CHAPTER

6

Support vector machine

Derek A. Pisner, David M. Schnyer

Department of Psychology, University of Texas at Austin, Austin, TX, United States

6.1 Introduction

As discussed in Chapter 1, there are four fundamental approaches that one can take in machine learning: supervised, unsupervised learning, semisupervised, and reinforcement learning. Classification is a form of supervised learning, which maps input data to output data based on many example input−output pairs determined during a training phase. Using classification, features related to a set of example observations can be used to train a decision function that generates class assignments (i.e., "labels") with a given accuracy. These features can be widely diverse based on anything from functional neuroimaging data to twitter posts. Once this decision function "classifier" has been created based on these features, it can then automatically attach class labels to new, unseen observations using the patterns established before. There are many types of machine learning algorithms that can perform classification, such as decision trees, naïve bayes, and deep learning networks. This chapter reviews Support Vector Machine (SVM) learning as one such algorithm. The power of an SVM stems from its ability to learn data classification patterns with balanced accuracy and reproducibility. Although occasionally used to perform regression (see Chapter 7), SVM has become a widely used tool for classification, with high versatility that extends across multiple data science scenarios, including brain disorders research.

An SVM decision function is more precisely an optimal "hyperplane" that serves to separate (i.e., "classify") observations belonging to one class from another based on patterns of information about those observations called features. That hyperplane can then be used to determine the most probable label for unseen data. The features used to infer the hyperplane are not typically raw data; rather, they are most often derivative data resulting from some kind of interpolation during the feature selection

Machine Learning
https://doi.org/10.1016/B978-0-12-815739-8.00006-7

stage, which is discussed later in this chapter. Features are further referenced by coordinates based on their relationships to each other and form the support vectors. As with other forms of machine learning, working with SVM involves balancing two complementary aims—(1) maximizing the percentage of correct labels assigned to new examples by the classifier (i.e., optimizing its accuracy) and (2) ensuring that the classifier is generalizable to new data (i.e., optimizing its reproducibility). While the former is bound by the informativeness of the features used (i.e., feature importance), the latter is bound by the number of unique examples used to train the model.

6.2 Method description

6.2.1 Overview

Fortunately, effective use of SVM in the neurosciences does not require an in-depth understanding of its mathematical foundation, but it does demand a clear conceptual understanding and conscientiousness with respect to application. The process of training an SVM decision function amounts to identifying a reproducible hyperplane that maximizes the distance (i.e., the "margin") between the support vectors of both class labels (Fig. 6.1).

Thus, the optimal hyperplane is that which "maximizes the margin" between classes. An SVM can be linear or nonlinear but is most commonly the former (nonlinear SVM is not covered in this chapter). Linear SVM problems range in their complexity depending on the number of features used. In the hypothetical case of two feature dimensions, for instance, the

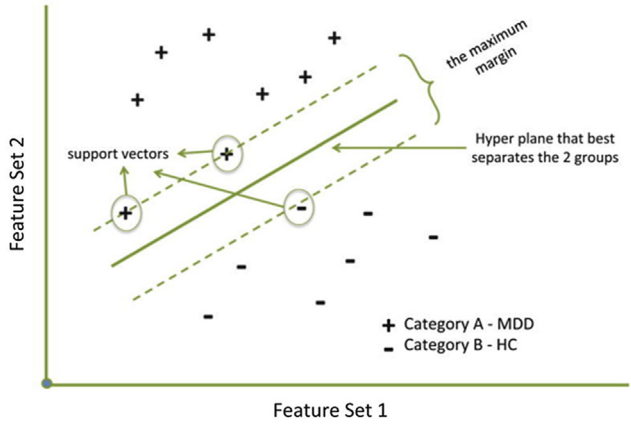

FIGURE 6.1 Illustration of the hyperplane that maximally separates the support vectors corresponding to each of the two to-be-predicted classes—here, major depressive disorder (MDD) and healthy controls (HC). *Taken from Schnyer, D. M., Clasen, P. C., Gonzalez, C., & Beevers, C. G. (2017). Evaluating the diagnostic utility of applying a machine learning algorithm to diffusion tensor MRI measures in individuals with major depressive disorder. Psychiatry research. Neuroimaging, 264, 1–9. http://10.1016/j.pscychresns.2017.03.003.*

hyperplane simply corresponds to a line, whereas in the case of three features, the hyperplane corresponds to a two-dimensional plane. Regardless of the SVM's level of complexity—that is, its dimensionality—classification problems are most often linear in the sense that the hyperplane used is straight and not curved. If we assume that the features that we use for SVM are linearly separable in this way, then we can readily draw a straight hyperplane (called a linear classifier) on a graph of the features that separates the two labels of the class of interest.

It follows that there are technically two types of margins to be maximized for a linear SVM. With a hard margin, no training errors are permitted. Although a hard margin may be the simplest and least computationally expensive, the linear separability of features rarely is this perfect in practice. Thus, a larger margin that allows for greater generalizability to new data can often be achieved by allowing the classifier to *misclassify*. Permitting misclassification can be achieved using what's referred to as a soft margin, which relies on the use of slack variables represented by ξ. These nonzero values $0 \leq \xi \leq 1$ in turn allow for classification error that can result when outliers in the training data lead to the hyperplane making mistakes—that is, misclassifying (Fig. 6.2). By that token, the hard margin becomes a special case of a soft margin where the slack variable is set to zero ($\xi = 0$). A penalty factor C, called the "soft-margin constant," is also introduced with the soft-margin approach to incur a penalty on slack variables. This parameter serves to control the

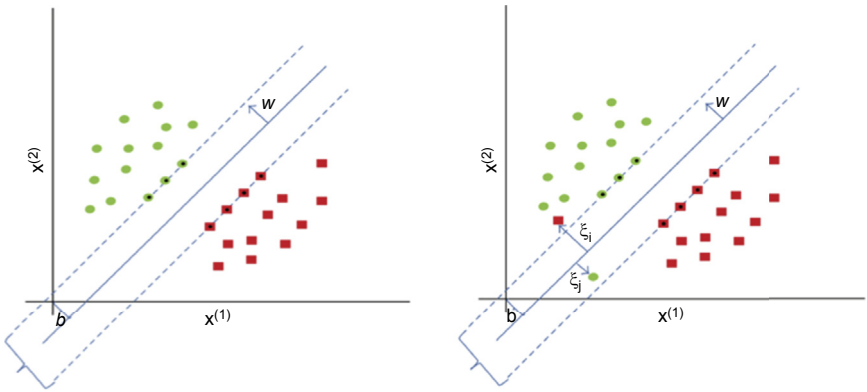

FIGURE 6.2 The left plot depicts a hard-margin hyperplane, where no training errors (i.e., misclassified support vectors) are permitted. The right plot depicts a soft-margin hyperplane, which allows for some degree of training error through the use of slack variables ξ. In both plots, w represents the margin, b represents the intercept between classes, and $x^{(i)}$ denotes the class labels, which are depicted in green (gray in print version) and red (dark gray in print version). *Adapted from Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., Brovelli, A. (2012). Multivoxel pattern analysis for fMRI data: A review.* Computational and Mathematical Methods in Medicine, 2012. *http://doi.org/10.1155/2012/961257.*

