NeuroImage: Clinical 10 (2016) 115–123

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Studying depression using imaging and machine learning methods

Meenal J. Patel [a],*, Alexander Khalaf [b], Howard J. Aizenstein [a,b,c]

[a] Department of Bioengineering, University of Pittsburgh, PA, USA
[b] University of Pittsburgh School of Medicine, PA, USA
[c] Department of Psychiatry, University of Pittsburgh School of Medicine, PA, USA

ARTICLE INFO

Article history:
Received 19 May 2015
Received in revised form 23 October 2015
Accepted 4 November 2015
Available online 10 November 2015

Keywords:
Depression
Machine learning
Treatment
Prediction
Review

ABSTRACT

Depression is a complex clinical entity that can pose challenges for clinicians regarding both accurate diagnosis and effective timely treatment. These challenges have prompted the development of multiple machine learning methods to help improve the management of this disease. These methods utilize anatomical and physiological data acquired from neuroimaging to create models that can identify depressed patients vs. non-depressed patients and predict treatment outcomes. This article (1) presents a background on depression, imaging, and machine learning methodologies; (2) reviews methodologies of past studies that have used imaging and machine learning to study depression; and (3) suggests directions for future depression-related studies.

Contents

* Corresponding author.
E-mail address: mnl1615@gmail.com (M.J. Patel).

http://dx.doi.org/10.1016/j.nicl.2015.11.003
2213-1582/

1. Introduction

Major depressive disorder has an estimated lifetime prevalence of approximately 17%, and has significant effects on quality of life, comorbid medical conditions, suicide risk, and healthcare utilization (Andrade et al., 2003). The diagnosis of MDD is largely based on application of criteria from the Diagnostic and Statistical Manual of Mental Disorders (DSM), and clinician judgment. Upon diagnosis most patients are started on first-line antidepressant agents which is largely a trial and error process, as initial pharmacotherapy treatment is only effective in approximately 50% of patients (Papakostas, 2009). Non-responding patients will often require multiple drug trials, which can result in persistence of symptoms for months.

This variability in treatment response is likely related to depression being an epiphenomenal manifestation of multiple neural pathological processes. This etiological heterogeneity means that objective biomarkers of disease are often less informative than the presence of depressive symptomatology itself. Therefore, diagnosis based on symptom-based criteria (i.e. DSM) becomes essential, as treating a patient's biomarkers makes little sense in the absence of distressing psychiatric symptoms. However, the study and application of such disease state biomarkers may be valuable in the development of methods that could be extended to other aspects of depression management. Specifically, the utility of biomarkers is likely more valuable in the early identification of treatment non-response, as symptom remission is known to lag behind changes in underlying neural function (Aizenstein et al, 2014; Hsieh et al., 2002; Lui et al., 2011; Mayberg et al., 1997).

Identified biomarkers representing the biological substrates of MDD have been derived from multiple domains including neuroimaging, neuropsychological testing, genetics, and proteomics (Douglas and Porter, 2009; Laje et al., 2007; Miller et al., 2009). With respect to neuroimaging, magnetic resonance imaging (MRI) in particular has demonstrated its capacity for non-invasively studying brain structure and function in depressed patients. The different underlying brain characteristics associated with depression are probed with various MRI modalities that can be broadly separated into structural and functional imaging methods, which differ with respect to numerous scanning parameters. Further differentiation of image analysis and processing methods allow for a wide range of data collection regarding anatomic volumes, hemodynamic response within various neural structures and circuits, demyelinating white matter hyperintensity lesions, and overall neuronal cellular integrity.

As each of these neuroimaging measures only likely represents one facet of depression's complex underlying biology, their collective assessment is much more informative than if performed individually. This has prompted many groups to simultaneously investigate multiple neuroimaging domains in depression for the increased perspective it can afford (de Kwaasteniet et al., 2013; Khalaf et al., 2015; Steffens et al., 2011). Moreover, with more formalized systematic aggregation of neuroimaging biomarkers, it may be possible to make significant advancements in our understanding and management of MDD. Multiple machine learning techniques have been utilized to classify patients based on disease state and treatment response. For example, one of the earliest reports demonstrating machine learning's potential in MDD by Haslam and Beck (1993) used a categorization algorithm to classify patients into syndromal subtypes using the Beck Depression Inventory item scores. Much of the previous work has been predicated on effectively applying machine-learning techniques to clinical challenges that have yet to be addressed by other means. Most prominently, prediction of treatment non-response has become an opportune target for such methods. A sensitive, specific, and reliable machine-learning technique which identifies MDD patients who are unlikely to respond to the current antidepressant agent being trialed, would allow clinicians to preemptively intervene in such patients by either switching agents or pursuing more definitive treatments such as electroconvulsive therapy (ECT). With such a valuable potential clinical application, the current state of neuroimaging-based machine learning in MDD warrants examination. It will be the objective of this article to (1) provide an overview of the various machine learning methodologies in use, (2) review existing literature employing machine-learning in MDD, (3) elucidate technical obstacles, and (4) explore future directions.

