Structural Feature Selection for Connectivity Network-Based MCI Diagnosis

Biao Jie[1,2], Daoqiang Zhang[1,2], Chong-Yaw Wee[2], and Dinggang Shen[2]

[1] Dept. of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[2] Dept. of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27599
{jbiao,dqzhang}@nuaa.edu.cn, dgshen@med.unc.edu

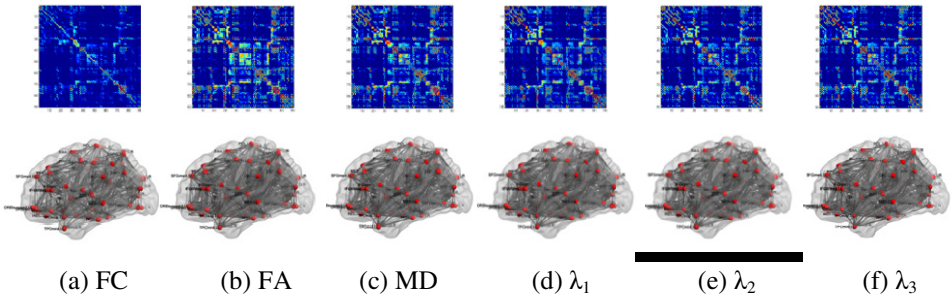
Abstract. Connectivity networks have been recently used for classification of neurodegenerative diseases, e.g., mild cognitive impairment (MCI). In typical connectivity network-based classification, features are often extracted from (multiple) connectivity networks and concatenated into a long vector for subsequent feature selection and classification. However, some useful network topological information may be lost in this type of approach. In this paper, we propose a new structural feature selection method which embeds the topological information of connectivity networks through graph kernel and then uses recursive feature elimination with graph kernel (RFE-GK) to select the most discriminative features. Furthermore, multiple kernel learning (MKL) is also adopted to combine multiple graph kernels for joint structural feature selection from multiple connectivity networks. The experimental results show the efficacy of our proposed method with comparison to the state-of-the-art method in MCI classification, based on the connectivity networks.

1 Introduction

Many methods have been developed for classification of Alzheimer's disease (AD) or its prodromal stage, i.e., mild cognitive impairment (MCI), based on either single or multiple modalities of biomarkers. Recently, connectivity networks have been used for diagnosis and classification of neurodegenerative diseases, e.g., schizophrenia and MCI [1, 2]. For example, Wee et al. proposed an effective network-based classification method to accurately identify MCI patients by using a collection of measures derived from white matter (WM) connectivity networks [1]. In their method, six types of connectivity networks as shown in Fig. 1 were first constructed from each subject by using six different physiological parameters, i.e., fiber count (FC), fractional anisotropy (FA), mean diffusivity (MD), and principal diffusivities (λ_1, λ_2, and λ_3), based on the parcellated 90 regions-of-interest (ROIs) of the brain, and then the clustering coefficient of each ROI in relation to the remaining ROIs is extracted from these connectivity networks as features for subsequent feature selection and classification.

In the conventional feature selection and classification methods, which solely rely on the feature vector obtained from previous feature extraction step (e.g., in [1]),

P.-T. Yap et al. (Eds.): MBIA 2012, LNCS 7509, pp. 175–184, 2012.



(a) FC (b) FA (c) MD (d) λ_1 (e) λ_2 (f) λ_3

Fig. 1. Six different connectivity matrices (networks) used in our structural feature selection method

some useful network topological information may be lost and thus the performance may be affected. In this paper, we address this problem by proposing a new structural feature selection method that preserves the network topological information to guide the feature selection. Specifically, graph kernel [3-5] is adopted to measure the topological similarity between a pair of graphs (i.e., connectivity networks) of individual subjects, and then support vector machine (SVM) with graph kernel is used to select the most discriminative features based on similar technique as recursive feature elimination (RFE) [6]. Moreover, to deal with multiple connectivity networks of each subject, we use multiple kernel learning (MKL) [7] to combine the graph kernels of each connectivity network for joint structural feature selection. The proposed method is evaluated on 10 MCI patients and 17 healthy controls, and promising experimental results are obtained.