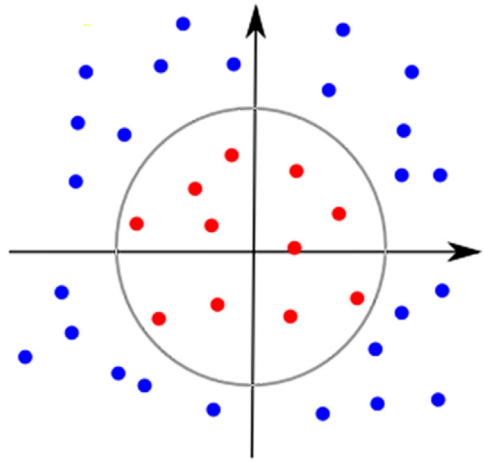

FIGURE 6.3 The plot above depicts a nonlinear classification problem with a curved hyperplane that separates the support vectors.

trade-off between hyperplane complexity and training errors (i.e., regularization) and reduces the chance of overfitting—i.e., fine-tuning the classifier for maximal accuracy in the training dataset so much that the model no longer generalizes to new data. In fact, some argue that soft-margin SVM is preferable even when training dataset is linearly separable, as the alternative could allow even a single outlier to determine the hyperplane boundary. Nevertheless, in high-dimensional datasets such as those found in neuroimaging where the dimensionality of the feature space almost always exceeds the number of study subjects to be classified, linear separability can usually be guaranteed through feature selection such that even a few outliers are usually nonthreatening.

In more difficult classification problems, however, optimal solutions may require outright curved hyperplanes—those which are nonlinear (Fig. 6.3). When the decision boundary of a classifier depends on the data in some nonlinear way (i.e., beyond the soft-margin case of having only a few outliers), the classifier is said to be nonlinear. In these cases, some type of kernel method is typically needed to transform the support vectors to a higher-dimensional input space. In other words, this additional step serves to convert a nonlinearly separable set of features to a set of linearly separable ones. As discussed in Section 6.3.2, kernel methods are also often used as a form of dimensionality reduction for linear SVM.

6.2.2 Stages of SVM analysis

There are essentially three stages to SVM analysis: (i) feature selection, (ii) training and testing the classifier, and (iii) performance evaluation.

It should be noted that these stages are not specific to SVM but are present in most machine learning methods, as discussed in Chapter 2.

6.2.2.1 Stage 1—feature selection

A prerequisite for training an SVM classifier involves the transformation of the original raw training data into a set of "features," which can be used as input for SVM. Most feature selection methods rank the features based on a specific criterion that reflects their degree of relevance. These feature selection methods can be divided into three main types: (1) embedded methods, (2) filter methods, and (3) wrapper methods (Bhaumik et al., 2017).

6.2.2.1.1 Embedded methods

With embedded methods, the feature selection is incorporated into the classifier itself, and the selection is performed automatically during the actual SVM training phase. To accomplish this, one can use what has been referred to as the "kernel trick"—i.e., a kernel method. In fact, the use of kernel functions can be found in almost all applications of SVM in neuroimaging (Cuingnet, Benali, & Chupin, 2010; Mahmoudi, Takerkart, Regragui, Boussaoud, & Brovelli, 2012; Pettersson-Yeo et al., 2014). Not only can kernel methods improve the computational efficiency of SVM training, but also they can be a convenient way to help prevent overfitting for the frequently ill-conditioned classification problems that emerge in neuroimaging experiments, where the dimensionality of a brain image typically far exceeds the number of examples available for training. In essence, a kernel function represents pairwise similarity measures between all example patterns, summarized in a kernel matrix with $N \times N$ dimensions, where N is the number of observations. Instead of relying on the raw feature vector as direct input to an SVM classifier, a kernel function allows one to train the SVM using the kernel matrix which, in both linear and nonlinear cases, maps the raw data to a higher-dimensional feature space (Fig. 6.4).

6.2.2.1.2 Filter methods

Filter methods perform feature reduction as a preprocessing step before classification and compute some relevance measure on the training set to remove the least important elements before fitting a hyperplane. The rationale for feature reduction is threefold: (1) it reduces redundancy in the raw data so that there will be a greater proportion of sample training data relative to the dimensionality of the features; (2) it aids in interpretation of the final classifier; for example, identifying the data that carry the most predictive information relevant to discriminating classes can help in focusing future efforts; and (3) for some classification algorithms, it can reduce computational load and accelerate the model

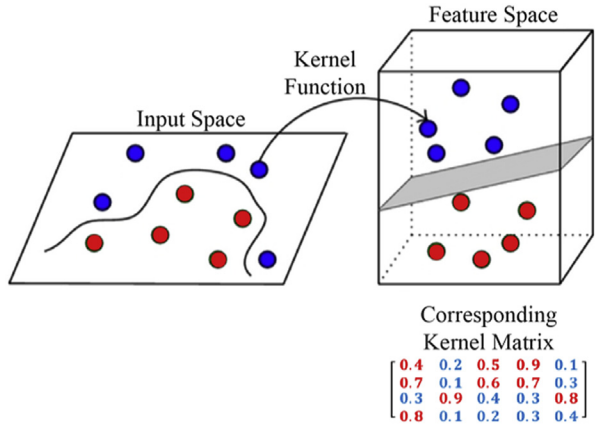

FIGURE 6.4 Above is a conceptual depiction of the "kernel trick," which involves transforming raw, input data to a high-dimensional feature space by way of a "similarity" function with corresponding kernel matrix.

training process. There are many forms of feature reduction that can be used in preparation for training an SVM classifier. Minimally, this usually involves removing features with near-zero variance and significantly correlated (i.e., "multicollinear") features because such features contribute complexity to the SVM without adding predictive power.

