Computer Methods and Programs in Biomedicine 161 (2018) 45–53

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome

Dongdong Sun[a,1], Ao Li[a,b,1,*], Bo Tang[a], Minghui Wang[a,b]

[a] School of Information Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China
[b] Research Centers for Biomedical Engineering, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China

ARTICLE INFO

Article history:
Received 5 January 2018
Revised 31 March 2018
Accepted 11 April 2018

Keywords:
Breast cancer survival prediction
Genomic data
Pathological image
Multiple kernel learning

ABSTRACT

Background and objective: Breast cancer is a leading cause of death from cancer for females. The high mortality rate of breast cancer is largely due to the complexity among invasive breast cancer and its significantly varied clinical outcomes. Therefore, improving the accuracy of breast cancer survival prediction has important significance and becomes one of the major research areas. Nowadays many computational models have been proposed for breast cancer survival prediction, however, most of them generate the predictive models by employing only the genomic data information and few of them consider the complementary information from pathological images.

Methods: In our study, we introduce a novel method called GPMKL based on multiple kernel learning (MKL), which efficiently employs heterogeneous information containing genomic data (gene expression, copy number alteration, gene methylation, protein expression) and pathological images. With above heterogeneous features, GPMKL is proposed to execute feature fusion which is embedded in breast cancer classification.

Results: Performance analysis of the GPMKL model indicates that the pathological image information plays a critical part in accurately predicting the survival time of breast cancer patients. Furthermore, the proposed method is compared with other existing breast cancer survival prediction methods, and the results demonstrate that the proposed framework with pathological images performs remarkably better than the existing survival prediction methods.

Conclusions: All results performed in our study suggest that the usefulness and superiority of GPMKL in predicting human breast cancer survival.

1. Introduction

Cancer is a class of diseases in which involves abnormal cell growth and body change [1]. In general, cancer is termed after the body part in which it originated, thus breast cancer is a kind of disease that develops from the breast tissue [2]. Breast cancer in women is still the most common malignancy, with a leading cause of cancer-related deaths throughout worldwide [3]. There are around 3.1 million breast cancer survivors in the United States (U.S.), and the chance of a woman dying from breast cancer is around 1 in 37, or 2.7% according to American Medical News Reports. This highlights the urgent need to design computational methods for a more precise survival prediction of breast cancer and may lead to the development of personalized treatment and management. Accordingly, this would ultimately contribute to reducing overall mortality rate of breast cancer and further improving the quality of life in breast cancer patients.

Towards this goal, during the past few years, many researches have adopted the microarray technology to study gene expression profiles in breast cancer, however only a small fraction shows clear prognostic significance [4,5]. For example, Van't Veer et al. use DNA microarray analysis on primary breast tumors from 117 patients and utilize a supervised classification method to recognize a 70-gene prognostic signature [4]. Further, they test these previously applicable prognostic markers in a series of 295 consecutive breast cancer patients and the results demonstrate the significance of 70-gene prognostic signature [6]. Wang et al. [7] reveal a 76-gene prognostic signature that can accurately predict distant tumor recurrence by clustering the gene expression profiles and correlating them with prognostic values. By using microarray markers, some machine learning classification methods, such as Support Vector Machine (SVM) [8], Bayes classifier [9], Random Forest (RF)

[*] Corresponding author at: School of Information Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China.

E-mail addresses: sddchina@mail.ustc.edu.cn (D. Sun), aoli@ustc.edu.cn (A. Li), tb214200@mail.ustc.edu.cn (B. Tang), mhwang@ustc.edu.cn (M. Wang).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cmpb.2018.04.008
0169-2607/© 2018 Elsevier B.V. All rights reserved.

[10] have also been applied to predict breast cancer survival. For instance, Nguyen et al. [10] propose to diagnose and prognosticate breast cancer based on random forest classifier and feature selection technique, which outperforms previously reported results.

Given the complexity and heterogeneity of breast cancer survival prediction, a more practical strategy, as proposed by Brenton et al. [11], is to use both clinical data and gene prognostic markers that may contain some complementary information. In addition, with the rapid development of new technologies in the area of medicine, a large amount of clinical data for breast cancer have been generated and collected. By combining both clinical data and microarray markers, different computational methods have been developed for the accurate survival prediction of breast cancer [9,12–14]. For example, Gevaert et al. develop Bayesian networks [9] to integrate both clinical and those 70 gene information by three different strategies including full, decision or partial integration, and demonstrate that the use of clinical and microarray data has better or comparable performance than the methods with clinical or microarray data, respectively. Khademi et al. [14] propose an interesting strategy to reduce the dimensionality of microarray data by applying manifold learning and deep belief network and integrate the clinical data by a probabilistic graphical model. The extensive experiments show a promising result compared to traditional classification methods. Except for microarray and clinical information, the reference human protein interactive network has also been explored to predict breast cancer survival. For example, Das et al. [15] design an elastic-net-based approach named ENCAPP by combining protein network with gene expression dataset to accurately predict survival for human breast cancer. However, the limitation of ENCAPP is that the accuracy of the survival prediction is highly dependent on the quality of the gene expression dataset. Thus there is still considerable room for the improvement of prediction performance of breast cancer survival by incorporating more cancer-related information.

Currently, with the advance of technology in medical imaging [16], there is a great opportunity to analyze pathological images and well study tumor morphology [17]. Previous studies show that some computational methods have been introduced to predict cancer clinical outcome based on pathological images by assuming that pathological images may provide complementary information related to tumor characteristics. Wang et al. [18] propose an integrated framework for non-small cell lung cancer computer aided diagnosis and survival analysis by using representative markers from histopathology images. Zhu et al. [19] design a prediction model to integrate pathological image features with gene expression signature for lung cancer survival prediction. By collecting 2186 Hematoxylin and Eosin pathological whole-slide images (WSIs) of non-small cell lung cancer [20], Yu et al. [21] further distill 9879 representative image features and employ common classification methods to distinguish shorter-term and longer-term survivors. Despite the good performance of the above mentioned approaches for lung cancer, there is still a lack of researches with pathological images for breast cancer clinical outcome analysis due to the complexity and heterogeneity of this serious disease. Meanwhile, the rapidly increasing number of features from different data sources and the use of heterogeneous features may bring a big challenge on how to effectively combine them to apply into breast cancer survival prediction.

