INVITED SPECIAL SECTION: METHODS IN EMOTION RESEARCH

emotion review

Emotion Review
Vol. 12, No. 4 (October 2020) 297–308

ISSN 1754-0739
DOI: 10.1177/1754073920930784
https://journals.sagepub.com/home/emr

Support Vector Machines and Affective Science

Chris H. Miller
Department of Psychology, California State University, USA

Matthew D. Sacchet
Center for Depression, Anxiety, and Stress Research, McLean Hospital, Harvard Medical School, USA

Ian H. Gotlib
Department of Psychology, Stanford Neurosciences Institute, Stanford University, USA

Abstract

Support vector machines (SVMs) are being used increasingly in affective science as a data-driven classification method and feature reduction technique. Whereas traditional statistical methods typically compare group averages on selected variables, SVMs use a predictive algorithm to learn multivariate patterns that optimally discriminate between groups. In this review, we provide a framework for understanding the methods of SVM-based analyses and summarize the findings of seminal studies that use SVMs for classification or data reduction in the behavioral and neural study of emotion and affective disorders. We conclude by discussing promising directions and potential applications of SVMs in future research in affective science.

Keywords
affective disorders, affective science, classification, functional magnetic resonance imaging (fMRI), machine learning, support vector machines (SVMs)

Introduction

Support vector machines (SVMs) are a type of classification method and machine learning algorithm increasingly used by affective scientists, and they offer an important alternative to traditional statistical methods typically utilized in the study of emotion and behavior. SVMs provide a powerful, empirically driven method to classify data, generate predictions, and explore structure in highly complex, multivariate data sets. In this review, we provide researchers with a framework for understanding the current methods and procedures of SVMs, review seminal studies that use SVMs in the behavioral and neural study of emotion and affective disorders, and suggest future directions and applications of SVMs in affective science. For an introduction to SVM methods, we refer readers to Casella, Fienberg, and Olkin (2015), which includes practical tutorials and exercises in R. Other available statistical packages include the MATLAB tool fitcsvm (MathWorks, 2017, Release 2017b), Python tool sklearn.svm (Pedregosa et al., 2011), and the specialized package LIBSVM (Chih-Chung & Chih-Jen, 2011).

SVMs and Related Methods

General Strengths of SVMs

An SVM is an empirically driven classification technique and type of supervised machine learning that is used most commonly to assign individual cases with high-dimensional data to two (or more) previously established groups. Whereas traditional statistical techniques designed to examine group differences, such as *t* tests and analysis of variance, typically compare group averages on selected dependent variables, SVMs use a predictive algorithm to learn multivariate patterns that optimally discriminate between groups. SVMs provide several important advantages over these statistical techniques by constructing a classifier to (a) analyze complex data sets composed of a large number of cases with many independent variables; (b) generate predictions about group membership for additional cases; (c) explore the relative and multivariate contributions of included features; (d) select a reduced number of features that most strongly influence group membership; and (e) formally compute the performance or generalizability of the classifica-

Corresponding author: Chris H. Miller, Department of Psychology, California State University Fresno, Fresno, CA 93619, USA.
Email: chmiller@csufresno.edu

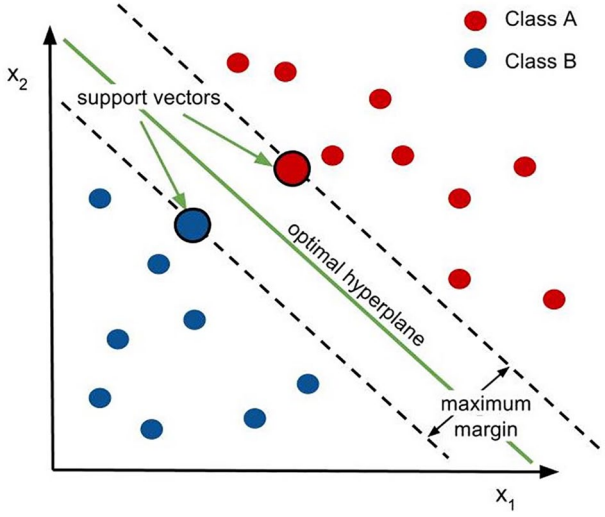

Figure 1. Using a hyperplane to separate labeled classes.
Note. In this simplified example, the optimal hyperplane is shown on a two-dimensional coordinate system as a separating line that maximizes the margin between the support vectors (shown as closed shapes) from two classes (shown as blue squares and red circles). In practice, training an SVM follows this intuition but does so across many predictor variables with complex interactions in multidimensional space.

