

Multi-Kernel LS-SVM Based Bio-Clinical Data Integration: Applications to Ovarian Cancer

Jaya Thomas^{a,b}, Lee Sael^{a,1,*}

^a*Department of Computer Science, State University of New York, Incheon, Korea*

^b*Department of Computer Science, Stony Brook University, New York, USA*

Abstract

The medical research facilitates to acquire a diverse type of data from the same individual for particular cancer. Recent studies show that utilizing such diverse data results in more accurate predictions. The major challenge faced is how to utilize such diverse data sets in an effective way. In this paper, we introduce a multiple kernel based pipeline for integrative analysis of high-throughput molecular data (somatic mutation, copy number alteration, DNA methylation and mRNA) and clinical data. We apply the pipeline on Ovarian cancer data from TCGA. After multiple kernels have been generated from the weighted sum of individual kernels, it is used to stratify patients and predict clinical outcomes. We examine the survival time, vital status, and neoplasm cancer status of each subtype to verify how well they cluster. We have also examined the power of molecular and clinical data in predicting dichotomized overall survival data and to classify the tumor grade for the cancer samples. It was observed that the integration of various data types yields higher log-rank statistics value. We were also able to predict clinical status with higher accuracy as compared to using individual data types.

Keywords: Integrative analysis, \LaTeX , Multiple kernel, Molecular data
Clinical data, Patient stratification

*Corresponding author

Email addresses: jaya.thomas@sunkorea.ac.kr (Jaya Thomas),
sael@cs.stonybrook.edu (Lee Sael)

1. Introduction

Cancer is a disease with extreme complexity which alters the function of combination of genes. It is believed to be an outcome of accumulated genetic changes [1]. Among various types of cancer, ovarian cancer is the fifth most common cancers diagnosed in females [2] with overall five year survival rate only around 44%[3]. The Cancer Genome Atlas (TCGA)[4] reports diverse genomic information with paired clinical information for more than 500 cases of ovarian serous cystadenocarcinoma. The genomic information includes copy number alteration (CNA), somatic mutation, gene expression, and DNA methylation. Understanding the genetic changes in cancer patients through this rich information allows for better diagnostics and treatment of cancer, including ovarian cancer.

Integrative analysis of multiple perspective of a patient helps in both patient stratification and clinical outcome prediction. Patient stratification and clinical outcome predictions both help the researchers in understanding and exploring the genomic characteristics in relationship with their current phenotypes and thus to recognizing opportunities for clinical improvement. In case of cancer data analysis, including ovarian cancer data, an improved stratification and clinical prediction can be achieved by integrative analysis of the multiple bio-clinical data. However, due to the complex relationship between the multiple data types, the integrative analysis is still a challenging task.

There are several works related to clinical outcome predictions. Wang *et al.* [5] have used gene expression data to predict distant metastasis of lymph-node-negative primary breast cancer. They identified a 76-gene signature consisting of 60 genes for patients positive for oestrogen receptors (ER) and 16 genes for ER-negative patients. Teschendorff *et al.* [6] proposed a gene expression classifier for ER positive breast cancer. Zhang *et al.* [7] used copy number alterations in combination with gene expression to identify the genomic loci and their mapped genes, having a high correlation with distant metastasis capability of human breast cancer. Deneberg *et al.* [8] used gene specific and global methylation

patterns predict outcome in patients with acute myeloid leukemia. They also concluded in their work that global and gene specific methylation patterns are independently associated with the clinical outcome in AML patients. Nair *et al.* [9] reported a comprehensive review on the clinical outcome prediction by the miRNA expression for numerous types of cancer. These approaches only integrated a smaller number of data types and failed to integrate with other levels of genomic data.

On the other hand, for the patient stratification, biomarkers, genetic profiles, research data along with clinical information are used to find a subgroup among the patients thereby making easier to detect and interpret relationships as well as predict outcomes in specific subgroup. Kim *et al.*[10] considers somatic mutation profile and exploited k-means clustering to identify the tumor subtypes. The sparsity of the mutation data was handled by applying Jaccard and Euclidean distance measures. Further, the Cox proportional hazards regression model was used to find the similarity between the derived subtypes and the patient survival time. In their recent work [11], a compressed somatic mutation profile was suggested for fast comparison. The profile utilized Gene-Ontology and non-negative matrix factorization for condensing the mutation profile. To verify their work, stratification was performed on various cancer types. Hofree *et al.* [12] has used genome-scale somatic mutation profiles in combination with a gene interaction network to carry out subgrouping of patients. Recently, Wang *et al.*[13] proposed a modified consensus clustering to carry out patient stratification for breast cancer patients. The approach considered both numerical and categorical data for mRNA and miRNA data set.

Analysis of one or few data types may not be sufficient for accurate predict or stratification. Thus, efforts to integrate the molecular data were carried out. Thomas *et al.*[14] work presents two general class of heterogeneous data integration, i.e., Multiple Kernel learning and Bayesian network, are detailed and discussed in the bioinformatics domain. Also, many problem specific integrative approaches have been proposed to associate the molecular data with the clinical outcome. These includes a software package implemented in R [15] to show

the effect of DNA methylation and copy number alterations in gene expression of several known oncogenes for two cancer type glioblastoma multiforme and ovarian. Kim *et al.* [16] proposed a graph based integrated framework using CNA, methylation, miRNA, and gene expression data to carry out molecular based classification of clinical outcomes. In this approach a single graph was constructed by determining the optimum linear combination coefficient from the multiple graph obtained at different genomic level. Sohn *et al.* [17] modeled the influence of multi-layered genomic features on gene expression traits by modeling an integrative statistical framework based on a sparse regression. The results showed that using CNA, miRNA, and methylation on gene expression in the predictive power for gene expression level is improved over a single data type based analysis. Schafer *et al.* ([18] approach integrated copy number and gene expression by a modified correlation coefficient and an explorative Wilcoxon test to find DNA regions of abnormalities. The recent work also includes model based prediction of clinical outcomes. Mankoo *et al.* [19] have applied multi-variate Cox Lasso model and median time-to-event prediction algorithm on data set integrated from the four genomic data types (CNA, methylation, miRNA, and gene expression data). Yuan *et al.* [20] evaluated the predictive power of patient survival and binary clinical outcome using clinical data in combination with one molecular data: somatic copy number alteration, DNA methylation, and mRNA, miRNA and protein expression. They showed slight improvement in some cases when clinical information was combined with one of the molecular. Although this paper showed the predictive power for clinical data in combination with a molecular data, all available molecular data was not used integratively.