6.2.2.1.3 Wrapper methods

With wrapper methods, the classifier is trained repeatedly using the feedback from every iteration to select a subset of features for the next iteration. Although more computationally expensive than embedded methods, wrapper methods can discard the data points that, when considered independently, do the worst job of discriminating between class labels. Perhaps the most common type of wrapper method used with traditional SVM is recursive feature elimination (RFE), which selects features by recursively ranking them among smaller and smaller subsets of features through cross-validation. As discussed in Chapter 2, cross-validation is a multipermutation technique for evaluating predictive models like SVM. It works by iteratively partitioning the original training dataset into new training and test sets, reevaluating model performance during each iteration.

6.2.2.2 Stage 2—training and testing the classifier

An SVM is trained using example observations where we already know the label assignments (e.g., patients and controls) of the examples in advance. Consequently, we can supervise SVM to exploit this a priori information for the purpose of predicting new label assignments.

Specifically, SVM projects each subject's feature coordinate onto the line defined by the decision function; if the projection falls on the positive side ($y > 0$), the pattern is classified as belonging to class A (e.g., major depressive disorder [MDD] group) or else class B (e.g., healthy controls [HCs] group). Training an SVM amounts to setting the parameters w and b in the decision function $f(x) = w \cdot x + b$ that orients the hyperplane in such a way that the resulting projection of points maximally separates the members of the two classes. For a linear classifier, the absolute value of the weights directly reflects the importance of a feature in discriminating the two classes. Importantly, this procedure assumes that no one class contains a lot more examples than the other because unbalanced classes would have a negative impact on classifier performance. Although methods for remedying unbalanced classes exist, these are beyond the scope of this chapter (see Batuwita & Palade, 2013).

Aside from tuning the core parameters w and b—a process that is carried out automatically when an SVM estimator is fit to the data—accuracy of the classifier can also critically depend on the choice of hyperparameter values. Hyperparameters are those variables which impact the fit of the decision function and are set before learning (i.e., training) begins. Although SVM typically has fewer hyperparameters than other types of machine learning algorithms (e.g., neural networks), any "free" parameters that result from feature selection (e.g., number of k-best features, soft-margin constant C, etc.) are generally treated as hyperparameters that require tuning (Ben-Hur & Weston, 2010).

6.2.2.3 Stage 3—evaluating SVM performance

SVM performance is commonly, although not exclusively, described by its sensitivity, specificity, and accuracy (see Chapter 2). In essence, these metrics provide information about both the accuracy and reproducibility of the SVM hyperplane in differentiating between classes. To jointly evaluate accuracy and reproducibility, however, permutation testing is needed, where a hyperplane is estimated iteratively with randomly permuted class labels, across a window of hyperparameter values, for several resampled versions of the dataset. That is, performance across these metrics is optimized through cross-validation. In fact, effective evaluation of any classification learning algorithm, especially SVM, should (always) involve multiple partitions of the data, as the hyperplane resulting from any single partition of the dataset might arbitrarily favor one classifier over another. Although the permutation testing of cross-validation is critical for training a reproducible SVM, the ultimate test of model performance is on unseen data. Although one would ideally like to train a classifier using as much of the available data as possible, that would leave insufficient data for testing the final (i.e., *learned*) SVM model. Thus, in addition to cross-validation, a preferred approach is to split the

dataset into train and test groups as an initial stage, where the latter group is "held out" for final evaluation of model performance. This step serves to confirm that the classifier can indeed generalize beyond the dataset used for training.

6.2.3 SVM in neuroimaging

The majority of existing applications of SVM in brain disorders research are based on neuroimaging data (Orrù, Pettersson-Yeo, Marquand, Sartori, & Mechelli, 2012). In the context of neuroimaging, SVM is typically used to perform multivoxel pattern analysis (MVPA), where sets of "voxels" of a structural and/or functional brain image are used as inputs to derive features for classification (Haxby, 2012; Haxby, Connolly, & Guntupalli, 2014) (Fig. 6.5). To review, the "images" produced in neuroimaging are comprised of three-dimensional units called voxels whose position can be presented in the form of a (x, y, z) coordinate in Euclidean space. The signal intensity at each voxel consists of values corresponding to any type of relevant brain measure such as blood-oxygen level dependence across voxels at discrete time points (or averages of time points) in the case of functional magnetic resonance imaging (fMRI) data, gray matter density in the case of structural T1/T2 MRI, or white matter characteristics such as fractional anisotropy in the case of diffusion MRI data. As we will see, SVM is well-suited for the high-dimensional "multivoxel" approach of MVPA, as its relative simplicity carries a lower risk of overfitting (e.g., as compared to neural networks). The application of SVM to neuroimaging need not be restricted to MVPA, however; derivative metrics from neuroimaging data

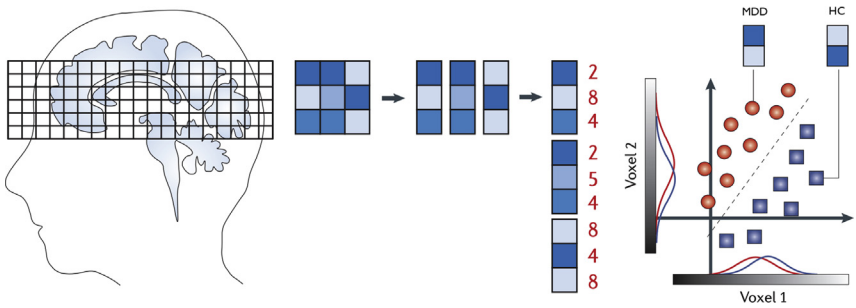

FIGURE 6.5 Above is a conceptual depiction of the transformation of raw voxel data from multivoxel pattern analysis to feature vectors with corresponding feature weights (red numbering) (dark gray numbering in print version), and the separation of these vectors by a hyperplane to classify major depressive disorder (MDD) versus healthy controls (HCs). *Adapted from Haynes and Rees, 2006.*

such as global graph measures that are not available on a "voxel-wise" basis can also be used as inputs to SVM (Rudd, 2017; Wang, Zuo, He, Bullmore, & Fornito, 2010).

With MVPA, a hyperplane is estimated based on an array of voxels that comprise a brain image. When SVM is applied through MVPA, the signal intensity of voxels represented in the array forms a corresponding feature vector with associated weights (w) that describe the contribution of each voxel to the linear decision function $f(x) = w \cdot x + b$ that defines the hyperplane. The features are further associated with one of the two classes, $y_i = -1$ or $+1$, each of which represent the labels of the *to-be-discriminated* class (e.g., HCs or depression diagnostic status). Although traditional feature reduction strategies for SVM like RFE can be effective on their own, several other methodological approaches have been developed uniquely for neuroimaging. These include feature selection by region of interest (ROI) (Poldrack, 2007; Mahmoudi et al., 2012) and the searchlight method (Kriegeskorte, Goebel, & Bandettini, 2006), both of which are discussed in greater detail later in this chapter.