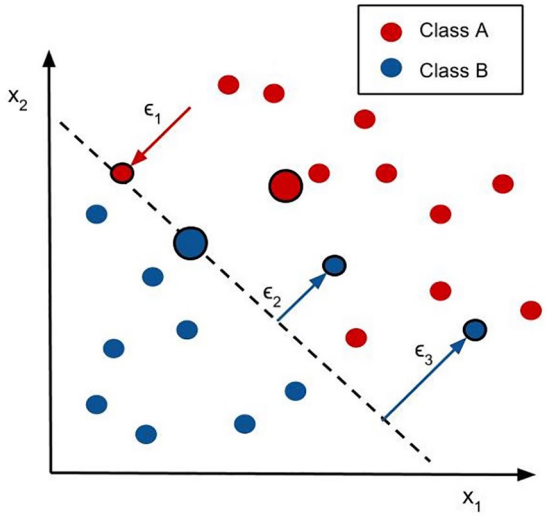

Figure 2. A soft-margin hyperplane.
Note. In this simplified example, a soft-margin hyperplane is permitted to misclassify three cases, each of which corresponds to a particular error term (ξ_i) and results in a penalty in the classifier's optimization problem through use of slack variables (ζ_i,) and a cost function, *C*.

tion solution in the form of predictive accuracy (Casella et al., 2015; Hastie, Tibshirani, & Friedman, 2009).

Using SVMs for Classification

SVMs are typically used as *classifiers* that categorize individual cases into two (or more) groups and characterize observed differences between these groups. As a form of supervised machine learning, SVMs require data sets composed of individual *cases* (e.g., human participants) associated with measured *features* (e.g., neural activation data from many brain regions) that have been sorted into *labeled classes* (e.g., depressed vs. nondepressed). Broadly, an SVM analysis typically consists of three phases: (a) a *training phase*, during which a predetermined portion of the data (i.e., *training set*) is used to construct a classifier or hyperplane capable of discriminating between classes by fitting model parameters; (b) a *validation phase*, during which another portion of the data (i.e., *validation set*) is used for making adjustments to the classifier by tuning its hyperparameters; and (c) a *testing phase*, in which the classifier attempts to sort a number of new cases (i.e., *testing set*) into the specified classes, and its performance is measured by comparing its class predictions to the actual group memberships. For example, when discriminating between depressed and nondepressed participants, researchers might begin by constructing an SVM that uses neural data from a randomly selected subset of individuals, and then use this classifier to predict, based on observed patterns in the neural data, whether each of the remaining cases belongs to the depressed or nondepressed group.

The training phase. First, the training phase involves building a classifier by using a predictive, iterative algorithm to locate a *hyperplane* that best separates the data into two labeled classes. Graphically, a hyperplane can be represented as a dividing line between the two classes of participants (see Figure 1). Because multiple hyperplanes can often separate the two classes, SVMs locate the observations from each class that occur closest to the margin, which serve as the eponymous *support vectors*, and find the maximum-margin hyperplane for which the distance between it and each support vector is as large as possible.

Alternatively, many SVMs implement a *soft-margin* (see Figure 2) that relaxes the requirement that the classifier perfectly separate marginal cases and instead allows it to misclassify individual cases at a specified penalty in order to minimize the possibility of overfitting peculiarities in the training set and improve the model's generalizability. This is accomplished by adding *slack variables* (ζ_n), which indicate the distance required to move a given observation so that it is correctly classified, and a *regularization constant* (i.e., cost function: *C*) that controls the trade-off between minimizing misclassification during training (higher values) versus reducing model complexity to improve generalizability to new data (lower values).