1.1. Magnetic resonance imaging modalities

This review focuses on MRI modalities as they are most extensively found to be incorporated in depression related machine learning literature. The most common MRI modalities used to study brain structure include T1-weighted imaging, T2-weighted imaging, and Diffusion Tensor Imaging (DTI). To study brain function, functional MRI (fMRI) is generally used. Each of the MRI modalities helps examine different aspects of the brain. T1-weighted images are used to study cortical regions because of its high gray–white tissue contrast that allows for more accurate labeling of gray matter regions. These images can be used to assess the severity of atrophy in cortical regions by studying regional volume differences and changes. T2-weighted images are used to study white matter hyperintensities (WMHs), which are indicative of ischemic or pre-ischemic white matter changes. Both local and global WMH volume measures have been implicated in the development and potentiation of depression, especially late life depression (Herrmann et al., 2008). DTI images are used to gain an understanding of the brain from a microscopic level and study the diffusion of molecules in brain tissues. Two important measures acquired from DTI images include mean diffusivity (MD) and fractional anisotropy (FA), which signify the displacement and directionally of diffusion in tissue, respectively. These measures help evaluate the tissue integrity by helping determine cortical regions where diffusion is significantly decreased and dispersed due to lesions. Finally, fMRI images are used to study brain activity as well as functional connectivity between different cortical regions by exploiting underlying blood flow (Le Bihan et al., 2001; Blink, 2004; Vink et al., 2007).

1.2. Machine learning

Machine learning is used to test the potential of each MRI measure as a relevant biomarker of depression. Machine learning consists of a group of methods used to develop prediction models from empirical data to make accurate predictions about new data. Depending on the data three possible types of learning can be employed, including supervised learning, semi-supervised learning, and unsupervised learning. Supervised learning is performed if all of the data is labeled; semi-supervised learning is performed when there is unlabeled data along with labeled data; and, unsupervised learning is performed when all of the data is unlabeled. Learning methods can be categorized into linear and nonlinear methods. Linear methods are simpler, while nonlinear methods are more flexible in nature. For supervised and semi-supervised learning, the methods can be further categorized as classification- or regression-based methods. Classification-based methods attempt to classify the data by discrete and categorical labels, while regression-based methods fit the data to a continuous function and thus work with continuous labels for the data. For unsupervised learning, the methods are primarily categorized as clustering methods—which group the data into clusters based on underlying similarities (Ghahramani, 2004; Kapitanova and Son, 2012; Muller et al., 2003). Since past studies have primarily used supervised learning methods, this paper will focus on reviewing supervised learning related studies.

1.3. Validation measures

Validation measures are used to assess how well a model developed by a learning method will perform on new unseen data. To compute these measures, the trained prediction model is first applied to a test data set and predictions of labels/categories for each instance are acquired. Then, the validation measures are computed by comparing these predictions with actual labels if they are available. The validation measures differ based on the type of framework used for the learning method.

For classification-based frameworks, some common validation measures include accuracy, specificity, sensitivity, and receiver operating characteristic curve—which consists of true positive rates (i.e. sensitivity) as a function of false positive rates (i.e. 1—specificity). The accuracy measure helps evaluate how accurately the prediction model classifies the test data overall. The specificity and sensitivity measures assess how accurately the prediction model classifies each label of the test data, and the receiver operating characteristic curve illustrates the overall performance of the learning method. Confusion matrices can also be used when labeled data is available, especially for models with more than two labels. A confusion matrix will help summarize how well the prediction model distinguishes samples with different labels. The confusion matrix is a K × K matrix for K labels, where one side of the matrix represents actual labels and the other side represents predicted labels (Baldi et al., 2000).

For regression-based frameworks, some common validation measures include correlation coefficients and mean squared error. The correlation coefficients and their corresponding significance values (i.e. *p-value*) help demonstrate how well the model predictions are correlated with the actual label values, and the mean squared error helps evaluate the level of error in the model predictions (Baldi et al., 2000; Meyer, 2012).

1.4. Machine learning with real-world data

Due to the nature of real-world data, several problems are encountered when trying to use learning methods to estimate an optimal prediction model that can generalize well to (i.e. make accurate predictions on) unseen data. More often than not, real-world data is high-dimensional (i.e. each sample has a large number of features) and limited in sample size, both of which can cause problems with estimating an accurate learning model. Also, learning methods often face a trade-off between high bias and high variance when attempting to estimate a model for empirical data. High bias indicates that the learning method is learning an incorrect model, while high variance indicates that the learning method is learning a random model. Each of these can cause the model to underfit (i.e. model is too simple) or overfit (i.e. model is too complex and does not generalize well for representing future unseen data points) the data (Domingos, 2012; Le Borgne, 2005). In order to estimate an optimal prediction model given these limitations of real-world data, studies utilize feature reduction and cross-validation methods.

1.4.1. Feature reduction

Feature reduction methods are used to reduce the number of features in high-dimensional data to a limited number of most relevant features for estimating a more accurate prediction model. These methods can be primarily categorized into supervised and unsupervised methods. Supervised methods require labeled data as they perform feature reduction with the help of the labels. These methods are primarily used to perform feature selection (i.e. select the most relevant features from a larger set of input features) and thus reduce the noise in the input data. Unsupervised methods, on the other hand, perform feature reduction based solely on information available in the features included in the data. These methods are primarily used for feature extraction where features are selected based on patterns found among the input features; this helps to reduce the dimensionality of the input data (Mwangi et al., 2013; Reif and Shafait, 2014).

1.4.2. Cross-validation

Cross-validation is used to estimate the accuracy of a prediction model created by the learning method(s) of choice. There are several techniques for performing cross-validation including k-fold cross-validation, holdout, and leave-one-out cross-validation. The latter two techniques can be essentially considered variants of k-fold cross-validation. This technique first divides the data into k equal sized sets. Then it performs the following: (1) classifies one of the k sets as the test set, while combining the others to form the training set; (2) uses the learning method to estimate a model that describes the data by training on the training set; (3) tests the estimated model on the test set; and (4) computes appropriate validation measure(s) to determine the model's precision. This process is reiterated for k-iterations, each time classifying a different set as the test set without repetition. Lastly, the validation measure(s) values from all the iterations are averaged to evaluate the overall performance of the learning method.