To address these issues, in this study, we conduct a new, powerful method named GPMKL for survival prediction of breast cancer by integrating genomic data (gene expression, copy number alteration (CNA), gene methylation and protein expression) and features distilled from pathological image. By employing those high-quality features, multiple kernel learning is further introduced to integrate and accurately predict survival time of breast cancer patients. To verify the effectiveness of pathological images, GPMKL is compared with different independent models that only use genomic data and the results indicate that pathological images could contribute to the remarkable prediction performance. Further, we also compare our proposed framework with other popular state-of-the-art survival models. The best performance achieved by GPMKL also demonstrates the feasibility of integration of genomic data and pathological images and the usefulness of GPMKL in breast cancer survival prediction.

Table 1
The properties of our breast cancer dataset.

Total population of patients	578
Cut-off (years)	5
Survival time	
Longer-term survivors	133
Shorter-term survivors	445
Average age at diagnosis	57.80
Median survival	40.46

2. Materials and methods

Fig. 1 shows the framework of the proposed method called GPMKL. The entire procedure consists of three steps: (A) data generation; (B) model optimization and feature selection; (C) prediction and evaluation. In detail, the proposed method first integrates genomic data including gene expression, gene methylation, CNA, protein and pathological images. Second, breast cancer patients are randomly divided into training and testing sets. The training set is used to train our model and tune parameters on ten-fold cross validation experiment. Finally, the trained model is utilized to do classification on the testing set. We describe each part of our framework further in the following sections.

2.1. Data preparation

We download the publicly available dataset of breast cancer samples from The Cancer Genome Atlas (TCGA) portal: https://portal.gdc.cancer.gov/. TCGA is a comprehensive resource including thousands of patients' data, which is consisted of gene expression, CNA, gene methylation, protein expression and pathological images. In this paper, the latest TCGA available data is downloaded to conduct our research (June 2017) and all above mentioned data types are retrieved from original dataset for further analysis. This downloaded dataset consists of five sub-data and each sub-data include different number of patients. For example, the gene expression and pathological images contain 1100 and 1054 patients, respectively. Then we use Venn diagram (Fig. 2) to vividly exhibit the detailed number of patients in different data types and finally obtain 585 valid patients which consist of those mentioned data types. This dataset contains 578 female and 7 male patients. Considering that the gender distribution is very biased in our dataset and some previous studies demonstrate numerous gender-specific differences for breast cancer [22,23], we further remove the 7 male patients. The average age at diagnosis is 57.80 and the average survival time of all these patients is 40.46 months. Similar to previous study, we define the breast cancer survival prediction in our study as a binary classification problem by the threshold of 5 years [14,24]. In detail, patients in our study are divided into two classes by their survival time, namely longer-term and shorter-term survivors. Among 578 patients, 445 patients are regarded as shorter-term survivors and 133 patients are regarded as longer-term survivors. Moreover, the shorter-term survivors are labeled as 0 while longer-term patients are labeled as 1. The detailed information about our dataset is illustrated in Table 1.

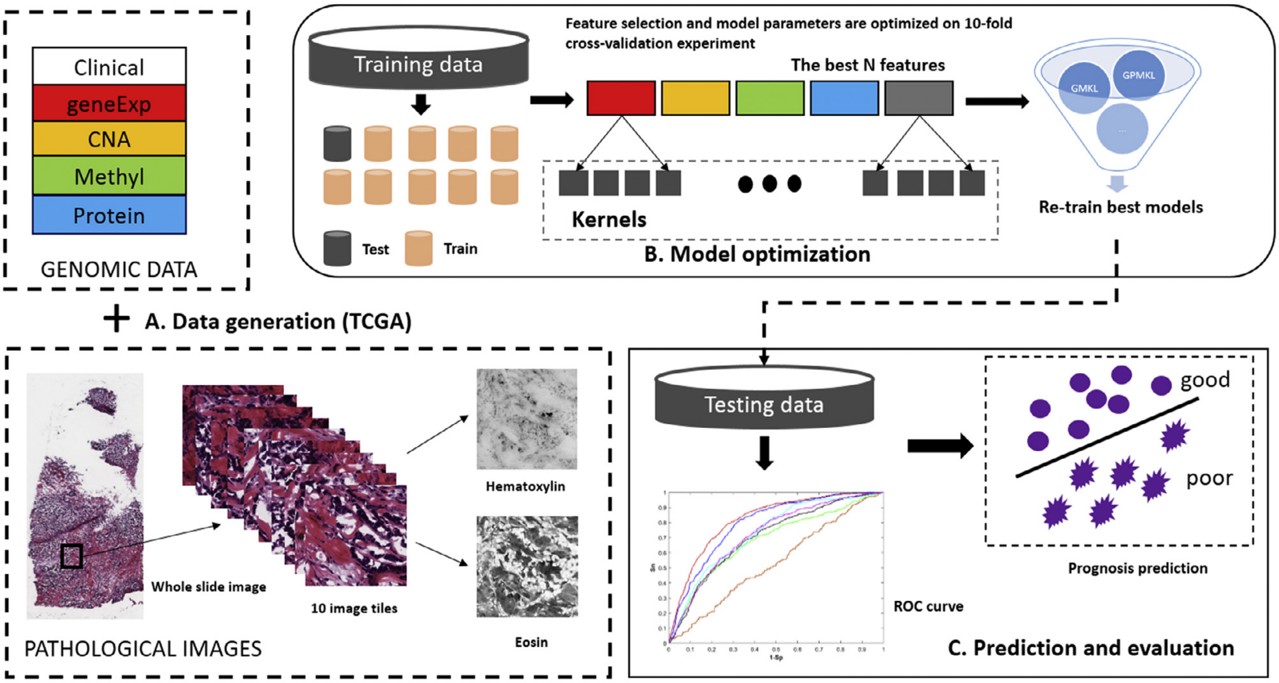

Fig. 1. Overview of the GPMKL method. (A) Genomic data and pathological images are obtained from TCGA. (B) Datasets are randomized, assigning patients to training and testing sets. Feature selection by FSelector package and model parameters are optimized from training sets by 10-fold cross-validation experiment. (C) The pre-trained models are used for precision prognostication with testing sets.