2 Materials and Method

2.1 Materials

In this study, 10 MCI patients and 17 socio-demographically matched healthy controls were recruited. Informed consent was obtained from all participants, and the experimental protocols were approved by the institutional ethics board. All the recruited subjects were diagnosed by expert consensus panels.

Data acquisition was performed using a 3.0-Tesla GE Signa EXCITE scanner. Diffusion-weighted images of each participant were acquired axially parallel to the anterior and posterior commissures (AC-PC) line with twenty-five-direction diffusion-weighted whole-brain volumes using diffusion weighting values: b=0 and 1000s/mm^2, flip angle=90°, TR/TE=17000/78ms. The imaging matrix=128×128 and FOV=256×256 mm^2 were used, leading to a voxel size of 2×2×2 mm^3. A total of 72 contiguous slices were acquired.

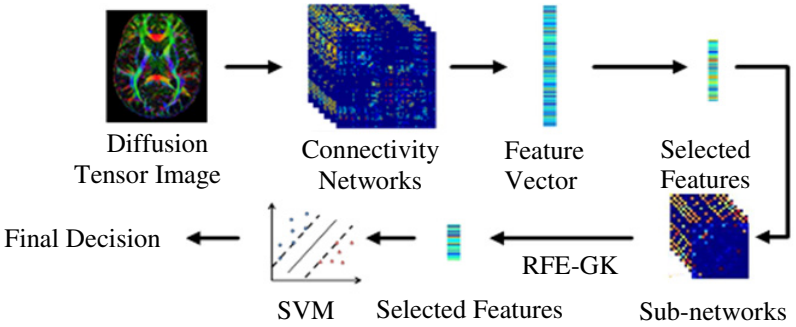


Fig. 2. Flowchart of the proposed method

2.2 Overview of Methodology

Fig. 2 shows the flowchart of the proposed method for connectivity network-based feature selection and classification, which contains three main steps:

i) Connectivity Network Construction and Feature Extraction. We followed the same procedure as in [1]. Specifically, each brain image was first parcellated into 90 ROIs by propagating the Automated Anatomical Labeling (AAL) ROIs [8] to each image using a deformable DTI registration algorithm called F-TIMER [9] with tensor orientation corrected using the method described in [10], and then six connectivity networks were constructed according to these ROIs and six physiological parameters, i.e., FC, FA, MD, and principal diffusivities ($\lambda 1$, $\lambda 2$, and $\lambda 3$). Note that each connectivity network is a 90×90 symmetric matrix. After we obtained the connectivity networks, we computed the clustering coefficient [1] of each ROI in relation to the remaining ROIs from connectivity networks as features for subsequent feature selection and classification. It is worth noting that each ROI in each connectivity network corresponds to one feature, and thus there are totally 540 features (90 ROIs × 6 connectivity networks) for each subject.

ii) Feature Selection. We adopted a two-stage feature selection strategy. Specifically, a standard t-test was first performed to screen out those features that are not significant for discrimination between MCI patients and healthy controls, i.e., those features with p-value larger than a given threshold (0.05 in this paper) will be omitted. Since each feature corresponds to a ROI or node in each of six different connectivity networks, six sub-networks can be constructed for each selected ROI or network node in each subject, after eliminating those non-significant features (ROIs or nodes) from all six original connectivity networks. Then, we measured the topological similarity by computing graph kernel between pair of sub-networks from same network type across different subjects. Furthermore, to deal with multiple connectivity networks available in each subject, we used MKL technique to combine the graph kernels from different connectivity networks. Finally, we used the RFE with the above learned

graph kernel, denoted as RFE-GK, to select the most discriminative features. It is worth noting that in each iteration step of RFE-GK, we need to re-compute the graph kernel according to the current remained feature subsets.

iii) Classification. A SVM classifier was adopted to identify the MCI patients from healthy controls by using the features selected in the previous steps. The classifier training of standard SVM is implemented using LIBSVM toolbox [11], with a linear kernel and a default value for the parameter C (i.e., C=1). Here, following [1], a nested Leave-One-Out (LOO) cross-validation strategy is used to enhance the generalization power of the classifier and to avoid the over-fitting on small sample dataset. The inner cross-validation loop was performed on the training data to decide the number of selected features and hyperparameter of the SVM models while the outer cross-validation loop was used to evaluate the generalizability of SVM models using unseen subjects.

