A $\ell_{2,1}$ norm regularized multi-kernel learning for false positive reduction in Lung nodule CAD

Peng Cao^{a,b,*}, Xiaoli Liu^{a,b}, Jian Zhang^c, Wei Li^b, Dazhe Zhao^{a,b}, Min Huang^d, Osmar Zaiane^e

^aComputer Science and Engineering, Northeastern University, Shenyang, China

^bKey Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, Shenyang, China

^cSchool of Computer & Software, Nanjing University of Information Science & Technology, Nanjing, China

^dInformation Science and Engineering, Northeastern University, Shenyang, China

^eComputing Science, University of Alberta, Edmonton, Alberta, Canada

Abstract

Objective: The aim of this paper is to describe a novel algorithm for False Positive Reduction in lung nodule Computer Aided Detection(CAD).

Methods: In this paper, we describes a new CT lung CAD method which aims to detect solid nodules. Specially, we proposed a multi-kernel classifier with a $\ell_{2,1}$ norm regularizer for heterogeneous feature fusion and selection from the feature subset level, and designed two efficient strategies to optimize the parameters of kernel weights in non-smooth $\ell_{2,1}$ regularized multiple kernel learning algorithm. The first optimization algorithm adapts a proximal gradient method for solving the $\ell_{2,1}$ norm of kernel weights, and use an accelerated method based on FISTA; the second one employs an iterative scheme based on an approximate gradient descent method.

Results: The results demonstrates that the FISTA-style accelerated proximal descent method is efficient for the $\ell_{2,1}$ norm formulation of multiple kernel learning with the theoretical guarantee of the convergence rate. Moreover, the experimental results demonstrate the effectiveness of the proposed methods in terms of Geometric mean(G-mean) and Area under the ROC curve (AUC), and significantly outperforms the competing methods.

Conclusions: The proposed approach exhibits some remarkable advantages both in heterogeneous feature subsets fusion and classification phases. Compared with the fusion strategies of feature-level and decision level, the proposed $\ell_{2,1}$ norm multi-kernel learning algorithm is able to accurately fuse the complementary and heterogeneous feature sets, and automatically prune the irrelevant and redundant feature subsets to form a more discriminative feature set, leading a promising classification performance. Moreover, the proposed algorithm consis-

*Corresponding author.

Email address: caopeng@cse.neu.edu.cn (Peng Cao)

tently outperforms the comparable classification approaches in the literature.

Keywords: Lung nodule detection, false positive reduction, classification, heterogeneous feature fusion, multi-kernel learning

1. Introduction

Lung cancer is one of the main public health issues in developed countries, and early detection of pulmonary nodules is an important clinical indication for early-stage lung cancer diagnosis [1]. Lung nodule refers to lung tissue abnormalities that are roughly spherical with round opacity and a diameter of up to 30 mm. Currently, nodules are mainly detected by one or multiple expert radiologists inspecting CT images of lungs. Recent research, however, shows that inter-reader variability in the detection of nodules by expert radiologists may exist. In addition, since three-dimensional (3D) image processing and analysis techniques become applicable in thin-section CT[3, 4], a thin-section CT scan includes hundreds of sections and requires considerable time and effort in image interpretation by radiologists. For more than a decade, significant effort has been focused on developing automated systems that detect/recognize suspicious lesions in thoracic CT imagery as well as other types of imagery.

It is therefore an important task to develop Computer Aided Detection (CAD) systems that can aid/enhance radiologist workflow and potentially reduce false negative findings. CAD is a scheme that automatically detects suspicious lesions (*i.e.* nodule, polyps and masses) in medical images of certain body part[5, 10, 11], and provides their locations to radiologists. It is because detection algorithms have high sensitivity that some non-nodule structures (e.g., blood vessels) are labeled as nodules inevitably in the initial nodule identification step. Since the radiologists must examine each identified object, it is highly desirable to eliminate these false positives (FPs) as much as possible while retaining the true positives (TPs). Therefore, significant effort is needed in order to improve the performance levels of current CAD schemes for nodule detection in thin-section CT [2].

The purpose of false positive reduction is to remove these false positives as much as possible while retaining a relatively high sensitivity. It is a binary classification between the nodule and non-nodule, aiming to learn a system capable of the prediction of the unknown output class of a previously unseen suspicious nodule with a good generalization ability, is a critical part in the Lung CAD system [24, 26, 25, 21, 68, 36]. Classification plays an important role in the reduction of false positives in lung computer aided detection and diagnosis methods. Regarding classification of suspicious lung nodule, a number of techniques have been presented using machine learning approaches to classify samples as true positive nodule and false nodule. One of the most frequently employed and simplest classifier is the rule-based classifier [20]; however it is hard to manually determine the selection of cut-off threshold to classify abnormal and normal. Since Linear discriminant analysis (LDA) offers simplicity in computation and effectiveness in classification, it is commonly used to discriminate

the potential nodule [21, 22]. Neural networks are often employed in nodule recognition tasks as nonlinear model[23, 25], which have the ability to automatically learn complex input-output relationships, and have low dependence on domain specific knowledge. However, in neural networks, it was difficult to determine the number of units in the hidden layer, and its gradient-based algorithm might be trapped in local minimal. In addition to neural networks, SVM[26, 27, 28, 29, 30, 31] is the most commonly employed for reduction of false positives. Compared with neural networks, SVM tends to find a global solution during the training as the model complexity has been taken into consideration as a structural risk in SVM training, hence SVM has a strong generalization classification ability. SVM utilizes nonlinear mapping to make the data linear separable, hence the kernel function is the key[35], which determines the mapping between the input space and the feature space. One possible way to solve this problem is to use cross-validation to select the optimal kernel, but it suffers from being time consuming and are easily overfitted.

In addition to the disadvantages of the popular classification methods, the use of heterogeneous features becomes more practical for suspicious nodules classification. As we known, the variance in the volumes, shapes, and appearances of the suspicious nodule objects is enormous. In order to achieve satisfactory results in nodule candidate classification, multiple heterogeneous feature subsets are extracted to describe the visual objects[36, 66, 26, 37]. Each representation of feature subset describes different aspects of the visual characteristics of nodule candidates, a joint analysis of these feature subsets can potentially exploit their complementary information and improve the prediction. A common practice is the concatenation of all feature subsets into a longer feature vector. However, this may be not enough for effective combination of features from different views. The disadvantage of this method is that it treats multiple features equally, being incapable of effectively exploring the complementary information. Thus, how to combining the multiple complementary feature subsets effectively based on different aspects is a critical challenge.

Indeed, there have been numerous reports on various ways of combining multi-modality data for efficient classification. For example, the dictionary learning for sparse representation based classification methods [38],[39] have been studied intensively for fusing multi-modality or heterogeneous feature subsets in recent years. Xu et al. extended multi-modality framework based on SRC (sparse representation-based classification), named as weighted multi-modality SRC (wmSRC), to combine multi-modality data (MRI,FDG-PET and florbetapir PET) for classification of Alzheimer’s disease[40]. SRC uses sparse coding to denote a test sample by a linear combination of some atoms from a given dictionary. Based on SRC, Xu et al. introduced weight parameters, which can be optimized using a grid search approach to fuse the multiple modalities. Taalimi et al. proposed a multi-modal (two modalities of SIFT and SURF from cell images) joint dictionary learning method, to obtain a discriminative and reconstructive dictionary while training a classifier simultaneously in the detection of Antinuclear Autoantibodies (ANAs)[41]. Based on the traditional dictionary learning, an $\ell_{2,1}$ regularization is introduced to allow the sparse

representations share the consistent pattern across all the modalities of the same class. Wang et al. proposed two different strategies to learn a fusion matrix for training face images with multiple features based on dictionary learning framework [42]. The first one is to learn a fusion matrix based on Fisher criterion from the training data to fuse the different features; The second strategy is to learn the fusion matrix and the core dictionary simultaneously, and use MMC to refine the fusion matrix. Most of those studies focused on modeling the classifiers in the original input space, and assumed linear relationship between the features and the class label. Unfortunately, the complex patterns in the inherent structure of suspicious nodule data are not able to be modeled in the input space very well. Many previous works have demonstrated that the class distribution of suspicious nodule data is more discriminative in the appropriate kernel space than the original input space[26][43].

In this work, we address the limitation of existing classification algorithms in handling heterogeneous features, and proposed a novel $\ell_{2,1}$ regularized multiple kernel learning method to accurately classify potential nodule objects. Multiple Kernel Learning (MKL) aims to seek a better result than single kernel learning by combining a compact set of sub-kernels[13]. MKL not only learns a optimal linear combination of given base kernels, but also can be used to combine heterogeneous sources of data. They are used as a framework aiming at learning the correct weighting of different features from training data. MKL have been successfully applied to different applications in computer vision and recently in medical imaging, such as for detection and Gleason grading of prostate cancer [14], gene prioritization [15] and Alzheimer’s disease diagnosis[16]. In order to effectively solve the specific data characteristic, we utilize multiple kernel learning with imposing the mixed $\ell_{2,1}$ norm constraint on the kernel weights, explicitly incorporating auxiliary information about structures in parameters, so as to exploit complementary heterogeneous feature subsets with different kernels and recognize the contribution of classification of different multiple feature subset at the same time. Moreover, we propose two efficient strategies to optimize the parameters of kernel weights in non-smooth $\ell_{2,1}$ regularized multiple kernel learning algorithm. The first optimization algorithm adapts a proximal gradient method for solving the $\ell_{2,1}$ norm of kernel weights, and use an accelerated method based on FISTA; the second one employs an iterative scheme based on an approximate gradient descent method. To perform a validation with our system, we empirically investigate and compare the proposed methods with the state-of-the-art approaches in the multi-source fusion, multiple kernel classification, well known linear and nonlinear classification and the false positive reduction. The experimental results show the unique properties of the proposed methods for overcoming the challenges in the lung nodule CAD and demonstrate the promising effectiveness of this method compared with other competitive learning algorithms.

The structure of the paper is as follows: in Section 2 we present the multiple kernel learning. In Section 3, we introduce the proposed method $\ell_{2,1}$ MKL and present the optimization procedure. In Section 4, the experimental results on LIDC dataset, a private lung dataset and public breast datasets are given, which

show excellent performance of the proposed method in comparison with other competitive methods. Finally, Section 5 concludes the paper.

2. Multiple kernel learning

Let $\mathbf{x}_i$ be a training instance, and the training set is denoted as $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N) \in \mathbb{R}^{P \times N}$, where N and P are the number of training instances and dimensionality of $\mathbf{X}$, respectively. Let $\mathbf{y} = (y_1, \dots, y_N)^T$ be the label vector, with $y_i \in \{+1, -1\}$ being the class label of the i -th training instance.

Support Vector Machines (SVM) can model the nonlinear distribution of the data by mapping the input data into a nonlinear feature space by kernel embedding[17, 18, 19]. The primal objective function of SVM is:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y_i (\mathbf{w}^T \mathbf{x}_i + b) + \xi_i \geq 1, \quad \xi_i \geq 0, \quad \forall i. \end{aligned}$$

where $\mathbf{w}$ is the normal of the separating hyperplane, b is the bias term, $\boldsymbol{\xi} = [\xi_1, \dots, \xi_N]$ is the vector of slack variables, and C is the regularization parameter. The optimal C can be obtained by cross-validation.

After optimization, SVM learned weighting $\alpha_i \in \mathbb{R}^N$ on training example in kernel feature space $\phi(\mathbf{x}_i)$, where $\phi : \mathcal{X} \rightarrow \mathcal{H}$ defines a (possible nonlinear) feature mapping from the original input space to a Hilbert space $\mathcal{H}$. The decision function can then be represented by

$$f(x) = \text{sign} \left(\sum_{i=1}^N y_i \alpha_i^* \mathbf{K}(\mathbf{x}, \mathbf{x}_i) + b^* \right) \quad (1)$$

where kernel function (kernel matrix) $\mathbf{K}(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}')$ is to assess the similarity of input samples, parameter $\boldsymbol{\alpha}^*$ and b^* are obtained by solving the dual of the optimization in (1).

SVM can model the nonlinear distribution of the data by mapping the input data into a nonlinear feature space by kernel embedding. However, the most suitable types and parameters of the kernels for a particular task is often unknown, and the selection of the optimal kernel by exhaustive search on a pre-defined pool of kernels is usually time-consuming, and sometimes causes over-fitting. Multiple kernel learning (MKL)[13], which learns the optimal kernel by a weighted, linear combination of predefined candidate kernels, has been introduced to handle the problem of kernel selection. Compared to a single kernel, such as SVMs, MKL attempts to achieve better results by combining several base kernels instead of using only one specific kernel. MKL assumes that $\mathbf{x}_i$ can be mapped to Q different Hilbert spaces, $\mathbf{x}_i \rightarrow \phi_q(\mathbf{x}_i)$, $q = 1, \dots, Q$, implicitly with Q nonlinear mapping functions, and the objective of MKL is to seek the optimal kernel combination $\tilde{\mathbf{K}}(x, x') = \sum_{q=1}^Q \mathbf{d}_q \mathbf{K}_q(x, x')$, $\mathbf{d}_q \geq 0$, $\sum_{q=1}^Q \mathbf{d}_q = 1$.

We obtain the primal objective function of multiple kernel learning:

$$\begin{aligned}
\min_{\mathbf{w}, \mathbf{d}, b, \xi} \quad & \frac{1}{2} \sum_{q=1}^Q \|\mathbf{w}_q\|^2 + C \sum_{i=1}^N \xi_i \\
\text{s.t.} \quad & y_i \left(\sum_{q=1}^Q \langle \mathbf{w}_q, \sqrt{\mathbf{d}_q} \phi_q(\mathbf{x}_i) \rangle + b \right) \geq 1 - \xi_i, \quad i = 1, \dots, N \\
& \sum_{q=1}^Q \mathbf{d}_q = 1, \mathbf{d}_q \geq 0, q = 1, \dots, Q
\end{aligned} \tag{2}$$

Since equation (2) is non-convex due to the products of $\mathbf{d}_q$ and $\mathbf{w}_q$, it can be resolved by applying the variable transformation $\mathbf{w}'_q = \sqrt{\mathbf{d}_q} \mathbf{w}_q$ as that in [44] and [45].

$$\begin{aligned}
\min_{\mathbf{w}', \mathbf{d}, b, \xi} \quad & \frac{1}{2} \sum_{q=1}^Q \frac{\|\mathbf{w}'_q\|^2}{\mathbf{d}_q} + C \sum_{i=1}^N \xi_i \\
\text{s.t.} \quad & y_i \left(\sum_{q=1}^Q \langle \mathbf{w}'_q, \phi_q(\mathbf{x}_i) \rangle + b \right) \geq 1 - \xi_i, \quad i = 1, \dots, n \\
& \sum_{q=1}^Q \mathbf{d}_q = 1, \mathbf{d}_q \geq 0, q = 1, \dots, Q
\end{aligned} \tag{3}$$

In order to obtain the best classification accuracy, MKL learns both the weights of the kernel combination and the parameters of the classifier by solving a single joint optimization problem. Accordingly, the decision function can be reformulated as:

$$f(x) = \text{sign} \left(\sum_{i=1}^N \alpha_i^* y_i \sum_{q=1}^Q \mathbf{d}_q^* \mathbf{K}_q(x, x_i) + b^* \right) \tag{4}$$

where $\mathbf{K}_q(\mathbf{x}, \mathbf{x}') = \phi_q(\mathbf{x}) \cdot \phi_q(\mathbf{x}')$, $\sum_{q=1}^Q \mathbf{d}_q^* = 1$

3. Proposed Method: $\ell_{2,1}$ norm regularized multiple kernel learning for heterogeneous feature subsets fusion

3.1. $\ell_{2,1}$ norm multiple kernel learning

In this section, we describe the proposed approach to fuse and select the representations or feature subset using the multiple kernel learning framework regularized by $\ell_{2,1}$ norm. An overview of the proposed methods is illustrated in Figure 1. From the preprocessed CT images, we first extract G heterogeneous feature subsets $\{\text{FeaSet}_1, \dots, \text{FeaSet}_G\}$ from the view of intensity, texture,

shape and gradient. Then, the nodule data with heterogeneous feature subsets are fused and selected by the $\ell_{2,1}$ MKL. Finally, the resulting kernel is used through SVM with the optimal kernel matrix $\mathbf{K}^*$ on the reduced feature sets to solve the classification problem.

