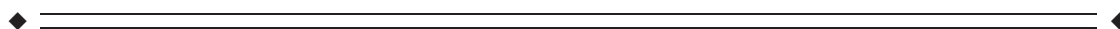


Multivariate Classification of Smokers and Nonsmokers Using SVM-RFE on Structural MRI Images

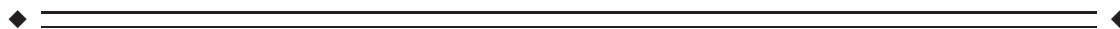
Xiaoyu Ding, Yihong Yang, Elliot A. Stein, and Thomas J. Ross*

Neuroimaging Research Branch, Intramural Research Program, National Institute on Drug Abuse, National Institutes of Health, Baltimore, Maryland



Abstract: Voxel-based morphometry (VBM) studies have revealed gray matter alterations in smokers, but this type of analysis has poor predictive value for individual cases, which limits its applicability in clinical diagnoses and treatment. A predictive model would essentially embody a complex biomarker that could be used to evaluate treatment efficacy. In this study, we applied VBM along with a multivariate classification method consisting of a support vector machine with recursive feature elimination to discriminate smokers from nonsmokers using their structural MRI data. Mean gray matter volumes in 1,024 cerebral cortical regions of interest created using a subparcellated version of the Automated Anatomical Labeling template were calculated from 60 smokers and 60 nonsmokers, and served as input features to the classification procedure. The classifier achieved the highest accuracy of 69.6% when taking the 139 highest ranked features via 10-fold cross-validation. Critically, these features were later validated on an independent testing set that consisted of 28 smokers and 28 nonsmokers, yielding a 64.04% accuracy level (binomial $P = 0.01$). Following classification, exploratory post hoc regression analyses were performed, which revealed that gray matter volumes in the putamen, hippocampus, prefrontal cortex, cingulate cortex, caudate, thalamus, pre-/post-central gyrus, precuneus, and the parahippocampal gyrus, were inversely related to smoking behavioral characteristics. These results not only indicate that smoking related gray matter alterations can provide predictive power for group membership, but also suggest that machine learning techniques can reveal underlying smoking-related neurobiology. *Hum Brain Mapp* 36:4869–4879, 2015. Published 2015. This article is a U.S. Government work and is in the public domain in the USA.

Key words: smoking addiction; structural MRI; voxel-based morphometry; support vector machine; recursive feature elimination; multivariate classification



Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: Intramural Research Program of the National Institute on Drug Abuse and FDA (to E.A.S.); Contract grant number: NDA13001-001-00000.

*Correspondence to: Thomas J. Ross, Neuroimaging Research Branch, Intramural Research Program, National Institute on Drug Abuse, National Institutes of Health, 251 Bayview Blvd., Baltimore, MD 21224. E-mail: tross@mail.nih.gov

Received for publication 31 December 2014; Revised 13 August 2015; Accepted 17 August 2015.

DOI: 10.1002/hbm.22956

Published online 24 October 2015 in Wiley Online Library (wileyonlinelibrary.com).

INTRODUCTION

Cigarette smoking is the most common form of recreational drug use despite its association with numerous negative health consequences on multiple organ systems [Fagerström, 2002; Yanbaeva et al., 2007]. Brain pathology, including stroke and silent brain infarcts, are also associated with chronic smoking [Swan and Lessov-Schlaggar, 2007; Whincup et al., 2004]. Despite these serious and often lethal consequences, about 20% of the United States population (and often higher percentages elsewhere) remain active smokers [Agaku et al., 2014]. Since treatment outcomes for smoking addiction is notoriously poor [Gifford et al., 2004; Reitzel et al., 2011; Rose et al., 1994], and

Published 2015. This article is a U.S. Government work and is in the public domain in the USA.

the negative consequences widely appreciated, better treatments are desperately needed. One reason for the poor treatment outcomes is that, as for many neuropsychiatric diseases, there are no clinically useful addiction disease biomarkers, which if available, could directly lead to improve treatment outcomes.

Magnetic resonance imaging (MRI) provides an effective and noninvasive approach to assess damage to the central nervous system. Voxel-based morphometry (VBM), which measures structural gray matter volume/density (GMV/GMD) [Ashburner and Friston, 2000], has consistently identified brain morphological abnormalities in smokers. These include reduced regional GMV/GMD in the prefrontal cortex (PFC), anterior cingulate cortex (ACC), thalamus, and the insula [Brody et al., 2004; Fritz et al., 2014; Gallinat et al., 2006; Liao et al., 2012; Zhang et al., 2011].

The above gray-matter alterations were identified using conventional univariate analysis in which a voxel-by-voxel comparison of GMV/GMD is applied to groups of smokers and controls in order to identify regions of statistical difference. While this type of statistical comparison can help to localize differences in brain regions as a function of smoking addiction, it cannot generally differentiate the samples from two or more groups (i.e. it is not very useful in detecting group membership). Moreover, univariate approaches treat each voxel independently, which is an overly simplistic assumption of brain structural organization [Bullmore and Sporns, 2012; He et al., 2007].

In contrast to univariate approaches, machine-learning-based pattern classification is a class of multivariate analyses. These types of analyses learn discriminative rules from an exemplar dataset that can subsequently automatically categorize group membership of a novel data sample. Machine learning techniques have been applied to structural MRI (sMRI) data in multiple brain disorders including schizophrenia [Kasperek et al., 2011; Nieuwenhuis et al., 2012], autism [Calderoni et al., 2012; Ecker et al., 2010; Jiao et al., 2010; Uddin et al., 2011], dementia [Chen and Herskovits, 2010; Oliveira et al., 2010], depression [Gong et al., 2011], supranuclear palsy/Parkinson syndrome [Focke et al., 2011] and borderline personality disorder [Sato et al., 2012]. Among these techniques, support vector machines (SVM) [Cortes and Vapnik, 1995], which determine a hyperplane that optimally distinguishes samples into two groups (e.g., patients and controls), have been widely used due to their reliable performance when handling high dimensional data. They are usually embedded in a cross-validation framework, in which samples are alternatively treated as testing data in order to validate the trained classification model.

Our group previously applied SVM-based classification to resting-state functional connectivity data from 21 smokers and 21 nonsmokers to predict their smoking status [Pariyadath et al., 2014]. Three network characteristics, including network representativeness, within network connectivity, and between network connectivity were tested.

TABLE I. Demographics of the participants

	Smokers	Healthy controls
Number	88	88
Age	32.3 ± 8.7	31.6 ± 8.0
Gender	44 M, 44 F	44 M, 44 F
FTND	5.4 ± 1.9	–
CPD	19.8 ± 7.2	–
Smoking years	14.0 ± 7.6	–
Lifetime usage	13.7 ± 9.4	–

FTND: Fagerström Test for Nicotine Dependence; CPD: cigarettes per day; Lifetime usage: measured in pack-years (=CPD × smoking years/20); M/F: male/female.