Integrative analysis method that can cover heterogeneity of data types in molecular data and clinical data can be beneficial in predicting the prognostics of patients via stratifying the patients in the different risk groups. Multiple kernel learning is well known for addressing various data heterogeneity. Moreover, Kernel methods, including multiple kernels, are well suited handling non-linearity of high dimensional data by mapping data to feature space [21].

In this paper, we make the following contributions:

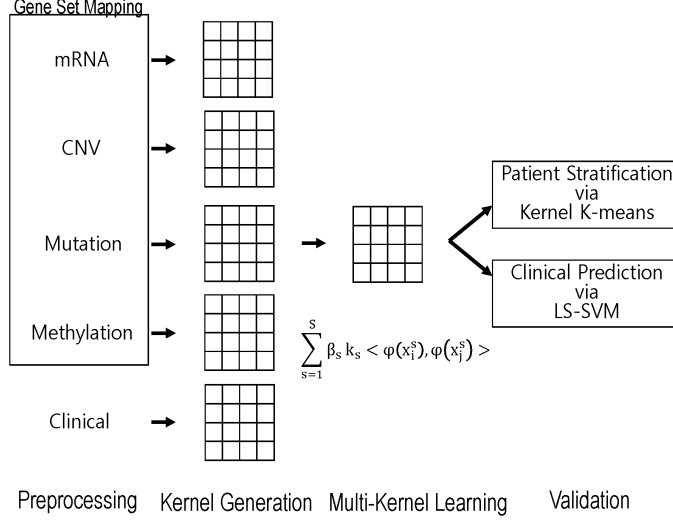


Figure 1: Multi kernel learning based integrative pipelined model

1. **Combines clinical data with multiple molecular data.** We examine how adding more molecular information increases the prediction performance in stratifying ovarian cancer patients, and predicting tumor grade and patient survival time.
2. **Propose a multiple kernel based pipeline model (Fig. 1) to integrate multiple heterogeneous data types.** The proposed model allows to analyze heterogeneous data i.e., combines data with diverse background distributions, relations, dimensions, and formats to enhance the statistical significance and thus, obtain more refined information.
3. **Propose the data pre-processing using patient-centered gene set analysis.** It allows to handle the large heterogeneous tumor data by grouping them into much smaller set of pathways and biologic processes.

2. Material and methods

2.1. Datasets and Raw Mutation Scores

Data are initially selected and downloaded 312 samples that contained all four genomic data types, i.e., copy number alternation, methylation, mRNA expression and the mutation information, from TCGA data portal [22] via TCGA assembler [23] and TCGA Firehose [24]. The summary of the genomic data types and number of associated genes for each data type in the 312 samples are shown in Table 1. Clinical information of the 312 samples is also downloaded from TCGA. The clinical data includes the survival time (days to death), age, tumor stage, tumor grade, vital status and neoplasm cancer status.

Table 1: Numbers of samples and features of data types for OV cancer.

Data Type	Platform	#Genes altered in 312 patient	Union of considered genes
Methylation	Illumina Human Meth.	13772	13772
CNA	Agilent 1M	16383	16070
mRNA expression	AgilentG4502A	18361	16070
Mutation	WUSM	9039	9039

Description of the data types and how each are further processed are provided in the following. For each DNA methylation sample, the percent signal that is methylated is described as beta value recorded for each sample locus. The beta values are continuous variable that range between 0 and 1 indicating the ratio of the intensity of the methylation [25]. After downloading level three data from TCGA assembler [23], the data is pre-processed and determined to be methylated, if they show a percentage of methylation (beta) greater than a certain threshold (0.3 for patient data) or unmethylated if the value fall below the threshold as discussed by Warden et al. [26]. Thus, the new data matrix constructed for the methylation has a score of 1, if gene is methylated else set

to 0. The level three data for copy number alternation (CNA) was obtained from the GISTIC [27] analysis. GISTIC identifies genomic regions that are significantly gained or lost across a set of tumors. It contains data about the significant regions of amplification and deletion as well as which samples are amplified or deleted in each of these regions. The matrix element with value of 0 indicates no amplification or deletion above the threshold. Amplifications are positive numbers: 1 denotes amplification above the amplification threshold; 2 denotes amplifications larger to the arm level amplifications observed for the sample. Deletions are represented by negative table values: -1 represents deletion beyond the threshold; -2 represents deletions greater than the minimum arm-level deletion observed for the sample. The data matrix generated from CNA data sets 1 if the gene is amplified or deleted and 0 if otherwise. The dataset downloaded for level 3 mRNA from TCGA Firehose [24] contains \log_2 ratio for the gene expression. The \log_2 ratio ranges from 0 to 16, representing relative gene expression levels. The level 2 somatic mutation data download from TCGA is already in the required matrix format with the entries showing either 0 or 1 indicating the presence or absence of mutation in the gene.

2.2. Gene Sets and Adjusted Mutation Scores

We group the genes based on involvement in same pathway or having similar molecular signature, thus some of the genes that do not fall in these categories were filtered out. The total number of genes considered in this study are summarized in the last column of Table 1.

Considering gene set takes into account the fact that genes do not act in isolation, but they interact with other genes through complex system. Also, cancer occurs not in a single gene, but rather, a group of genes that interact amongst each other in the complex biological network [10, 11]. Moreover, the biological significance can be better analyzed by considering the interaction with neighbouring genes. For the different data sources, this measure helps to construct a patient to geneset matrix containing the genomic information. We have considered, the functional group information of genes initially downloaded

from the Molecular Signatures Database (MSigDb) [28] and recreated to remove redundancy. In the MSigDB, we select the group information based on pathway (C2: 4722 gene sets) and based on motif (C3: 836 gene sets). MSigDB contains gene sets generated from KEGG [29], Canonical Pathway [30], BIOCARTA [31] and REACTOME [32]. The motif gene set contained in MSigDB are miRNA targets (MIR) and transcription factor target (TFT).