Figure 1: The procedure of our proposed method for false positive reduction

In [47], a base kernel is computed for each feature subset (representation), which may result in one important feature subset then removed if the kernel parameter is not set appropriately. In order to improve over it, there is a total of $G \times Q$ base kernels in our $\ell_{2,1}$ norm regularized MKL, G groups for G feature subsets and for each group (feature subset), we use Q kernel functions to produce Q embeddings as shown in Figure. 2. The $\ell_{2,1}$ norm is proposed in [46] to simultaneously estimate parameters and select factors with model, and commonly used to solve the problem of multi-task learning, where it can couple feature selection across tasks under the assumption that there lies a common low dimensional representation among different samples.

With these $G \times Q$ kernels, the corresponding coefficients are: $\mathbf{d} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_G]^T$, where $\mathbf{d}_g = [d_{g1}, d_{g2}, \dots, d_{gQ}]$. The $\ell_{2,1}$ norm of matrix $\mathbf{d}$ is defined as:

$$\|\mathbf{d}\|_{2,1} = \sum_{g=1}^G \sqrt{\sum_{q=1}^Q d_{gq}^2} = \sum_{g=1}^G \|\mathbf{d}_g\|_2 \quad (5)$$

Because $\ell_{2,1}$ enforces row sparsity on $\mathbf{d}$, the g -th row of $\mathbf{d}$ shrinks to zero if the corresponding feature subset is irrelevant to the output. In practice, the feature subset whose norm proportion is just less than a given threshold ratio (10%) is removed. The mixed $\ell_{2,1}$ constraint imposed on $\mathbf{d}$ will maintain sparsity between different group (feature subset) so as to generate a compact set of feature subset, while the associated values in each group need not be sparse, allowing our MKL algorithm to select more than one base kernel for each heterogeneous feature subset.

Figure 2: The $\ell_{2,1}$ norm of kernel weights in Multi-kernel learning for heterogeneous feature fusion

The objective function of $\ell_{2,1}$ norm regularized MKL is:

$$\begin{aligned}
 \min_{\mathbf{w}', \mathbf{d}, b, \xi} \quad & \frac{1}{2} \sum_{g=1, q=1}^{G, Q} \frac{\|\mathbf{w}'_{gq}\|^2}{d_{gq}} + C \sum_{i=1}^N \xi_i + C_\gamma \|\mathbf{d}\|_{2,1} \\
 \text{s.t.} \quad & y_i \left(\sum_{g=1, q=1}^{G, Q} \langle \mathbf{w}'_{gq}, \phi_{gq}(\mathbf{x}_i^{(g)}) \rangle + b \right) \geq 1 - \xi_i, \quad i = 1, \dots, N \quad (6) \\
 & d_{gq} \geq 0, \quad \forall g, q; \quad \xi_i \geq 0, \quad \forall i
 \end{aligned}$$

where $\mathbf{x}_i^{(g)}$ indexes the g -th feature set view of the instance of $\mathbf{x}_i$, C and C_γ are regularization parameters.

The primal formulation (6) can be seen as the following composite objective optimization problem:

$$\min_{\mathbf{d} \geq 0} J(\mathbf{d}) + C_\gamma \|\mathbf{d}\|_{2,1} \quad (7)$$

where

$$J(\mathbf{d}) = \begin{cases} \min_{\mathbf{w}', b, \xi} & \frac{1}{2} \sum_{g=1, q=1}^{G, Q} \frac{\|\mathbf{w}'_{gq}\|^2}{d_{gq}} + C \sum_{i=1}^N \xi_i \\ \text{s.t.} & y_i \left(\sum_{g=1, q=1}^{G, Q} \langle \mathbf{w}'_{gq}, \phi_{gq}(\mathbf{x}_i^{(g)}) \rangle + b \right) \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \end{cases} \quad (8)$$

The Lagrangian of problem (8):

$$\begin{aligned}
L = & \frac{1}{2} \sum_{g=1, q=1}^{G, Q} \frac{\|\mathbf{w}'_{gq}\|^2}{d_{gq}} + C \sum_{i=1}^N \xi_i \\
& + \sum_i^N \alpha_i (1 - \xi_i - y_i \sum_{g=1, q=1}^{G, Q} \mathbf{w}'_{gq} \phi_{gq}(\mathbf{x}_i^{(g)}) - y_i b) - \sum_i^N \beta_i \xi_i \quad (9)
\end{aligned}$$

where α_i and β_i are the Lagrange multipliers of the constraints.

To obtain the dual objective function we minimize L with respect to the variables $\mathbf{w}', b, \xi$ by setting the partial derivatives to 0,

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}'_{gq}} &= \mathbf{w}'_{gq} - d_{gq} \sum_i^N \alpha_i y_i \phi_{gq}(\mathbf{x}_i^{(g)}) = 0, \quad \forall g, q \\
\frac{\partial L}{\partial b} &= \sum_i^N \alpha_i y_i = 0 \\
\frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \beta_i = 0, \quad \forall i
\end{aligned} \quad (10)$$

By setting to zero the derivatives of this Lagrangian according to the primal variables, we derive the following associated dual problem:

$$\begin{aligned}
\max \sum_i^N \alpha_i - \frac{1}{2} \sum_{i, j}^N \alpha_i \alpha_j y_i y_j \sum_{g=1, q=1}^{G, Q} d_{gq} \mathbf{K}_{gq}(\mathbf{x}_i^{(g)}, \mathbf{x}_j^{(g)}) \\
C \geq \alpha_i \geq 0 \quad \forall i \\
\sum_i^N \alpha_i y_i = 0 \quad \forall i
\end{aligned} \quad (11)$$

Because of strong duality, $J(\mathbf{d})$ is also the objective value of the dual problem:

$$\begin{aligned}
J(\mathbf{d}) &= \sum_{i=1}^N \alpha_i^* - \frac{1}{2} \sum_{i, j=1}^N \alpha_i^* \alpha_j^* y_i y_j \sum_{g=1, q=1}^{G, Q} d_{gq} \mathbf{K}_{gq}(\mathbf{x}_i^{(g)}, \mathbf{x}_j^{(g)}) \\
&= \sum_{i=1}^N \alpha_i^* - \frac{1}{2} \sum_{i, j=1}^N \alpha_i^* \alpha_j^* y_i y_j \text{tr}(\mathbf{K}^{ij} \mathbf{d}) \quad (12)
\end{aligned}$$

where

$$\mathbf{K}^{ij} = \begin{bmatrix} \mathbf{K}_{11}(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(1)}) & \cdots & \mathbf{K}_{1Q}(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(1)}) \\ \vdots & \ddots & \vdots \\ \mathbf{K}_{G1}(\mathbf{x}_i^{(Q)}, \mathbf{x}_j^{(Q)}) & \cdots & \mathbf{K}_{GQ}(\mathbf{x}_i^{(Q)}, \mathbf{x}_j^{(Q)}) \end{bmatrix} \quad (13)$$

With a fixed $\mathbf{d}$, The objective value $J(\mathbf{d})$ can be obtained by any SVM algorithm using the combined kernel $\sum_{g=1, q=1}^{G, Q} d_{gq} \mathbf{K}_{gq}(\mathbf{x}_i^{(g)}, \mathbf{x}_j^{(g)})$. We propose two different optimization strategies to optimize the kernel weights $\mathbf{d}$: proximal gradient optimization algorithm and alternating iterative optimization algorithm.

3.2. Optimizing $\mathbf{d}$ with proximal gradient scheme, $\ell_{2,1}$ MKL-G

$J(\mathbf{d})$ in Eq.(7) is convex and differentiable, while $\|\mathbf{d}\|_{2,1}$ is convex but non-smooth. To optimize $\mathbf{d}$, we efficiently solve it by designing a new accelerated proximal gradient method with the idea of FISTA (fast iterative shrinkage-thresholding algorithm) [60] in this work. The $\mathbf{d}$ is iteratively updated by proximal gradient method with the following rule:

$$\begin{aligned} \mathbf{d}^{(t+1)} &= \arg \min_{\mathbf{d}} Q_L(\mathbf{d}, \mathbf{d}^{(t)}) \\ &= \arg \min_{\mathbf{d}} J(\mathbf{d}^{(t)}) + \langle \mathbf{d} - \mathbf{d}^{(t)}, \nabla J(\mathbf{d}^{(t)}) \rangle + \frac{L^{(t)}}{2} \|\mathbf{d} - \mathbf{d}^{(t)}\|^2 + C\gamma \|\mathbf{d}\|_{2,1} \\ &= \arg \min_{\mathbf{d}} \left\| \mathbf{d} - (\mathbf{d}^{(t)} - \frac{1}{L^{(t)}} \nabla J(\mathbf{d}^{(t)})) \right\|^2 + \frac{C\gamma}{L^{(t)}} \|\mathbf{d}\|_{2,1} \end{aligned} \quad (14)$$

where $L^{(t)}$ plays the role of a stepsize and is determined by the line search,

$$\nabla J(\mathbf{d}^{(t)}) = [\nabla J(d_{11}^{(t)}), \dots, \nabla J(d_{gq}^{(t)}), \dots, \nabla J(d_{GQ}^{(t)})] \quad (15)$$

and

$$\nabla J(d_{gq}^{(t)}) = -\frac{1}{2} \sum_{i,j=1}^N \alpha_i^* \alpha_j^* y_i y_j \mathbf{K}_{gq}(\mathbf{x}_i^{(g)}, \mathbf{x}_j^{(g)}) \quad \forall g, q \quad (16)$$

The differentiability of $J(d_{gq}^{(t)})$ is ensured by the uniqueness of α^* and by the differentiability of the objective function. Eq. (14) can be written as:

$$\mathbf{d}^{(t+1)} = \pi_L^{C\gamma}(\mathbf{d}^{(t)}) = \arg \min_{\mathbf{d}} \frac{1}{2} \|\mathbf{d} - \mathbf{u}^{(t)}\|_2^2 + \frac{C\gamma}{L^{(t)}} \|\mathbf{d}\|_{2,1} \quad (17)$$

where $\pi_L^{C\gamma}$ can be viewed as a proximal operator corresponding to the non-smooth penalty of $\|\mathbf{d}\|_{2,1}$, and $\mathbf{u}^{(t)} = \mathbf{d}^{(t)} - \frac{1}{L^{(t)}} \nabla J(\mathbf{d}^{(t)})$.

Thanks to the separability of $\mathbf{d}^{(t+1)}$ on each row, we can update the weights for each row $\mathbf{d}_g^{(t+1)}$ individually:

$$\mathbf{d}_g^{(t+1)} = \arg \min_{\mathbf{d}_g} \frac{1}{2} \|\mathbf{d}_g - \mathbf{u}_g^{(t)}\|_2^2 + \frac{C\gamma}{L^{(t)}} \|\mathbf{d}_g\|_2 \quad (18)$$

where $\mathbf{d}_g$ and $\mathbf{u}_g$ are the g -th rows of matrix $\mathbf{d}$ and $\mathbf{u}$ respectively.

Following [62], the row-wise updates can be done by soft-thresholding as

$$\mathbf{d}_g^{(t+1)} = \frac{\max \left\{ \left\| \mathbf{u}_g^{(t)} \right\|_2 - \frac{C_\gamma}{L^{(t)}}, 0 \right\}}{\left\| \mathbf{u}_g^{(t)} \right\|_2} \mathbf{u}_g^{(t)} \quad (19)$$

Moreover, a projection operator $\mathbf{P}$ is introduced ensuring that the non-negativity constraints on $\mathbf{d}^{(t+1)}$ ($\mathbf{d} \geq 0$) are satisfied.

$$\mathbf{P}(\mathbf{d}^{(t+1)}) = \max(\mathbf{0}, \mathbf{d}^{(t+1)}) \quad (20)$$

Meanwhile, in order to accelerate the proximal gradient method in Eq. (17), each point $\mathbf{u}^{(t)}$ uses a linear combination of the previous two points $\{\mathbf{d}^{(t-1)}, \mathbf{d}^{(t-2)}\}$, instead of at just the previous iterate $\mathbf{d}^{(t-1)}$ as follows:

$$\mathbf{u}^{(t)} = \mathbf{d}^{(t-1)} + \frac{a^{(t)} - 1}{a_{t+1}} (\mathbf{d}^{(t-1)} - \mathbf{d}^{(t-2)}) \quad (21)$$

The algorithm for updating kernel weighting $\mathbf{d}$ by the FISTA method is given in Algorithm 1, because the problem in Eq. (7) is a convex problem, the global optimum solution to the problem can be obtained. Moreover, the Algorithm 1 will monotonically decrease the objective of the problem Eq. (7) in each iteration and converge to the global optimum of the problem. As shown in [60], such an algorithm is guaranteed to converge at a rate $O(1/t^2)$.

Through the alternate optimization, the weights $\mathbf{d}$ can be learned to reflect the relative importance of different feature subsets. The final discriminant function is:

$$f(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^N \alpha_i^* y_i \sum_{g=1, q=1}^{G, Q} d_{gq}^* \mathbf{K}_{gq}(\mathbf{x}^{(g)}, \mathbf{x}_i^{(g)}) + b^* \right) \quad (22)$$

Since the regularization parameters C and C_γ need to be selected during the training, we use 5-fold cross validation on the training data to select the parameters.