2.3 Topology-Based Graph Kernel

Kernel-based learning methods work by first embedding the data into a higher dimensional feature space, and then searching for linear relations among the embedding data points. Given two subjects (vectors) x and x', the kernel can be defined as $k(x, x') = \langle \emptyset(x), \emptyset(x') \rangle$, where $\emptyset$ is a mapping function that maps data from original subject data space to feature space. Examples of common kernel function are linear function and Gaussian radial basis functions (RBF). Besides using feature vector, kernel can also be defined on more complex data types, e.g., graph, and the corresponding kernel is called graph kernel, which captures the semantics inherent in the graph structure. A number of methods have been proposed to define graph kernel [3-5], and have been successfully applied to a variety of problems such as image classification [3] and protein function prediction [5].

To define the graph kernel, some basic terms are first introduced. Here, a labeled graph G is defined as a triple (V, E, ℓ), where V is the set of vertices, E is the set of undirected edges, and $\ell: V \to L$ is a function that assigns labels from an alphabet L to nodes. A walk is a finite sequence of neighboring vertices, while a path is a walk such that all its vertices are distinct. A subtree is a subgraph of a graph, which has no cycles (i.e., any two vertices are connected by exactly one simple path). Subtree pattern extends the notion of subtree by allowing repetitions of nodes and edges. However, these same nodes (edges) are treated as distinct nodes (edges). Fig. 3 illustrates an example for subtree pattern. Comparing with path and walk, subtree pattern has better discriminative power to measure the similarity between graphs [3], thus it is used in this paper.

Given a pair of graph G and H, a graph kernel can be defined as $k(G, H) = \langle \emptyset(G), \emptyset(H) \rangle$, which takes into account the topology of the graph G and H. Generally, the computational complexity of graph kernel is very high. In order to improve the computational efficiency, Shervashidze and Borgwardt [4] proposed a new method to construct the subtree kernel based on the Weisfeiler-Lehman test of isomorphism. The basic process of the 1-dimensional Weisfeiler-Lehman test is as follows: First, every vertex of a graph is labeled with the number of edges connected to that vertex. Then,

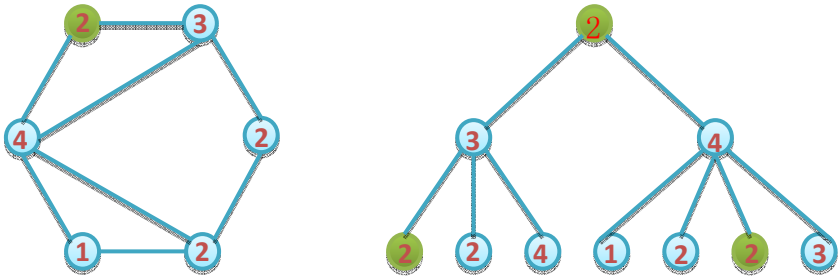


Fig. 3. A subtree pattern of height 2 rooted at the node 2

at each subsequent step (or iteration), the label of each vertex is updated based on its previous label and the label of its neighbors, i.e., parallelly augment the label of each vertex in graph by the sorted set of node labels of neighboring nodes, and compress these augmented labels into new, short labels. This process proceeds iteratively until the node label sets of two graphs differ, or the number of iteration reaches the maximum h. If the sets of new created labels are identical after h iteration, we cannot determine whether these two graphs are isomorphic or not.