When the available data set has a considerably large sample size, one way to perform cross-validation is to utilize the holdout approach. This is essentially a k-fold cross-validation technique where k equals one. When the available data set has too small of a sample size, a leave-one-out cross-validation method is used. With this method a k-fold cross-validation method is used where k is equal to the sample size of the data (Kohavi, 1995).

1.4.3. Integration of feature reduction and cross-validation

When using both feature reduction and cross-validation together, feature reduction should be performed at every iteration of cross-validation to avoid biasing the prediction model with information from the test set data. After performing feature reduction, the resulting dataset is used as an input to the learning method for estimating an optimal prediction model at every iteration of cross-validation. The average accuracy of prediction models from every iteration of the cross-validation determines the ability of the combined feature reduction and learning methods to estimating an optimal prediction model for a given dataset.

2. Past studies

Several past studies that have successfully explored predictive models for diagnosis and treatment response of depression. Below is a survey of studies found based on the following criteria: (1) the study focuses on studying depression and/or its treatment response; (2) the study uses magnetic resonance image related features for a supervised machine learning method; and (3) the study estimates a prediction model for depression, a measure of depression, and/or its treatment response. See Tables 1a–b and 2a–b for a brief summary of these studies and the methods they used.

3. Discussion

To the best of our knowledge the number of articles that use machine learning methods for studying depression is limited. Additionally, at a quick glance of the past studies presented in Tables 1 and 2, the methods used across these articles vary enough to make it difficult to draw comparisons simply based on the results. Thus, in this section we attempt to evaluate the methods used by the past studies.

3.1. Sample size

A common limitation across all past studies is sample size. The sample size used by past studies is small compared to what is optimal for machine learning methods to minimize the variance in assessments of accuracy, sensitivity, and specificity. This is especially true for the studies attempting to predict depression treatment response. Given the difficulty of recruiting depression patients for treatment protocols within a geographical area, the limitation of a small sample size is

Table 1
Past studies predicting depression diagnosis.

Author	Patient sample	Features [Imaging modality]	Feature reduction method	Cross-validation method	Machine learning method	Results (Note: highest accuracies presented)
Costafreda et al. (2009)	– 37 depressed – 37 non-depressed	– Smoothed gray matter voxel-based intensity values [T1-weighted]	– Voxel based morphometry – Filter method using ANOVA	– Leave-one-out cross-validation	– Support vector machines	– Accuracy[a]: 67.6% – Sensitivity[b]: 64.9% – Specificity[c]: 70.3%
Fu et al. (2008)	– 19 depressed – 19 non-depressed	– Smoothed whole brain voxel-based blood oxygen level dependent response during an implicit sad facial affect recognition task [fMRI]	– Filtering based on prior knowledge of anatomical regions that differ in activity between patient and controls during processing of emotional faces – Principal component analysis	– Leave-one-out cross-validation	– Support vector machines (linear kernel)	– Accuracy[a]: 86% – Sensitivity[b]: 84% – Specificity[c]: 89%
Hahn et al. (2011)	– 30 depressed – 30 non-depressed	– Smoothed whole brain voxel-based blood oxygen level dependent response during 3 depression related functional MRI tasks [fMRI]	– n/a	– Leave-one-out cross-validation	– Single-Gaussian process classification – Integration of Gaussian process classification and decision tree – Support vector machines (linear kernel) for comparison[d]	– Accuracy[a]: 83% – Sensitivity[b]: 80% – Specificity[c]: 87%
Marquand et al. (2008)	– 20 depressed – 20 non-depressed	– Smoothed whole brain voxel-based blood oxygen level response during a verbal working memory fMRI task [fMRI]	– Principal component analysis	– Leave-one-out cross-validation	– Support vector machines (linear kernel)	– Accuracy[a]: 68% – Sensitivity[b]: 65% – Specificity[c]: 70%
Mourao-Miranda et al. (2011)	– 19 depressed – 19 non-depressed	– Smoothed whole brain voxel-based and region-based blood oxygen level dependent response during implicit sad facial affect recognition task	– n/a	– Leave-one-out cross-validation	– One-class support vector machines (Non-linear kernel)	
Mwangi et al. (Jan 2012)	– 30 depressed	– Smoothed whole brain voxel-based intensity values [T1-weighted]	– Filtering out voxels from brain regions that differed significantly between patients recruited from different centers	– Leave-one-out cross-validation	– Relevance Vector Regression (Evaluation of BDI and HRSD scores)	Correlation Coefficient (*p*-values) – BDI scores: r = 0.694 ($p < 0.0001$) – HRSD scores: r = 0.34 ($p = 0.068$)