3.3. Optimizing $\mathbf{d}$ with alternating iterative scheme, $\ell_{2,1}$ MKL-A

We find that the result of taking derivative of $\|\mathbf{d}\|_{2,1}$ w.r.t. $\mathbf{d}$ is equivalent to the derivative of $\frac{1}{2} \text{Tr}(\mathbf{d}^T \mathbf{U} \mathbf{d})$ w.r.t. $\mathbf{d}$. Actually, the derivative of $\|\mathbf{d}\|_{2,1}$ w.r.t. $\mathbf{d}$, $\mathbf{U} \mathbf{d}$, is a sub-gradient because it is not smooth. $\mathbf{U} \in R^{G \times G}$ is a diagonal matrix with $U_{gg} = \frac{1}{\|\mathbf{d}_g\|_2}$, and it is also an unknown variable which is dependent on $\mathbf{d}$. Minimizing $\|\mathbf{d}\|_{2,1}$ in the objective function (7) is equivalent to minimizing $\text{Tr}(\mathbf{d}^T \mathbf{U} \mathbf{d})$ where:

$$\mathbf{U} = \begin{bmatrix} \frac{1}{2\|\mathbf{d}_1\|_2} & & \\ & \dots & \\ & & \frac{1}{2\|\mathbf{d}_G\|_2} \end{bmatrix} \quad (23)$$

Algorithm 1 FISTA Method ($\ell_{2,1}$ MKL-G) for optimizing $\mathbf{d}$ in Eq. (7)

Input: Training Data X , class label Y , $L_0 \geq 0$, $a^1 = 1$, $\mathbf{d}_0$, C and C_γ
Output: Weights of kernel $\mathbf{d}$

```

1: Initialize  $\mathbf{d}^1 = [d_{11}, \dots, d_{GQ}]$  with  $d_{gq} \in [0, 1]$ 
2:  $\mathbf{u}^1 = \mathbf{d}^1$ 
3: for  $t = 1$  to  $\dots$  do
4:   Find the smallest nonnegative integers  $i_t$  such that with  $L^{(t)} = 2^{i_t} L^{(t-1)}$ 
      $J(\pi_L^{C_\gamma}(\mathbf{u}^{(t)})) \leq Q_L(\mathbf{u}, \mathbf{u}^{(t)})$ , and set  $L^{(t)} = 2^{i_t} L^{(t-1)}$ 
5:   Calculate  $J(\mathbf{d}^{(t)})$  and  $\nabla J(\mathbf{d}^{(t)})$  according to Eq. (15)
6:    $\mathbf{u}^{(t)} = \mathbf{d}^{(t)} - \frac{1}{L^{(t)}} \nabla J(\mathbf{d}^{(t)})$ 
7:   Compute  $\mathbf{d}^{(t+1)} = \arg \min_{\mathbf{d}} \frac{1}{2} \|\mathbf{d} - \mathbf{u}^{(t)}\|_2^2 + \frac{C_\gamma}{L^{(t)}} \|\mathbf{d}\|_{2,1}$  by solving Eq. (19)
8:    $\mathbf{d}^{(t+1)} = \mathbf{P}(\mathbf{d}^{(t+1)})$ 
9:    $a^{(t+1)} = \frac{(1 + \sqrt{1 + 4(a^{(t)})^2})}{2}$ 
10:   $\mathbf{u}^{(t+1)} = \mathbf{d}^{(t+1)} + \frac{a^{(t)} - 1}{a^{(t+1)}} (\mathbf{d}^{(t+1)} - \mathbf{d}^{(t)})$ 
11:  if convergence then
12:     $\mathbf{d} = \mathbf{d}^{(t+1)}$ 
13:  end if
14: end for

```

Therefore, the objective function of (7) is transformed to:

$$\min_{\mathbf{d} \geq 0} J(\mathbf{d}) + C_\gamma \text{Tr}(\mathbf{d}^T \mathbf{U} \mathbf{d}) \quad (24)$$

In order to find out the global solution $\mathbf{d}$, We adopt the strategy of alternating minimization algorithm proposed in [61], in which to kernel weights $\mathbf{d}$ and $\mathbf{f}$ are alternatively minimized. The detailed procedure of updating $\mathbf{d}$ is listed in Algorithm 2. In each iteration, $\boldsymbol{\alpha}^{(t+1)}$ is obtained with $\mathbf{d}^{(t)}$ fixed according to Eq.(11), and $\mathbf{d}^{(t+1)}$ is calculated with the current $\mathbf{U}^{(t)}$ and $\boldsymbol{\alpha}^{(t+1)}$ by setting the derivative of Eq.(24) with respect to $\mathbf{d}$ to zero, $\mathbf{U}^{(t+1)}$ is updated with $\mathbf{d}$ at last. We have a closed-form solution $\mathbf{d}$ in Eq.(25). Then $\mathbf{U}$ is updated according to the current $\mathbf{d}$. The iteration procedure is repeated until the algorithm converges.

$$\mathbf{d}_{gq} = \frac{1}{2} \left(\sum_{i,j=1}^N \alpha_i^* \alpha_j^* y_i y_j \mathbf{K}_{gq}^{ij} \right) C_\gamma^{-1} \mathbf{U}_g^{-1} \quad (25)$$

where $\mathbf{K}_{gq}^{ij} = \mathbf{K}_{gq}(\mathbf{x}_i^{(g)}, \mathbf{x}_j^{(g)})$.

Next, we show that the iterative process shown in Algorithm 2 converges.

In the t -th iteration,

$$\mathbf{d}^{(t+1)} = \arg \min_{\mathbf{d}} \sum_{i=1}^N \alpha_i^{(t)} - \frac{1}{2} \sum_{i,j=1}^N \alpha_i^{(t)} \alpha_j^{(t)} y_i y_j \text{tr}(\mathbf{K}^{ij} \mathbf{d}^{(t)}) + C_\gamma \text{Tr}((\mathbf{d}^{(t)})^T \mathbf{U}^{(t)} \mathbf{d}^{(t)}) \quad (26)$$

Algorithm 2 Alternating iterative Method ($\ell_{2,1}$ MKL-A) for optimizing $\mathbf{d}$ in Eq. (24)

Input:

Training data $\mathbf{X}$, label $\mathbf{Y}$, C and C_γ

Output:

Weights of kernel $\mathbf{d}$

- 1: Initialize $\mathbf{d} = [d_{11}, \dots, d_{GQ}]$ with $d_{gq} \in [0, 1]$, $\mathbf{U}^{(1)}$ as an identity matrix and $t = 1$
 - 2: **repeat**
 - 3: Update $\alpha^{(t+1)}$ while keeping $\mathbf{d}^{(t)}$ fixed according to Eq.(11)
 - 4: Update $\mathbf{d}^{(t+1)}$ with $\alpha^{(t)}$ and $\mathbf{U}^{(t)}$ being fixed according to Eq.(25), and $\mathbf{d}^{(t+1)} = \mathbf{P}(\mathbf{d}^{(t+1)})$
 - 5: Update $\mathbf{U}^{(t+1)}$ with $\mathbf{d}^{(t+1)}$ according to Eq.(23)
 - 6: $t = t + 1$
 - 7: **until** Converges
-

which indicates

$$\begin{aligned} & \sum_{i=1}^N \alpha_i^{(t)} - \frac{1}{2} \sum_{i,j=1}^N \alpha_i^{(t)} \alpha_j^{(t)} y_i y_j \text{tr}(\mathbf{K}^{ij} \mathbf{d}^{(t+1)}) + C_\gamma \text{Tr}((\mathbf{d}^{(t+1)})^T \mathbf{U}^{(t)} \mathbf{d}^{(t+1)}) \\ & \leq \sum_{i=1}^N \alpha_i^{(t)} - \frac{1}{2} \sum_{i,j=1}^N \alpha_i^{(t)} \alpha_j^{(t)} y_i y_j \text{tr}(\mathbf{K}^{ij} \mathbf{d}^{(t)}) + C_\gamma \text{Tr}((\mathbf{d}^{(t)})^T \mathbf{U}^{(t)} \mathbf{d}^{(t)}) \end{aligned} \quad (27)$$

The above inequality can be extended as,

$$\begin{aligned} & \sum_{i=1}^N \alpha_i^{(t)} - \frac{1}{2} \sum_{i,j=1}^N \alpha_i^{(t)} \alpha_j^{(t)} y_i y_j \text{tr}(\mathbf{K}^{ij} \mathbf{d}^{(t+1)}) + C_\gamma \sum_{g=1}^G \frac{\|(\mathbf{d}_g^{(t+1)})\|_2^2}{2 \|(\mathbf{d}_g^{(t)})\|_2^2} \\ & \leq \sum_{i=1}^N \alpha_i^{(t)} - \frac{1}{2} \sum_{i,j=1}^N \alpha_i^{(t)} \alpha_j^{(t)} y_i y_j \text{tr}(\mathbf{K}^{ij} \mathbf{d}^{(t)}) + C_\gamma \sum_{g=1}^G \frac{\|(\mathbf{d}_g^{(t)})\|_2^2}{2 \|(\mathbf{d}_g^{(t)})\|_2^2} \end{aligned} \quad (28)$$

Because $(\|(\mathbf{d}_g^{(t+1)})\|_2 - \|(\mathbf{d}_g^{(t)})\|_2)^2 \geq 0$, we can obtain:

$$\|(\mathbf{d}_g^{(t+1)})\|_2 - \frac{\|(\mathbf{d}_g^{(t+1)})\|_2^2}{2 \|(\mathbf{d}_g^{(t)})\|_2^2} \leq \|(\mathbf{d}_g^{(t)})\|_2 - \frac{\|(\mathbf{d}_g^{(t)})\|_2^2}{2 \|(\mathbf{d}_g^{(t)})\|_2^2} \quad (29)$$

So the following inequality holds:

$$C_\gamma \sum_{g=1}^G \left(\left\| (\mathbf{d}_g^{(t+1)}) \right\|_2 - \frac{\left\| (\mathbf{d}_g^{(t+1)}) \right\|_2^2}{2 \left\| (\mathbf{d}_g^{(t)}) \right\|_2} \right) \leq C_\gamma \sum_{g=1}^G \left(\left\| (\mathbf{d}_g^{(t)}) \right\|_2 - \frac{\left\| (\mathbf{d}_g^{(t)}) \right\|_2^2}{2 \left\| (\mathbf{d}_g^{(t)}) \right\|_2} \right) \quad (30)$$

Combining Eq. (28) with Eq. (30) and using the definition of $\ell_{2,1}$ -norm, we arrive at

$$\begin{aligned} & \sum_{i=1}^N \alpha_i^{(t)} - \frac{1}{2} \sum_{i,j=1}^N \alpha_i^{(t)} \alpha_j^{(t)} y_i y_j \text{tr}(\mathbf{K}^{ij} \mathbf{d}^{(t+1)}) + C_\gamma \left\| (\mathbf{d}^{(t+1)}) \right\|_{2,1} \\ & \leq \sum_{i=1}^N \alpha_i^{(t)} - \frac{1}{2} \sum_{i,j=1}^N \alpha_i^{(t)} \alpha_j^{(t)} y_i y_j \text{tr}(\mathbf{K}^{ij} \mathbf{d}^{(t)}) + C_\gamma \left\| (\mathbf{d}^{(t)}) \right\|_{2,1} \end{aligned} \quad (31)$$

Thus the Algorithm 2 will monotonically decrease the objective of the problem in Eq.(7) in each iteration t .

3.4. Implementation details

For the two optimization algorithms, they are terminated when a stopping criterion is met based on the variation of $\mathbf{d}$ between two consecutive steps. Specifically, we stop when $\max(|\mathbf{d}^{(t+1)} - \mathbf{d}^{(t)}|) < 0.001$. Moreover, both methods take advantage of warm-start techniques for successive SVM retrains. The gradient based approach produces estimates of $\mathbf{d}$ on a smooth trajectory, so that the previous SVM solution provides a good guess for the current SVM training. The commonly used kernel functions are Radial basis function kernels, also called the RBF kernels, or Gaussian kernels, and we use RBF kernels with six different kernel bandwidths ($2^{-2}, 2^{-1}, \dots, 2^3$) for each feature subset, which totally yields 48 kernels. Kernel matrices were pre-computed and normalized to have unit trace. We use LIBSVM software[80] for SVM optimization. For all the kernel methods, the regularization parameters C and C_γ in Eq. (6)) are chosen by nested cross-validation strategy on the training data(trying values 0.01, 0.1, 1, 10, 100).

3.5. Discussions

We explain the key properties and possible extensions of the proposed algorithm in this section.

3.5.1. Computational complexity

Both two methods have the same asymptotic computation complexity. Obviously, in each iteration, both models solve a canonical SVM problem with the optimal kernel $\sum_{g=1, q=1}^{G, Q} d_{gq} \mathbf{K}_{gq}(\mathbf{x}_i^{(g)}, \mathbf{x}_j^{(g)})$ optimized by two different strategies. The optimization of standard SVM is a quadratic programming problem with a complexity of $O(n_{sv}^3)$ with n_{sv} being the number of support vectors.

The time complexity of the $\mathbf{d}$ calculation is negligible compared to the SVM solver. The computational complexity of our algorithm mainly depends on the complexity of the canonical SVM solver used in the main loop, which can be reduced by using a hot-start (i.e., giving previous α as input).

For both optimization strategies, the complexity of the computation of $\mathbf{d}$ is $O(G \times Q \times n_{sv}^2)$, where G is the size of feature sets and Q is the amount of the kernel functions for each feature set. Hence the computation of $\mathbf{d}$ depends on the size of feature subsets and base kernel functions. Moreover for $\ell_{2,1}$ MKL-G, the step size of each iteration, $L^{(t)}$, should be determined with a line search method which requires additional SVM optimizations for better convergence. Note that every function evaluation in the line search of $\ell_{2,1}$ MKL-G requires solving an SVM problem. The number of iterations before convergence clearly depends on the training data and the step size selection procedure.

3.5.2. Extensions to Other Kernel-Based Algorithms

Our $\ell_{2,1}$ MKL can also be applied to kernel-based algorithms other than binary classification, such as regression and multi-class classification. The proposed MKL method can be extended in several directions for solving different prediction tasks in the medical application, such as multi-task multi-kernel learning, cost sensitive multi-kernel learning and semi-supervised multi-kernel learning. We need to make two basic changes: (a) optimization problem and (b) gradient calculations from the objective value found. We shortly explain three variants for multi-kernel regression, multi-kernel multi-task learning and semi-supervised multi-kernel learning.