Let G_0 and H_0 represent the initial two labeled graphs respectively, G_i and H_i be the corresponding labeled graphs at i-th iteration $(i = 1, \dots, h)$, and $L_i = \{s_{i1}, s_{i2}, \dots, s_{i|L_i|}\}$ be the set of letters that occur as node labels in G_i or H_i. Assume all L_i are pairwise disjointed. Without loss of generality, assume every L_i is ordered, then the Weisfeiler-Lehman subtree kernel on two graphs is defined as:

$$k(G, H) = \langle \emptyset_{WL}^{(h)}(G), \emptyset_{WL}^{(h)}(H) \rangle, \tag{1}$$

where

$$\emptyset_{WL}^{(h)}(G) = (\sigma_0(G, s_{01}), \dots, \sigma_0(G, s_{0|L_0|}), \dots, \sigma_h(G, s_{h1}), \dots, \sigma_h(G, s_{h|L_h|}))$$

and

$$\emptyset_{WL}^{(h)}(H) = (\sigma_0(H, s_{01}), \dots, \sigma_0(H, s_{0|L_0|}), \dots, \sigma_h(H, s_{h1}), \dots, \sigma_h(H, s_{h|L_h|})),$$

with $\sigma_i(G, s_{ij})$ and $\sigma_i(H, s_{ij})$ are the number of occurrences of the letter s_{ij} in G and H, respectively. Intuitively, the Weisfeiler-Lehman subtree kernel counts the common original and compressed labels in two graphs. It can be proved that this kind of kernel is positive definite and the computational complexity for N graphs is $O(Nhm + N^2hn)$, where n and m are the numbers of nodes and edges of graphs, respectively [4]. In our method, we compute the graph kernel based on the above algorithm on a pair of same type sub-networks across different subjects, as shown in Fig. 2.

2.4 Recursive Feature Elimination with Graph Kernel (RFE-GK)

To deal with multiple connectivity networks, we adopt the MKL technique which is the process of learning a mixed kernel from multiple basis kernels. Given a series of training connectivity networks represented as $G^i = \{G_m^i\}_{m=1}^M$, with corresponding class labels $y^i, i = 1, \ldots, n$, where M is the number of connectivity networks of each subject, and n is the number of subjects, generally the mixed kernel can be learned through a linear combination of multiple basis kernels as below:

$$k(G^i, G^j) = \sum_{m=1}^{M} \mu_m k_m(G_m^i, G_m^j), \tag{2}$$

where $k_m(G_m^i, G_m^j)$ is the Weisfeiler-Lehman subtree kernel defined on the m-th connectivity network using Eq. 1, and μ_m is a nonnegative weight parameter with $\sum_{m=1}^{M} \mu_m = 1$. We adopt the simple MKL algorithm proposed in [7] to solve Eq. 2. Finally, we will perform structural feature selection using RFE based on graph kernel, which we call as RFE-GK, as shown in Algorithm 1.

It is worth noting that our RFE-GK method is different from the standard RFE framework. In the standard RFE-SVM, it eliminates features through ranking based on weight vector of a linear SVM, while in our method features are eliminated based on classification accuracy in a wrapper-like way. Moreover, there are two other key differences between the proposed feature selection method and the standard RFE-SVM, i.e., 1) the former uses graph kernel that preserves the topological (structural) information of data while the latter uses standard kernel on vector-type data without considering the structural information, and 2) the former uses MKL to combine multiple kernels from multiple connectivity networks to select features while the latter uses single kernel.

3 Experimental Results

3.1 Evaluating Classification Performance

A LOO cross-validation strategy was used to evaluate the classification performance. The performance of a classifier could be quantified by using accuracy, area under receiver operating characteristic curve (AUC), sensitivity and specificity, where the sensitivity represents the proportion of patients that are correctly predicted, and the specificity denotes the proportion of health controls that are correctly predicted. We compared our method with Wee's method and their classification performances are summarized in Table 1. Fig. 4 shows the ROC curves of compared methods. It is worth noting that in Table 1, 'All' denotes using all six connectivity networks, while FC, FA, MD, λ_1, λ_2 and λ_3 denote only using the individual connectivity network, respectively. Also, in Fig. 4 the proposed method and Wee's method [1] use all six connectivity networks, while the others using only single connectivity network.