6.3 Applications to brain disorders

Many would argue that the diagnosis and prognosis of brain disorders are to some extent classification problems in that they can be conceptualized in terms of Boolean class labels (e.g., "meets criteria" vs. "does not meet criteria"; "treatment responsive" vs. "treatment nonresponsive") (American Psychiatric Association, 2013). In clinical contexts, these classes might include MDD diagnostic status, vulnerability to Alzheimer's disease (AD), or treatment responsiveness. Recall too that a major appeal of classification learning methods is that they can be predictive; once a classifier such as an SVM has been generated, it can then be applied to new individuals to predict their class membership as well (Fig. 6.6). By this reasoning, the classification approach of SVM could perhaps offer a means for predicting diagnostic or prognostic class label algorithmically. These class predictions could in turn provide actionable information to clinicians about their patients, perhaps corroborating decision-related diagnosis, treatment planning, and even early intervention (Retico, Tosetti, Muratori, & Calderoni, 2014; Sundermann, Herr, Schwindt, & Pfleiderer, 2014).

Thus, a core motivation for including SVM within the larger repertoire of brain disorders research is that they carry translational potential for augmenting or perhaps one day even steering interventions for a variety of brain disorders (Huys et al., 2016; Sundermann et al., 2014). The past decade has produced a large corpus of machine learning studies that

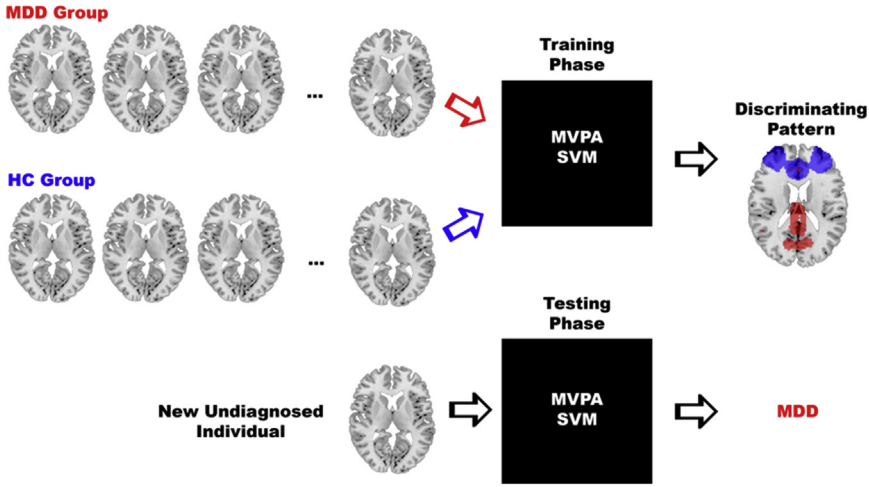

FIGURE 6.6 Simplified conceptual overview of how Support Vector Machine (SVM) classifiers might be used to aid diagnosis of major depressive disorder (MDD).

employ SVM or similar classification methods to predict diagnosis and prognosis for several types of brain disorders. As discussed in Chapter 3, these studies can be divided into three main categories: (1) studies that examine the diagnostic value of neuroimaging data by comparing patients with HCs; (2) studies which examine the potential of neuroimaging data for predicting the vulnerability to or onset of a disease by comparing the brain scans (acquired at baseline) of individuals with prodromal symptoms who subsequently did and did not develop the disorder; and (3) studies which examine the prognostic value of imaging data by comparing the brain scans obtained from patients before treatment onset who subsequently did and did not respond. In the sections that follow, we explore the use of SVM in clinical neuroimaging research covering three spectra of brain disorders—cognitive impairment, psychosis, and depression.

6.3.1 Predicting conversion from mild cognitive impairment to Alzheimer's disease

Early SVM studies that attempted to predict diagnosis of brain disease focused on mild cognitive impairment (MCI) and probable dementia of Alzheimer type (PDAT) (Orrù et al., 2012). Given the devastating impact of PDAT and its growing prevalence (Rizzi, Rosset, & Roriz-Cruz, 2014), there is increasing demand for approaches capable of predicting PDAT in its preceding prodromal MCI phase (Whitehead et al., 2004).

Such capability could facilitate early intervention with pharmaceuticals, to improve or stabilize the cognitive and behavioral symptoms of MCI. To date, however, no standard method exists to predict who will (or will not) develop dementia among those with MCI, and an extensive line of studies has used SVM with the hope of rectifying this (Bron, Smits, Niessen, & Klein, 2015; Cabra et al., 2015; Orrù et al., 2012). Davatzikos, Bhatt, Shaw, Batmanghelich, and Trojanowski (2011) and Nho et al. (2010) used MVPA, for instance, to train SVM classifiers that could predict MCI to PDAT conversion at 1-year follow-up with accuracies slightly above chance (60.8% and 65.0%, respectively). Distinctly, both of these studies employed feature selection based on neuroanatomically defined ROIs know to be affected by dementia, such as the temporal lobe, posterior cingulate/precuneus, and insula regions of the brain (Fig. 6.7, Box 6.1). Costafreda et al. (2011) likewise focused singularly on the hippocampus as a brain region for performing SVM. Using a fully automated prognostic procedure based on hippocampal morphometry, Costafreda and colleagues showed that they were able to predict MCI-PDAT conversion with an accuracy of 80% at 1-year follow-up.

Unlike prior studies that had relied almost entirely on T1/T2-weighted structural MRI, Haller et al. (2010) attempted to predict MCI-PDAT conversion at 1-year follow-up using diffusion tensor imaging. Their SVM-based analysis using white matter microstructural features yielded a high classification accuracy of 98.4%, discriminating between stable MCI patients and those who would likely develop dementia. In a more recent study, Cabral, Morgado, Campos Costa, and Silveira (2015) used SVM with data acquiring from fluorodeoxyglucose positron emission