- $\ell_{2,1}$ -norm multi-kernel regression

Regression models have been widely employed on the tracking disease progression or the prediction power of inferring cognitive outcomes in Alzheimer's disease[48][49]. We can also apply the $\ell_{2,1}$ -norm to ϵ -insensitive regression.

$$\begin{aligned}
 \min_{\mathbf{w}', \mathbf{d}, b, \xi} \quad & \frac{1}{2} \sum_{g=1, q=1}^{G, Q} \frac{\|\mathbf{w}'_{gq}\|^2}{d_{gq}} + C \sum_{i=1}^N \xi_i^+ + C \sum_{i=1}^N \xi_i^- + C_\gamma \|\mathbf{d}\|_{2,1} \quad (32) \\
 \text{s.t.} \quad & y_i - \sum_{g=1, q=1}^{G, Q} \langle \mathbf{w}'_{gq}, \phi_{gq}(\mathbf{x}_i^{(g)}) \rangle - b \leq \epsilon + \xi_i^+ \\
 & \sum_{g=1, q=1}^{G, Q} \langle \mathbf{w}'_{gq}, \phi_{gq}(\mathbf{x}_i^{(g)}) \rangle + b - y_i \leq \epsilon + \xi_i^- \\
 & d_{gq} \geq 0, \quad \forall g, q; \quad \xi_i^+, \xi_i^- \geq 0, \quad \forall i
 \end{aligned}$$

where $\epsilon \geq 0$ is the parameter in the ϵ -insensitive loss function and controls the accuracy of the regressor, $\xi^+ = \{\xi_1^+, \dots, \xi_N^+\}$ and $\xi^- = \{\xi_1^-, \dots, \xi_N^-\}$ are slack variables allowing for errors around the regression function.

- $\ell_{2,1}$ -norm multi-kernel multi-task learning

Multi-Task Learning (MTL) is a statistical learning framework which seeks at learning several models in a joint manner. The idea behind this paradigm is that, when the tasks to be learned are similar enough or are related in some sense, it may be advantageous to take into account these relations between tasks. Many medical diagnosis tasks have been benefited from multi-task learning[50][48]. The $\ell_{2,1}$ -norm multi-kernel multi-task formulation employs the $\ell_{2,1}$ -norm regularizer over the RKHS norms of the feature loadings corresponding to the given tasks and base kernels. The ℓ_2 -norm regularizer is applied across tasks. The ℓ_1 -norm regularizer is applied over the ℓ_2 norms leading to learning the shared kernel as a sparse combination of the given base kernels.

Suppose we are given T classification tasks to be learned from T different datasets $(x_{i,1}, y_{i,1})_{i=1}^{n_1}, \dots, (x_{i,T}, y_{i,T})_{i=1}^{n_T}$. The objective function of our problem is defined as:

$$\begin{aligned} \min_{\mathbf{w}', \mathbf{d}, b, \xi} \quad & \frac{1}{2} \sum_{q=1}^Q \sum_{t=1}^T \frac{\|\mathbf{w}'_{tq}\|^2}{d_{tq}} + C \sum_{t=1}^T \sum_{i=1}^{m_t} \xi_i \\ \text{s.t.} \quad & y_{ti} \left(\sum_{q=1}^Q \langle \mathbf{w}'_{tq}, \phi_{tq}(\mathbf{x}_i^{(g)}) \rangle - b_t \right) \geq 1 - \xi_{tq}, \xi_{tq} \geq 0, \quad i = 1, \dots, N \\ & d_{tq} \geq 0, \quad \forall t, q; \quad \sum_{q=1}^Q \left(\sum_{t=1}^T d_{tq}^2 \right)^{\frac{1}{2}} \leq 1 \end{aligned} \quad (33)$$

The formulation can be solved by an alternate optimization algorithm proposed in [51].

- $\ell_{2,1}$ -norm semi-supervised multi-kernel learning

In order to make the CAD systems perform well, a large amount of labeled instances are required for constructing a classifier model. However, the size of the labeled instances are usually insufficient since the task of providing label for each instance is quite time-consuming[52]. An effective strategy is needed to learn a model from a small amount of labeled instances combined with a large amount of unlabeled instances to enhance the performance of the learned model. We extend our methods to semi-supervised learning, which aims to enhance the performance of the classification model by using the labeled and unlabeled data together. In solving such problems, the given data set is often partially labeled, and one is required to make use of both the class information of the labeled data and the intrinsic relationships of the unlabeled ones to accomplish the learning tasks. Assume that we are given L labeled data and U unlabeled data. Generally speaking, semi-supervised kernel algorithms rely on the optimization of the following generic objective function in RKHS:

$$\min_f \sum_{i=1}^L \ell(f(\mathbf{x}_i), y_i) + C_1 \|f\|_H^2 + C_2 \mathcal{R}_f(\mathbf{f}) + C_3 \mathcal{R}_d(\mathbf{d}) \quad (34)$$

where $\mathbf{f} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_N))$, $N = L + U$. The first term in (34) represents the empirical loss on the labeled instance, the second term in (34) represents norm of f in $\mathcal{H}$ which aims at controlling the complexity of f in $\mathcal{H}$ for regularization purpose. The third term is a graph-based regularizer to preserve the similarities between samples in the whole $L + U$ dataset. We investigate a more common case of multi-kernel learning by incorporating the graph based regularization into our $\ell_{2,1}$ -norm multi kernel learning algorithm. We can rewrite the objective function as:

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \mathbf{d}, \boldsymbol{\xi}} \quad & \sum_{i=1}^L \boldsymbol{\xi}_i + \frac{C_1}{2} \boldsymbol{\alpha}^T K \boldsymbol{\alpha} + \frac{C_2}{2} \boldsymbol{\alpha}^T K \mathbf{L} K \boldsymbol{\alpha} + \gamma_3 \|\mathbf{d}\|_{2,1} \\ \text{s.t.} \quad & y_i \left(\sum_{j=1}^{L+U} \alpha_j K(\mathbf{x}_j, \mathbf{x}_i) + b \right) \geq 1 - \xi_i, \quad i = 1, \dots, L \\ & K = \sum_{g=1, q=1}^{G, Q} d_{gq} K_{gq}, \quad d_{gq} \geq 0, \quad \forall g, q; \quad \xi_i \geq 0, \quad \forall i \end{aligned} \quad (35)$$

where $\mathbf{L}$ is a Laplacian matrix.

3.5.3. Extension to nonlinear feature selection

Feature selection is an important problem in machine learning motivated by considerations of elimination of noisy and redundant features. In order to calculate the relative importance of each feature and remove some irrelevant and redundant features, the proposed multi-kernel learning methods should have an ability to perform both feature level and feature subset level analysis. There has been much recent progress in non-linear feature selection where the predictor is a non-linear function of the input features. In particular, MKL techniques have been shown to be amongst the most effective for nonlinear feature selection. They have been demonstrated to be superior to a number of linear and non-linear filter and wrapper methods. To perform nonlinear feature selection, a base kernel is computed for each feature (representation) in [47], which may result in one important feature then removed if the kernel function is not set appropriately. In the limit that only a single kernel is used per feature, we extend our proposed $\ell_{2,1}$ MKL for feature selection by associating multiple nonlinear base kernels with each individual feature, hence there is a total of $P \times Q$ base kernels in our $\ell_{2,1}$ norm regularized MKL, where P is the dimensionality of features and Q is the size of base kernel. The $\ell_{2,1}$ regularizer would enforce the sparsity among the features by ℓ_1 norm, while nonsparsity can be preserved for base kernels within the same feature by ℓ_2 norm.

4. Experimental study

4.1. CAD system

Unlike other publications in which candidate nodules were selected manually for classification evaluation, we obtained the appropriate candidate nodule

samples objectively using a candidate nodule detection algorithm. Our CAD scheme also contains lung segmentation, candidate nodule detection and VOI (Volume of Interest) segmentation. The sequence of steps for our scheme is shown in Fig. 3.

Figure 3: The flow cart of our Lung CAD system

4.2. Initial nodule detection

The database used in this work is the LIDC-IDRI[64]. In this database, four expert chest radiologists drew outlines for nodules having effective sizes of 3 mm or greater. The ground truth was then established by a blind reading and a subsequent unblinded reading. The LIDC Database contains 1012 cases, each of which includes images from a clinical thoracic CT scan and an associated XML file that records the results of a two-phase image annotation process performed by four experienced thoracic radiologists.

In the detection phase, we use the dot enhancement filter proposed by Li [65], which is aimed to simultaneously enhance objects of a specific structure (e.g. dot-like nodules) and suppress objects of other structures (e.g. line-like vessels) in multi-scale. If the nodule diameters are in the range $[d_{min}, d_{max}]$, the scales to be considered for the Gaussian smoothing is in the range $[d_{min}/4, d_{max}/4]$. The amount of scales is set to 5. After the 3D selective enhancement Hessian filter is applied on the original image, we use a thresholding technique and a 3D connected-component labeling technique to separate nodule candidates and identify all of the isolated objects [66]. The nodule-enhanced images is thresholded with a fixed, empirically selected value of 40 for separating nodule candidates from anatomic structures inside the lung regions. The value of the threshold T in the thresholding technique is empirically set to 40 since a good trade-off between sensitivity and FPs/I is obtained with 40. Moreover, the influence of the threshold value is shown in the Experiment IV. In general, a

lung nodule is defined as a focal opacity with a diameter between 3 and 30 mm[67], hence the objects with an effective diameter smaller than 2.5 mm are considered to be non-nodules and are eliminated [66]. Note that the effective diameter is defined as the diameter d of a sphere with the same volume of the segmented object ($d = \sqrt[3]{\frac{6V}{\pi}}$, where V is the volume of the VOI).

For each nodule candidate object, we developed a 3D constrained region-growing technique to segment it accurately in the original CT images. Fig. 4.2 shows some example result images of candidate VOI detection. The detected true nodule has different location and connection with the surrounding pulmonary structures.

Figure 4: Initial detection result of candidate nodules. TP indicated by arrow, other spots are FP

Due to the limitation that our detection methods only be effective on the solid nodule, we successfully detected 209 solid nodules along with 778 non-nodules from the 197 CT cases with 237 solid nodules. After obtaining the candidate nodule instances, we calculate the features for each of the nodule candidates from intensity, shape, texture and gradient distribution aspects. Then, we need to reduce the false positive nodule instances with our proposed classification model.

4.3. Feature extraction

Feature extraction plays an important role in the classification of suspicious nodules [21, 68]. However, there is not a single outstanding feature that can discriminate the nodule from non-nodule completely. This is due to the fact that the nodules vary enormously in volume, shape, and appearance, and the sources of false positives are different. The majority of false positives are mainly caused by blood vessels and other normal anatomic structures. Proposing an anatomical 3D feature leads to removing some blood vessels, segmental bronchus or bronchioles (small airways in the lung) which are not removed when only 2D features are considered. Some of the false positives can be easily distinguished

from true nodule, however, a large portion of them are difficult to distinguish. Therefore, for getting a high accuracy in candidate nodule classification, we should extract more features from many aspects, such as intensity, shape and gradient distribution. The extracted features are based on some appearance characteristics and clinical facts of nodules in CT scans:

- **Intensity features:** The nodules often have higher gray values than parts of vessels misidentified as nodules since a nodule typically appears as a relatively dense region in CT scans; and the intensity distribution of nodules CT appearance can be approximated by a Gaussian function.
- **Shape features:** Given the fact that pulmonary nodules do not have plate like or thin cylindrical structures. An isolated nodule or a nodule attached to a blood vessel is generally either depicted as a sphere or has some spherical elements, while a blood vessel is usually oblong, the shape features is to identify the objects that are elongated.
- **Texture features:** Texture properties of nodules allow us to characterize patterns based on their scale and orientation. Texture micropatterns like edges, lines, spots and flat areas in an image provide different kinds of useful discriminative information. Nodules do contain texture micropatterns at different scales and orientations. These micropatterns are helpful in the discrimination between nodules and non-nodules, and further between benign and malignant nodules.
- **Gradient features:** The true nodules have a high concentration because they grow from the center to the surround, thus nodules have high concentration of gradient vector while in a non-nodule, the gradient does not change dramatically as the vessel continues.

Our feature extraction process generated $P = 91$ image features from VOI for each potential nodule object. Using these features, we construct the input space for our classifiers. This section gives a brief introduction to the features we have collected for analysis and selection.

4.3.1. Intensity feature

We use some statistical gray feature to describe the global gray distribution within the candidate objects [26]. The gray value within the objects is characterized by the use of seven statistics (mean, variance, max, min, skew, kurtosis, entropy).

Meanwhile, for capturing the volume intensity distribution along the radial direction of the three dimension nodule candidates, we extract the feature of radial intensity distribution (RIS) [69]. The vector of radial intensity distribution feature for a nodule object should be decreasing, while for a non-nodule it changes irregularly. It is computed as follows:

$$RIS(r_{i+1}, r_i) = \sum_i I(x_i, y_i, z_i) \quad (36)$$

Where $I(x_i, y_i, z_i)$ denotes the intensity value in the position of (x_i, y_i, z_i) , (x_c, y_c, z_c) is the coordinate of the center,

$$(x_i, y_i, z_i) \in \left\{ (x_i, y_i, z_i) \mid r_i < \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2 + (z_i - z_c)^2} < r_{i+1} \right\}, \text{ and } i = [0, 1, 2, 3, 4].$$

4.3.2. Gabor texture feature

Gabor filters can be used effectively to extract these micro-patterns with different scales and orientations. A Gabor filter bank (GFB) consists of multiple filters having different scaling parameters, orientations and central frequency. Gabor filtering is a transform based method of extracting texture information. A Gabor filter is a sinusoid function modulated by a Gaussian:

$$\underline{\Psi}_{f,\theta,\varphi}(x, y, z) = \frac{1}{S} \times \exp(j2\pi(xf\sin\varphi\cos\theta + yf\sin\varphi\sin\theta + zf\cos\phi)) \quad (37)$$

$$\times \exp\left(\frac{(x - x_c)^2 + (y - y_c)^2 + (z - z_c)^2}{-2\sigma^2}\right) \quad (38)$$

where S is a normalization scale, f is the central frequency of the sinusoidal plane wave, ϕ and θ are the angles of the wave vector with w-axis and u v plane in the frequency domain (u, v, w) , and σ is the width of the Gaussian envelop in (x, y, z) domain. (x_c, y_c, z_c) is the position for signal analysis [73].

The response of $V(x, y, z)$ to $\Psi_{f,\theta,\varphi}(x, y, z)$ is represented as:

$$R_{f,\theta,\varphi}(x, y, z) = |(V \otimes \Psi_{f,\theta,\varphi})(x, y, z)| \quad (39)$$

where $V(x, y, z)$ is VOI, $\otimes$ is the convolution operation, $R_{f,\theta,\varphi}$ is the response of the VOI to a Gabor wavelet

We convolve the image with 48 Gabor filters tuned to four orientations of θ ($[0, \pi/4, \pi/2, 3\pi/4]$), four orientations φ ($[0, \pi/4, \pi/2, 3\pi/4]$) and three frequencies f ($[0.5, 0.25, 0.125]$). The mean value of each Gabor response VOI were calculated to be the texture features, which result in a total dimension of this feature as 48.

4.3.3. Shape feature

We attempt to distinguish true nodules from false ones by calculating the volumetric shape index (SI) and curvedness (CV) [70, 71, 72], as well as other explicit shape features to characterize the 3D shape [66].