We recreate the gene sets to generate unified gene groups with small overlaps while maximizing the number of genes covered. We filter out the gene sets with more than 85% overlap as described in Algorithm 1. After filtering, 2099 gene sets remained and the gene sets cover 16070 genes out of the initial 16095 genes.

Data: Gene sets in C2 and C3

Result: Selected gene set in F

$S = \{\text{all sets in C2 and C3 ordered by number of genes in the gene set, smallest to largest}\}$

$F = \{\}$ // empty set

```

foreach  $s1$  in  $S$  do
    SimSet = 0
    foreach  $s2$  in  $S - \{s1\}$  do
        dist =  $(s1 \cap s2) / (s1 \cup s2)$  // Jaccard similarity
        if  $dist \geq 0.85$  then
            SimSet++;
        end
    end
    if  $SimSet == 0$  then
        put  $s1$  to F
    end
    delete  $s1$  from S
end

```

Algorithm 1: Gene set selection.

Generated patient-to-gene set matrix contains gene sets as a new feature vector, where each entry is an aggregating values of the altered genes in the

gene set.

2.3. Kernel Matrix Representing Molecular Information

The patient-to-gene set matrix of the four data sources are used to create kernels using kernel functions. A feature function, $\phi(\mathbf{x})$, maps a the original data feature \mathbf{x} in the input space to a high-dimensional feature space. A Kernel function is a function that corresponds to the inner product in a expanded feature space: $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle$. A kernel matrix is formed by computing kernel functions between all pairs of data. Thus, the size of a kernel matrix is independent of the number of features and are solely dependent on the number of data. In practice, an explicit definition of feature function, $\phi(\mathbf{x})$, is not needed since they are tightly integrated in the definition of the kernel functions.

The kernel functions we used are linear and radial basis function (RBF). Details of linear and RBF kernel are as follows: Let i^{th} and j^{th} sample data be represented as vectors of adjusted mutation scores of each gene sets: \mathbf{x}_i and \mathbf{x}_j . A linear kernel of two samples is a dot product of their original feature vectors, \mathbf{x}_i and \mathbf{x}_j :

$$k_{linear}(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle . \quad (1)$$

A RBF kernel of two samples vectors \mathbf{x}_i and \mathbf{x}_j is defined as follows:

$$k_{RBF}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2), \quad (2)$$

where $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ is the squared Euclidean distance between the two original feature vectors and parameter σ controls the flexibility of the kernel. With smaller value for the parameter σ , the kernel matrix becomes closer to identity matrix while risking overfitting. On the other hand, larger values of parameter gradually reduce the kernel to a constant function, making it impossible to learn any non-trivial classifier [33]. In our experiment, we use a separate validation data sets consisting of 25% of total samples to determine the parameters of the kernels, such as the value of σ in RBF kernels. The choice of the kernel for different data sources is decided based on the existing study in the literature.

We explored the use of commonly used kernels including linear, sigmoid, polynomial and the radial basis function. We chose the kernel function that showed the best performance for each data type. For mRNA, we select RBF kernel as in [34], the authors reports in their work that the use of RBF kernel proved to be more effective as compared to linear, polynomial or other kernels. For different combination of geneset an accuracy of 92.59 % was observed. The methylation data analysis [35], shows that the use of SVM classifier with RBF kernel outperforms the k-nearest neighbors classifier (k-NN) and a naive Bayes classifier with an accuracy of 91.3 %. In [36], the performance of RBF kernel is compared to the clinical kernel for the clinical data source. Out of the five case study carried out, it was observed that the RBF kernel outperforms clinical in three with an average accuracy of 78.59 %. In case of CNV, we apply linear kernel, similar to [37] which uses linear kernel for CNV data source for classification and attain an accuracy of 61%. For mutation data source, we apply RBF kernel, as in [38] presents a detailed comparison result for linear, polynomial and RBF kernel functions for breast cancer mutation data. It is reported that the use of RBF kernel for classification achieved a higher accuracy as compared to other kernels. For BRCA1-BRCA2 dataset, RBF attained an accuracy of 100% as compared to 93.3% (linear) and 86.6 % (polynomial).

2.4. Multiple Kernel Learning for Cancer Classification

The kernel matrix constructed from each data types is further integrated to form a single kernel matrix using a multiple kernel learning approach. Several methods are suggested for integrating the kernels [39]. We take a two step approach that first combines the kernels in a weighted linear fashion and then perform learning on the combined kernel. The kernel combination is defined as follows:

$$K_{\beta}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^S \beta_s k_s(\phi(\mathbf{x}_i^s), \phi(\mathbf{x}_j^s)) \quad (3)$$

subjected to $\beta_s \geq 0$ and $\sum_{s=1}^S \beta_s = 1$,

where S is the number of kernels, \mathbf{x}_i^s is the original feature vector of kernel s of sample i , and β_n is the kernel coefficient of kernel s .

To obtain optimal weights for kernel combination, we take the optimization approach suggested by Zien *et al.* [40]. In their approach, the kernel coefficient is determined by the efficacy of each of the kernel matrix containing sets learned by Least Square Support Vector Machine (LS-SVM). LS-SVMs are closely related to regularization networks and Gaussian processes but additionally emphasize and exploit primal-dual interpretations from the optimization theory [41]. The primal form of a LS-SVM is optimized by the following minimization problem:

$$\min_{\mathbf{w}, b, err} \left(\frac{1}{2} \mathbf{w}^T \mathbf{w} + \gamma \sum_{i=1}^N err_i^2 \right) \quad (4)$$

$$\text{subjected to } y_i[\mathbf{w}^T \phi(\mathbf{x}_i) + b] = 1 - err_i^2 \text{ for } i = 1, 2, \dots, N$$

where \mathbf{w} is the weight vector we are trying to learn, err_s is the error variables that represent the value corresponding to misclassification in case of overlapping distribution, and γ is the regularization parameter that tackles data over fitting problem.

