, their work limited on the distribution of patients into different clusters and subtypes of cancers.

In this paper, we have proposed a novel computational framework based on fast multiple kernel learning and dimension reduction (fMKL-DR in short), addressing challenges in AD and cancer patient stratification. In the case study of AD patients, we enhanced significantly the performance of the multiple-ROIs approach when comparing with previous methods. Accuracy and Area Under a Curve (AUC) demonstrated that our proposed method was more accurate and robust than previous work. In the case study of cancer patients, the framework was applied to six types of cancers, i.e., glioblastoma multiforme cancer, ovarian cancer, lung cancer, breast cancer, kidney cancer, and liver cancer. We efficiently integrated gene expression, miRNA expression, and DNA methylation. The results showed that the classification model basing on integrated dataset was much more accurate than classification model basing on the single data type. The results obtained by both AD and cancer applications manifest that our developed model potentially stratifies patients and later aid disease prevention and prognosis.

Methods

The proposed framework is demonstrated in Fig. 1. There are five main steps. In the first step, we obtained biological data related to cancer and AD from various information sources. These data in different types (e.g. imaging data, numerical data, textual data) were pre-processed in the second step. Data transformation is required to obtain a complete matrix, in which each row is a sample data and each column represents a feature.



We employed a number of data pre-processing methods to eliminate redundant and noisy data. Furthermore, machine learning algorithms expect the scale of the training data to be equivalent, so we also used normalization to scale feature values to the range between -1 and 1 . We further employed a number of data pre-processing methods to eliminate redundant and noisy data. Specifically, we removed a feature if its data were missed in any subject (about 3% of the total number of features). Secondly, we generated kernel matrices from the above data matrix using different kernel functions. Each kernel matrix is a square symmetric and positive definite matrix, that represents the similarity of data samples based on a specific kernel function. In the next step, the kernel matrices are integrated into a final matrix using a multi-kernel learning framework. Since the kernel matrices have a large number of dimensions, we proposed the fast multi kernel learning framework combined with dimensional reduction algorithm, so-called fMKL-DR. Finally, we modelled predictive binary classifications with SVMs to stratify cancer and AD patients. Details of each step are described in the following sections.

Data curation and pre-processing

In this section, we present the methods for collecting and pre-processing MRI images and genomics data for Alzheimer's disease; proteomics and genomics data for cancer diseases.

Image data from AD patients

MRI images of Alzheimer's disease patients were extracted from Alzheimer's Disease Neuroimaging Initiative (ADNI).¹ The dataset consists of 710 T1-weighted subjects (data samples), including 200 subjects diagnosed with AD, 280 subjects with MCI, and 230 normal control (NC) subjects. Among 280 subjects with MCI, there are 120 subjects, that have MCI and convert to AD within 18 months (MCIC) and 160 subjects who have MCI and do not convert to AD (MCInc). To analyse the MRI images, we applied the six measures for cortical and sub-cortical regions proposed by Liu et al. [13]. More specifically, those six measures are Cortical Gray Matter Volume (CGMV), Cortical Thickness (CT), Cortical Surface Area (CSA), Cortical Curvature (CC), Cortical Folding Index (CFI) and Sub-cortical Volume (SV). We performed a complete procedure of image pre-processing, including spatial normalization, intensity normalization, skull stripping, segmentation and fill for obtaining the higher quality images. As the result, all images were registered with AAL atlas [31] before anatomic re-construction by FreeSurfer software.²

After calculating the six measures for cortical and sub-cortical regions, we represented the obtained data in terms of graphs. We generated six graphs G_k corresponding to six measures, namely as G_{CGMV} , G_{CT} , G_{CSA} , G_{CC} , G_{CFI} , G_{SV} . In a graph G_k , the set V_k denotes set of vertices (v_i), representing the regions. We denote E_k as

¹<http://adni.loni.usc.edu>

²<https://surfer.nmr.mgh.harvard.edu>

the set of edges (e_{ij}), consisting the weighted connections between two regions. The weight w_{ij} of the edge e_{ij} between two regions i and j is calculated as following:

$$w_{ij} = \frac{1}{d_{ij} + 1}$$

where d_{ij} is the distance between two regions and $d_{ij} = |m_i - m_j|$ (given m_i and m_j are measure values of region i and region j , respectively). For example, G_{CGMV} consists of 78 vertices and 3003 edges. In case of the sub-cortical region, G_{SV} was constructed by the set V_i of 12 vertices and the set E_i of 66 edges. Both V_i and set E_i are matrices $R[m,n]$, where m is the number of samples/subjects in our model, and n is the number of vertices or the number of edges, respectively. The illustration of the AD’s MRI images pre-processing is shown in Fig. 2.

Genomic data from AD patients

In addition to MRI images, gene expression data of AD patients is very interesting for stratifying AD patients. From the ADNI database,³ we downloaded a raw gene expression dataset of 442 subjects, which includes 43 AD subjects, 139 MCI subjects, and 260 NC subjects. We extracted 22,609 genes for each subject and represented the gene expression dataset as a matrix, in which columns are subjects and rows are genes.

Proteomic and genomic data from cancer patients

We extracted six cancer patient datasets from the TCGA database (The Cancer Genome Atlas, 2019)⁴ including Glioblastoma Multiforme (GBM), Ovarian Serous Cystadenocarcinoma (OV), Squamous Cell Lung Carcinoma (LUNG), Breast Invasive Carcinoma (BREAST), Kidney Renal Clear Cell Carcinoma (KIDNEY), and Liver Cancer (LIVER). In order to acquire multiple biomedical aspects related cancer, three data features were investigated in our model of cancer patient stratification, specifically gene expression, miRNA expression, DNA methylation. The statistics of cancer datasets are presented in Table 1. The raw data were pre-processed by removing the missing data and represented each data type as a matrix $R[m,n]$, in which columns are subjects, and rows are genes.

Data rescaling

The values in each dataset have different magnitudes, units, and ranges. This variety causes an issue that the higher magnitude dataset will have greater weight than the lower ones. Therefore, we rescaled all of data values into the same range, enabling the equity between

datasets. We used min-max scaling to scale the data into the range $[-1, 1]$.

Multiple kernel learning combined with dimensionality reduction (MKL-DR)

Multiple Kernel Learning (MKL) is a machine learning method, modelling a kernel ensemble from many kernel functions or kernel matrices. Recent research on MKL have shown that learning SVMs with multiple kernels not only increases the accuracy but also enhances the expandability of the classification [25]. The MKL framework aims to the optimal for linear combination from input kernels. MKL’s illustration is shown in Fig. 3.

In Lin et al. [28], multiple kernel learning was improved by combining it with dimensionality reduction algorithm, so-called the MKL-DR. This framework was developed basing on graph embedding [32]. Yan et al. constructed an ensemble model, which enabled the incorporation of several dimensional reduction methods. The method presented data in the form of graph and provided a unified framework for a broad set of DR algorithms. Moreover, the paper developed a new dimensional reduction method. Based on the input graph, the rejection vector was found to project the vertices of graph in new low-dimensional space so that it best characterised the similar relationship between pairs of training samples basing on the graph preserving criterion. MKL-DR integrated better data from different sources and reduced the data dimensions, enhancing accuracy and computational cost. In this paper, we embedded Linear Discriminant Analysis [33] into the MKL-DR framework.