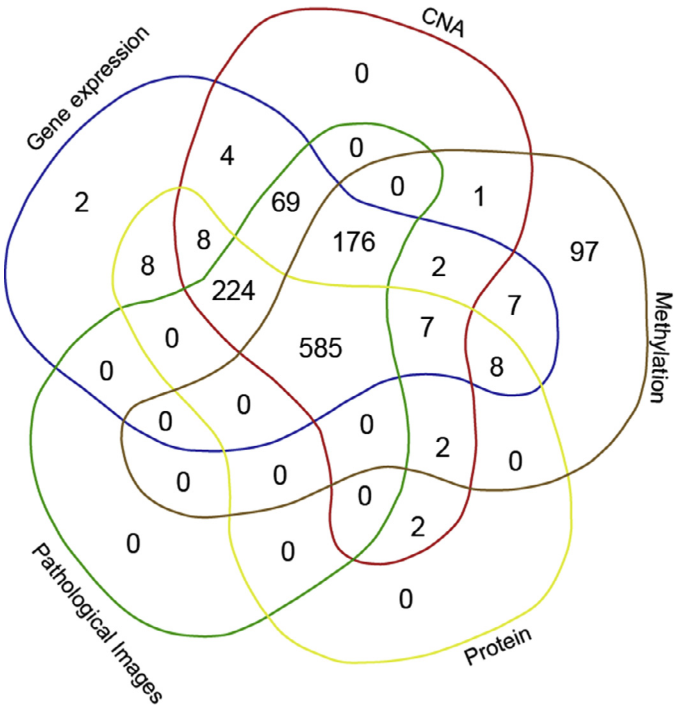

Fig. 2. Venn diagrams of intersections between the multi-data used in our study.

2.2. Genomic data

Similar to the previous work by Ding et al., we first delete the genes with missing values (NA) in more than 10% patients for gene expression, CNA, gene methylation and protein expression [25]. After that, the remaining missing values in each single data type are further estimated using a weighted nearest neighbors algorithm [26]. In addition, according to Gevaert et al. [6], the gene expression profiles are normalized and further discretized into three categories: under-expression (−1), over-expression (1) and baseline (0). For CNA features, we directly utilize the original data with relative linear copy-number values for each gene (from Affymetrix SNP6) [27]. As to gene methylation and protein expression, we also directly utilize the original data, which have been normalized by Z-score.

Table 2
The optimal number of features used for breast cancer survival prediction.

Data category	Number	Feature number
Gene expression	15,972	50
CNA	24,776	10
Methylation	16,474	40
Protein	215	120
Pathological image	1991	130

Through the aforementioned operations of data preprocessing, the gene expression, CNA and gene methylation still consist of approximately 16,000, 25,000 and 16,000 features, respectively and are shown in Table 2. However, increasing the number of features may lead to poor performance due to the curse of high dimensionality and small sample size problems [28,29]. Therefore, by following previous study published in nature communications [21], information gain ratio measure (*FSelector* package) is also applied in this work to obtain the most informative genes from our training dataset and avoid overfitting in model building. Here we use the AUC value (see Experimental Design) as the standard criteria to evaluate the performance of the optimal number of features from different sources listed in Table 2.

2.3. Pathological image data

The Hematoxylin and Eosin (H&E) pathological images corresponding to the genomic data are also obtained from TCGA. Since the original collected H&E images are with the extremely high resolution, the whole slide images with ×40 magnification are tiled into 1000 × 1000 pixels by using bftools [30] under the open microscopy environment. We extract 10 densest tiles from each original image followed previous study [21] since densest tiles include more cells for further investigations.

Motivated by Yao et al. [31], we extract image features with CellProfiler [21], which is a free open source tool developed to help researchers to quantitatively measure features from images. In our work, for each patient, three groups (from cell nuclei, cell cytoplasm and image tile) of features are obtained from pathological images resulting in 1991 image features. These types of image features include the number of cells in image tiles, cell shape, size, texture of the cells and nuclei, as well as the distribution of pixel intensity in the nuclei and cells. Here, 130 image features are also selected as input for our GPMKL model.

2.4. Multiple kernel learning

In our study, we target at integrating multi-dimensional data especially for pathological images. One of the most straightforward approaches for classification tasks is to merge multi-dimensional data as one type of features. Here different data may have different feature representation, combing directly these multiple sources of data as an input of one model would not be efficient [14]. Multiple kernel learning (MKL) has become a natural choice to solve this problem. It is anticipated that the optimal function can be learned by constructing a weighted linear combination of M basis kernels. The equation for combining kernels is represented as follows:

$$K\left(x_i, x_j\right) = \sum_{m=1}^{M} d_m K_m\left(x_i, x_j\right) \tag{1}$$

$$\text{s.t. } d_m \geq 0, \ \text{and} \ \sum_{m=1}^{M} d_m = 1 \tag{2}$$

where d_m is the weight for the mth basis kernel $K_m(x_i,x_j)$.

In fact, MKL framework combines kernels with many complex ways. Some methods have been proposed and many of them achieve better performance than uni-MKL consistently [32–34]. When using MKL, most of the weights with different kernels are 0, which is termed as sparse problem, meaning that some trained kernels do not make a contribution to final MKL model [35]. In our study, we employ simpleMKL [32] as our classification model. simpleMKL method is based on a weighted L2-*norm* regularization technique, which is more efficiently than other classification methods [35]. simpleMKL is also a dual problem implemented with multiple-kernel version of SVM, which can be defined as:

$$f(x) = \sum_{i=1}^{l} \alpha_i^* K\left(x_j, x_i\right) + b^* \tag{3}$$

The decision function for this dual problem is of the form:

$$\begin{aligned} &\min_{f,b,\varepsilon} \frac{1}{2} f_H^2 + C \sum_i \varepsilon_i \\ &\text{s.t.} \quad y_i(f(x_i) + b) \geq 1 - \varepsilon_i \quad \forall_i \\ &\varepsilon_i \geq 0 \qquad\qquad\qquad\qquad \forall_i \end{aligned} \tag{4}$$

where $\|f\|_H$ represents a kernel in Hilbert space associated with a kernel K_m. The overall kernel can be divided to different kernels, replace $\|f\|_H$ to $\sum_m \|f_m\|_{HM}$ then we obtain:

$$\begin{aligned} &\min_{\{f_m\},b,\varepsilon,d} \frac{1}{2} \sum_m \|f_m\|_{HM}^2 + C \sum_i \varepsilon_i \\ &\text{s.t.} \quad y_i \sum_m f_m(x_i) + y_i b \geq 1 - \varepsilon_i \quad \forall_i \\ &\varepsilon_i \geq 0 \quad \forall_i \\ &\sum_m d_m = 1, \ d_m \geq 0 \quad \forall_m \end{aligned} \tag{5}$$

This equation indicates a series of kernels in Hilbert space being formed in L2-*norm* formation. This optimization problem could be solved by the mathematical convex optimization algorithm. The detailed description could be found in [32].