The standard multiple kernel learning approach constructs the base kernels for each data type and determine their optimal kernel coefficient by solving Equation 5 [42, 43].

$$\min_{\beta, \mathbf{w}, b, err} \left(\frac{1}{2} \sum_{s=1}^S \beta \mathbf{w}^T \mathbf{w} + \gamma \sum_{i=1}^N err_i^2 \right) \quad (5)$$

$$\text{w.r.t. } w_k \in \mathbb{R}^{D_k}, err \in \mathbb{R}^N$$

$$\text{subjected to } \begin{cases} y_i \left(\sum_{s=1}^S \beta_s \mathbf{w}_s^T \phi_k(\mathbf{x}_i) + b \right) = 1 - err_i^2, \\ err_i \geq 0 \text{ for } i = 1, 2, \dots, N, \\ \sum_{s=1}^S \beta_s = 1, \beta_s \geq 0 \text{ for } s = 1, 2, \dots, S. \end{cases}$$

The derived dual for the problem in Equation 5 [44] is given as

$$\begin{aligned}
& \min \delta - \sum_{s=1}^S \alpha_i \\
& \delta \in \mathbb{R}, \alpha \in \mathbb{R}^N \\
& \text{subjected to } \begin{cases} 0 \leq \alpha \leq 1, \sum_{i=1}^S \alpha_i y_i = 0 \\ \frac{1}{2} \sum_{i,j=1}^S \alpha_i \alpha_j y_i y_j k_k(x_i, x_j) \leq \delta, \forall = 1, \dots, K \end{cases}
\end{aligned} \tag{6}$$

The standard multiple kernel learning approach constructs the base kernels for each data type, and determine their optimal kernel coefficient by solving Equation 2.4. Here, the optimization problem is solved using semi-defined linear programming. The dual for the problem is computed by considering the problem (D_k) , squaring the constraints δ , multiplying the constraints by $\frac{1}{2}$ and performing substitution as $\frac{1}{2}\delta^2 \mapsto \delta$ leads to dual form of multiple kernel learning Equation 6, here $k_k(x_i, x_j) = \langle \phi_k(x_i), \phi_k(x_j) \rangle$. This process uses transductive learning setting, where the kernel matrix is learned from data. Initially labeled training data is used to learn the good embedding, which is later applied to unlabeled test data. Considering semi-defined linear programming optimization using SVM enables to handle the optimization of convex cost functions and machine learning concerns, thus provides a powerful method for learning the kernel matrix [45].

2.5. Stratification Using Kernel K-means

Stratification of patients can be done with clustering methods. We use kernel K-means on the generated multiple kernel matrix for stratifying the Ovarian cancer to subtypes. The multiple kernel matrix contains the similarity information about pairs of data in the combined feature space. Thus, when we apply the kernel k-means to the multiple kernel matrix, data are clustered so that the clustering error is minimized in the combined feature space. The objective

function of kernel k-means is defined as follows:

$$D(\{\pi_c\}_{c=1}^k) = \sum_{c=1}^k \sum_{\mathbf{x}_i \in \pi_c} \|\phi(\mathbf{x}_i) - \mathbf{m}_c\|^2, \quad (7)$$

$$\text{where } \mathbf{m}_c = \frac{\sum_{\mathbf{x}_i \in \pi_c} \phi(\mathbf{x}_i)}{|\pi_c|},$$

where π_c denotes the clusters, $\{\pi_c\}_{c=1}^k$ denotes a partitioning of points, \mathbf{m}_c denotes the center of cluster π_c , and $|\pi_c|$ denote the size of the cluster π_c . The Euclidean distance between the data point, $\phi(\mathbf{x}_i)$, and the cluster center, \mathbf{m}_c , in the feature space is determined as follows [45]:

$$\|\phi(\mathbf{x}_i) - \mathbf{m}_c\|^2 = \phi(\mathbf{x}_i)\phi(\mathbf{x}_i) - \frac{\sum_{\mathbf{x}_j \in \pi_c} \phi(\mathbf{x}_i)\phi(\mathbf{x}_j)}{|\pi_c|} + \frac{\sum_{\mathbf{x}_j, \mathbf{x}_l \in \pi_c} \phi(\mathbf{x}_j)\phi(\mathbf{x}_l)}{|\pi_c|^2} \quad (8)$$

Here, the $\phi(x_i)\phi(x_j)$ is computed using appropriate selected kernel functions.

2.6. Clinical Feature Prediction Process

We also use the learned multiple kernel matrix for predicting clinical outcomes. For prediction, we again employ LS-SVM classifier. That is, using the multiple kernel as input, we run the LS-SVM to predict the survival time and the tumor grade of patients. The performance of the proposed model was evaluated on 312 samples in TCGA ovarian cancer data sets. Each of the sample contains sets of molecular data with matched clinical information. We split the 312 sample randomly so that 50% of the samples are assigned to the training set, 25% assigned to the validation set to learn the model parameters, and rest are assigned to the testing set to test the performance of the final model. For evaluations, we calculate the accuracy of survival prediction and area under the curve (AUC) of the receiver operating characteristic (ROC) curve for the tumor grade classification. The curves were constructed by plotting true positive rate (Sensitivity) in function of the false positive rate (100-Specificity) for different selected threshold for the tumor grade parameter. We selected the threshold range from 0.2 to 0.9. Each point on the ROC curve denotes a sensitivity/specificity pair corresponding to a selected decision threshold. The area

under the curve specify the ability of the test to correctly classify high grade and low grade tumor. The AUC value is computed by non-parametric method based on constructing trapeziods under the curve as an approximation of area.

3. Results

We report the results for validation and performance of combining the clinical features with the biological features by the multiple-kernel on two important translational bioinformatics tasks: patient stratification and clinical predictions.

3.1. Patient Stratification via *K*-means

We performed the kernel k-means clustering to stratify ovarian cancer patients using the generated kernel matrices. We compared four data type combinations as input to the k-mean clustering: the first multiple kernel is constructed from only the molecular data types listed in Table 1, the second is constructed from clinical information (i.e., age, stage, grade), the third is constructed by non-weighted linear combination of kernels of molecular as well as clinical data, and the fourth is construed by weighted linear combination of kernels of molecular and clinical data.

To evaluate the clustering result, we performed survival analysis on each clusters, or subgroups, using the Cox proportional hazards regression model in the R survival package [46] for each of the data type combinations. Out of 312 patient samples, the clustering was carried out for 75% (231) of the samples and 25% (81) to determine the number of clusters, k .