Shape index is a measure of the shape, which represents the local shape feature at each voxel while being less sensitive to the image intensity. Every distinct shape, except for the plane, corresponds to a unique SI. The volumetric shape index (SI) at each voxel p can be defined as:

$$SI(p) = \frac{1}{2} - \frac{1}{\pi} \arctan \frac{k_1(p) + k_2(p)}{k_1(p) - k_2(p)} \sqrt{\frac{k_1^2 + k_2^2}{2}} \quad (40)$$

$$CV(p) = \sqrt{\frac{k_1^2(p) + k_2^2(p)}{2}} \quad (41)$$

The SI and CV features are calculated based on every pixel, and they are further characterized by seven statistical operations (mean, variance, max, min, skew, kurtosis, entropy). Some other explicit shape features were also extracted from a candidate object, including the volume, the surface area and compactness.

4.3.4. Gradient distribution feature

We utilize the concentration feature to characterize the degree of convergence at voxel p based on the concept of gradient concentration (GC) [70, 74]. In short, the GC feature at a point p characterizes the overall direction of the gradient vectors around. It is defined by:

$$GC(p) = \frac{1}{D} \sum_{i=1}^D e_i^{max}(p) \quad (42)$$

$$e_i^{max}(p) = \max_{R_{min} \leq n \leq R_{max}} \left\{ \frac{1}{n+1} \sum_{l=1}^N \cos \theta_{il}(p) \right\} \quad (43)$$

where D is the number of the symmetrically direction vectors d^l originating from p . The angle is calculated between d^l and g_i^l , where g_i^l is the gradient vector located at distance l from p in direction d^l . We also calculate the gradient field strength at each pixel. The concentration and strength features based on every pixel are further characterized by seven statistical operations. We refer them as eight feature representations (Intensity, Sub-volume distribution, Gabor, SI, CV, Geometric, GC and GS), all of which are listed in Table 1.

4.4. Evaluating the effectiveness of $\ell_{2,1}$ MKL

To evaluate the classification performance of our proposed method in false positive reduction, we present several experiments to answer the following questions: how is the learning capacity of the $\ell_{2,1}$ norm as the regularization on fusing heterogeneous feature subsets and pruning the irrelevant and redundant feature subsets? what is the relative importance of the heterogeneous feature subsets and individual feature? what is the difference between the two optimization methods for $\ell_{2,1}$ MKL with respect to classification accuracy, convergence rate and computational complexity? how does the performance of our $\ell_{2,1}$ MKL compare to the well known classifiers and the state-of-the-art methods of false positive reduction? and how to apply our approaches to other CAD system?

We carry out extensive experiments to test the performance of $\ell_{2,1}$ MKL along various dimensions. The experiments consist of vertical comparison and horizontal comparison separately. The vertical evaluations in Experiment I, II, III and IV involve the comparison between two different strategies, the comparison among different fusion strategies of heterogeneous feature subsets, the

Table 1: Feature set for candidate nodule classification

ID	Category	Feature Set(<i>FeaSet</i>)	Description
F1-7	Intensity	Intensity statistical feature	The gray value within the objects was characterized by use of seven statistics (mean, variance, max, min, skew, kurt, entropy)
F8-12		Sub-volume distribution feature	The average intensity within each sub-volume along the radial directions
F13-60	Texture	Gabor feature	The mean value of each filtered image
F61-67	Shape	SI statistical feature	The volumetric shape index (SI) representing the local shape feature at each voxel was characterized by use of seven statistics.
F68-74		CV statistical feature	The volumetric curvedness (CV), which quantifies how highly curved a surface is, was characterized by use of seven statistics.
F75-77		Geometric feature	volume, surface area and compactness
F78-84	Gradient	GC statistical feature	The concentration characterizing the degree of convergence of the gradient vectors at each voxel, was characterized by use of seven statistics
F85-91		GS statistical feature	The gradient strength of the gradient vectors at each voxel, was characterized by use of seven statistics

performance of extension $\ell_{2,1}$ MKL to feature selection and the influence of the initial detection threshold parameter T .

The horizontal comparisons are a comparison with the state-of-the-art multiple kernel learning methods, a comparison with the widely used machine learning algorithms, and a comparison with the state-of-the-art methods for false positive reduction of Lung CAD in Experiment V. Moreover, Experiment VI and Experiment VII demonstrate the performance of our methods working on a private lung dataset and a public breast CAD dataset.

All the experiments are carried out by means of a 10-fold cross-validation. That is, the dataset was split into ten folds, each one containing 10% of the patterns of the dataset. For each fold, the algorithm is trained with the examples contained in the remaining folds and then tested with the current fold. We use metrics such as sensitivity, specificity, G-mean ($\text{G-mean} = \sqrt{\text{sensitivity} \times \text{specificity}}$) and AUC to evaluate the performance of the learning algorithm on imbalanced nodule candidate data. G-mean is calculated as following:

In each training set of 10-fold cross-validation, the class distribution between TPs and FPs is imbalance, which is a typical class imbalance problem [9, 33]. Class imbalanced data has detrimental effects on the performance of conventional classifiers. Typically classifiers attempt to reduce global error rate without taking the data distribution into consideration. As a result, standard classifiers usually tend to be overwhelmed by the majority class and ignore the minority class examples, resulting in poor performance. In order to address

Table 2: Average performance measures for the two $\ell_{2,1}$ MKL algorithms

Methods	Iterations	Time (s)	# SVM eval
$\ell_{2,1}$ MKL-G	34	168	184
$\ell_{2,1}$ MKL-A	38	125	79

the imbalanced data, we use Hybrid probabilistic sampling (HPS) method [36] adopts the combination of over-sampling and under-sampling, and incorporates probability to balance the original skewed class distribution. The Hybrid probabilistic sampling function in its data distribution re-sampling mechanism. It generates more accurate instances to generalize the decision region for the nodule class, and removes the redundant instances for the non-nodule class without destroying the structure of the data. The re-sampling ratio α in HPS is set to 70%.

4.4.1. Experiment I: The comparison between two proposed different optimization methods for $\ell_{2,1}$ MKL

To evaluate the quality of the solution provided by our two different optimization algorithm when considering a $\ell_{2,1}$ norm, we have compared their with respect to classification performance(sensitivity, specificity, G-mean and AUC), convergence rate and computational complexity.

Figure 5 illustrates how the global training error changed as the number of iteration increased. As can be seen, $\ell_{2,1}$ MKL-G converges faster than $\ell_{2,1}$ MKL-A. In practice, we see convergence in 35-40 iterations for both methods. The key issue for faster convergence is to select good gradient-descent step sizes at each iteration. The results demonstrate that the FISTA-style accelerated proximal descent method is more efficient for the $\ell_{2,1}$ norm formulation of multiple kernel learning with the theoretical guarantee of the convergence rate. The step size of each iteration should be determined with a line search method (e.g., Armijos rule whose search procedure allows backtracking and does not use any curve fitting method), which requires solving additional kernel machine problems. The computation time is shown in Table 2. Although $\ell_{2,1}$ MKL-G requires a smaller number of iterations, it takes significantly longer time to finish one iteration than the other approaches for $\ell_{2,1}$ MKL-A, due to the high computational cost of the line search. This is because $\ell_{2,1}$ MKL-A uses a closed-form to update the kernel weights, and call less the SVM function, while for $\ell_{2,1}$ MKL-G, the step size of each iteration, $L^{(t)}$, should be determined with a line search method which requires additional SVM optimizations for better convergence. However, using a hot-start may accelerate the training process, thus the computation time per SVM training is considerably reduced, since the gradient-based approach produces estimates of $\mathbf{d}$ on a smooth trajectory.

The $\ell_{2,1}$ norm formulation enforces group sparsity among different feature subsets, which actually performs as a role of feature subset selection, while at the same time exploiting complementary information among the different kernels. In order to visualize the contribution of each feature type in such a MKL-based

Figure 5: Convergence behavior of two optimization methods for $\ell_{2,1}$ MKL

fusion scheme, we plot the kernel weights of the base kernels of our multiple kernel learning in Fig.6(a) and Fig.6(b). Fig.6(a) and Fig.6(b) show the weight for each base kernel optimized by our two proposed different $\ell_{2,1}$ MKL. The weight values correspond to the contribution for representations. Kernels corresponding to the discriminative feature subset are assigned the highest weights. We found that the weights of the Gabor and gradient strength feature are very low (norm proportion is less than the given threshold ratio 10%). Henceforth, we retain the feature subsets with high weight as the most discriminative feature representation and remove the Gabor and gradient strength feature subsets, so as to achieve the heterogeneous feature subset fusion and selection.

The multi-scale and multi-orientation representation of Gabor filters is often demonstrated as a highly effective texture descriptor. However, being rotation-variant, the Gabor filters become not directly suitable for our problem. Furthermore, The Gabor features are extracted by calculating the statistical measures like mean and standard deviation from magnitudes of the filter responses. This way of features extraction does not keep the spatial locality of texture micropatterns, and some important information is lost. The results demonstrate that the texture features cannot represent the volume of nodule objects, which is consistent with the results in [34], where employing 3D haralick features did not show gain in discrimination performance of nodule classification compared with 2D.

Different kernel functions capture different, potentially useful, notions of similarity in the data. Our formulation can easily incorporate more kernels, thus we investigate the behaviour of the proposed algorithm ($\ell_{2,1}$ MKL-G) with varying kernel size ($Q = 1, 2, 4, 6, 8$) and kernel type (RBF, linear and polynomial) for each feature subset in this study. Fig. 7 shows the results for multiple kernel configurations with different kernel sizes and kernel type on the classification problem.

(a) $\ell_{2,1}$ MKL-G

(b) $\ell_{2,1}$ MKL-A

Figure 6: The kernel weights of heterogeneous *FeaSets* of two different proposed $\ell_{2,1}$ MKL algorithms on the nodule dataset. Different colors indicate the base kernel functions with different kernel bandwidths($2^{-2}, 2^{-1}, \dots, 2^3$)

Figure 7: The performance with different kernel function type and size

From Fig.7 we can observe that RBF kernels have more powerful performance than linear kernel and polynomial kernels, which demonstrates that the classes of nodule and non-nodule is more discriminative in the nonlinear space. The performance of AUC increases as more kernel function for each feature subset and more diverse features are included in the system. Moreover, we can see that few base kernels for each feature set resulting in important feature sets removed if the kernel parameter is not set appropriately. It thus justifies our assumption that a combination of different types of kernel functions can provide complementary information of kernel combination which is useful for classification. The AUC with a combination of three types kernel functions is nearly the same as the one with six RBF kernels. Considering the computational efficiency, we still choose the RBF kernels with six bandwidths as the base kernels.

At last, SVM classifiers are trained on the fused feature subsets by the kernel weights optimized by two different algorithms. The results are shown in Table 3. From Table 3, $\ell_{2,1}$ MKL-G demonstrates clear performance improvement over $\ell_{2,1}$ MKL-A on classifying suspicious nodule data except the metric of specificity. One of the key differences between $\ell_{2,1}$ MKL-G and $\ell_{2,1}$ MKL-A is the optimization method of kernel weights of heterogeneous feature subset, and this seems to lead to substantial differences in performance on both evaluation metrics.

Table 3: The comparison between our proposed methods $\ell_{2,1}$ MKL-G and $\ell_{2,1}$ MKL-A

Method	Sensitivity (%)	Specificity (%)	G-mean (%)	AUC (%)
$\ell_{2,1}$ MKL-G	85.23 $\pm$ 1.67	91.33 $\pm$ 2.08	88.23 $\pm$ 1.77	87.17 $\pm$ 2.41
$\ell_{2,1}$ MKL-A	84.18 $\pm$ 2.32	91.59 $\pm$ 2.49	87.81 $\pm$ 2.30	85.35 $\pm$ 2.27

4.4.2. Experiment II: The comparison with the individual feature subset and other feature fusion methods

Depending on the task, different fusion methods lead to different classification performances. There are three main types of fusion schemes, depending on the stage the information is combined. We investigate the three different types of data fusion methods, namely, early (feature-level fusion), late (decision-level fusion), and intermediate (kernel-level fusion) fusion in our experiment. The feature-level methods combine multiple *FeaSets* into a single fused one that is then used in a conventional SVM classifier, whereas decision-level ones combine several diverse classifiers to make a stronger final classifier.

We compare our proposed multiple kernel learning based method with the two other fusion strategies for heterogeneous feature subsets information. SVM-All (feature-level fusion) concatenate all the feature subsets from the different sources into a single feature vector. In SVM-All, the training data are normalized to have zero mean and unit variance, and the test data are then normalized using the mean and variance of the training data. Ensemble classifiers (decision-level fusion) construct multiple base classifiers, then combine the predicted results by different classifiers via voting to reach the final output, such as bootstrap (Bagging), random subspace method (RSM) and Adaboost. For

three ensemble methods, the amount of base SVM classifiers is fixed to 50. In addition, we also report the classification performances by SVM on each individual type of features (SVM-Inten, SVM-Text, SVM-Grad and SVM-Shape) as baselines.

Table 4: The comparison between our methods with other fusion strategies

Method	Sensitivity (%)	Specificity (%)	G-mean (%)	AUC (%)
SVM-Inten	72.61 $\pm$ 3.35	84.22 $\pm$ 1.89	78.20 $\pm$ 3.74	77.26 $\pm$ 3.76
SVM-Text	69.33 $\pm$ 4.11	83.39 $\pm$ 3.81	76.04 $\pm$ 4.46	74.81 $\pm$ 4.04
SVM-Grad	68.17 $\pm$ 3.69	81.60 $\pm$ 4.34	74.58 $\pm$ 3.92	75.73 $\pm$ 4.22
SVM-Shape	71.72 $\pm$ 4.50	85.63 $\pm$ 3.66	78.36 $\pm$ 4.19	77.96 $\pm$ 3.83
SVM-All	77.82 $\pm$ 5.27	89.18 $\pm$ 4.26	83.31 $\pm$ 4.19	80.36 $\pm$ 3.25
$\ell_{2,1}$ MKL-G	85.23 $\pm$ 1.67	91.33 $\pm$ 2.08	88.23 $\pm$ 1.77	87.17 $\pm$ 2.41
$\ell_{2,1}$ MKL-A	84.18 $\pm$ 2.32	91.59 $\pm$ 2.49	87.81 $\pm$ 2.30	86.35 $\pm$ 2.27
Bagging	79.89 $\pm$ 3.55	90.75 $\pm$ 4.29	85.15 $\pm$ 2.79	82.65 $\pm$ 2.77
RSM	78.28 $\pm$ 4.50	88.38 $\pm$ 3.11	83.17 $\pm$ 4.52	81.65 $\pm$ 3.49
Adaboost	81.53 $\pm$ 2.74	90.62 $\pm$ 2.68	84.89 $\pm$ 4.03	83.93 $\pm$ 3.13

From the results in Table 4, we can conclude that regardless of the fusion level (feature-level, kernel-level or decision-level), the methods combining features from multiple views consistently outperform the best single feature based classifier. It demonstrates that different source of feature set provide complementary information, which may be useful for classification of nodule candidates when used together. Among the different fusion methods, we empirically demonstrate that $\ell_{2,1}$ MKL outperforms the other two common fusion methods. For the feature-level fusion, features extracted from different views are highly correlated. Thus, simple concatenation method represents an equal confidence in each feature type, which is often not enough to effectively fuse features. MKL method successfully handles discrepancies in the discriminative power of different features by assigning lower weights to less discriminative feature kernels. For the decision-level fusion of ensemble classifiers, only the final decision classification outputs are used, resulting in information loss. Thanks to the integration of $\ell_{2,1}$ -norm into multiple kernel learning, the proposed algorithm is able to accurately fuse the complementary and heterogeneous feature sets, and automatically prune the irrelevant and redundant feature subsets to form a more discriminative feature set.