As can be seen from Table 1 and Fig. 4, our method performs the best in terms of classification accuracy and AUC values. Specifically, our method achieves a classification accuracy of 92.59% and a AUC of 0.965, which are higher than the state-of-the-art connectivity networks-based MCI classification method [1], which achieves

Algorithm 1. Recursive Feature Elimination with Graph Kernel (RFE-GK)
Input: Training connectivity networks represented as $G_i = \{G_m^i\}_{m=1}^M, i = 1, \ldots, n$, and corresponding class labels $y^i, i = 1, \ldots, n$.
Output: Ranked feature (ROI) list F
Initialize: Subset of surviving features (ROIs) $S = [1,2, \ldots, d]$, and ranked feature list $F = [\,]$, where d is the number of surviving features in previous feature selection step.
Repeat until $S = [\,]$
  For each $s \in S$
    For each pair of G^i and G^j, $i \neq j$, compute the combined graph kernel using Eq. 2 on sub-networks excluding feature (or ROI) s;
    Train Leave-One-Out SVM and get the accuracy
  End for
  Find s^* with corresponding maximum accuracy;
  Update ranked feature list $F = [s^*, F]$;
  Eliminate s^* from S
End repeat

the best classification accuracy of 88.89% and AUC of 0.929. Our proposed method successfully classified all the MCI patients while only misclassified 2 healthy controls. Table 1 also indicates that combining multiple connectivity networks will achieve much better performance than using single connectivity network alone.

Table 1. Comparison of classification performance of different methods using different connectivity networks

	Accuracy (%)		AUC		Sensitivity (%)		Specificity (%)	
	Wee's	Ours	Wee's	Ours	Wee's	Ours	Wee's	Ours
FC	70.37	70.37	0.653	0.565	-	60	-	76.47
FA	74.07	66.67	0.859	0.565	-	40	-	82.35
MD	59.26	74.07	0.647	0.641	-	40	-	94.12
λ_1	59.26	66.67	0.629	0.729	-	70	-	64.71
λ_2	55.56	62.96	0.594	0.582	-	30	-	82.35
λ_3	59.26	55.56	0.612	0.471	-	40	-	64.71
All	88.89	92.59	0.929	0.965	-	100	-	88.24

3.2 Evaluating Discriminative Power of Features

In this subsection, we evaluate the discriminative power of the selected features using the locality-preserving projection (LPP) approach [12]. Specifically, we used LPP to project the selected features by our method and Wee's method [1] into a 2-D space for visualization. For comparison, we also project the original features (i.e., 540 features) using LPP. Fig. 5 shows the 2-D visualization results of different methods. As can be

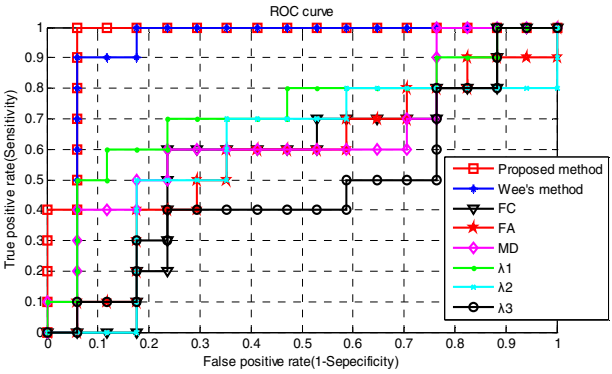


Fig. 4. ROC curves of different methods

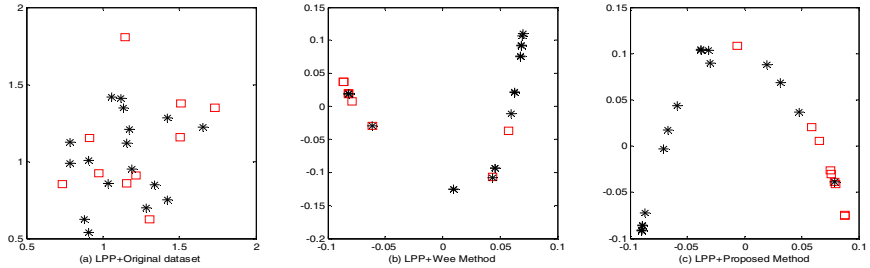


Fig. 5. The 2-D visualization results of different methods. Each point represents a subject, and different shapes (or colors) indicate different class labels.

seen from Fig. 5, both feature selection methods (including ours and Wee's) can enhance the discrimination ability of features compared with the original features, with the selected features by our method being more discriminative than those by Wee's method.