Age, FTND, CPD, smoking years and lifetime usage are calculated as mean ± SD.

Among these, within network connectivity offered maximal information for predicting smoking status with an accuracy of 78.6% using leave-one-out cross-validation (LOOCV). To the best of our knowledge, no study has yet utilized machine learning techniques with sMRI data from smokers (or any other drug dependent population) to identify disease characteristics. Characterizing useful biomarkers and developing effective diagnostic models will benefit not only clinical diagnoses but also treatment outcome by using discriminant features to identified potential novel treatment targets. Thus, the aim of our study was to: (1) classify smokers from nonsmokers using sMRI data with the aid of a SVM embedded cross-validation machine learning approach; and (2) determine those gray matter regions that are the most important discriminative features. In addition to the cross-validation method, we also validated our model on a completely independent dataset of smokers and nonsmokers, something rarely reported in the literature. We further explored the relationship between identified discriminative features and smoking characterization measures using regression analyses.

MATERIALS AND METHODS

Participants

Eighty-eight cigarette smokers and 88 nonsmoking healthy control participants (see Table I for demographics) were enrolled under several protocols approved by the Institutional Review Board of the National Institute on Drug Abuse Intramural Research Program (NIDA-IRP). All participants provided written informed consent and received monetary compensation for their participation. None of the smokers were currently trying to quit or seeking smoking cessation treatment. Controls were included if they had smoked fewer than 25 cigarettes in their lifetime and none in the past year. Potential participants were assessed with a comprehensive history and physical exam, general urine and blood laboratory panels, a computerized Structured Clinical Interview for DSM-IV with follow-up

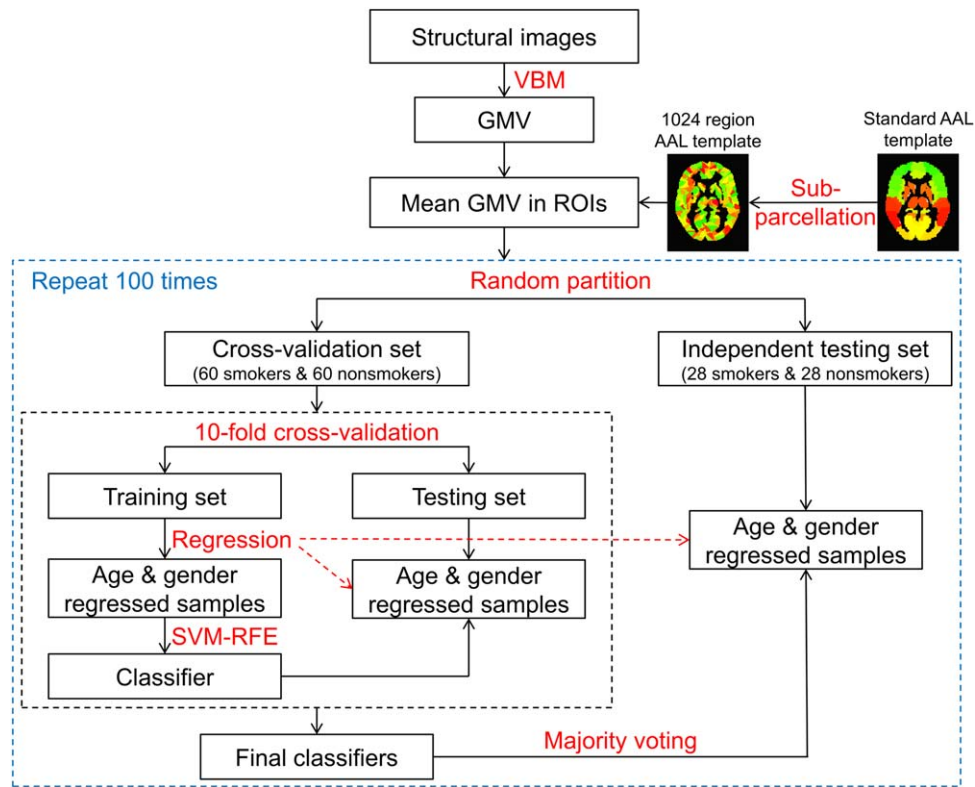


Figure 1.

Flow diagram of the classification approach employed in the study. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

clinical interview, and a drug use survey. Participants were excluded if they had any major medical illness, history of neurological or psychiatric disorders, or current or past dependence on any drug other than nicotine.

Data Acquisition

Structural MRI data were collected at the NIDA-IRP on a 3 T Siemens Allegra MRI scanner (Erlangen, Germany) equipped with a standard radio frequency birdcage head coil. High-resolution anatomical images were acquired using a three-dimensional (3D) magnetization prepared rapid gradient-echo (MPRAGE) T1-weighted sequence in 1mm^3 isotropic voxels (TR = 2,500 ms, TE = 4.38 ms, flip angle = 8° , FOV = 256×256 mm).

Analysis Overview

Figure 1 illustrates the overall analytical pipeline of our classification approach. VBM analysis was applied to the structural images to calculate the GMV in 1,024 cerebral cortical regions of interest (ROIs) defined using a sub-parcellated version of the Automated Anatomical Labeling (AAL) template [Cao et al., 2013; Tzourio-Mazoyer et al., 2002; Zalesky et al., 2010] described below, and whose

mean values were input to the classification procedure. Before classification, the entire dataset was randomly divided into a cross-validation set (60 smokers and 60 nonsmokers) and an independent testing set (28 smokers and 28 nonsmokers). To identify the set of features with the highest discriminative power, a SVM with recursive feature elimination (SVM-RFE) algorithm embedded in a balanced 10-fold cross-validation framework was performed on the cross-validation set. Before SVM-RFE, GMV in the training set was regressed against age and gender [Fjell and Walhovd, 2010; Xu et al., 2000]. The resulting beta-maps were applied to both the cross-validation and independent testing sets to calculate GMV residuals, which served as input features. The resulting models were then validated against the independent testing set. Final validation results for each independent testing sample was determined via majority voting of all of the SVM classifiers from the 10-fold cross-validation step. The entire procedure was repeated 100 times to avoid biased selection of the independent testing set.