Data are said to be noisy if they contain erroneous data values. These erroneous values can occur in the dependent (class noise) or independent (feature noise) variables in a data set [78]. In order to systematically investigate the robustness of $\ell_{2,1}$ MKL to noise, we manually introduced noise with different levels into the dataset, then assessed the performance of $\ell_{2,1}$ MKL and other comparative methods on the new datasets. We generated two types of noisy instances manually: class label noisy instances and attribute noisy instances, so as to simulate and evaluate the robustness of our method in both cases. We employ the same experimental procedure as in [78] to inject the label noisy instances into the dataset. Given a noise label l , each instance sees its label permuted with a probability of $l\%$. Moreover, we employ the same experi-

mental procedure as in [78] to generate the attribute noisy instances. Given a noise level l , each attribute in each instance is changed with a different value with a probability of $l\%$. This new value is selected uniformly among the other possible values for this attribute. Table 5 shows the AUC value of $\ell_{2,1}$ MKL and other comparative fusion algorithms under different class and feature levels. We find that the class noise is more detrimental to classification performance. The results clearly indicate the superiority of our method in the presence of noisy data, particularly the higher the noise level gets. Our $\ell_{2,1}$ MKL obtained the best performance except Bagging beat $\ell_{2,1}$ MKL under the 30% class noise. It is possible that noise in the training set do not show up in some certain subset, as the instances with noisy label may not be selected by the bootstrap sampling. All simulation results presented in this section illustrate the robustness of $\ell_{2,1}$ MKL to the negative influence of noise.

Table 5: Experimental results (average AUC) of tuning the noise level of class noise and feature noise

Noise type	Noise level	SVM-All	Adaboost	Bagging	RSM	$\ell_{2,1}$ MKL-G	$\ell_{2,1}$ MKL-A
class noise	10%	77.45	81.44	80.14	79.35	82.75	82.51
	20%	74.28	77.93	78.18	74.88	78.69	77.66
	30%	71.85	74.30	76.17	71.28	75.35	74.17
feature noise	10%	78.30	80.16	79.36	80.53	84.29	83.85
	20%	75.22	76.59	77.28	78.25	81.78	80.90
	30%	73.43	74.87	74.96	75.93	76.45	77.22

4.4.3. Experiment III: The evaluation of the performance of feature selection with $\ell_{2,1}$ MKL

In this experiment, we extend our proposed $\ell_{2,1}$ MKL for feature level to select discriminative features according to their predictive powers (scores) in the kernel space. The predictive powers of each feature determined by $\ell_{2,1}$ MKL is calculated as: $s_i = \|\mathbf{d}_i\|_2$. Then, We rank the contribution of each feature and identify which features are closely related to the discrimination of true nodule and false positives after the selection of heterogeneous feature subsets. To quantitatively evaluate the effect of the feature selection, we test the classification performance with respect to different numbers of the selected features in Fig.8. From the figure, it can be found that the classification performance of the proposed $\ell_{2,1}$ MKL reaches its peak at the number of 40. The performance of $\ell_{2,1}$ MKL with conducting the fusion of feature subset level and the selection of feature level is shown in Table 6. The feature selection with $\ell_{2,1}$ MKL not only can decrease the size of the feature vectors but also improves the classification performance. This further demonstrates that there exists irrelevant and redundant features in the heterogeneous feature subsets. Such irrelevant and redundant features may increase the computational complexity and lower the classification accuracy.

Moreover, we can obtain the relative importance of each feature from all the heterogeneous feature subsets without feature subset selection according

Figure 8: Classification performance (AUC) with respect to different number of selected features

Table 6: The performance of proposed $\ell_{2,1}$ MKL with respect to selection of feature subset & feature

	Method	Sensitivity (%)	Specificity (%)	G-mean (%)	AUC (%)
$\ell_{2,1}$ MKL-G	selection of feature subset	85.23 ± 1.67	91.33 ± 2.08	88.23 ± 1.77	87.23 ± 2.83
	selection of feature subset & feature	87.41 ± 1.81	92.58 ± 2.72	89.96 ± 3.09	91.39 ± 2.75
$\ell_{2,1}$ MKL-A	selection of feature subset	84.18 ± 2.32	91.59 ± 2.49	87.81 ± 2.30	85.35 ± 2.27
	selection of feature subset & feature	86.06 ± 2.52	91.98 ± 2.84	88.97 ± 3.82	88.10 ± 3.44

to the kernel weight. The value of the weight s for the top 10 of the most discriminative features is given in Table 7. We can find that the shape descriptor exhibited promising discriminating ability, and the Gabor feature appears only once, showing that the Gabor features provide less powerful discriminative information for discriminating suspicious nodules again.

4.4.4. Experiment IV: The influence of the threshold parameter T in initial detection

The choice of the threshold parameter T is important for the CAD system because it determines the size of true and false nodules, and affects directly the final performance. To observe performance with regard to the value of T , we varied the value of threshold in the nodule-enhanced images, and investigated the variation of performance with multiple threshold values in Table 8.

From Table 8, it can be found that the choice of T plays an important role in terms of the performance. With value of T decreasing, the sensitivity of initial nodule detection is increasing with higher imbalanced ratio. When $T = 30$, the sensitivity achieves to 99.2% with only two true nodules not being detected. However, the high imbalanced class distribution hinder the classification performance of our method in the phrase of false positive reduction. The best performance is obtained when T is set to 40, the sensitivity of our detection

Table 7: The top 10 features with their corresponding average kernel weights

$\ell_{2,1}$ MKL-G		$\ell_{2,1}$ MKL-A	
Feature name	Kernel weight	Feature name	Kernel weight
SI_max	0.0871	CV_min	0.0933
GC_entropy	0.0711	SI_mean	0.0914
SI_mean	0.0699	COMP	0.0881
VOL	0.0613	SI_max	0.0855
INTEN_mean	0.0609	INTEN_skew	0.0840
RIS(r_1, r_0)	0.0552	IRRE	0.0768
INTEN_skew	0.0449	RIS(r_2, r_1)	0.0722
COMP	0.0422	INTEN_var	0.0653
Gabor_mean($\pi/2.3\pi/4, 0.25$)	0.0408	CV_mean	0.0616
CV_mean	0.0387	GC_entropy	0.0532

Table 8: The performance with different value of threshold T in initial nodule detection

Threshold value	Number of true nodule	Number of false nodule	Imbalanced ratio	Sen(%) (Initial nodule detection)	AUC(%) (False positive reduction)
20	235	2132	1:9.1	99.2	73.47 ± 3.77
30	217	1358	1:6.1	90.8	83.66 ± 3.62
40	209	778	1:3.7	88.2	87.17 ± 2.41
45	151	634	1:3.5	63.7	85.23 ± 2.05
50	75	207	1:2.8	31.6	81.49 ± 2.59

method can achieve 89.55% in terms of sensitivity in initial nodule detection, and our proposed $\ell_{2,1}$ MKL-G combined with HPS achieve 87.17% in terms of AUC in false positive reduction. Moreover, when T is set to 30, it improved the detection sensitivity of true nodule by 2.95% with small decrease (4.03%) of the AUC.

4.4.5. Experiment V: The comparison with the state-of-the-art approaches

To show the superior performance of our algorithms, we selected several state-of-the-art approaches for comparison:

- (1) The comparison with the state-of-the-art multiple kernel learning methods;
- (2) The comparison with the widely used linear and nonlinear classifiers;
- (3) The comparison with the state-of-the-art methods for false positive reduction of Lung CAD.

- The comparison with the state-of-the-art multiple kernel learning methods

In this experiment we evaluate the effectiveness of our proposed $\ell_{2,1}$ norm regularized algorithm, and study how the behaviour of multiple kernel learning is affected by the norm used to regularize the kernel weights for suspicious nodule classification. In order to demonstrate the effectiveness, we also implemented the ℓ_1 norm, ℓ_2 norm, weighted-MKL proposed in [47] and uniformly weighted MKL for heterogeneous feature fusion, which are commonly used multiple kernel

learning methods. For both ℓ_1 norm and ℓ_2 norm, only the norm formulation in constraint is substituted by the $\|\mathbf{d}\|_1 \leq 1$ and $\|\mathbf{d}\|_2 \leq 1$, respectively.

In order to visualize the contribution of each feature type in different multi-kernel based fusion schemes, we plot the kernel weights of the base kernels for weighted-MKL, ℓ_1 norm MKL, ℓ_2 norm MKL, and our $\ell_{2,1}$ norm MKL in Fig. 9(a), 9(b), 9(c), 9(d)) and 9(e)), respectively. From Fig. 9, we see that our $\ell_{2,1}$ norm MKL selected more than one base kernel for each feature set, while the weighted MKL tends to select sparse base kernels for feature fusion, resulting in inaccurate feature subset selection.

We compare the four multi-kernel based fusion schemes comprehensively in Table 9. The ℓ_1 norm MKL performs worst since the information carried in the kernels that get zero weights is completely discarded, which may be explained by the fact that the combined kernels carry complementary information. Moreover, the ℓ_1 norm MKL and ℓ_2 norm MKL do not take the structure of feature subset into consideration, resulting in the prediction performance being dominated by our $\ell_{2,1}$ norm MKL. While some studies claimed that ℓ_1 regularization yields better performance for object recognition[59], our results show that ℓ_1 regularization may result in suboptimal performance due to the sparseness of the solutions. The experimental results demonstrated that the $\ell_{2,1}$ norm framework achieves the best. We obtained a more generalized performance for heterogeneous potential nodule data.

For further finding out whether the proposed $\ell_{2,1}$ MKL is significantly better than the other compared algorithms, we implemented a t -test on the classification results of the 10 runs to calculate the statistical significance of $\ell_{2,1}$ MKL. The t -test values are also listed in Table 9, from which we can clearly find that the $\ell_{2,1}$ MKL-G significantly outperforms the computing MKL methods except ℓ_2 MKL at the 5% significance level, and $\ell_{2,1}$ MKL-A significantly outperforms ℓ_2 MKL and weighted-MKL.

Table 9: The comparison of performance using multiple feature representation classification methods (Note that * indicates that $\ell_{2,1}$ MKL-G significantly outperform the competing method, and † indicates that $\ell_{2,1}$ MKL-A significantly outperform the competing method)

Method	G-mean	p -value	AUC	p -value
ℓ_1 MKL	$81.72 \pm 3.88^{*\dagger}$	0.044(0.049)	$81.13 \pm 3.41^*$	0.046(0.052)
ℓ_2 MKL	86.83 ± 5.95	0.088(0.091)	84.33 ± 4.74	0.075(0.082)
weighted-MKL	$81.53 \pm 3.16^{*\dagger}$	0.029(0.035)	$82.40 \pm 2.78^{*\dagger}$	0.038(0.044)
$\ell_{2,1}$ MKL-G	88.23 ± 1.77	-	87.23 ± 2.83	-
$\ell_{2,1}$ MKL-A	87.81 ± 2.30	0.078	85.35 ± 2.27	0.097

- The comparison with the widely used machine learning algorithms

We compared the performance of our method with three well-known classification approaches working on the hand-crafted VOI-based features and learned features by deep learning, respectively. The well-known methods for comparison are enumerated as follows:

Figure 9: Base kernel weights comparison of different MKL algorithms on the nodule dataset. Different colors indicate the base kernel functions with different kernel bandwidths ($2^{-2}, 2^{-1}, \dots, 2^3$)

k NN: The k NN classification works on the idea of calculating global distances between patterns, followed by ranking to determine the NNs that best represent the class of a test pattern[53]. We choose the Euclidean distance as distance metric, and the parameter of k is set to 5.

Multi-Kernel k -NN(MK- k NN): The accuracy of k -NN classifiers can depend very strongly on the specific selection of distance measures. We extend the traditional k -NN by substituting a kernel distance metric for the original one in Hilbert space, and the kernel is obtained by our $\ell_{2,1}$ MKL method.

Incomplete Source-Feature Selection (ISFS)[54]: ISFS performs simultaneous feature-level and source-level analysis. In the model of ISFS, α_i denotes the model learned using the i -th feature subset and γ is the weight that combines those learned models together, to induce sparsity on either the feature-level or the feature subset-level, thus achieving the goal of feature selection and feature subset selection simultaneously.

3D convolutional neural networks (CNN)[55]: Recent work on deep learning has examined how deep networks can be trained to produce useful representations for nodule[56][57]. CNN is used to learn highly discriminative features for nodule detection in lieu of hand-engineered ones such as geometric shape or texture. We employ a 3D CNN to examine how deep networks can be trained to produce useful representations for 3D suspicious nodules. The compared CNN was implemented based on CAFFE[58]. This comparison allows us to evaluate if the feature representations by deep learning can outperform the hand-crafted VOI-based features. We hierarchically constructed a deep 3D CNN model by stacking the C(convolutional layer), M(max-pooling layers) and FC (Fully connected layers), as shown in Fig.10. The last layers are fully-connected ones which were followed by logarithmic loss to be minimized. The kernel size is 5×5 in all convolutional layers. To capture a majority of nodule morphology, the input nodule patch size was set to $32 \times 32 \times 16$ voxels, and all the training 3D volumes were flattened into a feature vector as its input. The ultimate output layer employed the softmax activation to yield the prediction probabilities. An ℓ_2 norm weight decay was adopted to avoid overfitting during the training process. Moreover, we also compared a method incorporating CNN with MKL, named MK-CNN. In MK-CNN, the learned generic features by pre-training are inputted into a multi-kernel learning method and trained, in order to directly compare the two features calculated by different mechanisms based on the kernel classifier.

Experimental results are reported in Table 10 where the best results are boldfaced. A first glance at the results shows that our both $\ell_{2,1}$ MKL methods generally outperform all other compared methods on both metrics. Additionally, a statistical analysis is performed on the results and reported in Table 10. As can be seen, our proposed methods achieve statistically significant results compared to all other methods on most of the results. The ROC of the various approaches is shown in Fig.11(a). These results reveal several interesting points:

(1) We can observe that both $\ell_{2,1}$ MKL perform significantly better than the comparable classifiers except the CNN model. Although no significant difference occurs, our proposed $\ell_{2,1}$ MKL-G reports an increase of about 3.6% (2.6%) in

Figure 10: The structure of 3D-CNN

the G-mean and AUC compared to the CNN. The reason is that CNN requires a large amount of labeled training data, however the size of three dimensional data (VOI) from nodules is much less compared with two dimensional data (ROI) since one CT scan (patient) generally contains 1-2 nodules. The limited available training data degrades the performance of CNN, and leads to overfitting since the volume of interest from three dimensions introduce more parameters in the network. The provided training data are not sufficient for optimizing the parameters. Both of the above reasons hinders CNN from achieving best performance.