Mwangi et al. (May 2012)	– 30 depressed – 32 non-depressed	– Smoothed whole brain voxel-based intensity values [T1-weighted]	– Voxel based morphometry	– Leave-one-out cross-validation	– Relevance Vector Machines (Non-linear Gaussian Kernel) – Support vector machines (Non-linear Gaussian Kernel)	RVM: – Accuracy[a]: 90.3% – Sensitivity[b]: 93.3% – Specificity[c]: 87.5% SVM: – Accuracy[a]: 87.1% – Sensitivity[b]: 86.7% – Specificity[c]: 87.5%
Nouretdinov et al. (2011)	– 19 depressed – 19 non-depressed	– Smoothed whole brain voxel-based blood oxygen level dependent signal changes during observation of increasing levels of sadness [fMRI]	– n/a	– Leave-one-out cross-validation	– Support vector machines (linear kernel) with general probabilistic classification method (transductive conformal predictor)	– Accuracy[a]: 86.9% – Sensitivity[b]: 89.4% – Specificity[c]: 84.2%
Rondina et al. (2013)	– 30 depressed – 30 non-depressed	– Smoothed whole brain voxel-based blood oxygen level dependent response during passive viewing of emotional faces [fMRI]	– Survival Count on Random Subsamples (SCoRS) SCoRS is compared to other methods: – Recursive Feature Elimination[d] – Gini Contrast[d] – t-test[d]	– Leave-one-out cross-validation	– Support vector machines (linear kernel)	– Accuracy[a]: 72% – Sensitivity[b]: 77% – Specificity[c]: 67%
Rosa et al. (2015)	Dataset 1 Fu et al. (2008) – 19 depressed – 19 non-depressed Dataset 2 Hahn et al. (2011) – 30 depressed – 30 non-depressed	– Region-based functional connectivity (sparse compared to non-sparse network-based features) [fMRI]	– n/a	– Leave-one-subject-per-group-out cross-validation	– Sparse L1-norm support vector machines (linear kernel) – Non-sparse L2-norm support vector machines (linear kernel) for comparison[d]	Dataset 1 – Accuracy[a]: 78.95% – Sensitivity[b]: 68.42% – Specificity[c]: 89.47% Dataset 2 – Accuracy[a]: 85.00% – Sensitivity[b]: 83.33% – Specificity[c]: 86.67%
Zeng et al. (2012)	– 24 depressed – 29 non-depressed	– Region-based resting state functional connectivity [fMRI]	– Filter method using Kendall tau rank correlation coefficient	– Leave-one-out cross-validation	– Support vector machines (linear kernel)	– Accuracy[a]: 94.3% – Sensitivity[b]: 100% – Specificity[c]: 89.7%

BDI = Beck Depression Inventory (self-rated). fMRI = functional magnetic resonance imaging. HRSD = Hamilton Rating Scale for Depression (clinician-rated). RVM = relevance vector machines. SVM = support vector machines.

[a] Overall classification accuracy.

[b] Percent depressed patients identified.

[c] Percent non-depressed patients identified.

[d] Results of methods uses for comparison are not presented.

Table 2
Previous studies predicting depression treatment response.

Author	Patient sample	Features [Imaging modality]	Feature reduction method	Cross-validation method	Machine learning method	Results (Note: highest accuracies presented)
Costafreda et al. (2009)	– 9 responders – 9 non-responders	– Smoothed gray matter voxel-based intensity values [T1-weighted]	– Voxel based morphometry – Filter method using ANOVA	– Leave-one-out cross-validation	– Support vector machines	– Accuracy[a]: 88.9% – Sensitivity[b]: 88.9% – Specificity[c]: 88.9%
Liu et al. (2012)	– 17 responders – 18 non-responders	– Gray and white matter smoothed voxel-based intensity values [T1-weighted]	– Multivariate pattern analysis – Searchlight algorithm – Principal component analysis	– Leave-one-out cross-validation	– Support vector machines (linear kernel)	– Accuracy[a]: 82.9%
Marquand et al. (2008)	– 9 responders – 9 non-responders	– Smoothed whole brain voxel-based blood oxygen level dependent response during a verbal working memory fMRI task [fMRI]	– Principal component analysis	– Leave-one-out cross-validation	– Support vector machines (linear kernel)	– Accuracy[a]: 69% – Sensitivity[b]: 85% – Specificity[c]: 52%
Nouretdinov et al. (2011)	– 9 responders – 9 non-responders	– Smoothed voxel-based intensity values [T1-weighted]	– n/a	– Leave-one-out cross-validation	– Support vector machines (linear kernel) with general probabilistic classification method (transductive conformal predictor)	– Accuracy[a]: 83.3% – Sensitivity[b]: 77.8% – Specificity[c]: 88.9%

[a] Overall classification accuracy.
[b] Percent responders identified.
[c] Percent non-responders identified.

quite understandable. With respect to fMRI research, multiple efforts to create repositories to which individual groups can contribute data have attempted to address the broader issue of small sample sizes within neuroimaging research. These include the 1000 Functional Connectomes Project, the International Neuroimaging Data-Sharing Initiative, and the OpenfMRI Project, which are supported by the National Institutes of Health and the National Science Foundation. While these repositories are an encouraging step, lack of uniformity between contributing sites with respect to imaging parameters may introduce bias that impairs the sensitivity of studies which aggregate the data. Ideally, acquisition and processing parameters will progressively become more standardized throughout the neuroimaging research community, allowing for more meaningful data pooling. Nonetheless, in the meantime, it may be valuable to continue conducting machine learning studies with available sample sizes to refine prediction models.

3.2. Features

Features used by past studies have primarily focused on extracting information from T1-weighted imaging and fMRI. The use of these features in an initial attempt to model depression diagnosis and treatment response is in accordance with past neuroimaging-based studies that have predominantly found anatomical changes and altered brain activity to be valuable biomarkers of major depression (Dunlop and Mayberg, 2014; McGrath et al., 2013).