Voxel-Based Morphometry

GMV calculation was conducted using SPM8 software (Statistical Parametric Mapping, Wellcome Department of

Imaging Neuroscience, London, UK, <http://www.fil.ion.ucl.ac.uk/spm/software/spm8/>) and VBM8 toolbox (<http://dbm.neuro.uni-jena.de/vbm/>). The VBM8 toolbox in SPM8 extends the unified segmentation model [Ashburner and Friston, 2005] with Maximum A Posterior (MAP) [Rajapakse et al., 1997] and Partial Volume Estimation (PVE) [Tohka et al., 2004] techniques, which achieves a more accurate brain segmentation. Structural images were skull stripped and spatially normalized to MNI (Montreal Neurological Institute) space [Ashburner, 2007] at a resolution of $1.5 \times 1.5 \times 1.5 \text{ mm}^3$, then segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) using the following parameter settings: bias regularization = 0.0001, bias full width at half maximum (FWHM) cutoff = 60 mm, sampling distance = 3, hidden Markov random field (MRF) weighting = 0.15, and thorough cleanup of the partitions. The normalized segmented images were nonlinear modulated, which ensured that further statistical comparisons were made on relative rather than absolute measures of volume (i.e., corrected for individual brain size). The resulting GMV data were spatially smoothed with an 8 mm FWHM isotropic Gaussian kernel, and were resampled to a $2 \times 2 \times 2 \text{ mm}^3$ voxel size for further analysis.

Generation of the 1,024 Region AAL Template

Smoking related structural brain alterations have mostly been identified in the cerebral cortex at smaller spatial scales than the standard AAL atlas [Brody et al., 2004; Fritz et al., 2014; Gallinat et al., 2006; Zhang et al., 2011]. As such, and following the example of others [Cao et al., 2013; Hagmann et al., 2008; Zalesky et al., 2010], we randomly subdivided the cortex described in the standard AAL atlas into 1,024 equisized ROIs. The number of subdivisions was suggested by studies showing that a parcellation of about 1,000 regions provides a reasonable trade-off between spatial resolution and signal-to-noise ratio [Fornito et al., 2010; Zalesky et al., 2010]. Mean GMV values from the 1,024 ROIs then served as the inputs to the classification procedure. As a validation of this approach, we did a preliminary study that applied a SVM-RFE embedded leave-one-out cross-validation (LOOCV) procedure on raw GMVs and mean GMVs from the standard AAL atlas separately to demonstrate the efficiency of the subparcellated AAL approach over either a voxel-wise or a standard AAL approach (see Supporting Information).

SVM With Recursive Feature Elimination (SVM-RFE)

As a supervised machine learning algorithm, SVM [Cortes and Vapnik, 1995] performs pattern classification by finding a separating hyperplane defined by a weight vector w that maximizes the margin between samples. Those samples closest to the hyperplane are called support vectors (see [Burges, 1998] for a more detailed description of SVM).

In this study, SVM was performed with LIBSVM toolbox (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>) [Chang and Lin, 2011]. Since our feature dimension is much larger than the sample size, according to the LIBSVM manual, transforming these features to a higher dimensional space with a nonlinear kernel is unnecessary; thus a linear kernel SVM was adopted to reduce the risk of over-fitting. The linear kernel SVM has only one hyper-parameter: C , a trade-off between the margin width and the misclassification penalty. Data were scaled to [0, 1] in the SVM classifier. The value of the hyperparameter C was selected to be 1 using a grid-search method [Hsu et al., 2003] among the values [0.125, 0.25, 0.5, 1, 2, 4, 8] in a nested 10-fold cross-validation.

In order to identify the set of features with the greatest discriminative ability, we modified the SVM-RFE algorithm [De Martino et al., 2008; Guyon et al., 2002], which uses SVM iteratively to exclude noninformative features from the dataset while retaining discriminative features, into a cross-validation procedure since our purpose focuses on both feature selection and classifier evaluation. During each SVM-RFE iteration, an SVM classifier was trained and features were sorted according to a ranking criterion. For linear kernel SVM, a crucial advantage is that the importance of each original feature is directly related to its weight coefficient, which allows us to simply identify the most discriminative features in original space and perform the post hoc regression analysis described below. Here we used the square of the weight vector coefficients (w_i^2) as our ranking criterion. Features with the lowest ranking scores (low w_i^2) were removed from the dataset recursively.

10-Fold Cross-Validation

As illustrated in Figure 1 and the Appendix, the SVM-RFE algorithm was embedded in a balanced 10-fold cross-validation procedure to validate the classifier on the cross-validation set. In each trial, six smokers and six nonsmokers were removed for testing the classifier that was trained using all the other subjects in the cross-validation set, and the classification quality was assessed by the following five quantities:

$$\text{Sensitivity} = TP / (TP + FN) \quad (1)$$

$$\text{Specificity} = TN / (TN + FP) \quad (2)$$

$$\text{Accuracy} = (TP + TN) / (TP + FN + TN + FP) \quad (3)$$

$$\text{Precision} = TP / (TP + FP) \quad (4)$$

$$F \text{ score} = 2TP / (2TP + FP + FN) \quad (5)$$

Here, TP, FN, TN, and FP denote, respectively, the number of smokers correctly classified, the number of smokers predicted to be nonsmokers, the number of nonsmokers correctly classified, and the number of nonsmokers predicted to be smokers.

After cross-validation, accuracy was calculated over all cross-validation subjects for each feature set size (sizes

from 1 to 1,024), and the minimum feature set size at which the classifiers reached the highest accuracy was determined. Finally, the resulting SVM classifiers were validated on an independent testing set using the above-determined discriminative feature sets. Final validation results of each independent testing sample was determined via majority voting, i.e., creating a binary decision rule that selects the class that gets more than half of the votes of the SVM classifiers.

Regression Analyses Between Consistent ROIs and Smoking Measures

Candidate biomarkers are those brain regions that consistently appear discriminative during 10-fold cross-validation. To determine the threshold of consistent ROIs (i.e. the number of times a region had to be discriminative to be considered greater than chance), we performed a permutation analysis: We first randomly reassigned subject labels and then performed the 10-fold cross-validation classification. This procedure was repeated 100 iterations. We calculated the number of cross-validation folds in which an ROI was selected in random permutation among the previously determined minimum number of features from actual data. We then calculated the probability P that an ROI was selected in N folds ($N=1-900$) among all cross-validation folds (900 in total). The actual data were thresholded at $P < 0.001$ of the randomly permuted data, i.e., for any given ROI, there is a 0.1% probability that a region would occur by chance. We then investigated the relationship between GMVs within these consistent ROIs and smoking measures. Mean GMVs in significant ROIs were separately regressed against the Fagerström Test for Nicotine Dependence (FTND), a measure of dependence severity [Heatherton et al., 1991], and lifetime usage scores while controlling for age and gender.

RESULTS

Classification Performance

Over the 100 random iterations, the classifiers achieved sensitivity = $69.13\% \pm 5.34\%$ (mean \pm SD), specificity = $70.07\% \pm 4.96\%$, accuracy = $69.60\% \pm 4.15\%$, precision = $69.84\% \pm 4.27\%$, and F score = $69.42\% \pm 4.36\%$. The median (\pm MAD) number of features retained was 139 (± 84). Validation on the independent testing set via majority voting had a lower, but still statistically significant accuracy ($P = 0.01$, binomial test) of $64.04\% \pm 6.17\%$ (mean \pm SD) (sensitivity = $64.46\% \pm 9.64\%$, specificity = $63.61\% \pm 9.89\%$, precision = $64.29\% \pm 6.86\%$, F score = $64.01\% \pm 6.61\%$).