Fast MKL-DR using dynamic programming

There are three parameters that affect the performance of the MKL-DR, i.e., the number of samples (N), the number of data types (M), and the dimensions after being reduced (P). In case that the value of M is small, often between 3~10, the number of dimensions after being reduced is small (in our experiment, we chose $P = 5$). Therefore, the computation complexity is $O(N^3)$, which is polynomial time. The MKL-DR training algorithm calculates iterative equations:

$$S_W^A = \sum_{i,j=1}^N w_{ij} (\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^T A A^T (\mathbb{K}^{(i)} - \mathbb{K}^{(j)}) \tag{1}$$

$$S_{W'}^A = \sum_{i,j=1}^N w'_{ij} (\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^T A A^T (\mathbb{K}^{(i)} - \mathbb{K}^{(j)}) \tag{2}$$

$$S_W^\beta = \sum_{i,j=1}^N w_{ij} (\mathbb{K}^{(i)} - \mathbb{K}^{(j)}) \beta \beta^T (\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^T \tag{3}$$

³<http://adni.loni.usc.edu>

⁴<https://www.cancer.gov/tcga>

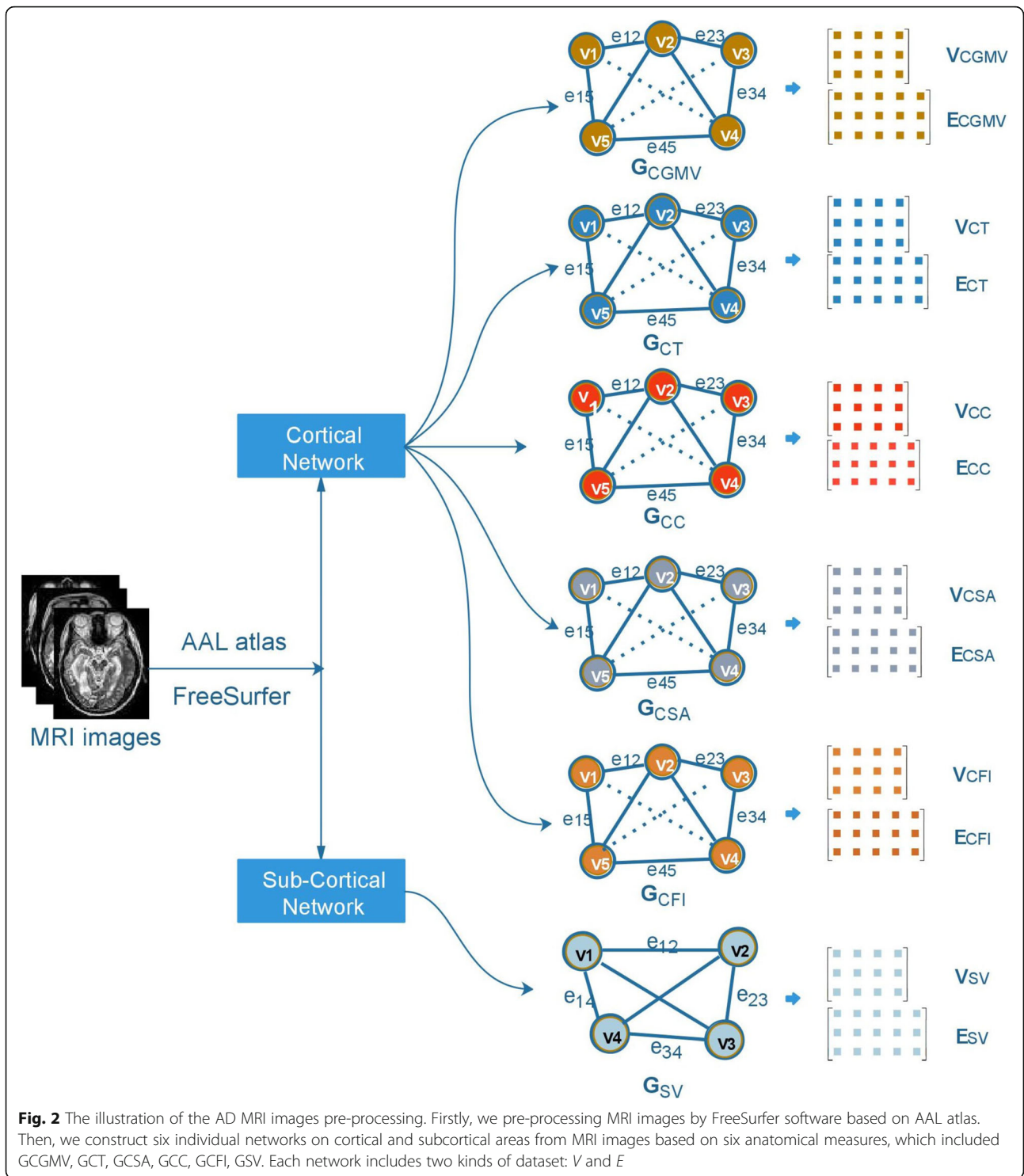


Fig. 2 The illustration of the AD MRI images pre-processing. Firstly, we pre-processing MRI images by FreeSurfer software based on AAL atlas. Then, we construct six individual networks on cortical and subcortical areas from MRI images based on six anatomical measures, which included GCGMV, GCT, GCSA, GCC, GCFI, GSV. Each network includes two kinds of dataset: V and E

$$S_{W'}^\beta = \sum_{i,j=1}^N w'_{ij} (\mathbb{K}^{(i)} - \mathbb{K}^{(j)}) \beta \beta^T (\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^T \quad (4)$$

Given $S_{W'}^A, S_{W'}^A, S_{W'}^\beta, S_{W'}^\beta$ are the matrices, which is used in the optimization problem [28]; w_{ip}, w'_{ij} are elements of

the similarity matrices aggregated from the kernel matrices; A is the sample coefficient matrix, and β is the kernel weight vector.

Each of four above Eqs. (1), (2), (3), (4) calculates the sum of the matrix chain multiplication (see more details in [28]). As the result, the MKL-DR will become

Table 1 Statistics of datasets in the cancer case study

Cancer	Number of Samples	Number of features (dimensions) per data type		
		Gene expression	DNA methylation	miRNA expression
LUNG	106	12,042	23,074	352
GBM	275	12,042	22,896	534
BREAST	435	12,042	24,978	354
OV	541	12,042	21,825	799
KIDNEY	122	17,899	24,960	329
LIVER	451	13,426	25,168	216

exhaustively time-consuming when increasing the number of samples.

Matrix chain multiplication has a combinatory property, meaning that changing calculation order between a pair of the matrices will affect the number of product operations without modifying the multiplication results. Therefore, we propose to use a dynamic-programming-based procedure to find the calculation sequence that gives minimum product operations. If the number of product operations is minimum, the computation time of the equations will be reduced.

The minimum product operation order problem based on dynamic programming:

Given N matrices A_1, A_2, \dots, A_N with size of A_i is $d_{i-1} \times d_i$.

Problem: Find the order of product matrix chain $A_1 \times A_2 \times \dots \times A_N$ to minimize the number of product operations.

Solution: Construct matrix F is a matrix $N \times N$, with element $F(i, j)$ is the total product operations to calculate matrix chain multiplication from A_i to A_j ($A_i \times A_{i+1} \times \dots \times A_j$). The formula to calculate $F(i, j)$ defined by:

$$F(i, i) = 0,$$

$$F(i, i + 1) = d_{i-1} \times d_i \times d_{i+1},$$

$$F(i, j) = \min (F(i, t) + F(t + 1, j) + d_{i-1} \times d_t \times d_j)$$

with $t = i + 1, i + 2, \dots, j - 1$. In other words, t is a mid-point to insert the parentheses to change the calculation order so that the number of product operations is minimum:

$$A_i \times A_{i+1} \times \dots \times A_j = (A_i \times A_{i+1} \times \dots \times A_t) \times (A_{t+1} \times A_{t+2} \times \dots \times A_j)$$

The dimension of the matrix of $(A_i \times A_{i+1} \times \dots \times A_t)$ is $d_{i-1} \times d_t$, and the dimension of the matrix of $(A_{t+1} \times A_{t+2} \times \dots \times A_j)$ is $d_t \times d_j$.