3.3. Learning method(s)

All past studies, except one which is a regression-based study (Mwangi et al., Jan 2012), have used support vector machines or a variant method as either their primary method or as a means of comparison with their primary method. Given the literature on machine learning methods, support vector machines are a popular method as observed among depression studies. Support vector machines draw its popularity from its useful strengths – especially when working with real-world data – including a reliable theoretical foundation and its insensitivity to high-dimensional data. Nevertheless, there are other methods that may perform equally well or better depending on the nature of the data (Wu et al., 2008).

However, there is variability as to whether a linear or non-linear learning method was used among these past studies. According to support vector machine literature, if the number of samples is significantly less than the number of features – which may be the case for several of these studies – non-linear learning methods do not significantly affect the results and it may be better to simply use linear learning methods (Hsu et al., 2010; Raudys and Jain, 1991). Thus, to avoid complexity and chances of overfitting, linear learning methods may be optimal.

3.4. Feature reduction method(s)

Compared to other methods, the past studies show the greatest heterogeneity in their selection of feature reduction methods. Nevertheless, the most commonly used methods fall into the category of supervised feature selection methods. Given the small sample sizes used by the past studies and the literature indicating that the performance of feature selection improves with an increase in sample size (Jain and Zongker, 1997), performing supervised feature selection methods on the small dataset may be suboptimal. Hypothetically, a more effective approach for obtaining a model that generalizes well over unseen data, as used by Fu et al. (2008), may be to perform feature selection based on biomarkers already shown to be associated with major depression and its treatment response by larger studies. However, this requires further testing with larger test sets to gain a more objective understanding of which feature reduction methods produce an optimal model given smaller datasets.

Some past studies have also used an unsupervised feature reduction method, namely principal component analysis. Once again, it may be more effective to perform unsupervised feature reduction methods on larger datasets as they provide more information for generalizing population trends accurately (Osborne and Costello, 2004).

3.5. Cross-validation method

Most past studies have used the leave-one-out cross-validation method. This is not surprising as it is the most popular method for studies with small sample sizes (e.g. proof-of-concept studies). Compared to several other cross-validation methods, the leave-one-out cross-validation method provides the learning method with more data for training a prediction model. Thus, it is utilized when the available data is limited due to a small sample size. Nevertheless, it is associated with high variance making it unreliable for obtaining accurate estimates of a prediction model for larger-scale studies (Elisseeff and Pontil, 2002; Refaeilzadeh, et al., 2009).

4. Future directions

Based on the discussion in the previous section of past studies, in this section we suggest some potential directions to explore for future studies.

4.1. Larger sample sizes

So far most past depression prediction studies have used a small sample size; especially when predicting depression treatment response. Even though a small sample size provides an initial direction for developing a prediction model, a larger sample size is valuable for developing a more robust prediction model that generalizes well to the wider population. By building a larger database for training a prediction model (such as collaborative databases like the Alzheimer's Disease Neuroimaging Initiative), the variations observed among depression patients can be more thoroughly incorporated which in the future may result in models with true clinical utility.

As the studies begin to use larger datasets, the methods employed will likely begin to vary and demonstrate improved validation measures. More specifically, the k-fold cross-validation method can be used with larger k-values instead of leave-one-out to allow for larger test sets on which to test prediction models and improve the generalizability of the models. The feature reduction methods will also improve in performance as they are shown to be affected by sample size. Last but not least, the selection of learning methods that work effectively with larger datasets is also more extensive. More details on selecting learning methods are provided under the Learning method(s) section below.

4.2. Learning method(s)

To accurately learn a framework or problem, it is not only important to select the right features, but also to select the right learning method. Sometimes, in order to create a larger sample size with a limited dataset, it may be useful to incorporate unlabeled data.

Thus, the first step is to determine whether the given data to be learned consists of labeled instances only, a mixture of labeled and unlabeled instances, or unlabeled instances only. Consequently, this will determine whether to use a supervised, semi-supervised, or unsupervised learning method, respectively. If data consist of a mixture of labeled and unlabeled instances, it would be beneficial to determine whether or not the unlabeled instances would help the learning method. If the unlabeled data does not sufficiently increase the overall sample size, it may be better to exclude it. The second step is to determine the goal (e.g. classification, regression, or clustering) of the learning method. The third step is to decide whether the nature of the data is linear or non-linear. Generally, when the data size is small, it is better to use a linear method to avoid overfitting. However, if the data size is sufficiently large, it may be beneficial to test non-linear methods to allow for more flexibility in the learning. The fourth step is then to select the learning method from the narrowed down options.

Since no one learning method is the best for all applications, it may be useful to test multiple methods. When selecting a learning method for a given framework or problem, one should consider evaluating several different aspects of the method including: computation time, underlying assumptions, interpretability, complexity, flexibility, optimization ability, and previous applications by other studies. If there are still too many options of methods to choose from, it may be helpful to use machine learning libraries (e.g. LIBSVM, LIBLINEAR) or software (e.g. MATLAB, Python scikit-learn, WEKA) for testing the performance of different methods on the data. Conversely, if there are too few options, it may be beneficial to modify (e.g. add constraints, regularize, combine methods (including learning, feature reduction, and/or boosting methods)), existing learning methods to make them more suitable for learning the given data. These and related techniques for selecting learning methods have been successfully utilized by Bibi and Stamelos (2006), Frank et al. (2004) and Kotthoff et al. (2012).

4.3. Parameter(s) selection

Parameter selection is a method that very few past studies (e.g. Mourao-Miranda et al., 2011; Mwangi et al., May 2012; Rosa et al., 2015) have utilized to optimize their results. However, this may relate to it being a more effective method on larger datasets, which would have greater variability in data than smaller datasets. In this section, we discuss the effects and implementation of parameter selection.