Map of Consistent ROIs

Using the permutation method described above, the threshold for a consistent ROI was determined to be 304

folds out of 100 rounds of 10-fold cross-validation. That is, if an ROI was selected more than 304 times in 100 rounds of 10-fold cross-validation (i.e. 900 chances), then that ROI is considered a consistent feature. Among the 1,024 ROIs, 150 ROIs surpassed that threshold and were considered to be consistent features. These ROIs are mainly located within the PFC, cingulate cortex, insula, caudate, putamen, cuneus, precuneus, thalamus, hippocampus, and parahippocampal, temporal, lingual, and pre-/postcentral gyrus (see Fig. 2).

Relationship Between Significant ROIs and Smoking Measures

We performed regression analyses between the mean GMVs in the 150 consistent ROIs and smoking measures (FTND and lifetime usage), controlling for both age and gender. Multiple regions showed a significant regression coefficient of $P < 0.05$ (Table II and Fig. 3). GMVs in the putamen, hippocampus, and the calcarine gyrus were inversely related to lifetime smoking history, while GMVs in the PFC, ACC, lingual gyrus, and the thalamus were inversely related to FTND scores. Finally, GMVs in the middle cingulate cortex, superior medial gyrus, pre-/postcentral gyrus, precuneus, caudate, and the parahippocampal gyrus were inversely related to both lifetime usage and FTND scores. However, due to the large number of regions, these results did not withstand rigorous multiple comparisons corrections and should be considered exploratory.

DISCUSSION

To the best of our knowledge, this is the first study that employed a machine learning approach to structural images in an addiction related clinical application. Given the massive medical and societal costs caused by cigarette use, we chose smoking as our model addiction system. We designed an SVM-RFE embedded 10-fold cross-validation framework to classify smokers from nonsmokers and validated our model on an independent test dataset. The whole procedure was repeated 100 times, and was followed by an exploration of the relationship between consistent features and smoking behavioral measures.

Feature extraction was carried out with the aid of a pre-generated 1,024 region AAL template, which increased the spatial specificity of the original 90 AAL cerebral cortical region template. The ROI approach also led to improved classification over using raw, voxel-wise, GMVs, as the spatial averaging improves signal-to-noise. Since our template was created by random parcellation of the standard AAL atlas [Cao et al., 2013; Hagmann et al., 2008; Zalesky et al., 2010], classifier performance might be further improved by applying various regional optimization algorithms (e.g., genetic algorithm) to the ROIs chosen for our template.

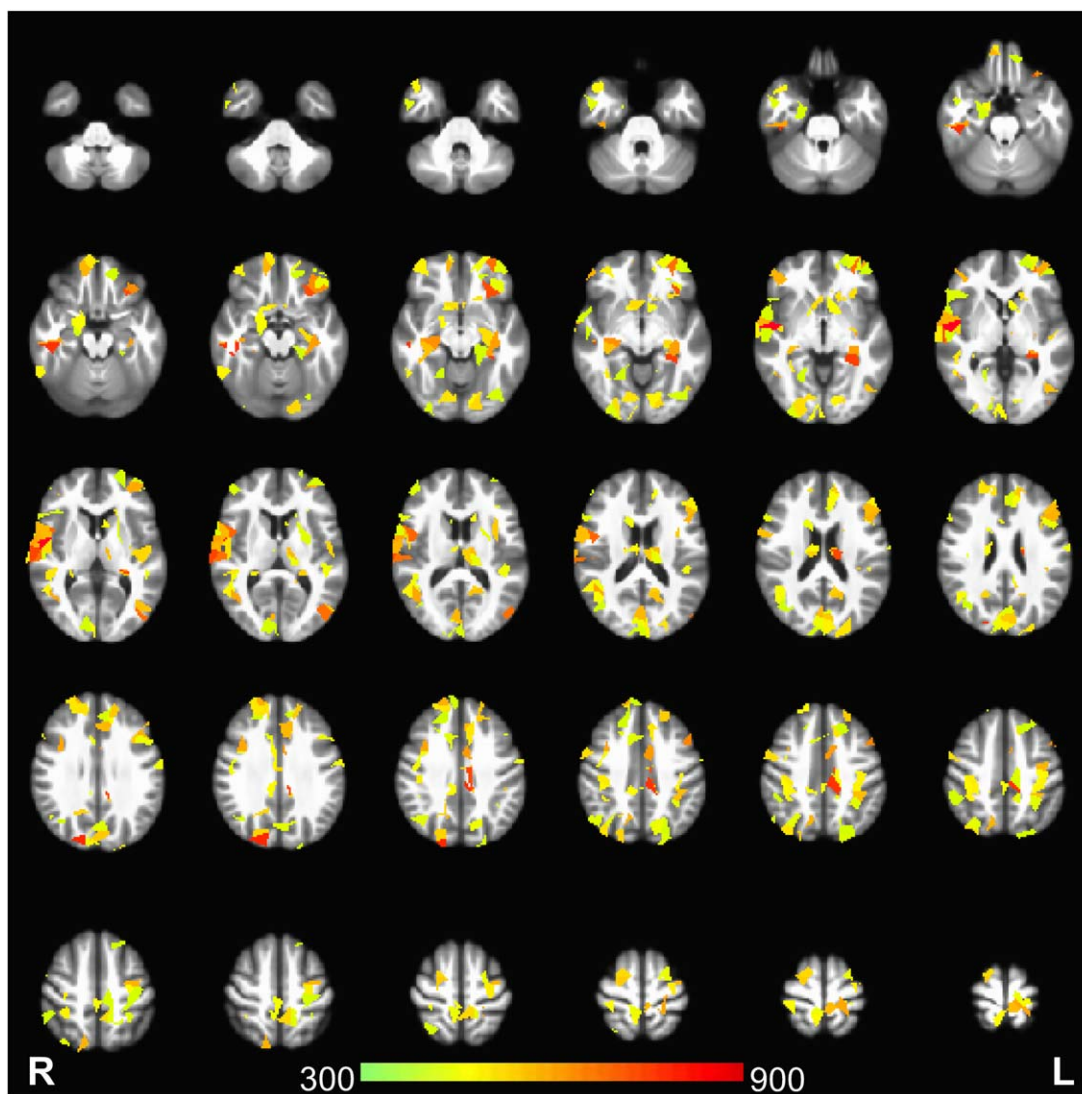


Figure 2.