2.5. MKL model on genomic data and pathological image

MKL has become a prior choice to integrate different data types in our study. Employing multiple kernels compared with one single kernel can make the decision function more powerful and enhance the predictive performance. Therefore, we propose a model called GPMKL by using simpleMKL method, which integrates the genomic data (gene expression, CNA, gene methylation and protein expression) and pathological images. Considering the fact that the data using in our study includes five data types, we construct 5 different kernels independently and further integrate them into a universal model. Each kernel corresponds each independent data type (gene expression, CNA, gene methylation, protein expression and pathological images). We choose all SVM kernel types to "Gaussian" (Eq. (6)) and all search range of the parameter δ is {0.25 0.5 1 2 5 7 10 12 15 17 20} [36].

$$K\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\delta^2}\right) \tag{6}$$

Finally, the simpleMKL method directly solves an integrated support vector machine optimization problem instead of learning kernel combination from independent kernels, which vastly reduces computation cost [36]. And simpleMKL uses a reduced gradient algorithm to find the best parameters [37]. In addition, we also develop two independent models named GMKL and PMKL for comparison, which employ genomic data and pathological images data, respectively.

2.6. Other popular state-of the art survival models

Furthermore, we compare the GPMKL with several existing survival models, which are utilized as baseline algorithms. A brief description for each model is given in the following:

Regular Cox models: The Cox proportion hazards model [38] is one of the commonly used semi-parametric model in survival analysis. The LASSSO-Cox [39] and elastic-net penalized Cox (EN-Cox) models [40] are used for comparison in this paper.

Parametric censored regression models (PCRM): PCRM survival model formulates the joint probability of the censored and uncensored instances as a product of death density function and survival functions, respectively. The likelihood function can be determined by combining these two components [41].

Random survival forests (RSF): Random survival forests are designed to enhance the survival prediction performance by ensemble base learning tree [42].

Boosting concordance index (BoostCI): BoostCI is based on the concordance index metric for survival data. It is also designed to an equivalent smoothed criterion by using the sigmoid function [43].

Supervised principal components regression (superPC): It is a model to overcome the issue of survival prediction which cannot guarantee the chosen principal components are related to the patients' label information. This method first chooses a subset of the features and then employs PCR to the chosen subset [44].

All other models except BoostCI used for comparison are implemented in R language. LASSO-Cox and EN-Cox are built using the *cv.glmnet* function from the *glmnet* package. PCRM is from the *survival* package and RSF is from the *randomForestSRC* package [42]. superPC is implemented by *superpc* package [44]. The implementation of BoostCI can be found in the supplementary materials of [43].

3. Experimental design

To comprehensively evaluate our proposed method, the collected dataset is randomly divided into training (80%) and testing sets (20%). Most informative features are selected by employing *FSelector* package and optimal model parameters are determined by 10 fold cross-validation experiment from the training set. Then the pre-trained model and optimal number of features are used to predict on testing set. Considering the robustness of the proposed method, this random dividing process is repeated 20 times [21]. In this manner we can estimate the performance of a specific method and make comparison with other methods.

To assess the relative performance of the GPMKL, we have applied receiver operating characteristic (ROC) curve, which is a commonly used way to show the overall performance. ROC curve is created by drawing the true positive rate (sensitivity, *Sn*) against false positive rate (1-specificity, 1-*Sp*) at various threshold settings. Also the corresponding area under ROC curve namely AUC is also computed, with AUC = 1 represents perfect performance and 0.5 means random guess. Besides, accuracy (*Acc*), precision (*Pre*) and Matthews correlation coefficient (*Mcc*) also have been utilized as prediction performance metrics for the prediction of breast cancer survival. The detailed definitions of those metrics are defined as below:

$$Sn = \frac{TP}{TP + FN} \tag{7}$$

$$Sp = \frac{TN}{TN + FP} \tag{8}$$

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{9}$$

$$Pre = \frac{TP}{TP + FP} \tag{10}$$

$$Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \tag{11}$$

where TP, FP, TN and FN stand for true positive, false positive, true negative and false negative, respectively.

The concordance index (C-index) is employed as our evaluation metric. The C-index is a nonparametric metric to quantify the ranking quality of rankings and is calculated as follows:

$$\mathrm{c} = \frac{1}{n} \sum_{i \in \{1 \ldots N | \delta_i = 1\}} \sum_{s_j > s_i} I\left[X_i\hat{\beta} > X_j\hat{\beta}\right] \tag{12}$$

where n is the number of comparable pairs, *I[.]* is the indicator function and $s.$ is the actual observation. The value of C-index ranges from 0 to 1. The larger C-index value means the better prediction performance of the model and vice versa. 0 is the worst condition, 1 is the best and 0.5 is the value as a random guess.

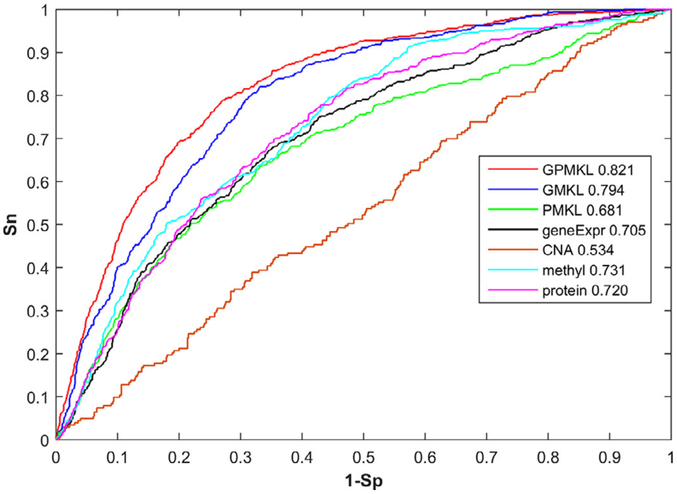

Fig. 3. ROC curves for classifying longer-term and shorter-term survivors from breast cancer dataset. GPMKL obtains an AUC value of 0.821.