It is useful to perform a parameter selection process because sometimes slight changes to a certain parameter's values of a given learning method cause considerable variability in the resulting prediction model. Selection of a parameter that somehow regulates the complexity (e.g. regularization parameters, which penalize complexity and target the overfitting problem) of the prediction model developed by the learning method is especially important. This is because, as discussed earlier, the complexity of a prediction model determines whether it overfits or underfits the data, which impacts its generalizability. The most common approach used for parameter selection is cross-validation to determine optimal parameter values (Lim and Yu, 2013).

However, a cross-validation technique is most-likely already being used to evaluate the overall generalization-based performance of a learning method. Thus to perform parameter selection, a nested inner cross-validation loop would need to be implemented. For this inner cross-validation loop, the training set at every iteration of the outer cross-validation loop is used as the full data set on which parameter selection is performed. This inner cross-validation loop would be implemented between steps one and two of the 4-step process of the k-fold cross-validation technique described in the Introduction section.

Any of the cross-validation techniques described in the "Cross-validation" section or any variant of these techniques (e.g. estimation stability with cross validation) can be used for the inner cross-validation loop to perform parameter selection. The only difference is that instead of iterating through different test sets, the parameter selection method iterates through each of the pre-defined set of possible parameter values. At every iteration, it uses the same training and test set to estimate a model that describes the data and assess the precision of the model by computing appropriate validation measures respectively. The parameter value that results in the most precise model is then selected as the optimal parameter value. The selected parameter value is then used to train the full data set (i.e. the training set of the outer cross-validation loop) for step 2 from the 4-step process of the k-fold cross-validation technique (Kohavi and John, 1995; Lim and Yu, 2013).

The process of selecting multiple parameters' optimal values is similar to the process used to select one parameter's optimal value. The only difference is that the cross-validation method iterates through each possible set of values from each parameter to find the optimal set. All possible combinations of a set of parameters' values are identified from pre-defined options of values for each parameter using the grid search technique (Bergstra and Bengio, 2012).

4.4. Variation in features

Although T1-weighted imaging and fMRI biomarker have been primarily associated with depression, there are more recent studies that have shown the relevance of DTI biomarkers (Dunlop and Mayberg, 2014; Schneider et al., 2011). Additionally, non-imaging measures have also been studied as potential biomarkers for depression and its treatment response (Lopresti et al., 2014; Thase, 2014). Thus it would be intriguing to study how multi-modal MRI features in conjunction with non-imaging measures affect prediction models of depression and its treatment response. Given the non-linearity of the brain's functionality, it is likely that a non-linear relation between these features may produce optimal prediction models as is shown by a preliminary study on late-life depression (Patel et al., 2015).

4.5. Clinical application

The longer term goal for this burgeoning field will be to identify and coalesce around a specific machine-learning technique or set of related techniques that have demonstrated significant accuracy, precision, sensitivity, and specificity in identification of treatment non-response at baseline or early in a treatment course. While such a scenario is unlikely to occur for a number of years, support vector machines and a few other supervised learning algorithms are currently showing promise in this area. Notwithstanding, when a sufficiently effective method has been thoroughly validated through preliminary studies, progressing to clinical trials will be necessary to demonstrate whether this technology will actually benefit patients. Specifically, future clinical trials will need to establish that machine-learning methods can successfully identify depressed patients unlikely to respond to the current agent being trialed, and that clinicians' use of this information results in improved outcomes for patients (i.e. decreased latency between diagnosis and remission). The latter will presumably result from clinicians either switching to other agents or attempting different treatment modalities, such as ECT or cognitive–behavioral therapy. With only 50% of patients experiencing treatment response to first-line agents, and an adequate antidepressant trial requiring several weeks before symptom improvement becomes evident, machine-learning methods have the potential to significantly reduce the duration of patient suffering. Outside the scope of depression, these techniques have shown promise in predicting treatment response in other various other psychiatric diseases, including schizophrenia and obsessive–compulsive disorder (Khodayari-Rostamabad et al., 2010; Salomoni et al., 2009). Therefore, advances in development of machine-learning are likely to have wide-ranging implications throughout psychiatry, regardless of the specific disease to which they are first applied.

Acknowledgments

This research was supported by NIH grant R01MH076079, University of Pittsburgh Clinical Scientist Training Program (UL1 TL1TR000005), and NIMH Medical Student Research Fellowship (R25 MH054318-18).

References

Aizenstein, H.J., Khalaf, A., Walker, S.E., Andreescu, C., 2014. Magnetic resonance imaging predictors of treatment response in late-life depression. J. Geriatr. Psychiatry Neurol. 27, 24–32.

Andrade, L., Caraveo-Anduaga, J.J., Berglund, P., et al., 2003. The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. Int. J. Methods Psychiatr. Res. 12, 3–21.

Baldi, Pierre, Brunak, Søren, Chauvin, Yves, Andersen, Claus A.F., Nielsen, Henrik, 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16 (5), 412–424.

Bergstra, James, Bengio, Yoshua, 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.

Bibi, S., Stamelos, I., 2006. Selecting the appropriate machine learning techniques for the prediction of software development costs. Artificial Intelligence Applications and Innovations. Springer, pp. 533–540.

Le Bihan, D., Mangin, J.F., Poupon, C., Clark, C.A., Pappata, S., Molko, N., Chabriat, H., 2001. Diffusion tensor imaging: concepts and applications. J. Magn. Reson. Imaging 13 (4), 534–546.

Blink, Evert J., 2004. mri: Physics. MRI-physics. net,(Nov 2004). second edition. pp. 4–8.