Consistent ROIs ($N = 150$) based on permutation analysis; the color of each ROI (see color bar for scale) denotes the number of times (hits) that it was selected by the 100 iterations of the 10-fold cross-validation procedure. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Supervised machine learning approaches are usually confirmed through cross-validation procedures, whose goal is to estimate the model fit to a dataset that is not used in training the model. In the present study, in addition to cross-validation, we critically tested our SVM classifiers on a completely independent dataset. Although the performance was consistently lower when the classifier was applied to the independent set, accuracy nevertheless remained statistically significant. To date, few neuroimaging machine learning studies have validated their model on a completely independent dataset [Kawasaki et al., 2007; Nieuwenhuis et al., 2012; Whelan et al., 2014]. The

accuracies of these studies were either increased, likely a consequence of small sample size as the cross-validation accuracy should asymptotically be maximal [Kawasaki et al., 2007] or, as with our study, decreased [Nieuwenhuis et al., 2012; Whelan et al., 2014]. The decrement in accuracy on the independent test set implies that an overfitting problem still exists in cross-validation, which is likely caused by a limited sample size. Since the ultimate goal of applying machine learning techniques to neuroimaging data is to build computer-assisted systems for clinical diagnoses of neuropsychological diseases, good generalization performance is required for classification

TABLE II. Consistent ROIs that are inversely related with smoking measures (uncorrected $P < 0.05$)

ROI	NmVx	Center coordinates	Nm folds	Smoking measure	r Score	P
Left middle frontal gyrus	162	(-38, 53, 5)	560	FTND	-0.253	0.023
Right lingual gyrus	154	(14, -80, -14)	450	FTND	-0.246	0.027
Left middle cingulate cortex	156	(-8, -35, 44)	752	FTND	-0.330	0.0026
				Lifetime usage	-0.242	0.025
Right precuneus	152	(6, -53, 73)	403	FTND	-0.221	0.048
Left putamen	156	(-24, 2, 9)	340	Lifetime usage	-0.232	0.031
Left middle orbital gyrus	155	(-38, 54, -9)	341	FTND	-0.254	0.022
Left middle cingulate cortex	133	(-10, -29, 53)	329	FTND	-0.260	0.019
				Lifetime usage	-0.256	0.017
Right paracentral lobule	158	(2, -36, 55)	380	FTND	-0.249	0.025
Left inferior frontal gyrus	147	(-51, 34, 21)	424	FTND	-0.242	0.029
Left anterior cingulate cortex	137	(-10, 36, 26)	394	FTND	-0.284	0.010
Left inferior frontal gyrus	152	(-46, 28, 25)	524	FTND	-0.265	0.017
Left precuneus	148	(-14, -50, 54)	346	FTND	-0.315	0.0042
Right postcentral gyrus	151	(63, 1, 20)	521	Lifetime usage	-0.254	0.018
Left parahippocampal gyrus	151	(-28, -41, -7)	735	FTND	-0.317	0.0040
Right caudate	140	(14, 11, 20)	369	FTND	-0.297	0.0070
Left caudate	138	(-16, 23, -3)	453	Lifetime usage	-0.212	0.049
Right hippocampus	155	(28, -24, -13)	603	Lifetime usage	-0.231	0.033
Right postcentral gyrus	151	(34, -41, 72)	403	FTND	-0.239	0.032
Left postcentral gyrus	152	(-34, -25, 53)	402	FTND	-0.251	0.024
				Lifetime usage	-0.348	0.0010
Left precentral gyrus	164	(-30, -22, 64)	432	FTND	-0.231	0.038
Right thalamus	156	(12, -13, 19)	413	FTND	-0.260	0.019
Left calcarine gyrus	158	(4, -81, 11)	534	Lifetime usage	-0.233	0.031
Right precentral gyrus	153	(57, -2, 48)	462	Lifetime usage	-0.212	0.049
Left middle cingulate cortex	143	(-8, -29, 35)	723	FTND	-0.278	0.012
				Lifetime usage	-0.343	0.0012
Left superior medial gyrus	148	(-6, 38, 30)	328	FTND	-0.269	0.015
				Lifetime usage	-0.228	0.035
Left precuneus	157	(-6, -41, 76)	527	Lifetime usage	-0.248	0.021
Right precuneus	157	(8, -44, 65)	363	FTND	-0.270	0.015
Right middle cingulate cortex	146	(10, 15, 42)	342	FTND	-0.279	0.012
Left precuneus	140	(-14, -52, 15)	486	FTND	-0.220	0.049
				Lifetime usage	-0.215	0.046
Left parahippocampal gyrus	151	(-20, -30, -18)	313	FTND	-0.226	0.042
				Lifetime usage	-0.250	0.020
Right middle cingulate cortex	149	(10, -39, 37)	405	FTND	-0.333	0.0024
				Lifetime usage	-0.256	0.017
Left postcentral gyrus	151	(-34, -30, 57)	315	Lifetime usage	-0.259	0.016
Left precuneus	157	(-6, -50, 58)	478	FTND	-0.267	0.016
Left middle cingulate cortex	157	(-6, 0, 44)	583	FTND	-0.252	0.023
Left postcentral gyrus	154	(-36, -31, 46)	421	FTND	-0.230	0.039

The regression results are described using the ROI's location, its size by the number of voxels, center voxel MNI coordinates, number of cross-validation folds in which the ROI was chosen, related smoking measures, regression score, and P value.

models. Due to the problem of over-fitting caused by limited sample size in cross-validation, the best approach to test a model's generalizability is validating the model on a new and independent dataset that was not used in any training iteration, thus we strongly suggest that machine learning models developed using neuroimaging data should be tested on an independent dataset not only with cross-validation [Gabrieli et al., 2015].

A previous work by our group achieved an accuracy of 78.6% via LOOCV using resting-state network characteris-

tics to predict smoking status on 21 smokers and 21 non-smokers [Pariyadath et al., 2014]. In contrast, this experiment was generated on a larger dataset of structural images, which makes the classification result more stable and reliable. Besides differences in classifiers (i.e., SVM vs. SVMRFE), the decrement in classification accuracy may also be caused by disparate feature characters (i.e., resting-state network characteristics vs. structural GMV) and different cross-validation procedures (i.e., LOOCV vs. 10-fold cross-validation).