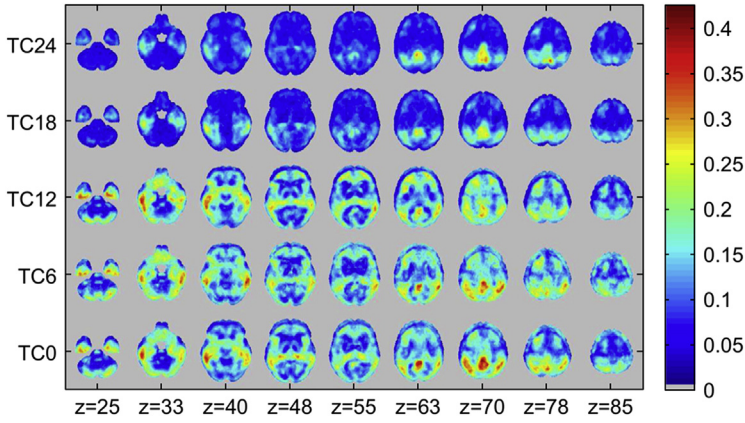

FIGURE 6.7 Spatial representation of FDG-PET feature importance from an SVM classifier used by Cabral et al. 2015. Along the x-axis, SVM features based on glucose metabolism are represented in terms of nine equally-spaced axial slices. Along the right side y-axis, feature voxel intensities are measured in terms of their mutual information (MI). In this context, MI quantifies how much knowing each feature reduces the uncertainty of the cognitive impairment classification by the SVM. Ultimately this study found that the brain regions with highest mean MI across cross-validation folds were the Posterior Cingulate Gyrus and the Precuneus.

BOX 6.1

Feature Reduction Using Regions of Interest

ROI-based feature reduction serves to reduce relevant brain voxels to only those known to be relevant to the to-be-predicted classes. This strategy can be approached in multiple ways. A researcher could, as in the aforementioned examples, preselect voxels only from one or more ROIs, for example, whose functional and/or structural properties are expected to differ between classes according to prior literature or resulting from cross-validated general linear model (GLM) analysis. Alternatively, ROIs could be selected because they have been predetermined to have the highest overall responsiveness to the relevant prediction class (i.e., PDAT-converted vs. MCI-stable). In either case, the ROIs will include spatially contiguous but not necessarily adjacent sets of voxels. Determining the highest overall responsiveness of voxels to the relevant prediction class can be accomplished using mass univariate testing with GLM such as the feature-wise t-test filter to discriminate features that have different group means or the selection of K best features from a multipermutation ANOVA F-test (De Martino et al., 2008). Although these approaches are widely used in neuroimaging studies with SVM, they should be approached with caution. When not employed within a cross-validation framework or without family-wise corrected *P*-values, mass univariate testing can easily lead to overfitting.

tomography—an imaging method that quantifies glucose metabolism in the brain—to study how MCI disease stage impacts diagnostic performance (Fig. 6.8). The authors found a decrease in SVM performance with longer temporal distance to conversion. However, with disease stage as a complementary feature, they achieved an accuracy of 85.1% at the time of conversion and 75% 2 years before it. Taken collectively, the results of these studies on aging populations are consistent with the idea that structural neuroanatomy of both gray and white matter is highly informative for predicting the progression of MCI to PDAT using SVM specifically.

6.3.2 Brain-based diagnostics of schizophrenia

At present, the diagnosis of schizophrenia and evaluation of its severity is carried out almost exclusively through clinical interview and self-report, without the use of biomarkers (Weiss, 2005). In the absence of more objective diagnostic information, schizophrenia diagnoses have

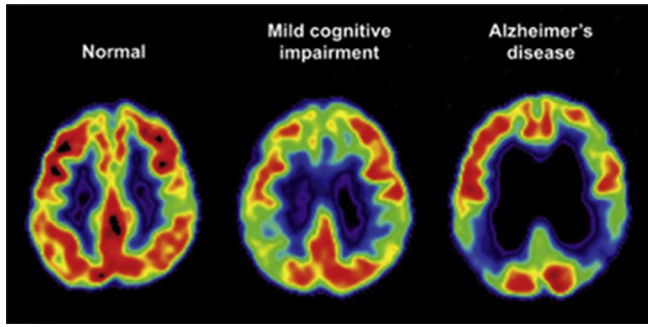

FIGURE 6.8 The image above taken from fluorodeoxyglucose positron emission tomography depicts neuroanatomically specific metabolic differences across a normal, mild cognitively impaired, and Alzheimer's brain.

therefore proved to be particularly difficult, especially when evaluation is performed by less experienced clinicians (Jablensky, 2010; Lee et al., 2018). More objective diagnostic approaches, such as those grounded in multimodal brain imaging, might therefore serve as a viable supplement or alternative for the assessment of schizophrenia and other forms of psychosis. Several studies have therefore tested SVM classifiers in the hope of filling this gap.

One recent study achieved a prediction accuracy of 88.4%, for instance, when employing SVM with RFE to classify schizophrenia based on white matter and gray matter volume measures (Lu et al., 2016). Another recent study found that SVM performance for diagnosing schizophrenia could be improved using fMRI activation patterns during performance of a task involving anticipation of monetary reward (Koch et al., 2015). Using a searchlight MVPA approach, constrained to frontal, temporal, occipital, and midbrain regions, the authors were able to predict schizophrenia diagnosis with peak accuracy of 93% for the right pallidum. Unlike predicting MCI-PDAT conversion, the inherent multimodality of schizophrenia biomarkers perhaps warrants other forms of feature reduction like the searchlight approach. Searchlight works by selecting fewer voxels (e.g., those within a sphere centered at a voxel) and then repeating the analysis (i.e., fitting a new hyperplane) across all voxels in the brain or within some prespecified area of interest. The result of this process is a multivariate information map where each voxel is assigned the classifier's performance (Fig. 6.9). The advantage of searchlight is that, such as RFE, it is viable even in the absence of a priori knowledge about underlying patterns in the data.

Given what is most likely a joint role for structural and functional brain biomarkers in schizophrenia (Birur, Kraguljac, Shelton, & Lahti, 2017; Karlsgodt, Sun, & Cannon, 2010), one promising avenue for SVM-based