The value of k (i.e., the number of clusters) was determined using the log rank statistics. Figure 2 shows the different log rank statistic values obtained for different number of clusters. Figure 2 (A) shows the plot for integrated molecular data indicating the best value for k being 5. Figure 2 (B) is a graph for determining the k (i.e., 5) value for clinical data. Similarly, figure 2 (C) and figure 2 (D) are plots when molecular data is integrated with clinical without and with weighted kernel coefficient, results in best clusters for $k=5$ and $k=6$

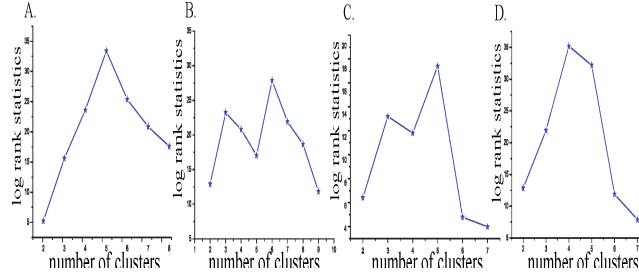


Figure 2: Log rank statistic to determine the number of clusters (A) Molecular data (B) Clinical data (C) Molecular and clinical data with non-weighted linear kernel coefficient (D) Molecular and clinical data with weighted kernel coefficient

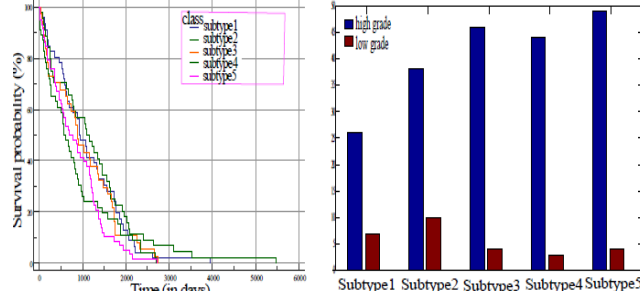


Figure 3: Kernel k-means clustering of the TCGA OV all molecular data reveals (a) five molecular subtypes (clusters) (b) tumor grade for each subtype.

respectively. We compared the survival times for these clusters using log-rank statistics and obtained the P-value. The P-value for all the above cases is less than 0.05. Thus, it shows that there exists a significant separation between the subgroups with respect to survival time.

Setting $k=5$ in kernel k-means clustering, the p-value of the subtype separation for survival analysis is 0.02 for all molecular data types (figure 3 (a)), 0.0079 for clinical data (figure 4 (a)), 0.009 for integrated molecular and clinical data with non-weighted kernel coefficient (figure 5 (a)), 0.0014 for integrated molecular and clinical data with weighted kernel coefficient (figure 6 (a)). It can be observed that the clusters identified by integrating the clinical data are more predictive with log-rank p-value of 1.4×10^{-3} . The size of the cluster formed is not uniform, however the method shows an ability to categorize the

Table 2: Patient stratification using molecular data

Cluster	Size	Avg. age	vital_status		neoplasm_cancer_status		
			Deceased	Living	with Tumor	Tumor Free	Missing
1	33	61.24	13 (39.39)	20 (60.61)	18 (54.54)	11 (33.33)	4 (12.12)
2	48	56.96	27 (56.25)	21 (43.75)	27 (56.25)	10 (20.83)	11 (22.92)
3	50	60.34	31 (62.0)	19 (38.0)	37 (74.0)	10 (20.0)	3 (6.0)
4	47	62.47	34 (72.34)	13 (27.66)	36 (76.59)	6 (12.77)	5 (10.64)
5	53	59.87	31 (58.49)	22 (41.51)	30 (56.60)	17 (32.08)	6 (11.32)

Table 3: Patient stratification using clinical data

Cluster	Size	Avg. age	vital_status		neoplasm_cancer_status		
			Deceased	Living	with Tumor	Tumor Free	Missing
1	50	49.28	30 (60.0)	20 (40.0)	30 (60.0)	14 (28.0)	6 (12.0)
2	64	71.22	38 (59.37)	26 (40.63)	39 (60.94)	17 (26.56)	8 (12.5)
3	45	68.24	27 (60.0)	18 (40.0)	30 (66.67)	9 (20.0)	6 (13.33)
4	72	52.61	51 (70.83)	31 (43.05)	49 (68.06)	14 (19.44)	9 (12.5)

Table 4: Patient stratification with linear kernel weights using molecular data and clinical data

Cluster	Size	Avg. age	vital_status		neoplasm_cancer_status		
			Deceased	Living	with Tumor	Tumor Free	Missing
1	46	55.07	0 (0.0)	46 (100)	15 (32.61)	25 (54.35)	6 (13.04)
2	30	59.23	0 (0.0)	30 (100)	12 (40.0)	15 (50.0)	3 (10.0)
3	108	60.74	108 (100)	0 (0.0)	91 (84.26)	4 (3.70)	13 (12.04)
4	24	63.29	24 (100)	0 (0.0)	22 (91.67)	0 (0.0)	2 (8.33)
5	23	64.87	4 (17.39)	19 (82.61)	8 (34.78)	10 (43.48)	5 (21.74)

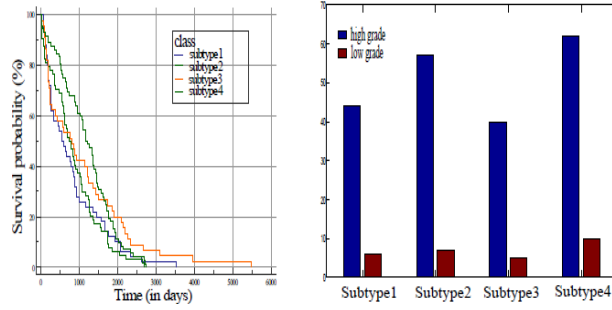


Figure 4: Kernel k-means clustering for clinical data

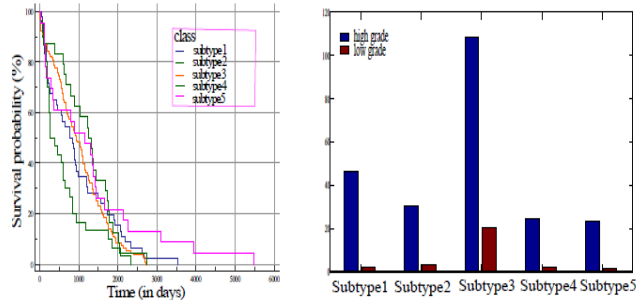


Figure 5: Kernel k-means clustering of the TCGA OV all molecular data with clinical data with linear kernel combination reveals (a) five molecular subtypes (clusters) (b) tumor grade for each subtype.

patient samples into sub groups that significantly differ in the survival time. In addition, to separating the patient according to survival time with significant statistics the subgroups are correlated to tumor grade.