In addition, complex data sets are sometimes not easily separable. In these cases, the *kernel trick* can be used to implicitly transform the training data from input space into higher dimensional feature space where the optimal decision boundary can be located with minimal computation costs (see Figure 3). Although a variety of kernels have been used, most investigators typically select among linear, polynomial, and radial basis function (RBF)/Gaussian kernels, and previous research provides some guidelines for doing so. For example, linear kernels appear to offer superior computational efficiency but inferior predictive performance, although the predictive performance of these types of kernels becomes increasingly comparable when working with a larger number of features. Thus, investigators gener-

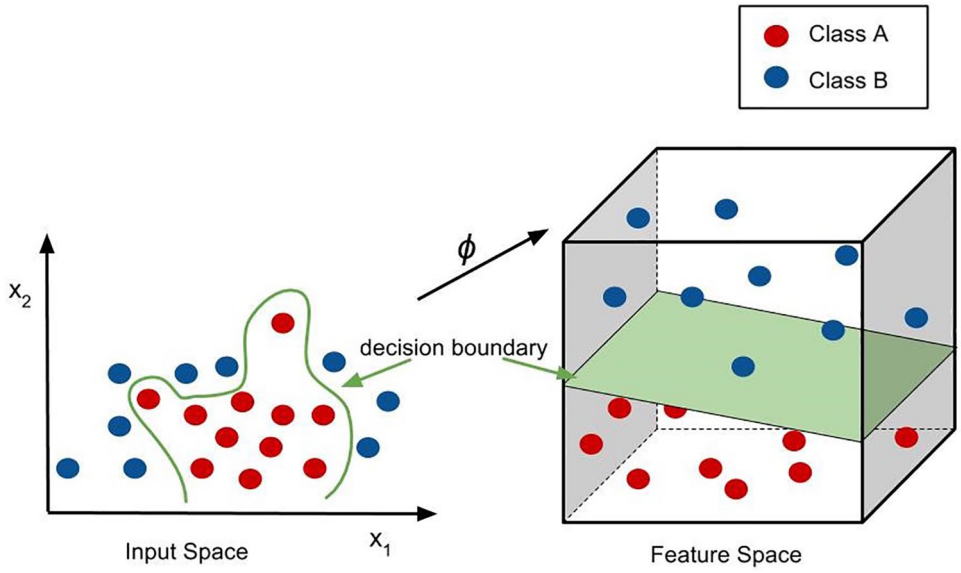

Figure 3. The kernel trick.
Note. In this example, the training data set is composed of two classes. However, the decision boundary between these two classes is nonlinear when represented in input space, so these data can be transformed into a higher dimensional space through use of a kernel function, which reveals a separating hyperplane.

ally recommend selecting a RBF kernel when working with a larger number of features and a linear kernel when working with fewer features (Gaspar, Carbonell, & Oliveria, 2012; Keerthi & Lin, 2003).

Classifier training also requires a *partitioning scheme* to determine the portion of data that will be selected for training. The simplest scheme, the split-half method, involves randomly splitting the study data into two equal-sized sets: one set used for training the classifier and the other set used for validating and testing classifier performance. Other approaches, however, such as 80–20 and 90–10 splitting, in which the majority of the data are used to train the classifier, often show superior performance and may be more suitable for certain data sets and analytic purposes (Hastie et al., 2009).

The validation phase. Although many early SVM analyses proceeded directly to the testing phase, it has now become common practice to include a validation phase that tunes the classifier's *hyperparameters*. These hyperparameters are user-specified values (e.g., regularization constant) or functions (e.g., kernel function) that can be manipulated systematically to construct multiple models and eventually select the classifier with the most desirable performance (i.e., *model selection*), in contrast to the parameters learned by the classifier itself (e.g., feature weights) during the training phase (i.e., *model fitting*).

Importantly, many contemporary SVMs use a *cross-validation* approach to classifier building (Kohavi, 1995). This approach involves iteratively establishing the classifier's initial parameters on a training set and then tuning its hyperparameters on a validation set. The most straightforward way of doing so involves partitioning study data into nonoverlapping sets for training and validation; however, this approach drastically reduces the data available for model building and can lead to a classifier that depends on an idiosyncratic selection of data for these partitions. Consequently, many investigators utilize some form of *resampling* in order to use the available study data as efficiently as possible and minimize the probability of overfitting (Esbensen & Geladi, 2010). For example, in *k-fold cross-validation*, the study data are divided repeatedly into *k* smaller sets, or *folds*; then *k-1* of these folds are used for training and validation, and the remaining fold is used for testing. This process is typically repeated such that every observation is used for training, validation, and testing in order to prevent a single, random partitioning of the data from biasing the resulting classifier. Each of these iterations generates a particular model along with corresponding performance results that are tabulated in order to select a final model that has been trained on the optimal hyperparameters (Varma & Simon, 2006).