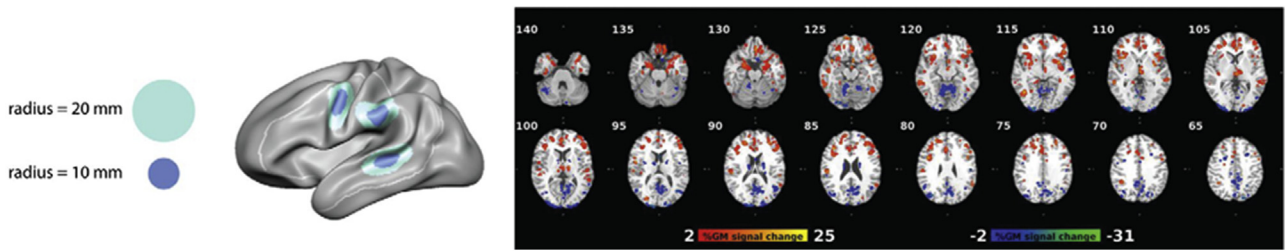

FIGURE 6.9 The left image here depicts the searchlight method, which involves using *spheres* of a given radius centered at each voxel to constrain feature selection while iteratively refitting hyperplanes to the features derived from those *spheres*. The result is a "Multivariate Information Map" as shown in the right mosaic. *Adapted from Cabral, C., Kambeitz-Ilankovic, L., Kambeitz, J., Calhoun, V. D., Dwyer, D. B., Von Saldern, S., et al. (2016). Classifying schizophrenia using multimodal multivariate pattern recognition analysis: Evaluating the impact of individual clinical profiles on the neurodiagnostic performance.* Schizophrenia Bulletin, 42*(1), S110—S117. http://doi.org/10.1093/schbul/sbw053.*

diagnosis of schizophrenia is to aggregate multimodal neuroimaging data using "ensemble feature selection" (Seijo-Pardo, Porto-Díaz, Bolón-Canedo, & Alonso-Betanzos, 2017). The core idea underlying this approach is that the combination of predictive information across multiple modalities will serve to increase classification accuracies. Cabral et al. investigated the joint effects of neurobiological and sociodemographic variables on SVM classification performance, for instance. Specifically, his team applied SVM to both gray matter volume and resting state fMRI (rsfMRI) (an approach to functional imaging where the participant is asked to lie still and do nothing) measures in patients with schizophrenia and HCs, but then further embedded the SVM in a cross-validated feature selection scheme that included moderation analysis using subjects' sociodemographic data, so as to generate an intricate "multimodal diagnostic system" (Cabral et al., 2016). Although the fMRI classifier showed a slightly higher accuracy (70.5%) compared to the structural classifier (69.7%), the combination of sMRI and rsfMRI outperformed single MRI modalities classification by reaching 75% accuracy. Furthermore, specific sociodemographic and clinical variables (e.g., age, difficulty in abstract thinking, emotional withdrawal, and self-reported negative symptoms) proved to be robust moderators of schizophrenia diagnosis ($r = 0.48$, $P < .001$), and when used to predict feature weights for the neuroimaging SVM, classification accuracy on an independent test set approached 100%. Another study similarly used a combination of scalp electrical activity (EEG) data, sMRI, and rsfMRI with an SVM classifier to achieve prediction accuracies nearing 100% (Sui et al., 2014). That study notably employed several sequential feature selection methods that included RFE, t-test filtering, and multiset canonical correlation analysis (MCCA). MCCA, which is intended for multimodal neuroimaging data, can uniquely discriminate between modality-common and modality-unique voxel patterns that can in turn be used as *very* high-dimensional feature inputs that demonstrate just how versatile SVM can be (Fig. 6.10).

6.3.3 Predicting treatment response in depression

Although SVM has been developed to predict MDD diagnosis and onset based on neuroimaging data, a smaller set of studies have sought to predict treatment outcome for MDD. As early as 2009, for instance, one study applied SVM to sMRI data to predict response to antidepressant medication (Costafreda, Chu, Ashburner, & Fu, 2009). The authors found that gray matter volume could be used to predict treatment response with 88.9% accuracy, but only in a small sample of $n = 37$. Gong et al. (2011) therefore attempted to replicate this finding in a slightly larger sample

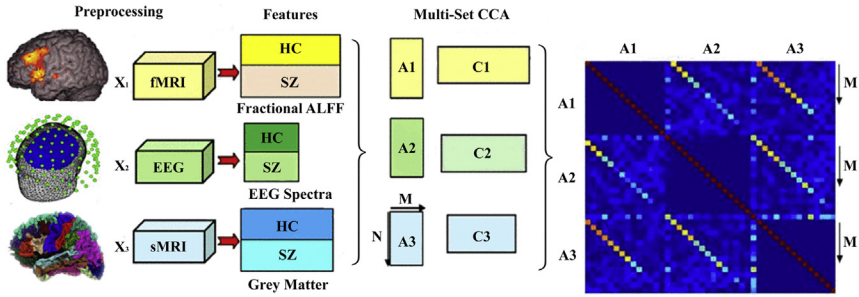

FIGURE 6.10 The above diagram depicts a fusion of features across three different modalities using multiset canonical correlation analysis (MCCA) as a form of ensemble feature selection for predicting Schizophrenia diagnosis using Support Vector Machine. *Taken from Sui, J., Castro, E., He, H., Bridwell, D., Du, Y., Pearlson, G. D., et al. (2014). Combination of FMRI-SMRI-EEG data improves discrimination of schizophrenia patients by ensemble feature selection. In Engineering in medicine and biology society (EMBC), 2014 36th annual international conference of the IEEE (pp. 3889–3892). IEEE. http://doi.org/10.1109/EMBC.2014.6944473.*

of $n = 61$ patients receiving antidepressant medication and including features from both gray matter and white matter volume. Nevertheless, Gong and colleagues found that SVM could only predict clinical outcome (at 3 months follow-up) with accuracies of <70%. SVM as applied to fMRI has revealed similar levels of performance for predicting MDD treatment response. Recently, a treatment study of $n = 80$ patients with MDD receiving antidepressants (randomly assigned to Selective Serotonin Reuptake Inhibitor (SSRI) or Serotonin-Norepinephrine Reuptake Inhibitor (SNRI)) developed an SVM capable of using task-based fMRI patterns to predict antidepressant success. The fMRI features were based on an ROI feature selection approach constrained to the lower amygdala. The study found that pretreatment amygdala activation (using fMRI) in response to subliminal presentation of sad facial expressions could classify treatments responders and nonresponders correctly with an accuracy of 75%.

At present, the translational utility of SVM for predicting MDD treatment response is unclear. In fact, the few existing studies that used SVM to predict treatment response for cognitive impairment and psychosis have similarly shown inferior levels of performance to those models used to predict diagnosis or disease trajectory (Khodayari-Rostamabad, Hasey, MacCrimmon, Reilly, & Bruin, 2010; Orrù et al., 2012). One explanation for this may be that SVM performance is limited when addressing classification problems involving higher complexity in the feature space. While diagnosis of brain disease perhaps largely depends on readily observable neural features already present in an individual's data, treatment success likely further depends on a highly complex variety of factors, such as neuroplasticity, treatment compliance, and social support (Martin Vazquez, 2016).