4. Results

4.1. Comparison of proposed method based MKL with multiple and single dimensional data

In order to examine the effectiveness of pathological image in breast cancer survival prediction, six different MKL based methods are designed, which include the use of single dimensional data namely the pathological image and the genomic data, respectively, and the joint use of genomic data and pathological images. For simplicity, the MKL based methods that use single dimensional data of gene expression, CNA, gene methylation, and protein expression are thereafter termed as geneExpr-kernel, CNA-kernel, methyl-kernel, and protein-kernel, respectively. It should be stated that the random dividing process is repeated 20 times for all those methods. The main difference between those methods and GPMKL is that they do not integrate multi-dimensional data while only use the single data type. The ROC curves are plotted for seven different methods to compare the predictive performance at each specificity level and displayed in Fig. 3. As shown in Fig. 3, the proposed method achieves significantly better overall performance than all other methods with single dimensional data. Besides ROC curve, the corresponding AUC value for each method is also calculated and displayed in Fig. 3. It is also suggested that GPMKL is consistently better than geneExpr-kernel, CNA-kernel, methyl-kernel, and protein-kernel, respectively. The AUC achieved by the proposed method is 11.6%, 28.7%, 9.0% and 10.1% higher than geneExpr-kernel, CNA-kernel, methyl-kernel, and protein-kernel, respectively. Meanwhile, the corresponding AUC value for GMKL trained with only genomic data is 0.794, and the AUC value for PMKL trained with only pathological image is 0.681, respectively. By incorporating pathological image, the corresponding AUC value is increased to 0.821. Taken together, it is indicated that the pathological image could improve performance for the prediction of breast cancer survival.

In addition, by following the study of Fan et al. [45], a threshold is set for each method such that the specificity of each method is equal to 90.0% (medium) or 95.0% (high). Then we calculate the corresponding *Pre, Acc, Sn* and *Mcc* values of each method and illustrated in Fig. 4. We can see that integration of genomic data and pathological image is better than only using the genomic data or the pathological images for classification, respectively. For example, at the high level of *Sp* (*Sp* = 95.0%), *Pre, Acc, Sn* and *Mcc* values of GPMKL are increased by 3.8%, 0.9%, 4.3% and 4.7% compared with

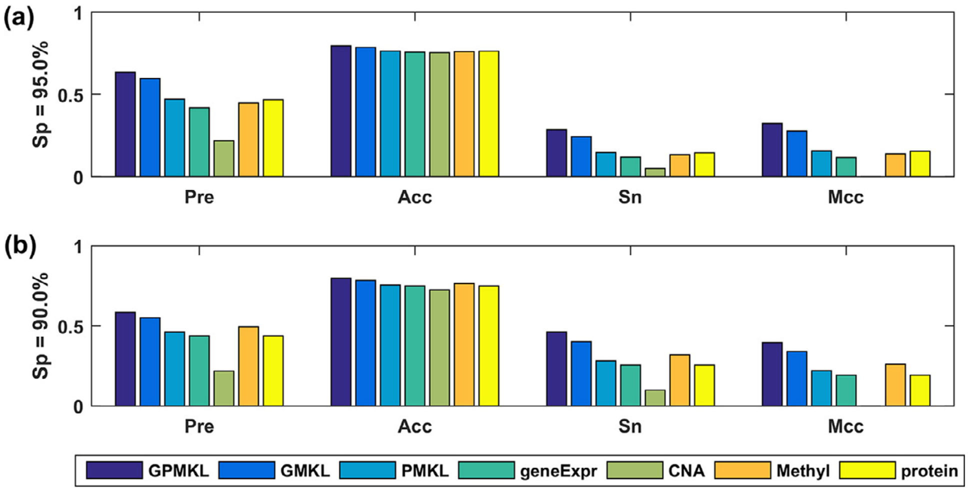

Fig. 4. Performance comparison between GPMKL and other models in different metrics. Pre, Acc, Sn and Mcc values at stringent levels of Sp = 95.0% (a) and Sp = 90.0% (b).

Table 3
Performance comparison of the proposed method and other existing related methods using AUC values.

Methods	Genomic data	Pathological image	Genomic data + pathological image
LASSO-Cox	0.697 ± 0.069	0.655 ± 0.059	0.698 ± 0.060
En-Cox	0.667 ± 0.080	0.649 ± 0.056	0.677 ± 0.067
PCRM	0.620 ± 0.067	0.608 ± 0.055	0.546 ± 0.043
RSF	0.722 ± 0.049	0.620 ± 0.067	0.718 ± 0.055
BoostCI	0.716 ± 0.037	0.622 ± 0.058	0.717 ± 0.037
superPC	0.659 ± 0.069	0.595 ± 0.056	0.698 ± 0.048
GPMKL	**0.802 ± 0.032**	**0.690 ± 0.046**	**0.828 ± 0.034**

GMKL and have an improvement of 16.4%, 3.2%, 13.9% and 16.7% compared with PMKL, respectively. In addition, when *Sp* decreases into 90.0%, the *Sn* value of the proposed method is 0.285, while the corresponding *Sn* values of geneExpr-kernel, CNA-kernel, methyl-kernel, and protein-kernel are 0.118, 0.049, 0.133 and 0.144, respectively. The precision values of GPMKL, GMKL, PMKL, geneExpr-kernel, CNA-kernel, methyl-kernel, and protein-kernel are 0.634, 0.596, 0.470, 0.418, 0.217, 0.447 and 0.467, respectively. In summary, the results demonstrate that GPMKL based on pathological images can successfully predict survival outcomes of breast cancer patients and is superior to current practice which only utilizes genomic data.

We implement an individual experiment to discuss which information plays an important role in improving the prediction results. Specifically, in each time we only remove one of gene expression, CNA, methylation and protein, and then draw the ROC curve (as shown in Supplementary Fig. S1) for comparison [46,47]. We find that the all the information are vital for breast cancer clinical outcome prediction, and gene expression and protein information play relatively more important role than others. In addition, we show the results of only using the pathological images and then coupling them with one omic data in Supplementary Fig. S2. From above results, we find that all omic data are valuable in coupling with the pathological images, and CNA make a little contribution to overall prediction performance.

4.2. Comparison with other prediction methods

As pointed out before, the proposed GPMKL by integrating heterogeneous features including genomic data and pathological images achieve the better performance than those methods with single dimensional data. To further verify the effectiveness of GPMKL, we also compare it with six popular state-of-the-art survival models, namely LASSO-Cox, En-Cox, PCRM, RSF, BoostCI and superPC. Three groups of experiments are conducted on three different combinations of pathological images and genomic data. Also, the AUC value (mean value and standard error) is utilized to evaluate prediction performance of different methods and the detailed results are listed in Table 3. From Table 3, it can be seen that GPMKL obtains the best performance in breast cancer survival prediction. In detail, the improvement by GPMKL can even be over 11.0% compared to the best performance (mean AUC value) achieved by RSF method on genomic data + pathological image group. For genomic data group, GPMKL obtains the mean AUC value of 0.802, which is 14.3%, 18.2% higher than superPC, PCRM, respectively. For pathological image group, GPMKL obtains the mean AUC value of 0.690, which is 9.5%, 8.2% higher than superPC, PCRM, respectively. Finally, it can be observed that the most of methods achieve better performance when employing both genomic data and pathological image than methods only using the single genomic data. To some extent, these results further demonstrates that the pathological images can provide enough complementary information for breast cancer survival prediction, and indicate the superiority of GPMKL.