In addition to mean survival time the clusters also show resembles in two other clinical features that are `vital_status` and `neoplasm_cancer_status`. The `vital_status` is categorized into two based on whether the patient current state is deceased or living. Similarly, the `neoplasm_cancer_status` is grouped into two patient samples with tumor or tumor free. Note that the missing field in `neoplasm_cancer_status` is indicating the unavailability of information.

It is observed that the combined molecular data and clinical data although are able to carry out clear distinction between clusters in terms of mean survival time but lack the distinction in terms of `vital_status` and `neoplasm_cancer_status`

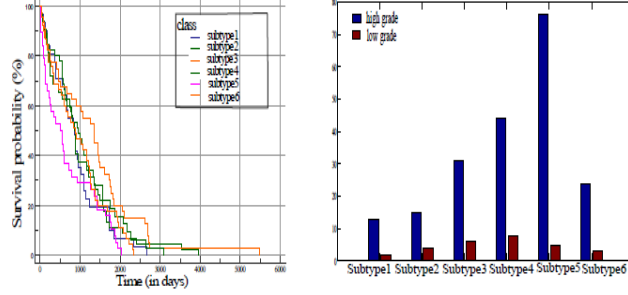


Figure 6: Kernel k-means clustering of the TCGA OV all molecular data with clinical data with optimized weights reveals (a) six molecular subtypes (clusters) (b) tumor grade for each subtype.

Table 5: Patient stratification with optimized kernel weights using molecular data and clinical data

Cluster	Size	Avg. age	vital_status		neoplasm_cancer_status		
			Deceased	Living	with Tumor	Tumor Free	Missing
1	15	59.53	1 (6.67)	14 (93.33)	4 (26.67)	10 (66.67)	1 (6.67)
2	19	63	19 (100)	0 (0.0)	17 (89.47)	0 (0.0)	2 (10.53)
3	37	67.38	37 (100)	0 (0.0)	33 (89.19)	0 (0.0)	4 (10.81)
4	52	56.60	52 (100)	0 (0.0)	41 (78.85)	3 (5.76)	8 (15.38)
5	81	58.09	0 (0.0)	81 (100)	29 (35.80)	40 (49.38)	12 (14.81)
6	27	61.11	27 (100)	0 (0.0)	24 (88.89)	1 (3.70)	2 (7.41)

summarize in Table 2 and Table 3 respectively. On the other hand, the results obtained when the clinical data is combined with molecular data are different as reported in Table 4 and Table 5. Table 4 shows results for linear combination of kernel matrix. It is observed that a distinct stratification can be obtained for the vital_status, but not in case of neoplasm_cancer_status. Table 5 reports the results with optimized kernel coefficient used to form the kernel matrix. It is observed that the stratification of the patients are more clear for both vital_status and neoplasm_cancer_status.

3.2. Clinical Outcome Prediction using Molecular Data

The observations from patient stratification section motivates us to consider integrated molecular data, clinical data and their combinations. The model is evaluated on a set of dichotomized overall prediction using a LS-SVM. We carry out prediction for two characteristic feature survival risk and tumor grade. The survival risk is divided into two based on high risk and low risk periods. The high risk is for which the survival time is lower than median survival time, whereas, low risk are once with value higher than median. For the selected samples from TCGA data the median survival time is set to 998 days. We also perform prediction on high and low grade tumor using molecular data, clinical data and their combinations. The low grade contains samples with tumor grade of type G1 or G2 where as, high grade contains samples corresponding to type G3 and G4 [47] .

Figure 7 shows the performance behavior of the model summary of the prediction when molecular data are considered in isolation and when all molecular data and clinical features are integrated. It is observed that a high AUC values of 0.8217, 0.8449, 0.8538, 0.8718 and 0.8937 are obtained for CNA, methylation, clinical, non-weighted integrated combination and weighted integrated data respectively. These observations also help us to infer some biological information. For the patient samples, in which the changes in tumor samples are due to the structural variation in the chromosome like a copy number variation or methylation seems to have a slightly higher influence on high or low grade clinical predictions. The tumor samples with functional changes like mutation and mRNA also directly relate and contribute towards the classification of clinical outcome. For the developed model, methylation performed better in comparison to other molecular data. Overall, the integration of the different data types improved the prediction accuracy.

The results for high risk and low risk survival are summarized in Table 6 and Table 7 with accuracy as the performance measure. The notation used in these tables are TP stands for true positive, FP stands for false positive, TN is true negative, FN is false negative, Spec. means specificity and Sens. means

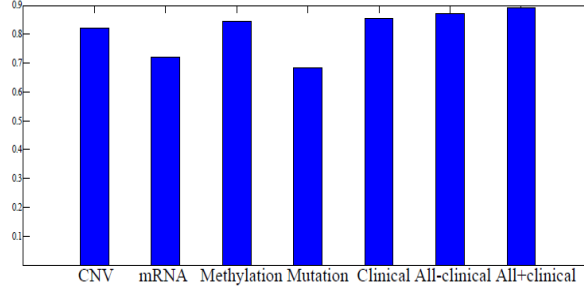


Figure 7: Area under curve for high vs low grade classification of OV

Table 6: Patient survival risk prediction with individual data types.