6.4 Conclusion

As we have seen, SVM is well-suited for addressing a range of classification problems, such as diagnosis or prognosis of brain disorders. The reader may nevertheless still wonder why more advanced classifiers, such as deep neural networks or decision tree learners, would not be superior in some way to SVM for addressing these problems. There may be a natural inclination to assume that model complexity implies model superiority, when in fact it is the *appropriateness of the model* for addressing the question at hand that should generally dictate model selection in supervised machine learning. Compared to other types of classifiers, the power and popularity of SVM largely stem from its ability to achieve balanced performance—high accuracies that are generalizable—even in cases where the dimensionality of the feature space greatly exceeds the number of sample observations available for training. Consequently, SVM has proven to be uniquely well-suited for the study of brain disorders with neuroimaging, where sample sizes are typically much smaller relative to the dimensionality of the feature space. In addition to their economy, SVM also offers *versatility*. As earlier examples in this chapter have shown, many different Kernel functions can be specified for SVM decision functions, and most software allows users to specify custom kernels. This capability facilitates the use of SVM classifiers for addressing linear classification problems, but without the burden of extensive hyperparameter tuning.

Like all machine learning algorithms, SVM is still susceptible to overfitting, perhaps especially due to the increased chance of model selection bias. In the context of neuroimaging, where the number of features typically exceeds observations, additional steps such as the use of nested cross-validation schemes may therefore be needed to avoid overfitting. Although its performance can vary across applications, SVM has proven to be a flexible, efficient, and convenient tool for clinical neuroimaging research that continues to make it a popular choice for classification learning.

6.5 Key points

- SVM is an optimal "hyperplane" which serves to separate (i.e., "classify") observations belonging to one class from another based on linearly or nonlinearly separable patterns of information about those observations called features.
- Compared to other types of classifiers, the power and popularity of SVM largely stem from its ability to achieve balanced performance—high accuracies that are generalizable—even in cases with high dimensionality.

- SVM is well-suited for addressing a range of classification problems, such as the diagnosis and prognosis of brain diseases such as AD, schizophrenia, and depression.
- When employed in neuroimaging analysis, SVM is typically used to perform MVPA, where sets of "voxels" of a structural and/or functional brain image are used as inputs to derive features for classification.

References

American Psychiatric Association. (2013). DSM-5's integrated approach to diagnosis and classifications. In *Diagnostic and statistical manual of mental disorders* (pp. 1–2). DSM-5.

Batuwita, R., & Palade, V. (2013). Class imbalance learning methods for support vector. *Imbalanced Learning: Foundations, Algorithms, Applications*, 83–100. http://doi.org/10.1002/9781118646106.

Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. *Methods in Molecular Biology(Clifton, N.J.), 609*, 223–239. http://doi.org/10.1007/978-1-60327-241-4_13.

Bhaumik, R., Jenkins, L. M., Gowins, J. R., Jacobs, R. H., Barba, A., Bhaumik, D. K., et al. (2017). Multivariate pattern analysis strategies in detection of remitted major depressive disorder using resting state functional connectivity. *NeuroImage: Clinical, 16*, 390–398. http://doi.org/10.1016/j.nicl.2016.02.018.

Birur, B., Kraguljac, N. V., Shelton, R. C., & Lahti, A. C. (2017). Brain structure, function, and neurochemistry in schizophrenia and bipolar disorder—a systematic review of the magnetic resonance neuroimaging literature. *Npj Schizophrenia, 3*(1), 15. http://doi.org/10.1038/s41537-017-0013-9.

Bron, E. E., Smits, M., Niessen, W. J., & Klein, S. (2015). Feature selection based on the SVM weight vector for classification of dementia. *IEEE Journal of Biomedical and Health Informatics, 19*(5), 1617–1626. http://doi.org/10.1109/JBHI.2015.2432832.

Cabral, C., Kambeitz-Ilankovic, L., Kambeitz, J., Calhoun, V. D., Dwyer, D. B., Von Saldern, S., et al. (2016). Classifying schizophrenia using multimodal multivariate pattern recognition analysis: Evaluating the impact of individual clinical profiles on the neurodiagnostic performance. *Schizophrenia Bulletin, 42*(1), S110–S117. http://doi.org/10.1093/schbul/sbw053.

Cabral, C., Morgado, P. M., Campos Costa, D., & Silveira, M. (2015). Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages. *Computers in Biology and Medicine, 58*, 101–109. http://doi.org/10.1016/j.compbiomed.2015.01.003.

Costafreda, S. G., Chu, C., Ashburner, J., & Fu, C. H. Y. (2009). Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PLoS One, 4*(7), e635. http://doi.org/10.1371/journal.pone.0006353.

Costafreda, S. G., Dinov, I. D., Tu, Z., Shi, Y., Liu, C. Y., Kloszewska, I., et al. (2011). Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. *NeuroImage, 56*(1), 212–219. http://doi.org/10.1016/j.neuroimage.2011.01.050.

Cuingnet, R., Benali, H., & Chupin, M. (2010). Spatial and anatomical regularization of SVM for brain image analysis. In *Advances in neural information processing systems* (pp. 460–468).

Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N., & Trojanowski, J. Q. (2011). Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging, 32*(12), 2322. e19. http://doi.org/10.1016/j.neurobiolaging.2010.05.023

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*(1), 44–58. http://doi.org/10.1016/j.neuroimage.2008.06.037.

Gong, Q., Wu, Q., Scarpazza, C., Lui, S., Jia, Z., Marquand, A., et al. (2011). Prognostic prediction of therapeutic response in depression using high-field MR imaging. *NeuroImage, 55*(4), 1497–1503. http://doi.org/10.1016/j.neuroimage.2010.11.079.

Haller, S., Nguyen, D., Rodriguez, C., Emch, J., Gold, G., Bartsch, A., et al. (2010). Individual prediction of cognitive decline in mild cognitive impairment using support vector machine-based analysis of diffusion tensor imaging data. *Journal of Alzheimer's Disease, 22*(1), 315–327. http://doi.org/10.3233/JAD-2010-100840.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage, 62*(2), 852–855. http://doi.org/10.1016/j.neuroimage.2012.03.016.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience, 37*, 435–456. http://doi.org/10.1146/annurev-neuro-062012-170325.

Haynes, J. D., & Rees, G. (2006). Decoding Mental States from Brain Activity in Humans. *Nature Reviews Neuroscience, 7*, 523–534.

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience, 19*(3), 404–413. http://doi.org/10.1038/nn.4238.

Jablensky, A. (2010). The diagnostic concept of schizophrenia: Its history, evolution, and future prospects. *Dialogues in Clinical Neuroscience, 12*(3), 271. http://doi.org/10.1097/ALN.0b013e318212ba87.