The testing phase. The testing phase of classification involves determining the predictive performance of the final model, which is intended to reflect the generalizability of the classifier to subsequent data sets. Accordingly, some investigators recommend using an independent *holdout set* that has not been previously seen by the classifier in order to prevent information learned during training from leaking into the final testing of the classifier and to gain *external* out-of-sample estimates of the ability of the model to classify entirely new data. Other researchers, however, recommend reporting *internal* cross-validation performance, or the mean performance of the final model obtained during cross-validation, in order to better reflect its stability across many possible random partitions of the available data (Cawley & Talbot, 2010).

In either case, the final performance of the classifier is typically expressed as some variation of the percent of cases correctly classified; it is often adjusted for the chance level of performance and should also reflect any necessary considerations for unbalanced groups. For example, in the simplest scenario, a two-class SVM with balanced groups, there is a 50% chance level of performance that can be subtracted from the observed classification accuracy to indicate performance above chance. Overall accuracy can also be decomposed into more specific types of accuracy to reflect the number of true positive, false positive, true negative, and false negative predictions as well as recall and precision scores that can be combined as weighted averages to generate an F_1-score. Another approach to this problem involves using permutation testing to measure the likelihood of obtaining the observed accuracy by chance, which is popular among researchers but can be adversely affected by interdependency among features (Ojala & Garriga, 2010).

Many investigators use a receiver operating characteristic (ROC) curve that plots the classifier's true positive (i.e., sensitivity) versus false positive (i.e., 1 – specificity) rates, and then summarizes performance as a single metric in the form of area under the curve (Hastie et al., 2009). Unfortunately, many measures of predictive performance are adversely influenced by unbalanced groups, and although ROC curves appear least affected by this data skew, they can mask other forms of poor performance under some circumstances (Jeni, Cohn, & de la Torre, 2013; Rakotomamonjy, 2004). Therefore, some investigators have used random subsampling (Foland-Ross et al., 2015) to maintain comparison groups of equal size while still capturing the benefits of using all available data.

The Problem of Overfitting

Although SVMs are useful in analyzing complex data sets, they are highly susceptible to the problem of *overfitting*, which occurs when a particular classifier appears to accurately categorize individual cases during training, but it achieves much lower levels of performance with new data sets. This problem occurs when a classifier is trained on a particular data set and becomes overly sensitive to peculiarities in the training data such that it does not generalize well to other samples. Moreover, classifiers that employ complex decision boundaries or are trained on data sets with a large number of features, both of which are particular strengths of SVMs, are especially susceptible to overfitting—a problem referred to as the *curse of dimensionality*. Consequently, many of the design features and analytic tools used in SVM-based analyses, including soft margins, cross-validation, and holdout sets have been developed to address this fundamental problem and improve generalizability to new data sets (Casella et al., 2015; Hastie et al., 2009).

Using SVMs for Feature Selection

In addition to their utility in classification and prediction, SVMs are often used as a *data reduction* strategy to identify the most discriminative features from a much larger data set, either for theoretical purposes or to minimize problems associated with overfitting (Huang & Wang, 2006). Many feature selection procedures are based on applying a predetermined cutoff to a *rank-ordered list* of variables sorted by their relative contributions to the classifier. For example, studies of resting-state fMRI may examine functional connectivity between every pairwise combination of voxels throughout the whole brain, resulting in hundreds of thousands of neural features that would be theoretically uninterpretable and would likely produce serious overfitting problems in even the largest available data sets. By initially fitting a classifier to a training set composed of this full set of features, however, researchers can compute *feature weights* that reflect the relative ability of each feature to discriminate between classes and then retain either a predetermined number of features or only those features that reach a particular feature weight threshold. It is important to note, however, that each of these feature weights reflects the contribution of a single variable in the multivariate environment used during classifier training rather than the independent contribution of each individual variable. Alternatively, recursive feature elimination enables researchers to include feature selection as a part of the model-building process and determine the optimal number of features to include. In particular, this algorithm begins by fitting a model using all available features and then iteratively removing the lowest ranked features and refitting a model with the top predictors, based on their absolute discriminative weights; this process continues until it reaches an empty set, at which point the highest performing model is selected, or until a termination condition such as number of desired features is reached (De Martino et al., 2008).