Data types	TP	FP	TN	FN	Spec.	Sens.	Accuracy
CNA	0.6904	0.3095	0.7368	0.2632	74.36	68.29	71.25
Mutation	0.6667	0.3333	0.7105	0.2895	71.79	65.85	68.75
Methylation	0.6905	0.3095	0.7105	0.2895	72.5	67.5	70
mRNA	0.6428	0.3571	0.6842	0.3158	69.23	63.41	66.25
Clinical	0.7143	0.2857	0.73684	0.2632	75.0	70.0	72.5
CNA+Mutation+mRNA Methylation	0.7381	0.2619	0.7368	0.2632	75.61	71.79	73.75

Table 7: Patient survival risk prediction with individual data types and clinical.

Data types	TP	FP	TN	FN	Spec.	Sens.	Accuracy
CNA+clinical	0.7143	0.2857	0.7368	0.2632	75	70	72.5
Mutation+clinical	0.6667	0.3333	0.7368	0.2632	73.68	66.67	70
Methylation+clinical	0.7143	0.2857	0.7105	0.2895	73.17	69.23	71.25
mRNA+clinical	0.6429	0.3571	0.7368	0.2632	72.97	65.12	68.75
CNA+Mutation+Methylation +mRNA+Clinical	0.7619	0.2381	0.7632	0.2368	78.05	74.36	76.25

sensitivity. The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. Table 6 reports the results for molecular data and their integration. It summarizes the behavior of individual data set in prediction accuracy. Amongst the individual molecular data CNA is able to predict low risk and high risk patient for nearly 71% of the

samples. Next, in order to determine the behavior of data type when integrated with clinical data experimentation were carried out, the results are reported in Table 7. The results show a overall increase in accuracy by integration: for the low risk vs. high risk survival classification. These findings are useful as they suggest that some biological information may be fused to various data sources from different genomic levels. Thus, integration of these independent data types increases the chances of success in cancer outcome predictions.

4. Conclusion

In this paper, we have developed a multiple kernel learning based pipeline for integrative analysis of heterogenous data types and apply it on ovarian cancer data. The data types we look at are molecular data and clinical data. The model is used to carry out patient stratification and clinical outcome prediction. We use kernel k-means to perform stratification of patient for survival time, vital_status and neoplasm_cancer_status. Stratification is done considering different test cases including integrated molecular data, clinical data, integration using linear non-weighted combination and integration using weighted kernel coefficient combination. The patient stratification results for different test cases show that the integration of molecular and clinical data result in better pattern forming relation. The clinical outcome prediction is done for tumor grade and survival risk. In case of tumor grade, a better AUC of 0.8937 was achieved for weighted kernel combination in comparison to 0.8538 considering only clinical data. For survival risk prediction it was observed that when molecular data are integrated with clinical data the overall prediction of the system is improved. This work concludes that integration of molecular data along with clinical data not only helps in carrying out better patient stratification but also improves the prediction accuracy of the model.

5. Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2015R1C1A2A01055739) and by the KEIT (Korea Institute for Industrial Economics and Trade), Korea, under the "Global Advanced Technology Center" (10053204).

References

References

- [1] B. Aunoble, R. Sanches, E. Didier, et al, Major oncogenes and tumor suppressor genes involved in epithelial ovarian cancer (review)., *Int J Oncol.* 16 (2000) 567–576.
- [2] L. A. G. Ries, J. L. Young, G. E. Kee, M. P. Eisner, Y. D. Lin, M. Horner, SEER Survival Monograph Cancer Survival Among Adults U.S. SEER Program Patient and Tumor Characteristics., SEER Program, NIH Pub. No. 07-6215, National Cancer Institute, Bethesda, MD, 2007.
- [3] C. E. Board", Ovarian Cancer: Statistics, <http://www.cancer.net/cancer-types/ovarian-cancer/statistics>.
- [4] T. The Cancer Genome Atlas Research Network, Integrated genomic analyses of ovarian carcinoma., *Nature* 474(7353) (2011) 609–615.
- [5] Y. Wang, J. G. Klijn, Y. Zhang, A. Sieuwerts, M. Look, F. Yang, D. Talandov, M. Timmermans, M. E. M. van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, J. A. Foekens, Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer., *Lancet.* 365 (2002) 671–679.
- [6] A. E. Teschendorff, A. Naderi, N. L. Barbosa-Morais, S. E. Pinder, I. O. Ellis, et al., A consensus prognostic gene expression classifier for er positive breast cancer., *Genome Biology* 7 (2006) R101.

- [7] Y. Zhang, J. W. Martens, J. X. Yu, J. Jiang, A. M. Sieuwerts, et al., Copy number alterations that predict metastatic capability of human breast cancer., *Cancer res.* 69 (2009) 3795–3801.
- [8] S. Deneberg, M. Grovdal, M. Karimi, M. Jansson, H. Nahi, et al., Gene-specific and global methylation patterns predict outcome in patients with acute myeloid leukemia., *Leukemia* 24 (2010) 932–941.
- [9] V. S. Nair, L. S. Maeda, J. P. Ioannidis, Clinical outcome prediction by mi-crnas in human cancer: a systematic review., *J Nat Cancer Institute* 104 (2012) 528–540.
- [10] S. Kim, L. Sael, H. Yu, Identifying cancer subtypes based on somatic mutation profile, in: *Proceedings of the ACM 8th International Workshop on Data and Text Mining in Bioinformatics, DTMBIO '14*, ACM, New York, NY, USA, 2014, pp. 19–22. doi:10.1145/2665970.2665980. URL <http://doi.acm.org/10.1145/2665970.2665980>
- [11] S. Kim, L. Sael, H. Yu, A mutation profile for top- k patient search exploiting Gene-Ontology and orthogonal non-negative matrix factorization, *Bioinformatics* (2015) btv409.
- [12] M. Hofree, J. Shen, H. Carter, A. Gross, T. Ideker, Network-based stratification of tumor mutations., *Nat Methods* 10(11) (2013) 1108–1115.
- [13] C. Wang, R. Machiraju, K. Huang, Breast cancer patient stratification using a molecular regularized consensus clustering method., *Methods* 67(3) (2014) 304–312.
- [14] J. Thomas, L. Sael, Overview of integrative analysis methods for heterogeneous data, in: *The 2015 International Conference on Big Data and Smart Computing (BigComp 2015)*, 2015, pp. 266–270.
- [15] R. Louhimo, S. Hautaniemi, Cnamet: an r package for integrating copy number, methylation and expression data., *Bioinformatics* 27 (2011) 887–8.