Meanwhile, the C-index of each method is also calculated and illustrated in Fig 5. Undoubtedly, GPMKL outperforms other algorithms consistently in breast cancer survival prediction. It shows that the average C-index value of GPMKL is 0.643, while the corresponding average C-index values of LASSO-Cox, En-Cox, PCRM, RSF, BoostCI and superPC are 0.612, 0.602, 0.531, 0.617, 0.598 and 0.605, respectively. The reason why GPMKL achieves the best performance is that genomic data and pathological images can provide predictive powers and could have complementary relationship, and multiple kernel learning is very efficient in predicting breast cancer survival time.

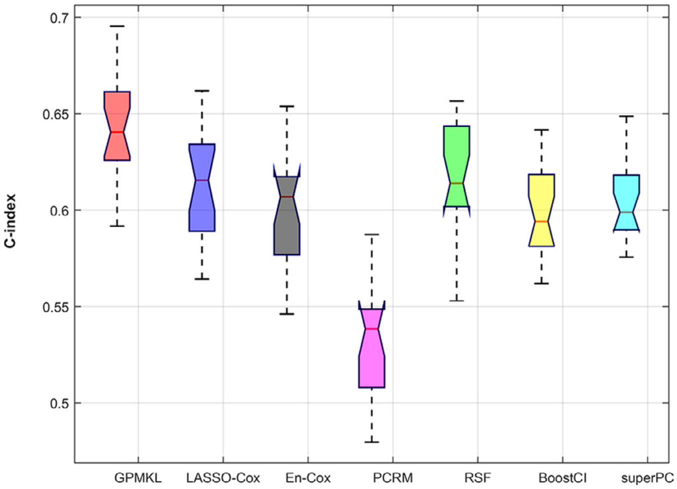

Fig. 5. The C-index between different methods and the random dividing process is repeated 20 times.

4.3. Investigation of quantitative features from pathological images

In our study, we first use feature extraction pipeline which is built from CellProfiler [48,49]. For each patient in the experiment, we extract 1991 features obtained from original pathological images and select 130 features (Supplementary Table S1) as input to our GPMKL model. Types of these features are divided into cell nuclei-level features (98 features), cell cytoplasm-level features (12 features) and image tile-level features (20 features). Thus cell nuclei-level features are the main features types, which is reasonable and consistent with previous work studying in non-small cell lung cancer survival [21]. Further, we also investigate the 130 features associated with breast cancer survival prediction. The frequent prognostic features that distinguish longer-term survivors and shorter-term survivors include distribution of pixel intensity of the nuclei, nuclei and cytoplasm texture features. In addition, as to each tumor cell, Zernike shape features in cell nucleus are extracted frequently for identifying longer-term survivor and shorter-term survivors. The most important features also include texture features which quantifies the correlations between nearby pixels in the regions of interest. This indicates that both local anatomical structures including shape of cell nuclei and cytoplasm, and global modes of the tumor cell nucleus such as texture of the nuclei are related to clinical survival outcomes.

4.4. Analysis of novel BRCA-related genes

Additionally, a simple analysis of novel genes related to breast cancer is conducted. In detail, we search the corresponding top 10 ranked genes on genomic data. It should be noted that for genomic data, we only take gene expression and methylation dataset as examples to further verify the effectiveness of GPMKL and find that some genes have been indicated to have effects on breast cancer survival prediction. These genes and their linked references are listed in Table 4. For gene expression dataset, by consulting literature we find that 5 of top 10 feature names ranked in candidate genes are reported to have functions in breast cancer. For example, Connett et al. find that the IRF2 expression is proved to be associated with breast cancer (PMID: 16241857). Meanwhile, Jiang et al. point out that SEMA3A has also been considered as a candidate tumor suppressor not only in breast cancer, but also in prostate cancer and glioma (PMID: 25812535). On the other hand, the 3 genes from top 10 on the methylation dataset are also found to play critical roles in human breast cancer survival prediction. For example, miR-185 directly targets E2F6 and DNMT1 and indirectly upregulates BRCA1 in triple-negative breast cancer (PMID: 25956916). All these studies highlight the importance of the top ranked candidate genes in breast cancer.

Table 4
The genes that have been reported before.

Genes/gene ID	Evidence	Paper refer
IRF2/3660	Interferon regulatory factor 1 (IRF-1) and IRF-2 expression in breast cancer tissue microarrays	PMID: 16241857
RAB8B/51762	A miR-200b-binding site was found in the 3′-UTR of RAB21, RAB23, RAB18, RAB3B, RAB37, RAB8B, RAB7A, RAP1B, RAP2C	PMID: 24477653
ZNF438/220929	transcription factors (FOXN3, NFIB, TAF6, TCF12, ZNF295, ZNF438)	PMID: 25956916
SEMA3A/10371	Recently, SEMA3A has also been considered as a candidate tumor suppressor, since it is often downregulated in numerous types of cancer, including prostate cancer, breast cancer and glioma.	PMID: 25812535
PIP4K2A/5305	Amplification of PIP4K2A was only observed in a small fraction of breast cancers	PMID: 24209622
PCNXL3/399909	However, recent results from the genome-wide scan of 1140 breast cancer cases and 1140 controls that genotyped 5 markers to capture the majority of common variation in SIPA1 as well as the neighboring regions (that included the genes PCNXL3 and MAP3K11)	PMID: 19089925
E2F6/1876	miR-185 suppresses tumor proliferation by directly targeting E2F6 and DNMT1 and indirectly upregulating BRCA1 in triple-negative breast cancer	PMID: 25956916
REXO4/57109	Downregulation of hPMC2 imparts chemotherapeutic sensitivity to alkylating agents in breast cancer cells	PMID: 25849309

5. Discussions and conclusions

As breast cancer is the most common and malignant disease in the world, in our presented research, we focus on improving prediction performance of breast cancer survival time. Considering the pathological images have become a very active field in healthcare research, we propose a method named GPMKL by integrating genomic data and pathological images for breast cancer survival prediction. The comparison with other survival prediction methods demonstrates that with the help of pathological images and multiple kernel learning techniques, the proposed GPMKL model is very excellent in breast cancer survival prediction work. There are several factors leading to the success of GPMKL. First, multiple kernel learning technique is good at capturing heterogeneous features from different types of data by using different kernels. Second, we consider the amount of valuable information from pathological images and use them into our proposed models.