3.3 The Most Discriminative Regions

To determine the most discriminative regions, we counted the frequency of features selected at different LOO folds, and ranked them according to their frequencies. The experimental results show that the most frequently selected ROIs include right rectus gyrus, right insuta, right superior temporal gyrus, and left Inferior frontal gyrus (triangular), which are in agreement with the previously reported brain regions related to MCI [1, 13, 14].

Furthermore, the "exclude-one" strategy was adopted to estimate the contribution of each selected ROI. That is, each time one ROI was excluded from the selected ROIs, and the classification performance was investigated using the remaining ROIs; the experiment was repeated for all 4 most frequently selected ROIs. The classification

accuracies and AUC values are summarized in Table 2. As can be seen from Table 2, right rectus gyrus and right insula have the most discriminative power, because excluding either of them will cause the most apparent decline in both classification accuracy and AUC.

Table 2. The classification performance under the "exclude-one" strategy

Excluded ROI	Accuracy (%)	AUC
Right rectus gyrus	66.67	0.670
Right insula	70.37	0.759
Right superior temporal gyrus	81.48	0.865
Left Inferior frontal gyrus (triangular)	81.48	0.894

4 Conclusion

In this paper, we have proposed a novel structural feature selection method based on graph kernel, for enhancing the connectivity networks-based classification. Specifically, graph kernel was used to measure the topological similarity between two subnetworks of subjects. Furthermore, multiple kernel learning was used to combine multiple graph kernels obtained from multiple connectivity networks. Experimental results on MCI classification show not only significant improvement of classification performance in terms of accuracy and AUC value, but also show great potential of our method in detecting sensitive ROI regions for MCI. In the future, we will investigate combining the graph kernel used in the current study with the existing linear or Gaussian kernels for further improving performance.

Reference

1. Wee, C.Y., Yap, P.T., Li, W., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D.: Enriched white matter connectivity networks for accurate identification of MCI patients. Neuroimage 54, 1812–1822 (2011)
2. Shen, H., Wang, L.B., Liu, Y.D., Hu, D.W.: Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. Neuroimage 49, 3110–3121 (2010)
3. Harchaoui, Z., Bach, F.: Image classification with segmentation graph kernels. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, Vols. 1-8, pp. 612–619 (2007)
4. Shervashidze, N., Borgwardt, K.M.: Fast subtree kernels on graphs. In: Advances in Neural Information Processing Systems 22, pp. 1660–1668 (2009)
5. Borgwardt, K.M., Kriegel, H.-P.: Shortest-path kernels on graphs. In: Fifth IEEE International Conference on Data Mining, pp. 74–81 (2005)
6. Rakotomamonjy, A.: Variable selection using SVM based criteria. Journal of Machine Learning Research 3, 1357–1370 (2003) (special issue on special feature)
7. Rakotomamonjy, A., Bach, F.R., Canu, S., Grandvalet, Y.: MKLsimple. Journal of Machine Learning Research 9, 2491–2521 (2008)

8. Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M.: Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273–289 (2002)
9. Yap, P.T., Wu, G., Zhu, H., Lin, W., Shen, D.: F-TIMER: fast tensor image morphing for elastic registration. IEEE Trans. Med. Imaging 29, 1192–1203 (2010)
10. Xu, D., Mori, S., Shen, D., van Zijl, P.C., Davatzikos, C.: Spatial normalization of diffusion tensor fields. Magn. Reson. Med. 50, 175–182 (2003)
11. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001)
12. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. Adv. Neur. In. 14, 585–591 (2002)
13. Lenzi, D., Serra, L., Perri, R., Pantano, P., Lenzi, G.L., Paulesu, E., Caltagirone, C., Bozzali, M., Macaluso, E.: Single domain amnestic MCI: a multiple cognitive domains fMRI investigation. Neurobiol. Aging 32, 1542–1557 (2011)
14. Schroeter, M.L., Stein, T., Maslowski, N., Neumann, J.: Neural correlates of Alzheimer's disease and mild cognitive impairment: A systematic and quantitative meta-analysis involving 1351 patients. Neuroimage 47, 1196–1206 (2009)