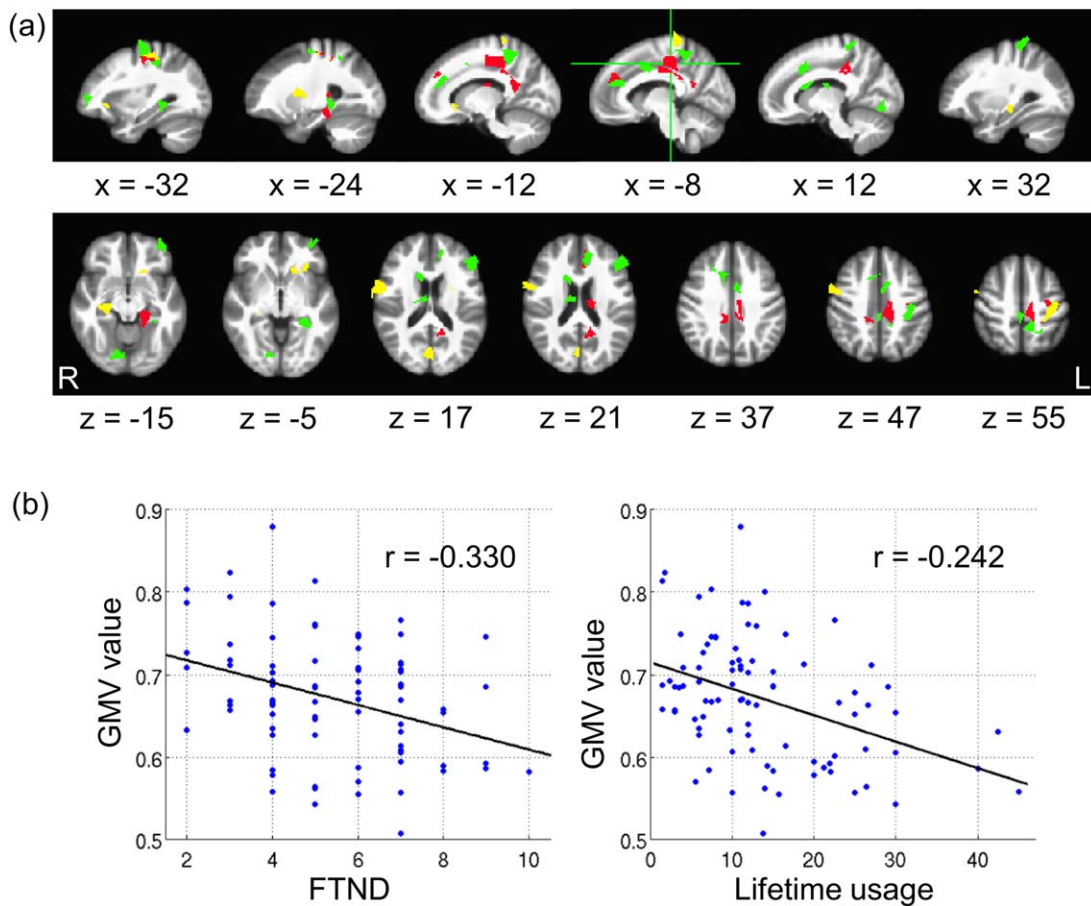


Figure 3.

ROIs that are inversely related with smoking measures: (a) Green color: ROIs having negative regression coefficients with FTND scores; yellow color: ROIs having negative regression coefficients with lifetime cigarette usage; red color: ROIs having negative regression coefficients with both smoking measures. (b)

An example of the inverse relationship between an ROI in middle cingulate cortex (red region with green cross in (a)) and smoking measures. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Previous studies analyzed with traditional univariate methods have reported that smokers have gray matter alterations in the PFC, cingulate cortex, lingual gyrus, cuneus, precuneus, thalamus, pre-/postcentral gyrus, temporal lobe including parahippocampal gyrus and the insula [Brody et al., 2004; Fritz et al., 2014; Gallinat et al., 2006; Liao et al., 2012; Zhang et al., 2011]. In the present multivariate study, ROIs within these regions were also consistently selected as discriminative features. Further, our subsequent regression analyses indicated a significant inverse relationship between the GMV in multiple discriminative ROIs and smoking intensity measures. Fritz et al. [2014] also found GMV loss in the cingulate and prefrontal cortex that was inversely related with lifetime usage. However, like the results from our study, these data need to be interpreted with caution as they did not pass stringent, multiple comparisons correction [Fritz et al., 2014].

Overall, our findings demonstrate that smoking is associated with GMV alterations in a wide range of cortical regions. Standing out among these regions, the insula has been shown to be a critical neural substrate for nicotine addiction, and has been implicated in multiple addictive behaviors, including craving and a lack of inhibitory control [Naqvi et al., 2007, 2014]. Being a locus for interoceptive representation, the posterior insula is thought to process low-level sensory signals and sequentially pass the information towards the anterior insula for higher-level awareness and affective processing and integration [Craig, 2010]. Indeed, the insula, along with the ACC, as part of the so called salience network (SN) [Seeley et al., 2007], is thought to be involved in switching activity between brain networks processing introspection (default mode network) and executive processing (executive control network) [Ding and Lee, 2013; Naqvi et al., 2014; Sutherland et al.,

2012]. Critically, traumatic alteration to insular functioning has been shown to dramatically affect craving and smoking behavior [Naqvi et al., 2007].

In addition to the insula, other cortical areas also showed GMV alterations. The ACC, involved in interconnecting functional circuits with the striatum, as well as being constituent of the SN, is strongly associated with nicotine addiction. Nicotine enhances the functional connectivity coherence strength of various cingulate-neocortical circuits, while critically, and in a double dissociation fashion, the strength of a functional circuit between the dorsal ACC and striatum is negatively correlated with nicotine addiction severity (FTND) and not altered by acute nicotine administration [Hong et al., 2009]. Results of an exploratory VBM analysis showed that striatal and hippocampal GMV are associated with a smoking cessation treatment outcome [Froeliger et al., 2010]. Compared with individuals who relapsed, those who quit had significantly higher GMV in the putamen and occipital lobe, while demonstrating lower GMV in hippocampus and cuneus. GMV alterations observed in the PFC and the thalamus have also been related to the neurobiology of substance addiction including smoking [Goldstein and Volkow, 2011; Sutherland et al., 2012; Zhang et al., 2011]. Smaller GMVs in these regions may be causally related to the fact that smokers have less efficient processing ability in working memory [Xu et al., 2005], faster cognitive decline in executive function [Sabia et al., 2012], and worse mood and interpersonal skills [Lyvers et al., 2008].

This study represents an important first step toward making clinical diagnoses and matching addiction treatment options to individuals with the aid of machine learning techniques. The limitations of this study include a limited data size and a single imaging modality. Future studies will need to address these two issues by increasing the number of participants and introducing machine learning techniques to multi-modal neuroimaging data. Further, given the cross-sectional nature of this design, we cannot determine if the smaller GMVs in more addicted/longer using individuals was a consequence of use or an antecedent of use.

CONCLUSION

We investigated anatomical abnormalities in smokers using a multivariate classification approach and explored the predictive value of GMVs in smokers using an SVM-RFE embedded framework. The SVMs achieved an accuracy of 69.60% in 10-fold cross-validation and provided good group separation even on an independent test set. Thus, either through years of cigarette smoking or as a consequence of predisposing genetic based antecedents of addiction or some gene \times environment interaction, smokers have sufficient gray matter alterations to be distinguished from nonsmokers at greater than chance levels. Instantiating this finding, subsequent regression analyses

showed that greater nicotine addiction/duration of use is associated with smaller GMVs in multiple brain regions. Our study not only identified structural biomarkers of smokers, but also revealed their discriminative power in predicting group memberships. It further highlights the potential of using machine learning approaches to aid the clinical diagnoses of drug addiction.