Based on the above assumption, we developed an improved algorithm to matrix chain multiplication to minimize product operations shown in Algorithm 1.

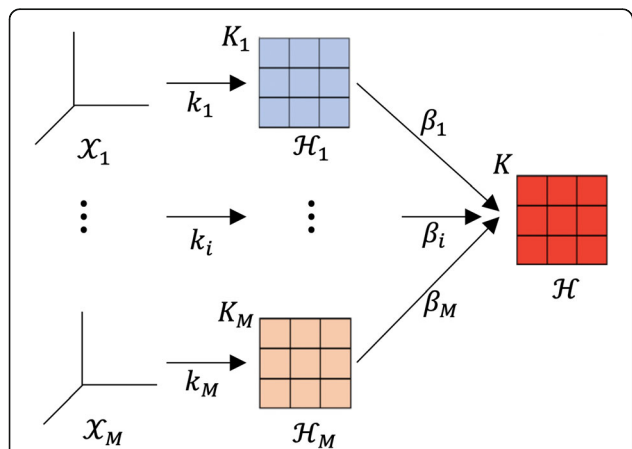


Fig. 3 The illustration of Multiple Kernel Learning given that X_i is original dataset, K_i is kernel matrix that is constructed by kernel function k_i , β_i is weighted coefficient combining K_1 to K_M to unify a final kernel matrix K , \mathcal{H}_i is the Hilbert space of i th dataset in the kernel method

Algorithm 1. The Matrix Chain Multiplication Ordering Procedure (MCMO)

```

Input       $N$  matrices sizes  $d_1, d_2, \dots, d_N$ ;
Output    The multiplication ordering  $O = [o_1, o_2, \dots, o_r]$  so the number of product
              operations is smallest
1   $F(i, i) = 0$  with  $i = 1, \dots, N$ ;
2   $F(i, i + 1) \leftarrow d_{i-1} \times d_i \times d_{i+1}$  with  $i = 1, \dots, N - 1$ ;
3   $O = []$ ;
4  for  $i \leftarrow N$ 
5    for  $j \leftarrow N$ 
6      if  $(i \neq j)$ 
7         $F(i, j) = \min (F(i, t) + F(t + 1, j) + d_{i-1} \times d_t \times d_j)$ 
8        if  $(i = 0)$ 
9          Add  $t$  to  $O$ ;
10 Return  $O$ ;
    
```

The MCMO complexity is $O(N^3)$. However, in the above equations, N (the number of matrices in chain) is small and equals to 4. Consequently, the time consumption of MCMO is trivial. Moreover, MCMO only calls 2

times in fMKL-DR, which is built as demonstrated in Algorithm 2.

The fast MKL-DR algorithm is described below.

Algorithm 2. The fMKL-DR Training Procedure

Input	The DR method specified by W and W' matrices (described in [28]) M base kernel matrices $\{K_n\}_{n=1}^M$ corresponding to the M datasets;
Output	Sample coefficient matrix $A = [a_1, a_2, \dots, a_p]$; Kernel weight vector β ;
1	Step 1: Initialization value for A or β
2	Step 2: Compute the best ordering:
3	$O_A = \text{MCMO}(\text{size}(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}), \text{size}(\beta), \text{size}(\beta^\sim), \text{size}(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}))$
4	$O_\beta = \text{MCMO}(\text{size}(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}), \text{size}(A^\sim), \text{size}(A), \text{size}(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}))$
5	Step 3: Repeat
6	Compute S_W^A by (1) and $S_{W'}^A$ by (2) based on the order O_A ;
7	Solve the optimization problem (28) [28] to get β ;
8	Compute S_W^β by (3) and $S_{W'}^\beta$ by (4) based on the order O_β ;
9	Solve the generalized eigenvalue problem (31) [28] to get A ;
10	Until converge or reach maximum number of iteration (T)
11	Step 4: Return A and β

The matrices $S_W^A, S_{W'}^A, S_W^\beta, S_{W'}^\beta$ are calculated in the lines 6th and 8th of Algorithm 2 based on the ordering of O_A and O_β . These matrices have the same values as MKL-DR ones.

Experimental design

We carried out three main experiments to investigate the performance of our method. The first experiment was done on the AD patients’ the MRI image dataset, comparing our proposed method to previous work. In the second experiment, we tested our method on the AD patients’ gene expressions. The last experiment was designed based on the six cancer patient datasets. In all of the experiments, we evaluated three measurements including accuracy, AUC, and computational time.

E1: Experiment based on MRI images dataset of Alzheimer’s disease patients

After pre-processing the Alzheimer’s disease patients MRI images datasets, we generated 12 individual networks. We designed the experiment as follows:

- Step 1. We built multiple kernel matrices by using different parameters from 12 original datasets. Specifically, from each dataset, we used the Gaussian kernel function $k(x, x') = \exp(-\|x - x'\|^2 / 2\sigma^2)$. We run the experiments with 5 different parameters $\sigma \in \{10^{-6}, 10^{-3}, 1, 10^3, 10^6\}$ to generate 5 kernel matrices. We constructed 60 kernel matrices from 12 original datasets, that were used as input data for the next step.

- Step 2. Develop four models for the four binary classification problems, AD and NC, AD and MCI, NC and MCI, and MCIc and MCIInc, denoted by $C_{AD/NC}, C_{AD/MCI}, C_{NC/MCI}$ and $C_{MCIc/MCIInc}$ correspondingly. We employed libSVM library⁵ with 60 kernel matrices generated in Step 1. The set of parameters is the default one in libSVM with $-\text{svm_type} = 0$ (C-SVC), $-\text{c}$ (cost) = 1, $-\text{wi}$ (weight) = 1.
- Step 3. For each classification problems ($C_{AD/NC}, C_{AD/MCI}, C_{NC/MCI}$ and $C_{MCIc/MCIInc}$), 20 experiments were performed. In each experiments, 2/3 dataset were randomly selected from the original dataset for training, and the rest was used for testing. The best result among the 20 experiments is reported and statistically tested.
- Step 4. Compare the results between our proposed method and the recent other methods.

E2: Experiment based on Alzheimer’s disease patient gene expression dataset

We designed the experiment based on AD patient gene expression dataset as follows:

- Step 1. Generate four different kernel matrices by four different kernel functions including Gaussians, Polynomial, Linear, Sigmoid kernel function, which were developed in the dimensionality reduction library.⁶ Default parameters were set as below.
 Gaussian Kernel: t is number of samples.
 Polynomial: $d = 2$
 Linear: $c = 0$
 Sigmoid: $\alpha = 1/D$, D is the number of dimensions of dataset, and $c = 0$.
- Step 2. Integrate and reduce dimensions from the four kernel matrices from Step 1 by fMKL-DR into a unified kernel.
- Step 3. Develop five classification models based on the libSVM library, i.e., C_{Gaussian} with Gaussian kernel matrix), $C_{\text{Polynomial}}$ with Polinomial kernel matrix), C_{Linear} with Linear kernel matrix), C_{Sigmoid} with Sigmoid kernel matrix), and $C_{\text{fMKL-DR}}$ with unified kernel matrix that is got in Step 2). The libSVM parameters were set as default values (showed in Step 1 of E1).
- Step 4. Carry out the same procedure as Step 3 in E1.
- Step 5. Compare the results between the classification model using single gene expression

⁵<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

⁶<http://www.cad.zju.edu.cn/home/dengcai/Data/data.html>

Table 2 Accuracy of the previous methods and the proposed method for different AD patient groups

Method	Classification Model Accuracy (%)			
	AD/NC	AD/MCI	NC/MCI	MCIc/MCIInc
Chupin et al., 2009 [8]	80.51	73.48	71.94	64.21
Ahmed et al., 2015 [11]	86.40	74.51	76.29	68.72
Khedher et al., 2015 [12]	88.96	84.59	82.41	70.11
Dai et al., 2013 [9]	90.81	85.92	81.92	71.04
Suk et al., 2014 [10]	93.05	88.98	83.67	72.86
Liu et al., 2018 [13]	95.24	90.85	86.35	74.28
Proposed method (the best result among 20 runs)	96.50	91.25	87.65	78.49
Proposed method (at 90% confidence level of <i>t</i> -test)	95.80	90.63	86.47	77.42

dataset with different kernel functions (Step 1) and the unified kernel (Step 2).