Costafreda, S.G., Chu, C., Ashburner, J., Fu, C.H., 2009. Prognostic and diagnostic potential of the structural neuroanatomy of depression. PLoS One 4 (7), e6353. http://dx.doi.org/10.1371/journal.pone.0006353.

de Kwaasteniet, B., Ruhe, E., Caan, M., Rive, M., Olabarriaga, S., Groefsema, M., Heesink, L., vanWingen, G., Denys, D., 2013. Relation between structural and functional connectivity in major depressive disorder. Biol. Psychiatry 74, 40–47.

Domingos, Pedro, 2012. A few useful things to know about machine learning. Commun. ACM 55 (10), 78–87.

Douglas, K.M., Porter, R.J., 2009. Longitudinal assessment of neuropsychological function in major depression. Aust. N. Z. J. Psychiatry 43, 1105–1117.

Dunlop, B.W., Mayberg, H.S., 2014. Neuroimaging-based biomarkers for treatment selection in major depressive disorder. Dialogues Clin. Neurosci. 16 (4), 479–490.

Elisseeff, A., Pontil, M., 2002. Leave-one-out error and stability of learning algorithms with applications. NATO-ASI Series on Learning Theory and Practice.

Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I.H., 2004. Data mining in bioinformatics using Weka. Bioinformatics 20 (15), 2479–2481.

Fu, C.H., Mourao-Miranda, J., Costafreda, S.G., Khanna, A., Marquand, A.F., Williams, S.C., Brammer, M.J., 2008. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. Biol. Psychiatry 63 (7), 656–662. http://dx.doi.org/10.1016/j.biopsych.2007.08.020.

Ghahramani, Zoubin, 2004. Unsupervised learning. Advanced Lectures on Machine Learning. Springer, pp. 72–112.

Hahn, T., Marquand, A.F., Ehlis, A.C., Dresler, T., Kittel-Schneider, S., Jarczok, T.A., ... Fallgatter, A.J., 2011. Integrating neurobiological markers of depression. Arch. Gen. Psychiatry 68 (4), 361–368. http://dx.doi.org/10.1001/archgenpsychiatry.2010.178.

Haslam, Beck, 1993. Categorization of major depression in an outpatient sample. J. Nerv. Ment. Dis. 181 (12), 725–731.

Herrmann, L.L., Le Masurier, M., Ebmeier, K.P., 2008. White matter hyperintensities in late life depression: a systematic review. J. Neurol. Neurosurg. Psychiatry 79, 619–624.

Hsieh, M.H., McQuoid, D.R., Levy, R.M., Payne, M.E., MacFall, J.R., Steffens, D.C., 2002. Hippocampal volume and antidepressant response in geriatric depression. Int. J. Geriatr. Psychiatry 17, 519–525.

Hsu, C., Chang, C., Lin, C., 2010. A practical guide to support vector classification. Tech. rep. Department of Computer Science, National Taiwan University.

Jain, A., Zongker, D., 1997. Feature selection: evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. 19 (2), 153–158.

Kapitanova, Krasimira, Son, Sang H., 2012. Machine learning basics. *Intelligent Sensor Networks: The Integration of Sensor Networks, Signal Processing and Machine Learning*p. 13.

Khalaf, A., Edelman, K., Tudorascu, D., Andreescu, C., Reynolds, C.F., Aizenstein, H., 2015. White matter hyperintensity accumulation during treatment of late life depression. Neuropsychopharmacology http://dx.doi.org/10.1038/npp.2015.158.

Khodayari-Rostamabad, A., Reilly, J.P., Hasey, G., Debruin, H., 2010. Using pre-treatment EEG data to predict response to SSRI treatment for MDD. Conf. Proc. IEEE Eng. Med. Biol. Soc. 2010, 6103–6106.

Kohavi, Ron, 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Paper Presented at the IJCAI.

Kohavi, Ron, John, George H., 1995. Automatic parameter selection by minimizing estimated error. Paper Presented at the ICML.

Kotthoff, L., Gent, I.P., Miguel, I., 2012. An evaluation of machine learning in algorithm selection for search problems. AI Communications 25 (3), 257–270.

Laje, G., Paddock, S., Manji, H., Rush, A.J., Wilson, A.F., Charney, D., McMahon, F.J., 2007. Genetic markers of suicidal ideation emerging during citalopram treatment of major depression. Am. J. Psychiatry 164, 1530–1538.

Le Borgne, Y., 2005. Bias-variance Trade-off Characterization in a Classification Problem: What Differences with Regression? Machine Learning Group, Univ. Libre de Bruxelles, Belgium

Lim, Chinghway, Yu, Bin, 2013. Estimation Stability with Cross Validation (ESCV) (arXiv preprint arXiv:1303.3128).

Liu, F., Guo, W., Yu, D., Gao, Q., Gao, K., Xue, Z., ... Chen, H., 2012. Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. PLoS One 7 (7), e40968. http://dx.doi.org/10.1371/journal.pone.0040968.

Lopresti, A.L., Maker, G.L., Hood, S.D., Drummond, P.D., 2014. A review of peripheral biomarkers in major depression: the potential of inflammatory and oxidative stress biomarkers. Biol. Psychiatry 48 (3), 102–111.

Lui, S., Wu, Q., Qiu, L., et al., 2011. Resting-state functional connectivity in treatment-resistant depression. Am. J. Psychiatry 168, 642–648.