APPENDIX

Algorithm 1: SVM-RFE embedded N-fold cross-validation

Randomly partition subjects into equal-sized N folds, each fold contains the same number of smokers and nonsmokers

For $S = 1 \dots N$ -fold

Exclude subjects in S -fold for testing

Initialize FeatureSet to all features

While FeatureSet is not empty

Train SVM using FeatureSet

Test SVM on S -fold

Compute weight vector of SVM

Rank features according to w_i^2

Remove one feature with the lowest ranking

End While

End For

Compute accuracy over all subjects for each FeatureSet size

Find the minimum FeatureSet size on which SVM get the highest accuracy

REFERENCES

- Agaku IT, King BA, Dube SR (2014): Current cigarette smoking among adults-United States, 2005–2012. *MMWR Morb Mortal Wkly Rep* 63:29–34.
- Ashburner J (2007): A fast diffeomorphic image registration algorithm. *Neuroimage* 38:95–113.
- Ashburner J, Friston KJ (2000): Voxel-based morphometry—the methods. *Neuroimage* 11:805–821.
- Ashburner J, Friston KJ (2005): Unified segmentation. *Neuroimage* 26:839–851.
- Brody AL, Mandelkern MA, Jarvik ME, Lee GS, Smith EC, Huang JC, Bota RG, Bartzokis G, London ED (2004): Differences between smokers and nonsmokers in regional gray matter volumes and densities. *Biol Psychiatry* 55:77–84.
- Bullmore E, Sporns O (2012): The economy of brain network organization. *Nat Rev Neurosci* 13:336–349.
- Burges CJ (1998): A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discover* 2:121–167.
- Calderoni S, Retico A, Biagi L, Tancredi R, Muratori F, Tosetti M (2012): Female children with autism spectrum disorder: an insight from mass-univariate and pattern classification analyses. *Neuroimage* 59:1013–1022.
- Cao Q, Shu N, An L, Wang P, Sun L, Xia MR, Wang JH, Gong GL, Zang YF, Wang YF, He Y (2013): Probabilistic diffusion tractography and graph theory analysis reveal abnormal white

- matter structural connectivity networks in drug-naive boys with attention deficit/hyperactivity disorder. *J Neurosci* 33:10676–10687.
- Chang C-C, Lin C-J (2011): LIBSVM: A library for support vector machines. *ACM Trans Intelligent Syst Technol* 2:27.
- Chen R, Herskovits EH (2010): Machine-learning techniques for building a diagnostic model for very mild dementia. *Neuroimage* 52:234–244.
- Cortes C, Vapnik V (1995): Support-vector networks. *Mach Learn* 20:273–297.
- Craig AD (2010): The sentient self. *Brain Struct Funct* 214:563–577.
- De Martino F, Valente G, Staeren N, Ashburner J, Goebel R, Formisano E (2008): Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage* 43:44–58.
- Ding X, Lee SW (2013): Changes of functional and effective connectivity in smoking replenishment on deprived heavy smokers: A resting-state fMRI study. *PLoS One* 8:e59331.
- Ecker C, Rocha-Rego V, Johnston P, Mourao-Miranda J, Marquand A, Daly EM, Brammer MJ, Murphy C, Murphy DG, Consortium MA (2010): Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach. *Neuroimage* 49:44–56.
- Fagerström K (2002): The epidemiology of smoking: Health consequences and benefits of cessation. *Drugs* 62(Suppl2):1–9.
- Fjell AM, Walhovd KB (2010): Structural brain changes in aging: Courses, causes and cognitive consequences. *Rev Neurosci* 21:187–221.
- Focke NK, Helms G, Scheewe S, Pantel PM, Bachmann CG, Dechent P, Ebentheuer J, Mohr A, Paulus W, Trenkwalder C (2011): Individual voxel-based subtype prediction can differentiate progressive supranuclear palsy from idiopathic Parkinson syndrome and healthy controls. *Hum Brain Mapp* 32:1905–1915.
- Fornito A, Zalesky A, Bullmore ET (2010): Network scaling effects in graph analytic studies of human resting-state fMRI data. *Front Syst Neurosci* 4:22.
- Fritz HC, Wittfeld K, Schmidt CO, Domin M, Grabe HJ, Hegenscheid K, Hosten N, Lotze M (2014): Current smoking and reduced gray matter volume—a voxel-based morphometry study. *Neuropsychopharmacology* 39:2594–2600.
- Froeliger B, Kozink RV, Rose JE, Behm FM, Salley AN, McClernon FJ (2010): Hippocampal and striatal gray matter volume are associated with a smoking cessation treatment outcome: Results of an exploratory voxel-based morphometric analysis. *Psychopharmacology (Berl)* 210:577–583.
- Gabrieli JD, Ghosh SS, Whitfield-Gabrieli S (2015): Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* 85:11–26.
- Gallinat J, Meisenzahl E, Jacobsen LK, Kalus P, Bierbrauer J, Kienast T, Witthaus H, Leopold K, Seifert F, Schubert F, Staedtgen M (2006): Smoking and structural brain deficits: A volumetric MR investigation. *Eur J Neurosci* 24:1744–1750.
- Gifford EV, Kohlenberg BS, Hayes SC, Antonuccio DO, Piasecki MM, Rasmussen-Hall ML, Palm KM (2004): Acceptance-based treatment for smoking cessation. *Behav Ther* 35:689–705.
- Goldstein RZ, Volkow ND (2011): Dysfunction of the prefrontal cortex in addiction: Neuroimaging findings and clinical implications. *Nat Rev Neurosci* 12:652–669.
- Gong Q, Wu Q, Scarpazza C, Lui S, Jia Z, Marquand A, Huang X, McGuire P, Mechelli A (2011): Prognostic prediction of therapeutic response in depression using high-field MR imaging. *Neuroimage* 55:1497–1503.
- Guyon I, Weston J, Barnhill S, Vapnik V (2002): Gene selection for cancer classification using support vector machines. *Mach Learn* 46:389–422.
- Hagmann P, Cammoun L, Gigandet X, Meuli R, Honey CJ, Wedeen VJ, Sporns O (2008): Mapping the structural core of human cerebral cortex. *PLoS Biol* 6:e159.
- He Y, Chen ZJ, Evans AC (2007): Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. *Cereb Cortex* 17:2407–2419.
- Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO (1991): The Fagerström test for nicotine dependence: A revision of the Fagerström tolerance questionnaire. *Br J Addict* 86:1119–1127.
- Hong LE, Gu H, Yang Y, Ross TJ, Salmeron BJ, Buchholz B, Thaker GK, Stein EA (2009): Association of nicotine addiction and nicotine's actions with separate cingulate cortex functional circuits. *Arch Gen Psychiatry* 66:431–441.
- Hsu C-W, Chang C-C, Lin C-J (2003): A Practical Guide to Support Vector Classification, Technical report, Department of Computer Science, National Taiwan University.
- Jiao Y, Chen R, Ke X, Chu K, Lu Z, Herskovits EH (2010): Predictive models of autism spectrum disorder based on brain regional cortical thickness. *Neuroimage* 50:589–599.
- Kasperek T, Thomaz CE, Sato JR, Schwarz D, Janousova E, Marecek R, Prikryl R, Vanicek J, Fujita A, Ceskova E (2011): Maximum-uncertainty linear discrimination analysis of first-episode schizophrenia subjects. *Psychiatry Res* 191:174–181.
- Kawasaki Y, Suzuki M, Kherif F, Takahashi T, Zhou SY, Nakamura K, Matsui M, Sumiyoshi T, Seto H, Kurachi M (2007): Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. *Neuroimage* 34:235–242.
- Liao Y, Tang J, Liu T, Chen X, Hao W (2012): Differences between smokers and non-smokers in regional gray matter volumes: A voxel-based morphometry study. *Addict Biol* 17:977–980.
- Lyvers M, Thorberg FA, Dobie A, Huang J, Reginald P (2008): Mood and interpersonal functioning in heavy smokers. *J Subst Use* 13:308–318.
- Naqvi NH, Gaznick N, Tranel D, Bechara A (2014): The insula: A critical neural substrate for craving and drug seeking under conflict and risk. *Ann NY Acad Sci* 1316:53–70.
- Naqvi NH, Rudrauf D, Damasio H, Bechara A (2007): Damage to the insula disrupts addiction to cigarette smoking. *Science* 315:531–534.
- Nieuwenhuis M, van Haren NE, Hulshoff Pol HE, Cahn W, Kahn RS, Schnack HG (2012): Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *Neuroimage* 61:606–612.
- Oliveira PP, Nitri R, Busatto G, Buchpiguel C, Sato JR, Amaro E (2010): Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease. *J Alzheimers Dis* 19:1263–1272.
- Pariyadath V, Stein EA, Ross TJ (2014): Machine learning classification of resting state functional connectivity predicts smoking status. *Front Hum Neurosci* 8:425.
- Rajapakse JC, Giedd JN, Rapoport JL (1997): Statistical approach to segmentation of single-channel cerebral MR images. *IEEE Trans Med Imaging* 16:176–186.
- Reitzel LR, McClure JB, Cofta-Woerpel L, Mazas CA, Cao Y, Cinciripini PM, Vidrine JI, Li Y, Wetter DW (2011): The efficacy of computer-delivered treatment for smoking cessation. *Cancer Epidemiol Biomarkers Prev* 20:1555–1557.
- Rose JE, Behm FM, Westman EC, Levin ED, Stein RM, Ripka GV (1994): Mecamylamine combined with nicotine skin patch