E3: Experiment based on cancer patient datasets

We carried out the following four steps on each dataset to evaluate the proposed method.

- Step 1. Run the fMKL-DR algorithms by initializing A or β . Both of initializations produce similar results, even though the first initializing A obtains faster convergence. Integrate multiple data sources and reduce the number of data dimensions by fMKL-DR.
- Step 2. Develop four classification models based on three single datasets and one combined dataset and SVMs, i.e., C_{GE} for the gene expression dataset, C_{DNA} for the DNA methylation dataset, C_{miRNA} for the mirRNA expression dataset), and $C_{fMKL-DR}$ for the unified kernel matrix obtained in Step 1. For the three models (C_{GE} , C_{DNA} , and C_{miRNA}), we used the default values in libSVM with Gaussian kernel function ($\gamma = 1/\text{num_features}$, num_features is the number of features of data). In the case of the model $C_{fMKL-DR}$, the kernel is the one obtained unified kernel matrix and the other parameters are set as the default values in libSVMs.
- Step 3. Perform the same procedure as Step 3 in E1.

- Step 4. Compare the results between the classification model using a single data type and the one with data integration.

These abovementioned experimental steps were run on the six patient datasets of cancer patients to evaluate the efficiency of the methods.

Statistical test

We performed statistical tests to assess the robustness of the obtained results, specifically one sample *t*-test with $n = 20$. At 95% confidence level, the hypothesis tests (on mean of accuracy and AUC) were done to evaluate whether they are statistically significant.

Results

Application of stratifying Alzheimer’s disease patients

To investigate performance of our classification model, we carried out four experiments for four classification subgroups including AD/NC, AD/MCI, NC/MCI, MCIc/MCIInc (MCI and converted to AD/ MCI and not converted to AD). We evaluated our method by comparing accuracy and area under curve (AUC) of our classification model to previous works, which used the same dataset of the MRI images from ADNI.

Table 3 Comparative AUC results of the previous methods and the proposed method for different AD patient groups

Method	AUC			
	AD/NC	AD/MCI	NC/MCI	MCIc/MCIInc
Chupin et al., 2009 [8]	0.7851	0.7328	0.7155	0.6638
Ahmed et al., 2015 [11]	0.8487	0.7562	0.7677	0.6814
Khedher et al., 2015 [12]	0.9256	0.8859	0.8134	0.7076
Dai et al., 2013 [9]	0.9429	0.8743	0.8118	0.7086
Suk et al., 2014 [10]	0.9475	0.9007	0.8203	0.7123
Liu et al., 2017 [13]	0.9754	0.9355	0.9107	0.7885
Proposed method (the best result among 20 runs)	0.9786	0.9412	0.9151	0.8024
Proposed method (at 90% confidence level of <i>t</i> -test)	0.9705	0.936	0.911	0.7945

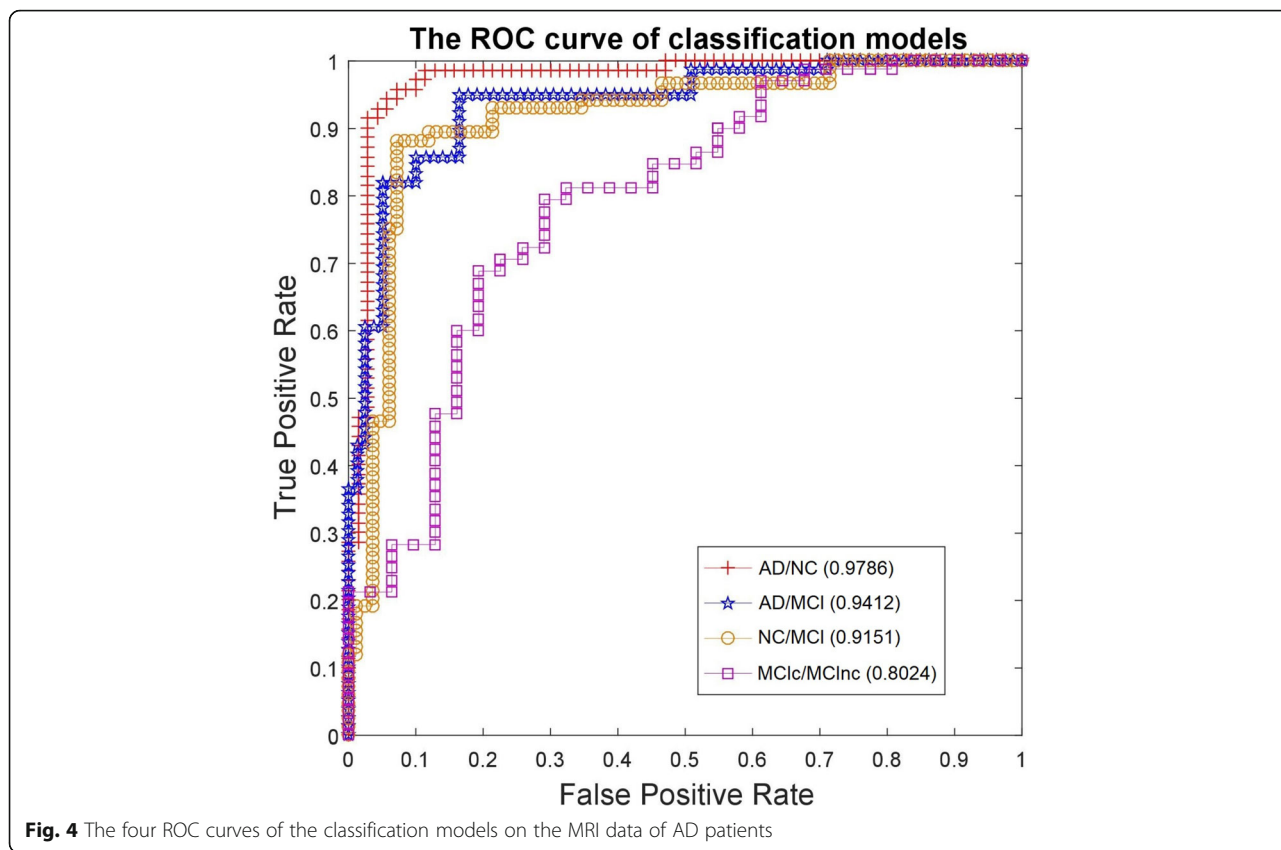


Fig. 4 The four ROC curves of the classification models on the MRI data of AD patients

Tables 2 and 3 showed accuracy and AUC of the previous work and our method applied to the four patient subgroups, respectively. In term of both accuracy and AUC, the whole brain-based method achieved significantly better results than single ROI-based method [8] or multiple ROIs-based method proposed by [11]. The results showed that the multiple ROIs-based approach was more appropriate than the other. It is well aligned with clinical practice that AD related to all of brain regions, rather than a specific one or some regions. In comparison with the whole brain-based methods proposed by Khedher et al. [12], and Suk et al. [10], and Dai et al. [9], or Liu et al. [13], we obtained better accuracy and AUC than they did. AUC of our proposed method is close to 1, demonstrating the accurateness and robustness of the method.