Despite the efficiency of GPMKL, it still has some limitations in survival prediction of breast cancer. For example, this work can be extended and validated by employing a larger population of breast cancer patients. The current sample size is limited by the availability of genomic data and pathological image data. It is expected that the performance of the proposed method would be enhanced when more samples become available in the future. In addition, we also think that it will be more meaningful for cancer researchers if GPMKL built for each subtype since clinical outcome prediction is heavily affected by the subtypes [50,51]. Here, the TCGA breast

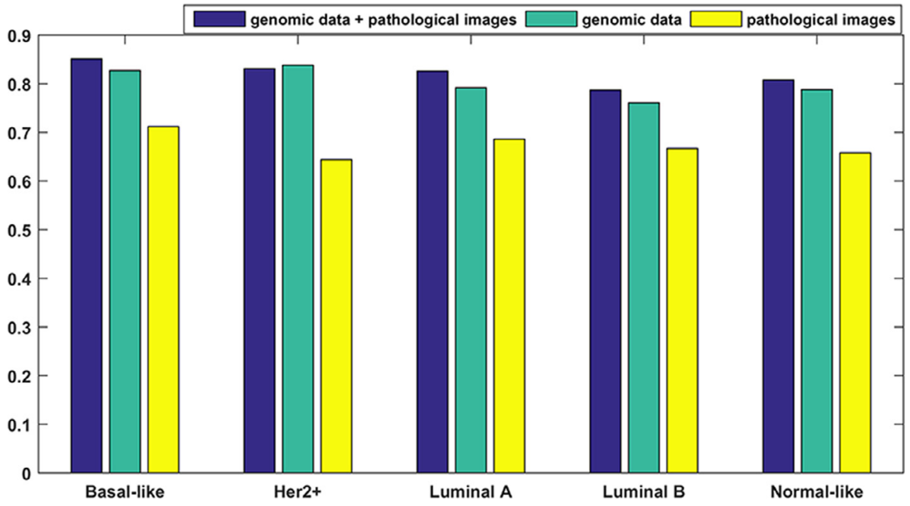

Fig. 6. The breast cancer subtypes show different prediction performance by GPMKL.

cancer dataset in our study involves five major molecular subtypes: luminal (A or B), HER2-enriched, basal-like and normal-like. Similar to previous studies [52,53], we use the efficient classifier called PAM50 [54] to distinguish these five subtypes from our dataset, while the corresponding AUC values of luminal A, luminal B, HER2-enriched, basal-like and normal-like are shown in Fig. 6, respectively. These results also suggest that integrating genomic data and pathological images could effectively improve the prediction performance in each breast cancer subtype. Furthermore, a promising expansion to the GPMKL in the future work would be to employ deep learning models for multimodal data fusion and feature extraction. We will also try to extract comprehensive valuable information directly on WSIs and extend our framework to other cancers. Finally, another research direction is to construct a multi-task learning system aiming to cancer susceptibility, cancer recurrence, and cancer treatment.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China through Grants No. 61571414, No. 61471331 and No. 31100955.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2018.04.008.

References

[1] D. Hanahan, R.A. Weinberg, The hallmarks of cancer, Cell 100 (2000) 57–70.
[2] M.F. Akay, Support vector machines combined with feature selection for breast cancer diagnosis, Expert Syst. Appl. 36 (2009) 3240–3247.
[3] J. Ferlay, C. Héry, P. Autier, R. Sankaranarayanan, Global burden of breast cancer, in: Breast Cancer Epidemiology, Springer, 2010, pp. 1–19.
[4] L.J. Van't Veer, H. Dai, M.J. Van De Vijver, Y.D. He, A.A. Hart, M. Mao, et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (2002) 530–536.
[5] D.M. Abd El-Rehim, G. Ball, S.E. Pinder, E. Rakha, C. Paish, J.F. Robertson, et al., High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses, Int. J. Cancer 116 (2005) 340–350.
[6] M.J. Van De Vijver, Y.D. He, L.J. Van't Veer, H. Dai, A.A. Hart, D.W. Voskuil, et al., A gene-expression signature as a predictor of survival in breast cancer, N. Engl. J. Med. 347 (2002) 1999–2009.
[7] Y. Wang, J.G. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang, et al., Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, Lancet North Am. Ed. 365 (2005) 671–679.
[8] X. Xu, Y. Zhang, L. Zou, M. Wang, A. Li, A gene signature for breast cancer prognosis using support vector machine, in: Biomedical Engineering and Informatics (BMEI), 2012 5th International Conference on, 2012, pp. 928–931.
[9] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, B. De Moor, Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks, Bioinformatics 22 (2006) e184–e190.
[10] C. Nguyen, Y. Wang, and H.N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," 2013.
[11] J.D. Brenton, L.A. Carey, A.A. Ahmed, C. Caldas, Molecular classification and molecular forecasting of breast cancer: ready for clinical application? J. Clin. Oncol. 23 (2005) 7350–7360.
[12] Y. Sun, S. Goodison, J. Li, L. Liu, W. Farmerie, Improved breast cancer prognosis through the combination of clinical and genetic markers, Bioinformatics 23 (2006) 30–37.
[13] A.-L. Boulesteix, C. Porzelius, M. Daumer, Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value, Bioinformatics 24 (2008) 1698–1706.
[14] M. Khademi, N.S. Nedialkov, Probabilistic graphical models and deep belief networks for prognosis of breast cancer, in: Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on, 2015, pp. 727–732.
[15] J. Das, K.M. Gayvert, F. Bunea, M.H. Wegkamp, H. Yu, ENCAPP: elastic-net-based prognosis prediction and biomarker discovery for human cancers, BMC genomics 16 (2015) 263.
[16] W.K. Moon, Y.-W. Lee, Y.-S. Huang, S.H. Lee, M.S. Bae, A. Yi, et al., Computer-aided prediction of axillary lymph node status in breast cancer using tumor surrounding tissue features in ultrasound images, Comput. Methods Programs Biomed. (2017).
[17] J.T. Kwak, S.M. Hewitt, Multiview boosting digital pathology analysis of prostate cancer, Comput. Methods Programs Biomed. 142 (2017) 91–99.
[18] H. Wang, F. Xing, H. Su, A. Stromberg, L. Yang, Novel image markers for non-small cell lung cancer classification and survival prediction, BMC Bioinform. 15 (2014) 310.
[19] X. Zhu, J. Yao, X. Luo, G. Xiao, Y. Xie, A. Gazdar, et al., Lung cancer survival prediction from pathological images and genetic data—an integration study, in: Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on, 2016, pp. 1173–1176.
[20] K. Tomczak, P. Czerwinska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, Contemp Oncol (Pozn) 19 (2015) A68–A77.
[21] K.-H. Yu, C. Zhang, G.J. Berry, R.B. Altman, C. Ré, D.L. Rubin, et al., Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features, Nature Commun. 7 (2016).
[22] M.T. Dorak, E. Karpuzoglu, Gender differences in cancer susceptibility: an inadequately addressed issue, Front. Gen. 3 (2012) 268.
[23] J.M. Greif, C.M. Pezzi, V.S. Klimberg, L. Bailey, M. Zuraek, Gender differences in breast cancer: analysis of 13,000 breast cancers in men from the National Cancer Data Base, Ann. Surg. Oncol. 19 (2012) 3199–3204.
[24] M. Lundin, J. Lundin, H. Burke, S. Toikkanen, L. Pylkkänen, H. Joensuu, Artificial neural networks applied to survival prediction in breast cancer, Oncology 57 (1999) 281–286.
[25] Z. Ding, S. Zu, J. Gu, Evaluating the molecule-based prediction of clinical drug responses in cancer, Bioinformatics 32 (2016) 2891–2895.