Many works have shown the strong representation and classification performance of deep learning compared with the hand-crafted features, which is not completely fair for the hand-crafted features since they didn't extract sufficient and complementary features from multiple views. Even if a more complete representation of nodules is provided, the simple concatenation of features is applied without being fused in a principled way in prior work [26][36][66]. Moreover, training a deep CNN requires extensive computational and memory resources compared with the traditional extraction of hand-crafted features and classification. Compared with the deep features automatically learned by the trained CNN from the original image, the traditional feature extraction has less complexity. Furthermore, CNN has less interpretability, which is critical for the medical domain. Therefore, our $\ell_{2,1}$ MKL combined with heterogeneous hand-crafted feature subsets has advantages with respect to the effectiveness and practicality for the classification of nodule candidates.

(2) We found that both CNN and $\ell_{2,1}$ MKL outperform MK-CNN, which indicates that the learned deep features at pre-training stage without fine-tuning cannot represent the nodule candidates well.

(3) The proposed method improved on average by 19.94% (G-mean) and 14.00% (AUC) compared to k NN, and by 4.72% (G-mean) and 2.63% (AUC) compared to MK- k NN. Additionally, MK- k NN outperforms the original k NN, which indicates that the instances of nodule and non-nodule have more clearly

separable properties in the kernel space than the original input space.

(4) Both of $\ell_{2,1}$ MKL and ISFS are developed for fusing multiple heterogeneous feature subsets(sources). From Table 10, we find that $\ell_{2,1}$ MKL-G, on average, improved by 6.44%/7.96% compared to ISFS in terms of G-mean and AUC. It reveals the effectiveness of our multi-kernel learning over the linear method.

Table 10: The comparison between our methods with the widely used classification approaches (Note that * indicates that $\ell_{2,1}$ MKL-G significantly outperform the competing method, and † indicates that $\ell_{2,1}$ MKL-A significantly outperform the competing method)

Method	G-mean	p -value	AUC	p -value
$\ell_{2,1}$ MKL-G	88.23 $\pm$ 3.16	-	87.17 $\pm$ 2.41	-
$\ell_{2,1}$ MKL-A	87.81 $\pm$ 2.30	0.078	85.35 $\pm$ 2.27	0.097
k NN	73.56 $\pm$ 4.39* †	0.006(0.008)	76.46 $\pm$ 2.32* †	0.029(0.030)
MK- k NN	84.25 $\pm$ 5.21* †	0.039(0.043)	81.19 $\pm$ 3.41* †	0.044(0.049)
ISFS	82.89 $\pm$ 2.60* †	0.017(0.031)	80.74 $\pm$ 1.58* †	0.033(0.038)
CNN	85.16 $\pm$ 4.42	0.066(0.076)	84.93 $\pm$ 2.87	0.054(0.062)
MK-CNN	84.99 $\pm$ 3.15* †	0.022(0.027)	80.55 $\pm$ 3.41* †	0.039(0.041)

- The comparison with the state-of-the-art for false positive reduction of Lung CAD

In this experiment, we compare our proposed methods against some current methods for false positive reduction. Linear discriminant analysis (LDA) technique is most frequently used in removing false positives [21]. Random forest (RF) is an effective ensemble classification of nodule and non-nodule patterns [32]. The parameters of random forest are set as follows: the number of trees is set to 50 and the number of features at each split is set to 5. Asymmetric AdaBoost (AsyAda) sets different weights for two different errors [37]. The asymmetry parameter is set to 10 in the AsyAda algorithm, and SVM is chosen as the base classifier.

Cost sensitive SVM (CS-SVM) is a good solution on the unbalanced data sets [30, 43]. In CS-SVM, We fix $C- = C$ and $C+ = C \times C_{rf}$, where C and C_{rf} are respectively the regularization parameter and the ratio misclassification cost factor. The ranges for C and γ are based on a grid search for SVM parameters as recommended in [80]. The ratio cost C_{rf} is set to be the ratio of the amounts of two classes. All the experiments are conducted by 10-fold cross validation.

Table 11: The comparison between our methods with other approaches for false positive reduction (Note that * indicates that $\ell_{2,1}$ MKL-G significantly outperform the competing method, and $\dagger$ indicates that $\ell_{2,1}$ MKL-A significantly outperform the competing method)

Method	G-mean	p -value	AUC	p -value
$\ell_{2,1}$ MKL-G	88.23 ± 3.16	-	87.17 ± 2.41	-
$\ell_{2,1}$ MKL-A	87.81 ± 2.30	0.078	85.35 ± 2.27	0.097
RF	$78.24 \pm 4.39^*$	0.044(0.052)	$79.43 \pm 3.48^{*\dagger}$	0.039(0.045)
AsyAdaboost	$79.25 \pm 5.21^{*\dagger}$	0.035(0.042)	$80.74 \pm 3.39^{*\dagger}$	0.038(0.043)
CS-SVM	$77.23 \pm 6.46^{*\dagger}$	0.026(0.031)	$81.57 \pm 2.91^{*\dagger}$	0.013(0.015)
LDA	$75.64 \pm 6.27^{*\dagger}$	0.017(0.018)	$72.82 \pm 2.04^{*\dagger}$	0.009(0.012)

Experiment results (Table 11) demonstrate the benefit of $\ell_{2,1}$ MKL compared to other commonly used methods for false positive reduction in Lung nodule CAD. At confidence level $\alpha = 0.05$, the results in Table 11 show that the proposed methods both significantly outperform AsyAdaboost, CS-SVM and LDA with respect to AUC and G-mean, and $\ell_{2,1}$ MKL-G significantly outperform random forest with respect to AUC and G-mean while $\ell_{2,1}$ MKL-A significantly outperform random forest only with respect to AUC. We also showed the ROC analysis of the various approaches in Fig. 11(b). By the horizontal comparison, our methods have been shown to improve on the performance of reducing false positives while maintaining the high recognition of true nodules i.e. our methods do not jeopardize the high sensitivity of the initial detection.

(a) The comparison with the state-of-the-art classification methods

(b) The comparison with the state-of-the-art false positive reduction methods

Figure 11: ROC analysis of the various approaches

4.4.6. Experiment VI: Evaluation on other real lung databases

As a further independent data set, we adopted the scans collected from the Guangzhou hospital and the Shenyang hospital in China. These databases included nodules of different sizes (3-30mm). A database of 108 thoracic CT scans from two hospitals from the Guangzhou and Shenyang was used to evaluate the effectiveness of the proposed method. Slice thickness varied from 0.5mm to 2.0mm and the total slice number for each scan varied from 75 to 226. The nodule locations of these scans are marked consensually by at least two experienced expert radiologists. Our database consists of 165 thin section CT scans with 192 solid nodules, obtained from both hospitals. Figure 12 shows some example result images of candidate VOI detection. The detected true nodule has different location and connection with the surrounding pulmonary structures.

Figure 12: Initial detection result of candidate nodules. TP indicated by arrow, other spots are FP

Details of the performance test for the comparative methods are shown in Table 12. Using the private dataset, we showed that the proposed methods convincingly outperforms the state-of-the-art methods on overall classification performance in terms of G-mean and AUC.

Multi-kernel Learning in Lung nodule CAD

Table 12: The comparison between our methods with other approaches for false positive reduction in the private dataset (Note that * indicates that $\ell_{2,1}$ MKL-G significantly outperform the competing method, and $\dagger$ indicates that $\ell_{2,1}$ MKL-A significantly outperform the competing method)

Method	G-mean	p -value	AUC	p -value
$\ell_{2,1}$ MKL-G	90.88 $\pm$ 3.35	-	91.30 $\pm$ 2.15	-
$\ell_{2,1}$ MKL-A	89.81 $\pm$ 2.15	0.078	89.85 $\pm$ 2.46	0.097
k NN	74.88 $\pm$ 3.95* $\dagger$	0.004(0.008)	78.87 $\pm$ 2.10* $\dagger$	0.012 (0.015)
MK- k NN	84.91 $\pm$ 4.83* $\dagger$	0.033(0.039)	86.56 $\pm$ 3.81* $\dagger$	0.041 (0.048)
ISFS	78.15 $\pm$ 2.13* $\dagger$	0.021(0.028)	85.30 $\pm$ 1.99* $\dagger$	0.043 (0.048)
CNN	85.70 $\pm$ 3.53	0.079(0.086)	86.52 $\pm$ 3.34	0.066 (0.069)
MK-CNN	81.19 $\pm$ 3.76* $\dagger$	0.013(0.016)	84.60 $\pm$ 3.08* $\dagger$	0.033(0.035)
RF	83.74 $\pm$ 4.16* $\dagger$	0.024(0.026)	85.69 $\pm$ 4.04* $\dagger$	0.027(0.028)
AsyAdaboost	80.66 $\pm$ 4.72* $\dagger$	0.015(0.019)	84.59 $\pm$ 3.77* $\dagger$	0.025 (0.031)
CS-SVM	82.23 $\pm$ 4.55* $\dagger$	0.011(0.015)	84.54 $\pm$ 3.32* $\dagger$	0.017 (0.021)
LDA	78.35 $\pm$ 5.15* $\dagger$	0.009(0.013)	79.35 $\pm$ 2.33* $\dagger$	0.012(0.016)

4.4.7. Experiment VII: The extension to other CAD

The proposed methods could be applied for the detection of many other potential lesions, such as mass, polyp, and microcalcification. After detecting and segmenting the suspicious objects, one or multiple features are extracted to represent the objects based on the appearance properties of lesions. The common difficulty of classification in CAD systems is to balance the number of samples used and fuse the heterogeneous feature subset extracted, both of which can confuse the classifier. Hence, our method can solve both of the major issues, so as to improve the false positive reduction. In this study, to ensure generalizability of our findings, we considered another common computer aided detection: breast CAD. Moreover, our classification method has been built in a commercial breast CAD system shown in Fig.14. The high number of false positives and the resulting number of avoidable breast biopsies are also the major problems faced by current mammography Computer Aided Detection (CAD) systems[81].

For evaluating the proposed methods, two sets of suspicious mammogram regions have been used. The first one, obtained from the Digital Database for Screening Mammography (DDSM), contains 1365 regions (989 normal and 376 abnormal samples). The second set of suspicious regions was obtained from the database of Mammographic Image Analysis Society (mini-MIAS) and contains 266 (136 normal and 55 abnormal) samples. Masses have a high concentration value of gradient vector since the masses of cancers grow from the center to the surround. Therefore, we use iris filter to detect the potential mass. Iris filter is a filter which can adaptively obtain the concentration of gradient vectors[82]. Threshold value in our experiment was calculated adaptively for each mammogram. It was estimated as the gray level value at the 95% of the cumulative distributive function. Potential masses can be detected and segmented from the background using a threshold value in the enhanced image processed by iris filter. For the ROIs of masses candidates, we extracted intensity, shape, tex-

ture and gradient features on the two-dimensional (2D) data. Instead of SI and CV, we employed six morphologic features (Form Factor, Roundness, Aspect Ratio, Convexity, Solidity and Extent) as shape descriptors, which have been demonstrated to be effective in classifying breast tumors [83]. Fig.13 shows the weight of each base kernel optimized by our $\ell_{2,1}$ MKL-G. From Fig.13, it has been observed that Gabor feature and shape feature have more important contribution of discriminating suspicious mass samples. It also demonstrates that the texture features are more effective representation for 2D lesions.

Figure 13: The kernel weights of heterogeneous *FeaSets* of proposed $\ell_{2,1}$ MKL-G algorithm on the breast dataset. Different colors indicate the base kernel functions with different kernel bandwidths($2^{-2}, 2^{-1}, \dots, 2^3$)

The Table 13 shows the classification performance of the competing methods. The results strongly suggest that the proposed multi-kernel method can aid in the fusion of heterogeneous feature subset, diagnosis of diseases and can be very helpful to the physicians for their final decision on their patients. Moreover, CNN outperforms our methods in terms of G-mean and AUC since there are sufficient training samples for 2D instances. However, our methods take less computation time for training than CNN.

Multi-kernel Learning in Lung nodule CAD

(a)

(b)

Figure 14: The breast computer aided detection system

Table 13: The comparison between our methods with the competing approaches for false positive reduction in breast CAD (Note that * indicates that $\ell_{2,1}$ MKL-G significantly outperform the competing method, and † indicates that $\ell_{2,1}$ MKL-A significantly outperform the competing method)

Method	G-mean	p -value	AUC	p -value
$\ell_{2,1}$ MKL-G	87.23 ± 2.57	-	90.58 ± 2.19	-
$\ell_{2,1}$ MKL-A	85.66 ± 2.73	0.078	89.19 ± 2.48	0.075
k NN	$79.22 \pm 3.76^{*\dagger}$	0.014(0.018)	$80.87 \pm 2.31^{*\dagger}$	0.024(0.029)
MK- k NN	$82.86 \pm 4.83^{*\dagger}$	0.026(0.035)	$85.56 \pm 3.34^{*\dagger}$	0.033(0.035)
ISFS	$84.60 \pm 2.74^{*\dagger}$	0.028(0.040)	$86.98 \pm 2.37^*$	0.043(0.052)
CNN	89.59 ± 3.28	0.097(0.101)	92.97 ± 3.80	0.088(0.092)
MK-CNN	$83.44 \pm 3.76^{*\dagger}$	0.037(0.045)	$82.58 \pm 3.66^{*\dagger}$	0.011(0.015)
RF	$81.67 \pm 4.05^{*\dagger}$	0.017(0.025)	$80.95 \pm 3.15^{*\dagger}$	0.030(0.031)
AsyAdaboost	$80.52 \pm 4.14^{*\dagger}$	0.031(0.033)	$81.13 \pm 2.78^{*\dagger}$	0.027(0.035)
CS-SVM	$79.77 \pm 4.80^{*\dagger}$	0.020(0.023)	$81.79 \pm 3.27^{*\dagger}$	0.016(0.022)
LDA	$74.36 \pm 5.19^{*\dagger}$	0.011(0.015)	$77.60 \pm 3.52^{*\dagger}$	0.007(0.010)

5. Conclusion

An automated lung nodule detection system can help detect lung abnormalities in CT lung images. Focusing on improving the performance of false positive reduction effectively, we investigated and presented a multiple kernel learning framework for the lung nodule classification problem. Specifically, we have proposed a kernel method for integrating heterogeneous feature subsets for false positive reduction, and optimization is done using proximal gradient optimization algorithm and alternating iterative optimization algorithm, respectively. Through theoretical justifications and empirical studies, we demonstrated the effectiveness of the method $\ell_{2,1}$ MKL on the performance of reducing false positives vertically and horizontally. In the future, we will extend our methods to the detection and false positive reduction of other type of nodule, such as GGO.