Figure 4 shows the ROC curves of four analysis groups. We achieved high AUCs for all of the groups. The AD/NC model had the highest value equal to 0.978, followed by AD/MCI equal to 0.941, NC/MCI equal to 0.915, and MCIc/MCIc equal to 0.802. The ROC confirmed that our method was highly efficient in stratifying AD patients for all phases.

Gene expression data of the AD patients was processed and run with different kernels functions in our framework, specifically $k1 = \text{Gaussians}$, $k2 = \text{Polynomial}$, $k3 = \text{Linear}$, $k4 = \text{Sigmoid}$, and our integrated kernel function (see more in Table 4). Accuracy of the integrated kernel function method was higher than all of single-kernel function methods on the same dataset.

Table 4 Case study of the AD patient stratification, accuracy between gene expression and proposed method integrated four kernels (different kernel functions: $k1 = \text{Gaussians}$, $k2 = \text{Polynomial}$, $k3 = \text{Linear}$, $k4 = \text{Sigmoid}$). fMKL-DR is the best accuracy among the 20 runs, and fMKL-DR^a is accuracy tested at 95% confidence level

Tasks	# Subjects	Accuracy (%)					
		Gaussian	Polynomial	Linear	Sigmoid	fMKL-DR	fMKL-DR ^a
AD/NC	303	88.12	87.13	88.19	87.13	91.09	90.1
AD/MCI	182	83.33	81.67	83.33	80.00	85.00	83.33
NC/MCI	399	70.68	69.92	69.92	69.17	75.94	75.19

Application of stratifying cancer patients

The accuracy and reliability of classification models is shown in Fig. 5 and Table 5. The results demonstrated that classification models based on a single data type were less accurate than the one basing on integrated data. Specifically, testing the method on all cancer datasets, the fMKL-DR model produced the best accuracy.

Table 5 shows accuracy of the classifiers on the different datasets. In this table, the fourth, fifth, and sixth

columns represent accuracy of the classifiers based on single datasets, such as Gene expression, DNA methylation, and miRNA expression. The last two columns show accuracy of the classifier based on the integrated dataset with fMKL-DR. The classifier on the integrated dataset has a high accuracy ranging from ~ 72% to ~ 94%. Especially, accuracy on BREAST and KIDNEY datasets were very high, values of 94.29 and 87.50%, respectively. Thus, in all cancer patient datasets, the classifiers based on

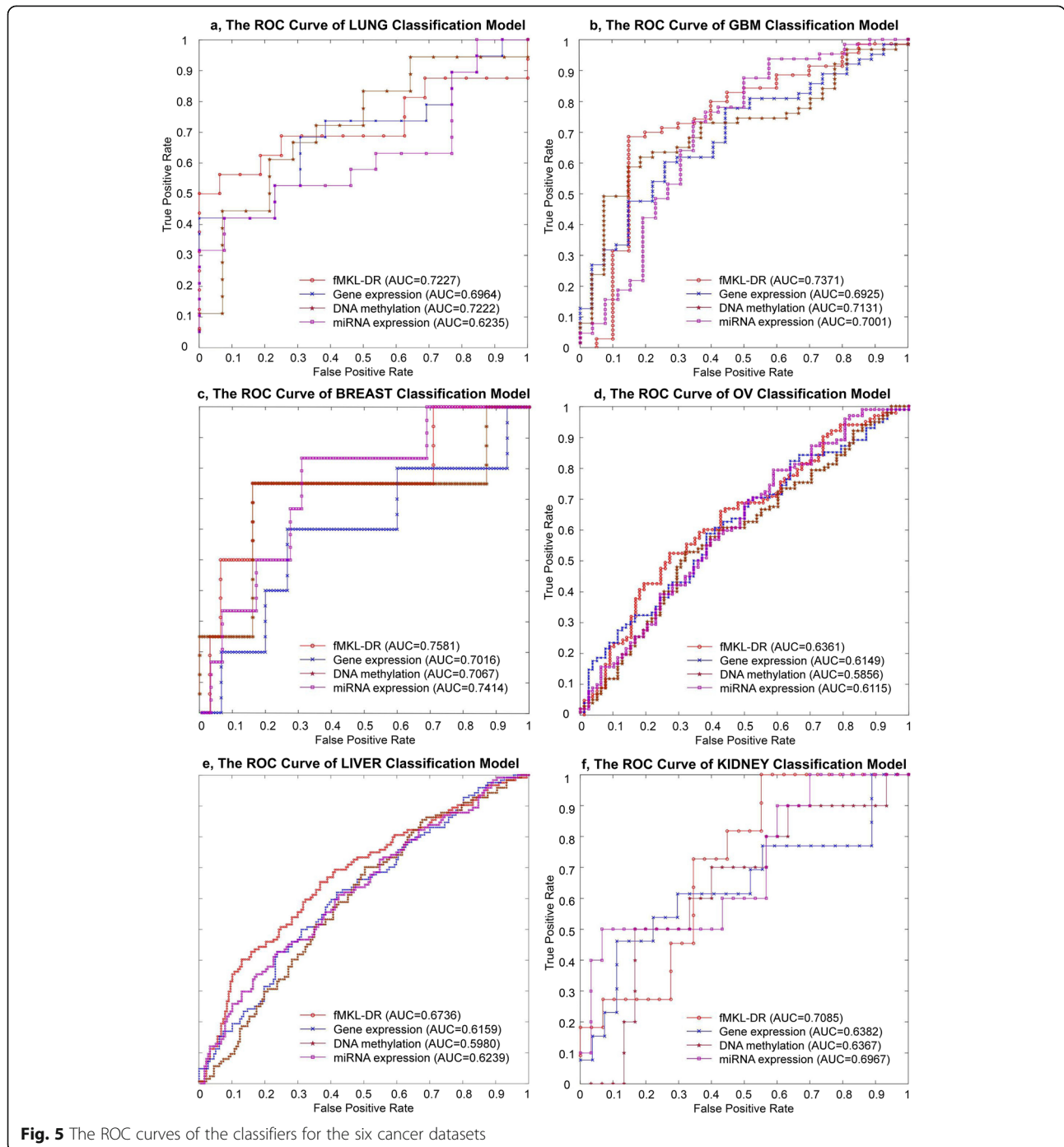


Table 5 Case study of the cancer patient stratification: accuracy obtained for each dataset and data integration. fMKL-DR is the best accuracy among the 20 runs, and fMKL-DR^a is accuracy tested at 95% confidence level

Cancer	Number of Samples	Alive/Dead	Gene expression	DNA methylation	miRNA expression	fMKL-DR	fMKL-DR ^a
LUNG	106	42/64	62.50	65.63	71.88	78.13	75.00
GBM	275	202/73	77.44	75.56	76.67	81.11	80.00
BREAST	435	360/75	88.57	88.57	91.43	94.29	93.57
OV	541	258/283	59.44	58.33	55.00	62.22	61.67
KIDNEY	122	90/32	81.25	81.25	78.13	87.50	85.00
LIVER	451	277/174	66.00	69.00	65.00	72.00	71.33

fMKL-DR have demonstrated to be effective for patient stratification since it obtained better results than the other classifiers based on each single dataset.

Figure 5 shows the ROC curve of the classifiers on the different cancer datasets, including lung cancer, GBM, breast cancer, OV cancer, liver cancer, and kidney cancer. For each data type of cancers, we drew four ROC curves corresponding to the four classifiers training on Gene Expression, DNA Methylation, miRNA Expression and the integrated dataset. The classifier implemented by our method (fMKL-DR) obtained the best AUC value when compared to classifiers on each individual dataset. In addition, these AUC values are relatively high, ranging from 0.63 to 0.75, this implies that the predicted results of our models are reliable.