[26] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, et al., Missing value estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520–525.

[27] J. Xi, A. Li, Discovering recurrent copy number aberrations in complex patterns via non-negative sparse singular value decomposition, IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) 13 (2016) 656–668.

[28] J. Tan, J.H. Hammond, D.A. Hogan, C.S. Greene, ADAGE analysis of publicly available gene expression data collections illuminates Pseudomonas aeruginosa-host interactions, bioRxiv (2015) 030650.

[29] A. Jain, D. Zongker, Feature selection: Evaluation, application, and small sample performance, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 153–158.

[30] M. Linkert, C.T. Rueden, C. Allan, J.-M. Burel, W. Moore, A. Patterson, et al., Metadata matters: access to image data in the real world, J. Cell Biol. 189 (2010) 777–782.

[31] J. Yao, D. Ganti, X. Luo, G. Xiao, Y. Xie, S. Yan, et al., Computer-assisted diagnosis of lung cancer using quantitative topology features, in: International Workshop on Machine Learning in Medical Imaging, 2015, pp. 288–295.

[32] A. Rakotomamonjy, F.R. Bach, S. Canu, Y. Grandvalet, SimpleMKL, J. Mach. Learn. Res. 9 (2008) 2491–2521.

[33] M. Kloft, U. Brefeld, S. Sonnenburg, A. Zien, Lp-norm multiple kernel learning, J. Mach. Learn. Res. 12 (2011) 953–997.

[34] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, J. Mach. Learn. Res. 12 (2011) 2211–2268.

[35] M. Kloft, U. Brefeld, P. Laskov, and S. Sonnenburg, "Non-sparse multiple kernel learning," 2008.

[36] Y. Zhang, A. Li, C. Peng, M. Wang, Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning, IEEE/ACM Trans. Comput. Biol. Bioinf. 13 (2016) 825–835.

[37] Y. Gu, Q. Wang, X. Jia, J.A. Benediktsson, A novel MKL model of integrating LiDAR data and MSI for urban area classification, IEEE Trans. Geosci. Remote Sens. 53 (2015) 5312–5326.

[38] D.R. Cox, Regression models and life-tables, in: Breakthroughs in Statistics, Springer, 1992, pp. 527–541.

[39] R. Tibshirani, The lasso method for variable selection in the Cox model, Stat. Med. 16 (1997) 385–395.

[40] Y. Yang, H. Zou, A cocktail algorithm for solving the elastic net penalized Cox's regression in high dimensions, Statist. Interface 6 (2013) 167–173.

[41] J.D. Kalbfleisch, R.L. Prentice, The Statistical Analysis of Failure Time Data, vol. 360, John Wiley & Sons, 2011.

[42] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, Random survival forests, Ann. Appl. Statist. (2008) 841–860.

[43] A. Mayr, M. Schmid, Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations, PLoS One 9 (2014) e84483.

[44] E. Bair, T. Hastie, D. Paul, R. Tibshirani, Prediction by supervised principal components, J. Am. Statist. Assoc. 101 (2006) 119–137.

[45] W. Fan, X. Xu, Y. Shen, H. Feng, A. Li, M. Wang, Prediction of protein kinase-specific phosphorylation sites in hierarchical structure using functional information and random forest, Amino acids 46 (2014) 1069–1078.

[46] Y. Dou, B. Yao, C. Zhang, PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine, Amino Acids 46 (2014) 1459–1469.

[47] F. Zhang, M. Wang, J. Xi, J. Yang, A. Li, A novel heterogeneous network-based method for drug response prediction in cancer cell lines, Sci. Rep. 8 (2018) 3355.

[48] A.E. Carpenter, T.R. Jones, M.R. Lamprecht, C. Clarke, I.H. Kang, O. Friman, et al., CellProfiler: image analysis software for identifying and quantifying cell phenotypes, Genome Biol. 7 (2006) R100.

[49] L. Kamentsky, T.R. Jones, A. Fraser, M.-A. Bray, D.J. Logan, K.L. Madden, et al., Improved structure, function and compatibility for CellProfiler: modular high--throughput image analysis software, Bioinformatics 27 (2011) 1179–1180.

[50] Z. Liu, X.-S. Zhang, S. Zhang, Breast tumor subgroups reveal diverse clinical prognostic power, Sci. Rep. 4 (2014) 4002.

[51] C. Desmedt, B. Haibe-Kains, P. Wirapati, M. Buyse, D. Larsimont, G. Bontempi, et al., Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes, Clin. Cancer Res. 14 (2008) 5158–5165.

[52] C.M. Kelly, P.S. Bernard, S. Krishnamurthy, B. Wang, M.T. Ebbert, R.R. Bastien, et al., Agreement in risk prediction between the 21-gene recurrence score assay (Oncotype DX®) and the PAM50 breast cancer intrinsic classifier^{TM} in early-stage estrogen receptor–positive breast cancer, Oncologist 17 (2012) 492–498.

[53] M. Gnant, M. Filipits, R. Greil, H. Stoeger, M. Rudas, Z. Bago-Horvath, et al., Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone, Ann. Oncol. 25 (2013) 339–345.

[54] J.S. Parker, M. Mullins, M.C. Cheang, S. Leung, D. Voduc, T. Vickery, et al., Supervised risk predictor of breast cancer based on intrinsic subtypes, J. Clin. Oncol. 27 (2009) 1160–1167.