6. Acknowledgments

This research was supported by the National Science Foundation for Distinguished Young Scholars of China under Grant (No.71325002), the National Key Technology Research and Development Program of the Ministry of Science and Technology of China under grant (No.2014BAI17B01), the National High Technology Research and Development Program of China (No.2015AA020106), the National Natural Science Foundation of China (No.61502091), the Fundamental Research Funds for the Central Universities (No.N140403004), and the Postdoctoral Science Foundation of China (No.2015M570254).

- [1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. *CA: a cancer journal for clinicians* 2015;65(1):5–29.
- [2] Li Q. Recent progress in computer-aided diagnosis of lung nodules on thin-section CT. *Computerized medical imaging and graphics* 2007;31(4-5):248–57.
- [3] Valente IRS, Cortez PC, Neto EC, Soares JM, de Albuquerque VHC, Tavares JMR. Automatic 3D pulmonary nodule detection in CT images: A survey. *Computer methods and programs in biomedicine* 2016;124:91-107.
- [4] Choi WJ, Choi TS. Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor. *Computer methods and programs in biomedicine* 2014;113(1):37–54.
- [5] Kaya A, Can AB. A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics. *Journal of biomedical informatics* 2015;56:69-79.
- [6] Temesguen M, Russell CH, Steven KR. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. *Medical Image Analysis* 2010;14:390-406.
- [7] Gomathi M, Thangaraj P. Computer aided medical diagnosis system for detection of lung cancer nodules: a survey. *International Journal of Computational Intelligence Research* 2009;5(4):453–62.
- [8] Dhara AK, Mukhopadhyay S, Khandelwal N. Computer-aided detection and analysis of pulmonary nodule from CT images: A survey. *IETE Technical Review* 2012;29(4):265–75.
- [9] Chawla NV, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter* 2004;6(1):1–6.
- [10] Yang X, Zheng Y, Siddique M, Beddoe G. Learning from imbalanced data: a comparative study for colon CAD. In: *Proceedings of the SPIE*. 2008;6915.
- [11] Li C, Shi C, Zhang H, Chen Y, Zhang S. Multiple instance learning for computer aided detection and diagnosis of gastric cancer with dual-energy CT imaging. *Journal of biomedical informatics* 2015;57:358–368.
- [12] Zhang D, Shen D. MultiCost: Multi-stage Cost-sensitive Classification of Alzheimers Disease. In: *Machine Learning in Medical Imaging*. 2011:344–351.
- [13] Gönen Mehmet, Alpayd E. Multiple Kernel Learning Algorithms. *Journal of Machine Learning Research* 2011;12:2211–2268.

- [14] Tiwari P, Kurhanewicz J, Madabhushi A. Multi-kernel graph embedding for detection, Gleason grading of prostate cancer via MRI/MRS. *Medical image analysis* 2013;17(2):219–235.
- [15] Y S, Falck T, Daemen A, Tranchevent LC, Suykens JA, Moor BD, Moreau Y. L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC bioinformatics* 2010;11(1):309–33.
- [16] Zhang D, Wang D, Zhou Y, Yuan H, Shen D. Multimodal classification of Alzheimer’s disease and mild cognitive impairment. *Neuroimage* 2011;55(3):856–867.
- [17] Gu B, Sheng VS, Wang Z, Ho D, Osman S, Li S. Incremental learning for ν -support vector regression. *Neural Networks* 2015;67:140–150.
- [18] Gu B, Sheng VS. A robust regularization path algorithm for ν -support vector classification. *IEEE Transactions on Neural Networks and Learning Systems* 2016.
- [19] Gu B, Sheng VS, Tay KY, Romano W, Li S. Incremental support vector learning for ordinal regression. *IEEE Transactions on Neural networks and learning systems* 2015;26(7):1403–1416.
- [20] Li Q, Doi K. Analysis and minimization of overtraining effect in rule-based classifiers for computer-aided diagnosis. *Medical physics* 2006;33:320–8.
- [21] Ge Z, Sahiner B, Chan HP, Hadjiiski LM, Wei J , Bogot N, et al. Computer aided detection of lung nodules: false positive reduction using a 3-D gradient field method. In: *Proceedings of SPIE* 2004;5370:1076–82.
- [22] Tao Y, Lu L, Dewan H, Chen A, Corso J, Xuan J, et al. Multi-level ground glass nodule detection and segmentation in CT lung images. *Medical Image Computing and Computer-Assisted Intervention–MICCAI*. 2009, p. 715–23.
- [23] Jiang J, Trundle P, Ren J. Medical image analysis with artificial neural networks. *Computerized Medical Imaging and Graphics* 2010;34(8):617–31.
- [24] Choi WK, Choi TS. Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images. *Information Sciences* 2012;212(1):57-78.
- [25] Suzuki K, Armato SG, Li F, Sone S, Doi K. Massive training artificial neural network for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. *Medical physics* 2003;30:1602–17.
- [26] Boroczky L, Zhao LZ, Lee KP. Feature subset selection for improving the performance of false positive reduction in lung nodule CAD. *IEEE Transactions on Information Technology in Biomedicine* 2006;10(3):504–11.

- [27] da Silva Sousa JRF, Silva AC, Paiva AC, Nunes RA. Methodology for automatic detection of lung nodules in computerized tomography images. *Computer methods and programs in biomedicine* 2010;98(1):1–14.
- [28] Keshani M, Azimifar Z, Tajeripour F, Boostani R. Lung nodule segmentation and recognition using SVM classifier and active contour modeling: A complete intelligent system. *Computers in biology and medicine* 2013;43(4):287–300.
- [29] Netto SMB, Silva AC, Nunes RA, Gattass M. Automatic segmentation of lung nodules with growing neural gas and support vector machine. *Computers in biology and medicine* 2012;42(11):1110–1121.
- [30] Campadelli P, Casiraghi E, Valentini G. Support vector machines for candidate nodules classification. *Neurocomputing* 2005;68:281–8.
- [31] Wang Q, Kang W, Wu C, Wang B. Computer-aided detection of lung nodules by svm based on 3D matrix patterns. *Clinical Imaging* 2013;37(1):62–9.
- [32] Lee SLA, Kouzani AZ, Hu EJ. Random forest based lung nodule classification aided by clustering. *Computerized medical imaging and graphics* 2010;34(7):535–42.
- [33] He H, Garcia E. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 2009;21(9):1263–84.
- [34] Han F, Wang H, Zhang G, Han H, Song B, Li L, Moore W, Lu H, Zhao H, Liang Z. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. *Journal of digital imaging* 2015;28(1):99–115.
- [35] Caicedo JC, González FA, Romero E. Content-based histopathology image retrieval using a kernel-based semantic annotation framework. *Journal of biomedical informatics* 2011;44(4):519–528.
- [36] Cao P, Yang JZ, Li W, Zhao DZ, Zaiane O. Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD. *Computerized Medical Imaging and Graphics* 2014;38(3):137–150.
- [37] Dolejsi M, Kybic J, Tuma S, Polovincák M. Reducing false positive responses in lung nodule detector system by asymmetric Adaboost newblock. In: *Proceedings of fifth IEEE International Symposium on Biomedical Imaging ISBI*. 2008, p. 656–9.
- [38] Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence* 2009;31(2):210–227.
- [39] Javidi M, Pourreza HR, Harati A. Vessel segmentation and microaneurysm detection using discriminative dictionary learning and sparse representation. *Computer Methods and Programs in Biomedicine* 2017;139:93–108.

- [40] Xu L, Wu X, Chen K, Yao L. Multi-modality sparse representation-based classification for Alzheimer’s disease and mild cognitive impairment. *Computer Methods and Programs in Biomedicine* 2015;122(2):182–190.
- [41] Taalimi A, Ensafi S, Qi H, Lu S, Kassim AA, Tan CL. Multimodal Dictionary Learning and Joint Sparse Representation for HEP-2 Cell Classification newblock In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2015, p. 308–315.
- [42] Wang D, Wang X, Kong S. Integration of multi-feature fusion and dictionary learning for face recognition. *Image and Vision Computing* 2013;31(12):895–904.
- [43] Cao P, Zhao DZ, Zaiane O. Measure oriented cost-sensitive SVM for 3D nodule detection. In: *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2013, p.3981–3984.
- [44] Marius K, Brefeld U, Sonnenburg S, Laskov P, Müller KR, Zien A. Efficient and accurate lp-norm multiple kernel learning. In: *Advances in neural information processing systems*. 2009, p. 997–1005.
- [45] Zien A, Ong CS. Multiclass Multiple Kernel Learning. In: *Proceedings of the 24th international conference on Machine learning*. 2007, p. 1191–1198.
- [46] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2006;68(1):49–67.
- [47] Dileep AD, Sekhar CSC. Representation and feature selection using multiple kernel learning. In: *International Joint Conference on Neural Networks(IJCNN 2009)*. 2009, p. 717–722.
- [48] Yan J, Li T, Wang H, Huang H, Wan J, Nho K, Risacher SL, Saykin AJ, Shen L. Cortical surface biomarkers for predicting cognitive outcomes using group l2,1 norm. *Neurobiology of aging* 2015;36:185–193.
- [49] Wan J, Zhang Z, Yan J, Li T, Rao BD, Fang S. Sparse Bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer’s disease. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012, p. 940–947.
- [50] Ye J, Farnum M, Yang E, Verbeeck R, Lobanov V, Raghavan N, Novak G, DiBernardo A, Narayan VA. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. *BMC neurology* 2012;12(1):185–193.
- [51] Rakotomamonjy A, Flamary R, Gasso G, Canu S. lp-lq penalty for sparse linear and sparse multiple kernel multitask learning. *IEEE Transactions on Neural Networks* 2011;22(8):1307–20.

- [52] Li M, Zhou Z. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 2012;37(6):1088–1098.
- [53] James AP, Dimitrijevic S. Nearest neighbor classifier based on nearest feature decisions. *The Computer Journal* 2012.
- [54] Xiang S, Yuan L, Fan W, Wang Y, Thompson PM, Ye J. Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage* 2014;102:192–206.
- [55] Anirudh R, Thiagarajan JJ, Bremer T, Kim H. Lung nodule detection using 3D convolutional neural networks trained on weakly labeled data. In: *SPIE Medical Imaging*. 2016, p. 978532–978532.
- [56] Van G, Bram S, Arnaud AA, Jacobs C, Ciompi F. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In: *SPIE Medical Imaging*. 2015, p. 286–289.
- [57] Setio AAA, Ciompi F, Litjens G, Gerke P, Wille MMW, Naqibullah M, Sánchez CI, van Ginneken B. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging* 2016;35(5):1160–1169.
- [58] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In: *Proceedings of the 22nd ACM international conference on Multimedia* 2014:675–678.
- [59] Rakotomamonjy A, Bachin FC, Canu S, Grandvalet Y. SimpleMKL. *Journal of Machine Learning Research* 2008;9:2491–2521.
- [60] Beck A, Marc T. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. In: *SIAM journal on imaging sciences* 2009:183–202.
- [61] Rakotomamonjy A, Flamary R, Gasso G, Canu S. lp-lq Penalty for sparse linear and sparse multiple kernel multitask learning. *IEEE Transactions on Neural Networks* 2011;22(8):1307–1320.
- [62] Liu J, Ji S, Ye J. Multi-task feature learning via efficient l2, 1-norm minimization. In: *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence* 2009:339–348.
- [63] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 2002;16(1):321–57.
- [64] Armato III, Samuel G, McLennan G, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics* 2011;38(2):915–931.

- [65] Li Q, Sone S, Doi K. Selective enhancement filters for nodules, vessels, and airway walls in two-and three-dimensional CT scans. *Medical physics* 2003;30:2040–51.
- [66] Li Q, Li F, Doi K. Computerized detection of lung nodules in thin-section ct images by use of selective enhancement filters and an automated rule-based classifier. *Academic radiology* 2008;15(2):165–75.
- [67] Hansell DM, Bankier AA, MacMahon HM, Theresa C, Muller NL, Remy J. Fleischner society: glossary of terms for thoracic imaging. *Radiology* 2008;246(3):697–722.
- [68] Yang M, Periaswamy S, Wu Y. False positive reduction in lung ggo nodule detection with 3d volume shape descriptor. In: *IEEE International Conference on Acoustics, Speech and Signal Processing* 2007;1:437–40.
- [69] Lu X, Wei GQ, Qian J, Jain AK. Learning-based pulmonary nodule detection from multislice CT data. In: *Proceedings of CARS 2004 Computer Assisted Radiology and Surgery*. 2004, p.1356–63.
- [70] Yoshida H, Nappi J. Three-dimensional computer-aided diagnosis scheme for detection of colonic polyps. *IEEE Transactions on Medical Imaging* 2001;20(12):1261–74.
- [71] Koenderink JJ. *Solid shape*. Cambridge Univ Press; 1990.
- [72] Kawata Y, Niki H, Ohmatsu H, Kakinuma R, Eguchi K, Kaenko M, et al. Quantitative surface characterization of pulmonary nodules based on thin-section CT images. *IEEE Transactions on Nuclear Science* 1998;45(4):2132–8.
- [73] Zhu Z, Shen L, Sun Y, He S, Ji Z. Memetic three-dimensional gabor feature extraction for hyperspectral imagery classification. In: *Advances in Swarm Intelligence*. 2012:479–488.
- [74] Kobatake H, Murakami M. Adaptive filter to detect rounded convex regions: Iris filter. In: *Proceedings of the 13th International Conference on Pattern Recognition*. 1996, p. 340–4.
- [75] Fan W, Stolfo SJ, Zhang J, Chan PK. AdaCost: misclassification cost-sensitive boosting. In: *Proceedings of international conference machine learning*. 1999, p. 97–105.
- [76] Domingos P. MetaCost: a general method for making classifiers cost-sensitive. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. 1999, p. 155–64.
- [77] Tomek I. Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics* 1976;6:769–72.

- [78] Chen S, He H, Garcia EA. RAMOBoost: Ranked minority oversampling in Boosting. *IEEE Transactions on Neural Networks*. 2010;21(10):1624–1642.
- [79] Chawla NV, Lazarevic A, Hall LO, Bowyer KW. SMOTEBoost: Improving prediction of the minority class in Boosting. *Knowledge Discovery in Databases: PKDD*. 2003, p. 107–19.
- [80] Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. 2003.
- [81] Zyout I, Czajkowska J, Grzegorzec M. Multi-scale textural feature extraction and particle swarm optimization based model selection for false positive reduction in mammography. *Computerized Medical Imaging and Graphics*. 2015, **46**: 95–107.
- [82] Terada T, Fukumizu Y, Yamauchi H, Chou H, Kurumi Y. Detecting mass and its region in mammograms using mean shift segmentation and Iris Filter. In *2010 International Symposium on Communications and Information Technologies (ISCIT)*, 2010, p.1176–1179.
- [83] Wu WJ, Lin SW, Moon WY. Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. *Computerized Medical Imaging and Graphics*. 2012;36(8):627–633.