Discussion

The fMKL-DR framework has been showed the robustness and accurateness in both applications of AD and cancer patient stratification. We have tested different datasets of heterogenous data types, from imaging data to numerical data. The numbers of dimensions are also varied. The classification models were built for various cancer types and all of them achieved significant results, showing the high generalization of the model. When stratifying AD patients, the method could classify even the different phases of AD patients, not only AD or non-AD patients. Therefore, it is very promising for early diagnosis and follow up of effective treatment for AD patients, especially since the late phase of AD is untreatable.

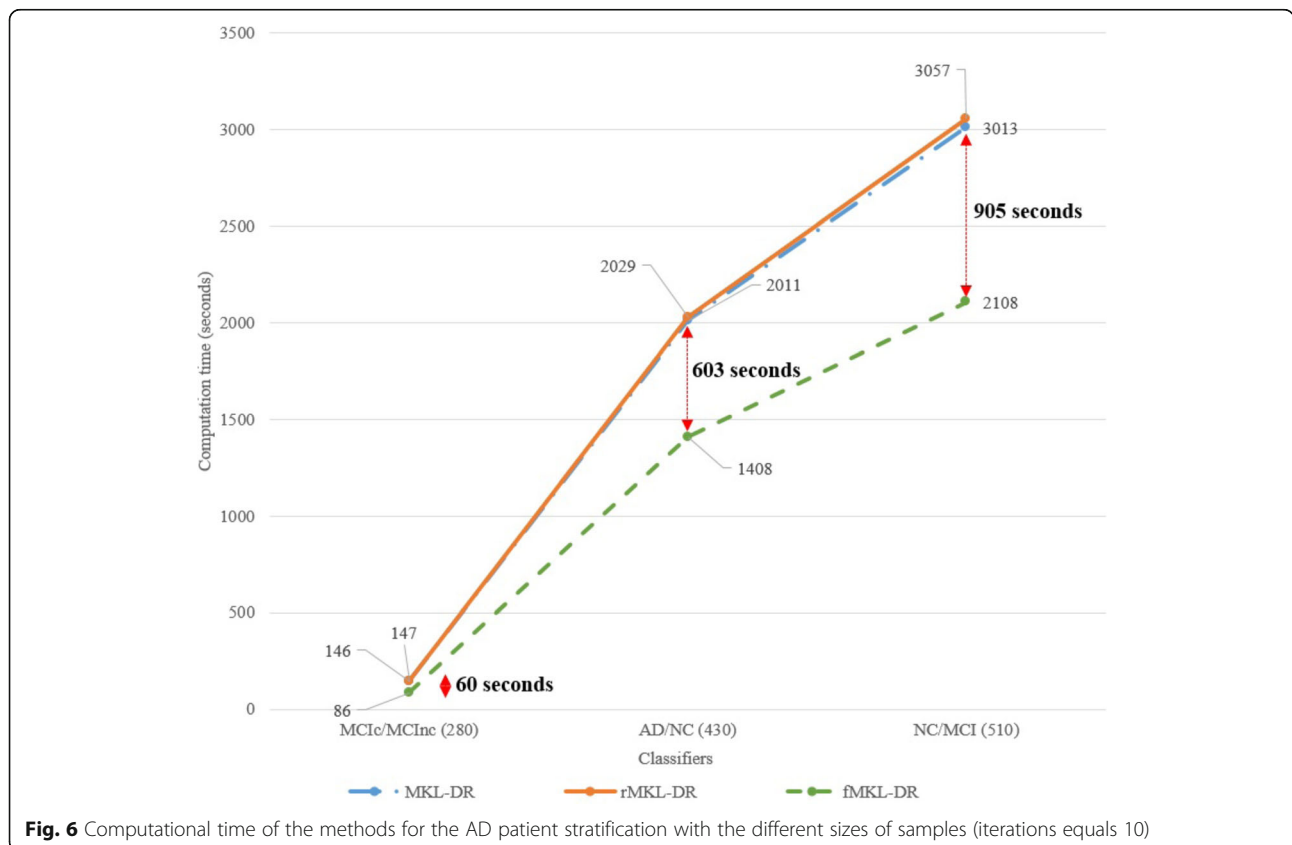


Fig. 6 Computational time of the methods for the AD patient stratification with the different sizes of samples (iterations equals 10)

Table 6 The comparison of the computational time of MKL-DR, rMKL-DR and the proposed method (in bold) on Alzheimer’s Disease MRI dataset

Tasks	#Samples	Computation time (seconds)								
		5 iterations			10 iterations			20 iterations		
		MKL-DR	rMKL-DR	fMKL-DR	MKL-DR	rMKL-DR	fMKL-DR	MKL-DR	rMKL-DR	fMKL-DR
AD/NC	430	1005	1014	702	2011	2029	1408	4022	4060	2817
AD/MCI	480	1322	1337	843	1646	1675	1687	3292	3351	3374
NC/MCI	510	1506	1528	996	3013	3057	2108	6027	6115	4217
MCIc/MCInc	280	73	73	43	146	147	86	293	294	173

In addition to high accuracy for both cancer and AD, our proposed method is much faster than the previous work. This advantage is of great significance because there are more and more data available, and data will be more and more complex. The problems of data sparsity, and heterogeneity requires cost-effective and accurate methods. Figure 6 and Table 6 show the computation time of two previous methods, MKL-DR, rMKL-DR and the proposed method fMKL-DR on Alzheimer’s Disease dataset. Figure 7 and Table 7 demonstrate the computational time when running the experiments on the cancer datasets. In all cases, our method reduced notably the computational cost. Especially, we saved more time than previous method did when the scales of datasets were increased. For example, we saved 702 s for the breast cancer dataset of 435 subjects, and 958 s for the ovarian

cancer dataset of 541 subjects. Our proposed method is very beneficial when analysing huge data cohorts.

The fMKL-DR method is advantageous in optimizing the best number of dimensions/features after reduction, denoted as f . Feature selection and feature engineering are crucial steps to acquire the best performance. We set f from 100 to 1000 by step of 50, and run all the tests with set values of f . The results of AD patient stratification with different numbers of features are showed in Fig. 8. The highest accuracy is obtained with about 400 features for AD/NC classification, about 450 AD/MCI and NC/MCI classifications, and about 500 features for MCIc/MCInc classification. It turns out two main points. Firstly, the optimal number of features is classifier-independent, meaning that there is no common set up of dimension reduction for all classifiers.