Karlsgodt, K. H., Sun, D., & Cannon, T. D. (2010). Structural and functional brain abnormalities in schizophrenia. *Current Directions in Psychological Science, 19*(4), 226–231. http://doi.org/10.1177/0963721410377601.

Khodayari-Rostamabad, A., Hasey, G. M., MacCrimmon, D. J., Reilly, J. P., & Bruin, H. de (2010). A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy. *Clinical Neurophysiology, 121*(12), 1998–2006. http://doi.org/10.1016/j.clinph.2010.05.009.

Koch, S. P., Hägele, C., Haynes, J. D., Heinz, A., Schlagenhauf, F., & Sterzer, P. (2015). Diagnostic classification of schizophrenia patients on the basis of regional reward-related fMRI signal patterns. *PLoS One, 10*(3), 1–18. http://doi.org/10.1371/journal.pone.0119089.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America, 103*(10), 3863–3868. http://doi.org/10.1073/pnas.0600244103.

Lee, J., Chon, M. W., Kim, H., Rathi, Y., Bouix, S., Shenton, M. E., et al. (2018). Diagnostic value of structural and diffusion imaging measures in schizophrenia. *NeuroImage: Clinical, 18*, 467–474. http://doi.org/10.1016/j.nicl.2018.02.007.

Lu, X., Yang, Y., Wu, F., Gao, M., Xu, Y., Zhang, Y., et al. (2016). Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images. *Medicine (United States), 95*(30). http://doi.org/10.1097/MD.0000000000003973.

Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., & Brovelli, A. (2012). Multivoxel pattern analysis for fMRI data: A review. *Computational and Mathematical Methods in Medicine, 2012*. http://doi.org/10.1155/2012/961257.

Martin Vazquez, D. M. J. (2016). Adherence to antidepressants: A review of the literature. *Neuropsychiatry, 6*(5), 236–241. http://doi.org/10.4172/Neuropsychiatry.1000145.

Nho, K., Shen, L., Kim, S., Risacher, S. L., West, J. D., Foroud, T., et al. (2010). Automatic prediction of conversion from mild cognitive impairment to probable Alzheimer's disease using structural magnetic resonance imaging. In *AMIA annual symposium proceedings* (Vol. 2010, p. 542). American Medical Informatics Association.

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience and Biobehavioral Reviews, 36*(4), 1140–1152. http://doi.org/10.1016/j.neubiorev.2012.01.004.

Pettersson-Yeo, W., Benetti, S., Marquand, A. F., Joules, R., Catani, M., Williams, S. C. R., et al. (2014). An empirical comparison of different approaches for combining multimodal neuroimaging data with support vector machine. *Frontiers in Neuroscience, 8*, 189. http://doi.org/10.3389/fnins.2014.00189.

Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Social Cognitive Affective Neuroscience, 2*, 67–70.

Retico, A., Tosetti, M., Muratori, F., & Calderoni, S. (2014). Neuroimaging-based methods for autism identification: A possible translational application? *Functional Neurology, 29*(4), 231.

Rizzi, L., Rosset, I., & Roriz-Cruz, M. (2014). Global epidemiology of dementia: Alzheimer's and vascular types. *BioMed Research International, 2014*. http://doi.org/10.1155/2014/908915.

Rudd, J. M. (2017). Application of support vector machine modeling and graph theory metrics for disease classification. *Model Assisted Statistics and Applications, 13*(4), 341–349.

Schnyer, D. M., Clasen, P. C., Gonzalez, C., & Beevers, C. G. (2017). Evaluating the diagnostic utility of applying a machine learning algorithm to diffusion tensor MRI measures in individuals with major depressive disorder. *Psychiatry research. Neuroimaging, 264*, 1–9. https://doi.org/10.1016/j.pscychresns.2017.03.003.

Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V., & Alonso-Betanzos, A. (2017). Ensemble feature selection: Homogeneous and heterogeneous approaches. *Knowledge-Based Systems, 118*, 124–139. http://doi.org/10.1016/j.knosys.2016.11.017.

Sui, J., Castro, E., He, H., Bridwell, D., Du, Y., Pearlson, G. D., et al. (2014). Combination of FMRI-SMRI-EEG data improves discrimination of schizophrenia patients by ensemble feature selection. In *Engineering in medicine and biology society (EMBC), 2014 36th annual international conference of the IEEE* (pp. 3889–3892). IEEE. http://doi.org/10.1109/EMBC.2014.6944473.

Sundermann, B., Herr, D., Schwindt, W., & Pfleiderer, B. (2014). Multivariate classification of blood oxygen level-dependent fMRI data with diagnostic intention: A clinical perspective. *American Journal of Neuroradiology, 39*(5), 848–855. http://doi.org/10.3174/ajnr.A3713.

Wang, J., Zuo, X., He, Y., Bullmore, E. T., & Fornito, A. (2010). Graph-based network analysis of resting-state functional MRI. *Neuroscience, 4*, 1–14. http://doi.org/10.3389/fnsys.2010.00016.

Weiss, K. M. (2005). A computerized self-report symptom distress inventory: For use as a routine clinical interview in schizophrenia. *Psychiatry, 2*(10), 47.

Whitehead, A., Perdomo, C., Pratt, R. D., Birks, J., Wilcock, G. K., & Evans, J. G. (2004). Donepezil for the symptomatic treatment of patients with mild to moderate Alzheimer's disease: A meta-analysis of individual patient data from randomised controlled trials. *International Journal of Geriatric Psychiatry, 19*(7), 624–633. http://doi.org/10.1002/gps.1133.

Further reading

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Muller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics* (Vol. 8,(February), 1–10. http://doi.org/10.3389/fninf.2014.00014.

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Transactions On Intelligent Systems and Technology (TIST)*, *2*(3), 27. http://doi.org/10.1145/1961189.1961199.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). Pymol: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, *7*(1), 37–53. http://doi.org/10.1007/s12021-008-9041-y.

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 1489–1507. http://doi.org/10.1162/jocn.2007.19.9.1498.

Sabuncu, M. R., & Konukoglu, E. (2014). Clinical prediction from structural brain MRI scans: A large-scale empirical study. *Neuroinformatics*, *13*(1), 31–46. http://doi.org/10.1007/s12021-014-9238-1.

Schrouff, J., Rosa, M. J., Rondina, J. M., Marquand, A. F., Chu, C., Ashburner, J., et al. (2013). PRoNTo: Pattern recognition for neuroimaging toolbox. *Neuroinformatics*, *11*(3), 319–337. http://doi.org/10.1007/s12021-013-9178-1.