- facilitates smoking cessation beyond nicotine patch treatment alone. *Clin Pharmacol Ther* 56:86–99.
- Sabia S, Elbaz A, Dugravot A, Head J, Shipley M, Hagger-Johnson G, Kivimaki M, Singh-Manoux A (2012): Impact of smoking on cognitive decline in early old age: The Whitehall II cohort study. *Arch Gen Psychiatry* 69:627–635.
- Sato JR, de Araujo Filho GM, de Araujo TB, Bressan RA, de Oliveira PP, Jackowski AP (2012): Can neuroimaging be used as a support to diagnosis of borderline personality disorder? An approach based on computational neuroanatomy and machine learning. *J Psychiatr Res* 46:1126–1132.
- Seeley WW, Menon V, Schatzberg AF, Keller J, Glover GH, Kenna H, Reiss AL, Greicius MD (2007): Dissociable intrinsic connectivity networks for salience processing and executive control. *J Neurosci* 27:2349–2356.
- Sutherland MT, McHugh MJ, Pariyadath V, Stein EA (2012): Resting state functional connectivity in addiction: Lessons learned and a road ahead. *Neuroimage* 62:2281–2295.
- Swan GE, Lessov-Schlaggar CN (2007): The effects of tobacco smoke and nicotine on cognition and the brain. *Neuropsychol Rev* 17:259–273.
- Tohka J, Zijdenbos A, Evans A (2004): Fast and robust parameter estimation for statistical partial volume models in brain MRI. *Neuroimage* 23:84–97.
- Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002): Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15:273–289.
- Uddin LQ, Menon V, Young CB, Ryali S, Chen T, Khouzam A, Minshew NJ, Hardan AY (2011): Multivariate searchlight classification of structural magnetic resonance imaging in children and adolescents with autism. *Biol Psychiatry* 70:833–841.
- Whelan R, Watts R, Orr CA, Althoff RR, Artiges E, Banaschewski T, Barker GJ, Bokde AL, Büchel C, Carvalho FM, Conrod PJ, Flor H, Fauth-Bühler M, Frouin V, Gallinat J, Gan G, Gowland P, Heinz A, Ittermann B, Lawrence C, Mann K, Martinot JL, Nees F, Ortiz N, Paillère-Martinot ML, Paus T, Pausova Z, Rietschel M, Robbins TW, Smolka MN, Ströhle A, Schumann G, Garavan H, Consortium I (2014): Neuropsychosocial profiles of current and future adolescent alcohol misusers. *Nature* 512: 185–189.
- Whincup PH, Gilg JA, Emberson JR, Jarvis MJ, Feyerabend C, Bryant A, Walker M, Cook DG (2004): Passive smoking and risk of coronary heart disease and stroke: Prospective study with cotinine measurement. *BMJ* 329:200–205.
- Xu J, Kobayashi S, Yamaguchi S, Iijima K, Okada K, Yamashita K (2000): Gender effects on age-related changes in brain structure. *AJNR Am J Neuroradiol* 21:112–118.
- Xu J, Mendrek A, Cohen MS, Monterosso J, Rodriguez P, Simon SL, Brody A, Jarvik M, Domier CP, Olmstead R, Ernst M, London ED (2005): Brain activity in cigarette smokers performing a working memory task: Effect of smoking abstinence. *Biol Psychiatry* 58:143–150.
- Yanbaeva DG, Dentener MA, Creutzberg EC, Wesseling G, Wouters EF (2007): Systemic effects of smoking. *Chest* 131:1557–1566.
- Zalesky A, Fornito A, Harding IH, Cocchi L, Yücel M, Pantelis C, Bullmore ET (2010): Whole-brain anatomical networks: Does the choice of nodes matter? *Neuroimage* 50:970–983.
- Zhang X, Salmeron BJ, Ross TJ, Geng X, Yang Y, Stein EA (2011): Factors underlying prefrontal and insula structural alterations in smokers. *Neuroimage* 54:42–48.