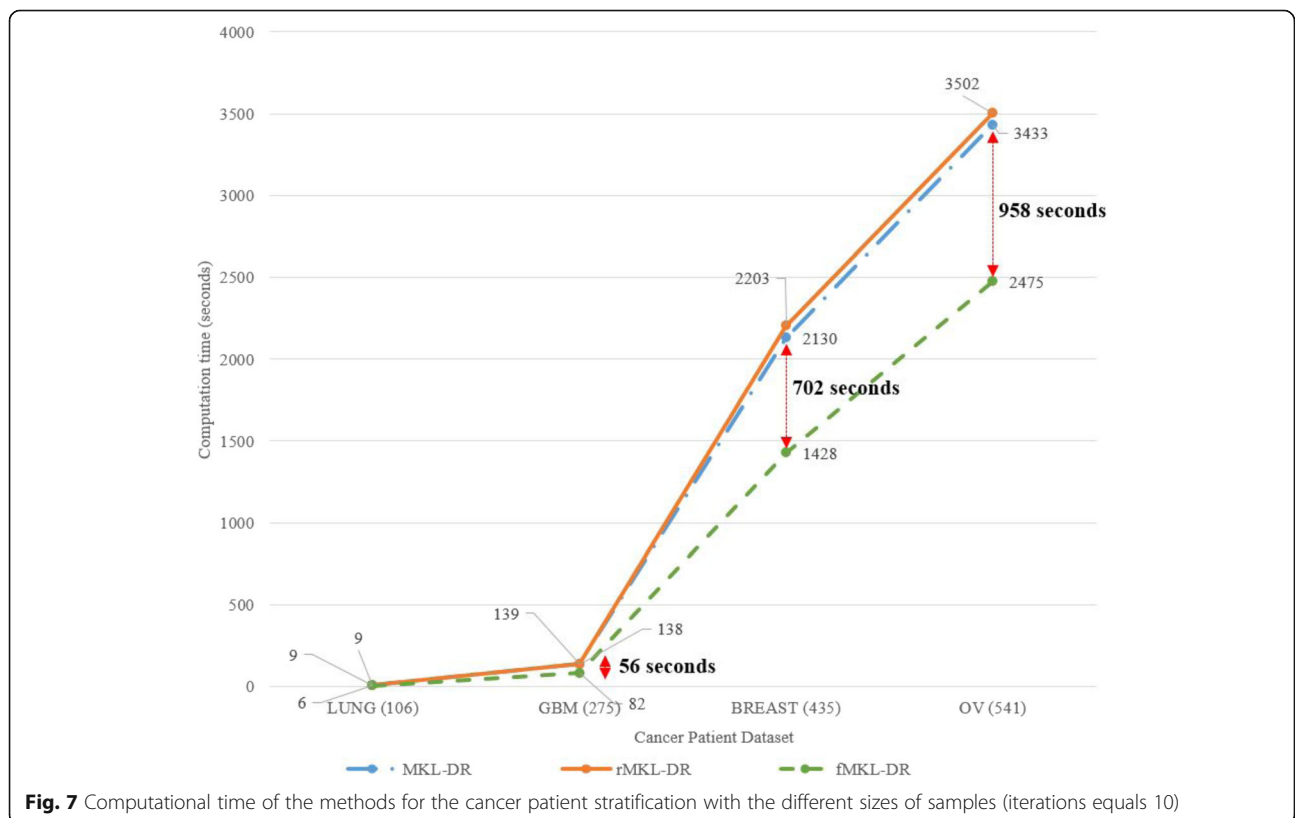


Table 7 Computation time (in second) in comparing to previous methods for the cancer datasets

Cancer	Num of Samples	Computation time (seconds)								
		5 iterations			10 iterations			20 iterations		
		MKL-DR	rMKL-DR	fMKL-DR	MKL-DR	rMKL-DR	fMKL-DR	MKL-DR	rMKL-DR	fMKL-DR
LUNG	106	4	4	3	9	9	6	18	19	13
KIDNEY	122	5	5	4	12	12	8	36	37	27
GBM	275	69	69	41	138	139	82	276	279	165
BREAST	435	1064	1074	714	2130	2149	1428	4262	4298	2857
LIVER	451	1183	1195	752	2367	2391	1508	4734	4782	3017
OV	541	1716	1750	1237	3433	3502	2475	6867	7005	4952

Secondly, increasing number of dimensions does not always improve performance of the model, even decrease accuracy. As the result, our framework optimised the numbers of dimensions to ensure the best performance of the classifier.

There are some limitations that will be improved in the future work. Firstly, data preprocessing, including missing data handling, will be better performed. In this paper, we removed a feature if its data were missed in any subject. Therefore about 3% of the total number of features (e.g., 20,000 genes of the whole genome) were removed. The missing mechanism will be deeply analysed to define the type of missing data, missing completely at random or missing at random or missing not at random. Depending on the types of missing data, a appropriate technique, such as data imputation, will be applied. Secondly, parameter optimisation will be of interest. We set the default values in our experiments to ensure the comparative evaluation with the other related work, however the better performance can be obtained by testing several sets of parameters and parameter optimisation algorithms.

The proposed framework is very useful when analysing high dimensional genomics data. Dimension reduction can be applied in the significance analysis on gene expression data. In multi-assay data exploration, dimension reduction facilitates downstream gene set, pathway and

network analysis of variables. Integrative data analysis is of great interest due to the complexity of biology and medicine. Because multi-omics data, such as DNA sequence, epigenome, transcriptome, protein, metabolites, is more and more available, there is an increasing need for developing methods with high performance. Single kernel methods are not suitable to encode heterogenous or multi-modal datasets. The proposed framework can be easily adjusted to ingrate new data types, making it flexible in other applications, such as biomarker identification. The framework is adaptable for other disease studies, in addition to cancer and AD. Precision medicine is required big data integration, especially electronic health records and whole genome sequencing data. Therefore, our method will be very beneficial to precision medicine.

Conclusions

In this paper, we have proposed the accurate and fast kernel learning framework. We employed the framework to stratify AD patients and cancer patients. We handled a wide range of data, including the MRI image data for AD, the gene expression data for both AD and cancer, miRNA expression, DNA methylation for cancer. By carrying out a number of testing strategies, the results showed that our model performed better than previous

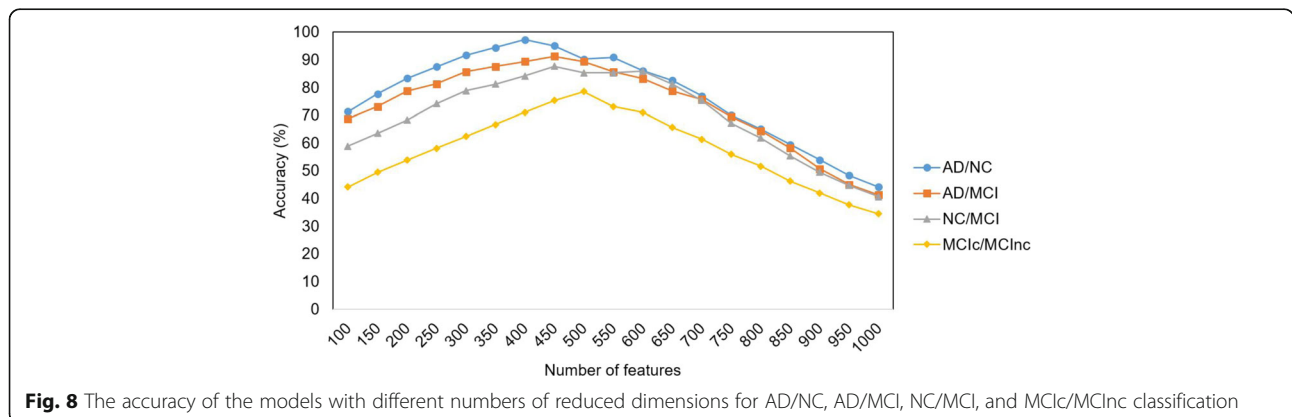


Fig. 8 The accuracy of the models with different numbers of reduced dimensions for AD/NC, AD/MCI, NC/MCI, and MCIc/MCInc classification

work, in term of accuracy, AUC, and computational time. In fact, the electronic health records will be more and more available, offering better insights of patient studies. As there are more and more data available from diverse sources, it is emerging to develop computational method to combine and mine those heterogenous data. Our proposed framework is very promising to handle high dimensional data and to aid precision medicine.

Abbreviations

fMKL-DR: Fast Multiple Kernel Learning Framework; AD: Alzheimer's disease; MCI: Mild Cognitive Impairment; ROI: Regions of Interest; MRI: Magnetic Resonance Imaging; NC: Normal Control; AUC: Area Under a Curve; MCIc: MCI and converted to AD; MCInc: MCI and not converted to AD; GBM: Glioblastoma Multiforme; CGMV: Cortical Gray Matter Volume; CT: Cortical Thickness CT; CSA: Cortical Surface Area; CC: Cortical Curvature; CFI: Cortical Folding Index; SV: Sub-cortical Volume; OV: Ovarian Serous Cystadenocarcinoma; MKL: Multiple Kernel Learning; MKL-DR: Multiple Kernel Learning and Dimensionality Reduction; MCMO: Matrix Chain Multiplication Ordering Procedure

Acknowledgements

The results published here are in whole or part based upon data generated by the ADNI (<https://adni.loni.usc.edu>), and TCGA Research Network (<https://www.cancer.gov/tcga>). This research was supported by the Vietnam Ministry of Education and Training, Project No. B2020-SPH-11.