Marquand, A.F., Mourao-Miranda, J., Brammer, M.J., Cleare, A.J., Fu, C.H., 2008. Neuroanatomy of verbal working memory as a diagnostic biomarker for depression. Neuroreport 19 (15), 1507–1511. http://dx.doi.org/10.1097/WNR.0b013e328310425e.

Mayberg, H.S., Brannan, S.K., Mahurin, R.K., et al., 1997. Cingulate function in depression: a potential predictor of treatment response. Neuroreport 8, 1057–1061.

McGrath, C.L., Kelley, M.E., Holtzheimer III, P.E., Dunlop, B.W., Craighead, W.E., Franco, A.R., Craddock, R.C., Mayberg, H.S., 2013. Toward a neuroimaging treatment selection biomarker for major depressive disorder. JAMA Psychiatry 70 (8), 821–829. http://dx.doi.org/10.1001/jamapsychiatry.2013.143.

Meyer, David, 2012. Support vector machines. The Interface to libsvm in Package e1071. e1071 Vignette.

Miller, A.H., Maletic, V., Raison, C.L., 2009. Inflammation and its discontents: the role of cytokines in the pathophysiology of major depression. Biol. Psychiatry 65, 732–741.

Mourao-Miranda, J., Hardoon, D.R., Hahn, T., Marquand, A.F., Williams, S.C., Shawe-Taylor, J., Brammer, M., 2011. Patient classification as an outlier detection problem: an application of the One-Class Support Vector Machine. Neuroimage 58, 793–804.

Muller, K.R., Anderson, C.W., Birch, G.E., 2003. Linear and nonlinear methods for brain–computer interfaces. IEEE Trans. Neural Syst. Rehabil. Eng. 11 (2), 165–169. http://dx.doi.org/10.1109/TNSRE.2003.814484.

Mwangi, B., Matthews, K., Steele, J.D., 2012. Prediction of illness severity in patients with major depression using structural MR brain scans. J. Magn. Reson. Imaging 35 (1), 64–71. http://dx.doi.org/10.1002/jmri.22806.

Mwangi, Benson, Tian, Tian Siva, Soares, Jair C., 2013. A review of feature reduction techniques in neuroimaging. Neuroinformatics 1–16.

Nouretdinov, I., Costafreda, S.G., Gammerman, A., Chervonenkis, A., Vovk, V., Vapnik, V., Fu, C.H., 2011. Machine learning classification with confidence: application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. Neuroimage 56 (2), 809–813. http://dx.doi.org/10.1016/j.neuroimage.2010.05.023.

Osborne, J.W., Costello, A.B., 2004. Sample size and subject to item ratio in principal components analysis. Pract. Assess. Res. Eval. 9 (11).

Papakostas, G.I., 2009. Managing partial response or nonresponse: switching, augmentation, and combination strategies for major depressive disorder. J. Clin. Psychol. 70 (Suppl. 6), 16–25.

Patel, M.J., Andreescu, C., Price, J.C., Edelman, K.L., Reynolds, C.F., Aizenstein, H.J., 2015. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. Int. J. Geriatr. Psychiatry 30 (10), 1056–1067.

Raudys, S.J., Jain, A.K., 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans. Pattern Anal. Mach. Intell. 13 (3), 252–264.

Refaeilzadeh, P., Tang, L., Lu, H., 2009. Cross-validation. Encycl. Database Syst. 532–538. http://dx.doi.org/10.1007/978-0-387-39940-9_565.

Reif, Matthias, Shafait, Faisal, 2014. Efficient feature size reduction via predictive forward selection. Pattern Recogn. 47 (4), 1664–1673.

Rondina, J., Hahn, T., de Oliveira, L., Marquand, A., Dresler, T., Leitner, T., Fallgatter, A., Shawe-Taylor, J., Mourao-Miranda, J., 2013. SCoRS — a method based on stability for feature selection and mapping in neuroimaging. IEEE Trans. Med. Imaging 33 (1), 85–98.

Rosa, M.J., Portugal, L., Hahn, T., Fallgatter, A.J., Garrido, M.I., Shawe-Taylor, J., Mourao-Miranda, J., 2015. Sparse network-based models for patient classification using fMRI. Neuroimage 105, 493–506.

Salomoni, G., Grassi, M., Mosini, P., Riva, P., Cavedini, P., Bellodi, L., 2009. Artificial neural network model for the prediction of obsessive–compulsive disorder treatment response. J. Clin. Psychopharmacol. 29, 343–349. http://dx.doi.org/10.1097/JCP.0b013e3181aba68f.

Schneider, B., Prvulovic, D., Oertel-Knochel, V., Knochel, C., Reinke, B., Grexa, M., Weber, B., Hampel, H., 2011. Prog. Neurobiol. 95, 703–717.

Steffens, D.C., Taylor, W.D., Denny, K.L., Bergman, S.R., Wang, L., 2011. Structural integrity of the uncinate fasciculus and resting state functional connectivity of the ventral prefrontal cortex in late life depression. PLoS One 6, e22697.

Thase, M., 2014. Using biomarkers to predict treatment response in major depressive disorder: evidence from past and present studies. Dialogues Clin. Neurosci. 16 (4), 539–544.

Vink, M., Raemaekers, M., van der Schaaf, A., Mandl, R., Ramsey, N., 2007. Pre-processing and Analysis.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z., Steinbach, M., Hand, D.J., Steinberg, D., 2008. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14, 1–37. http://dx.doi.org/10.1007/s10115-007-0114-2.

Zeng, L.L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., ... Hu, D., 2012. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. Brain 135 (Pt 5), 1498–1507. http://dx.doi.org/10.1093/brain/aws059.