- [16] D. Kim, H. Shin, Y. S. Song, J. H. Kim, Synergistic effect of different levels of genomic data for cancer clinical outcome prediction., *J Biomed Inform* 45 (2012) 1191–1189.
- [17] K. A. Sohn, D. Kim, J. Lim, J. H. Kim, Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors., *BMC systems biology* 7 (2013) S9.
- [18] M. Schafer, H. Schwender, S. Merk, C. Haferlach, K. Ickstadt, M. Dugas, Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities., *Bioinformatics* 25 (2009) 3228–3235.
- [19] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine, C. Sander, Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles., *PLoS One* 6 (2011) e24709.
- [20] Y. Yuan, E. M. V. Allen, L. Omberg, et al., Assessing the clinical utility of cancer genomic and proteomic data across tumor types., *Nature Biotechnology* 32 (2014) 644–652.
- [21] S. S. Bucak, R. Jin, A. K. Jain, Multiple kernel learning for visual object recognition: A review., *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(7) (2014) 1354–1369.
- [22] TCGA, The Cancer Genome Atlas, <http://tcga-data.nci.nih.gov/>.
- [23] Y. Zhu, P. Qiu, Y. Ji, Tcga-assembler: Open-source software for retrieving and processing tcga data., *Nature Methods* 11(6) (2008) 599–600. doi: 10.1038/nmeth.2956.
- [24] B. I. T. G. D. A. Center”, Firehose Broad GDAC, <http://gdac.broadinstitute.org/>.
- [25] Y. Liu, Y. Ji, P. Qiu, Identification of thresholds for dichotomizing dna methylation data., *EURASIP J Bioinform Syst Biol.* 2013(1): 8. doi: 10.1186/1687-4153-2013-8.

- [26] C. D. Warden, H. Lee, J. D. Tompkins, X. Li, C. Wang, A. D. Riggs, H. Yu, R. Jove, Y. C. Yuan, Cohcap: an integrative genomic pipeline for single-nucleotide resolution dna methylation analysis., *Nucleic Acids Res.* 41(11) (2013) e117. doi:10.1093/nar/gkt242.
- [27] R. Beroukhi, G. Getz, L. Nghiemphu, et al., Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma, in: *Proceedings of the National Academy of Sciences of the United States of America*, 2007, pp. 20007–20012.
- [28] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles., *Proc Natl Acad Sci USA* 102 (2005) 15545–15550.
- [29] M. Kanehisa, S. Goto, Kegg: Kyoto encyclopedia of genes and genomes., *Nucl. Acids Res.* 28(1) (2000) 27–30.
- [30] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, J. P. Mesirov, Molecular signatures database (msigdb) 3.0., *Bioinformatics* 27(12) (2011) 1739–1740. doi:10.1093/bioinformatics/btr260.
- [31] D. Nishimura, The view from web biocarta., *Biotech Software and Internet Report* 2(3) (2001) 117–120.
- [32] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, P. D’Eustachio, The reactome pathway knowledgebase., *Nucl. Acids Res.* 42(D1) (2014) D472–D477.
- [33] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*,

Cambridge university press, The Edinburgh Building, Cambridge, UK, 2004.

- [34] A. L. V. Gomes, L. J. K. Wee, A. M. Khan, L. H. V. G. Gil, E. T. A. Marques, C. E. Calzavara-Silva, T. W. Tan, Classification of dengue fever patients based on gene expression data using support vector machines., *PLoS One* 5(6) (2010) e11267. doi:doi.org/10.1371/journal.pone.0011267.
- [35] W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell, B. E. Engelhardt, Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements., *Genome Biol.* 16(1) (2015) 14. doi:10.1186/s13059-015-0581-9.
- [36] M. Thomas, K. D. Brabanter, J. A. Suykens, B. D. Moor, Predicting breast cancer using an expression values weighted clinical classifier., *BMC Bioinformatics* 15:411. doi:10.1186/s12859-014-0411-1.
- [37] J. A. Seoane, I. N. M. Day, T. R. Gaunt, C. Campbell, A pathway-based data integration framework for prediction of disease progression., *Bioinformatics* 30(6) (2014) 838–845. doi:10.1093/bioinformatics/btt610.
- [38] M. Pirooznia, Y. Deng, Svm classifier – a comprehensive java interface for support vector machine classification of microarray data., *BMC Bioinformatics* 7(Suppl 4) (2006) S25. doi:10.1186/1471-2105-7-S4-S25.
- [39] M. Gonen, E. Alpaydm, Multiple kernel learning algorithms (2011).
- [40] A. Zien, C. S. Ong, Multiclass multiple kernel learning, in: *Proceedings of the 24th Int. Conf. Mach. Learn.*, Corvallis, OR, 2007, pp. 1191–1198.
- [41] J. K. Suykens, J. Vandewalle, Least squares support vector machine classifiers., *Neural Processing Lett* 9 (1999) 293–300.
- [42] S. Yu, L.-C. Tranchevent, X. Liu, W. Glanzel, J. A. Suykens, B. D. Moor, Y. Moreau, Optimized data fusion for kernel k-means clustering., *IEEE*

Transactions on Pattern Analysis and Machine Intelligence 34(5) (2012)
1031–1039.

- [43] Y. R. Yeh, T. C. Lin, Y. Y. Chung, Y. C. F. Wang, A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection., IEEE Transactions on Multimedia 14 (2012) 563–574.
- [44] F. R. Bach, G. R. G. Lanckriet, M. I. Jordan, Multiple kernel learning, conic duality, and the SMO Algorithm, in: ICML '04 Proceedings of the twenty-first international conference on Machine learning, 2004, p. 6.
- [45] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, M. I. Jordan, Learning the kernel matrix with semi-definite programming, The Journal of Machine Learning Research 5 (2004) 27–72.
- [46] T. M. Therneau, A Package for Survival Analysis in S, r package version 2.37-7 (2014).
URL <http://CRAN.R-project.org/package=survival>
- [47] National Cancer Institute, Tumor grade, <http://www.cancer.gov/cancertopics/factsheet/Detection/tumor-grade>.