Authors' contributions

TTG, TPN, DHT conceived and designed the study; TTG performed computational analyses; TTG, DHT collected data. TPN wrote the first draft of the manuscript. All authors contributed to writing the paper, read and approved the final manuscript.

Funding

This research was supported by the Vietnam Ministry of Education and Training, project No. B2020-SPH-11.

Availability of data and materials

The datasets were curated from public databases, ADNI (<https://adni.loni.usc.edu>), and TCGA Research Network (<https://www.cancer.gov/tcga>). The processed data along with codes are available upon request.

Ethics approval and consent to participate

Not applicable. The study does not involve human subjects, only use public data.

Consent for publication

Not applicable.

Competing interests

The authors declare that there is no competing interest in relation to the publication of this article.

Author details

¹VNU University of Engineering and Technology, Hanoi, Vietnam. ²TayBac University, Son La, Vietnam. ³Life Sciences Research Unit, Belval, University of Luxembourg, Luxembourg City, Luxembourg. ⁴Megeno S.A., Belval, Esch-sur-Alzette, Luxembourg. ⁵Hanoi National University of Education, Hanoi, Vietnam.

Received: 14 September 2019 Accepted: 28 May 2020

Published online: 16 June 2020

References

- Valdes G, Luna JM, Eaton E, Simone CB, Ungar LH, Solberg TD. MediBoost: a patient stratification tool for interpretable decision making in the era of precision medicine. *Sci Rep*. 2016;6(1):37854.

- Zhong X, Yang H, Zhao S, Shyr Y, Li B. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. *BMC Genomics*. 2015;16(7):57.
- Alderdice M, Richman SD, Gollins S, et al. Prospective patient stratification into robust cancer-cell intrinsic subtypes from colorectal cancer biopsies. *J Pathol*. 2018;245(1):19–28.
- Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol*. 2011;7(8):e1002141. <https://doi.org/10.1371/journal.pcbi.1002141>.
- Doerr M, Edelman E, Gabitzsch E, Eng C, Teng K. Formative evaluation of clinician experience with integrating family history-based clinical decision support into clinical practice. *J Pers Med*. 2014;4(2):115–36.
- Duthey B. Background paper 6.11: Alzheimer disease and other dementias. *A Public Health Approach Innov*. 2013;(February):1–74. https://www.who.int/medicines/areas/priority_medicines/prior_med_ch6_12/en/.
- Nettiksimmons J, DeCarli C, Landau S, Beckett L. Biological heterogeneity in ADNI amnesic mild cognitive impairment. *Alzheimers Dement*. 2014;10(5):511–521.e1.
- Chupin M, Gérardin E, Cuingnet R, et al. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. *Hippocampus*. 2009;19(6):579–87.
- Dai D, He H, Vogelstein JT, Hou Z. Accurate prediction of ad patients using cortical thickness networks. *Mach Vis Appl*. 2013;24(7):1445–57.
- Suk HI, Lee SW, Shen D, et al. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. *Neuroimage*. 2014;101:569–82.
- Ahmed OB, Benois-Pineau J, Allard M, Amar CB, Catheline G, et al. ADNI. Classification of alzheimers disease subjects from mri using hippocampal visual features. *Multimed Tools Appl*. 2015;74(4):1249–66.
- Khedher L, Ramírez J, Gorriiz JM, Brahim A, Segovia F, Initiative DN. Early diagnosis of alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented mri images. *Neurocomputing*. 2015;151:139–50.
- Liu J, Wang J, Tang Z, Hu B, Wu FX, Pan Y. Improving Alzheimer's disease bioinform. 2017;15(5):1649–59.
- Linn KA, Gaonkar B, Satterthwaite TD, Doshi J, Davatzikos C, Shinohara RT. Control-group feature normalization for multivariate pattern analysis of structural mri data using the support vector machine. *Neuroimage*. 2016;132:157–66.
- Liu J, Li M, Lan W, Wu FX, Pan Y, Wang J. Classification of Alzheimer's disease using whole brain hierarchical network. *IEEE/ACM Trans Comput Biol Bioinform*. 2016;15(2):624–32.
- Liu J, Li M, Pan Y, Wu FX, Chen X, Wang J. Classification of schizophrenia based on individual hierarchical brain networks constructed from structural MRI images. *NanoBioscience IEEE Trans*. 2017;16(7):600–8.
- Hernando E. microRNAs and cancer: role in tumorigenesis, patient classification and therapy. *Clin Transl Oncol*. 2007;9(3):155–60.
- Vanneschi L, Farinaccio A, Mauri G, Antoniotti M, Provero P, Giacobini M. A comparison of machine learning techniques for survival prediction in breast cancer. *BioData Min*. 2011;4(1):12.
- Cosgun E, Karaagaoglu E. The new hybrid method for classification of patients by gene expression profiling. *J Integr Des Process Sci*. 2010;14(2):27–42.
- Fleischer T, Frigessi A, Johnson KC, Edwardsen H, Touleimat N, Klajic J, Riis ML, Haakensen VD, Wämberg F, Naume B, Helland Å. Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biol*. 2014;15(8):435.
- Kwon YJ, Lee SJ, Koh JS, et al. Genome-wide analysis of DNA methylation and the gene expression change in lung cancer. *J Thorac Oncol*. 2012;7(1):20–33.
- Udali S, Guarini P, Ruzzenente A, Ferrarini A, Guglielmi A, Lotto V, Tononi P, Pattini P, Moruzzi S, Campagnaro T, Conci S. DNA methylation and gene expression profiles show novel regulatory pathways in hepatocellular carcinoma. *Clin Epigenetics*. 2015;7(1):43.
- Bach FR, Lanckriet GR, Jordan MI. Multiple kernel learning, conic duality, and the SMO algorithm. In: *Proceedings of the twenty-first international conference on Machine learning*; 2004. p. 6.
- Liang M, Li Z, Chen T, Zeng J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Trans Comput Biol Bioinform*. 2015;12(4):928–37.

25. Gönen M, Alpaydin E. Localized multiple kernel learning. In: Proceedings of the 25th international conference on Machine learning; 2008. p. 352–9.
26. Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. 2014;11(3):333–7.
27. Giang TT, Nguyen TP, Tran DH. Stratifying cancer patients based on multiple kernel learning and dimensionality reduction. In: 2017 9th international conference on Knowledge and Systems Engineering (KSE). New York: IEEE; 2017. p. 106–11. <https://doi.org/10.1109/KSE.2017.8119443>.
28. Lin YY, Liu TL, Fuh CS. Multiple kernel learning for dimensionality reduction. *IEEE Trans Pattern Anal Mach Intell*. 2011;33(6):1147–60. <https://doi.org/10.1109/TPAMI.2010.183>.
29. Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop: IEEE; 2004. p. 178–8.
30. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*. 2015;31(12):i268–75.
31. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. Automated anatomical labeling of activations in SPM using a macroscopic anatomical Parcellation of the MNI MRI single-subject brain. *NeuroImage*. January 2002;15(1):273–89.
32. Yan S, Dong X, Zhang B, Zhang H-J, Yang Q, Lin S. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans Pattern Anal Mach Intell*. 2007;29(1):40–51.
33. Fisher RA. The statistical utilization of multiple measurements. *Ann Eugenics*. 1938;8:376–86.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