Comparison of SVMs and Other Methods

SVMs are most frequently used for classification problems and offer an alternative to other statistical techniques such as *logistic regression* (LR) and *linear discriminant analysis* (LDA), which are used for similar purposes but involve a different set of statistical assumptions and mathematical operations. Although each of these classification methods offers a somewhat unique set of advantages and disadvantages, in practice, the solutions derived from each method frequently resemble each other.

Logistic regression. LR is frequently used to classify individual cases into two groups and serves a similar purpose to that of SVMs. Its mathematic model, the *logistic function*, can be viewed as a special case of the generalized linear model and is a continuous function that resembles an "S" shape (sigmoid curve). It is composed of a linear combination of predictor variables and regression coefficients and is used to estimate the probability of group membership; individual cases are assigned to a particular group according to whether their resulting probability values fall above or below a threshold, which by convention is typically set at $p = .50$ (see Figure 4). The model-fitting process is often based on maximum likelihood estimation, which involves an iterative process that begins with a tentative solution and continues with incremental revisions until it converges on an open-form expression that appears to maximize the *likelihood function*, or the agreement between the observed data and selected model (Cox, 1958; Menard, 2002).

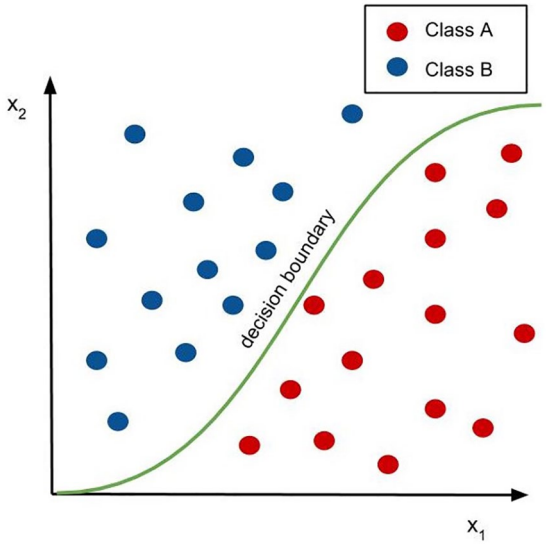

Figure 4. Using a logistic function to separate labeled classes.
Note. In this example, a logistic function is used to separate cases into two classes based on the probability of membership in two classes.

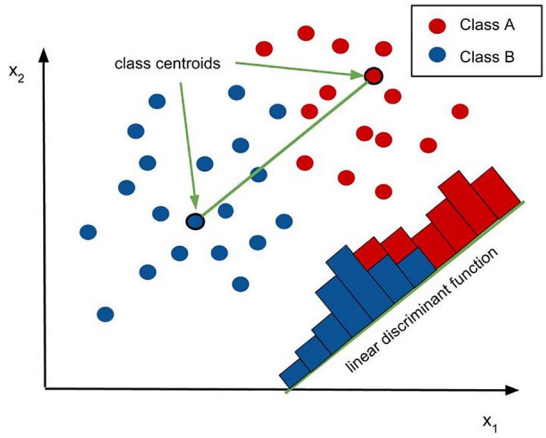

Figure 5. Using a linear discriminant function to separate labeled classes.
Note. In this example, a linear discriminant function is used to separate cases into two classes based on their values along a one-dimensional subspace that passes through the centroids of each class.

LR attempts to optimally fit the training set under the assumption that the predictor variables are not sufficient to determine the response variable. It performs well with data sets that are characterized by a small number of noisy predictor variables that only provide a probabilistic estimate of the outcome variable, but it tends to perform poorly with data sets that have a more determinate response variable. It is also susceptible to overfitting and poor generalizability when using a large number of predictor variables, and it performs poorly in high-dimensional spaces, particularly near the margin. LR also often fails to select the optimal solution in cases where multiple separating hyperplanes are possible because solutions with low predictive power at the margins may still achieve maximum likelihood. In contrast, SVMs tend to generalize better to testing sets, particularly when regularization parameters are used to minimize model complexity and prevent overfitting. They are also more appropriate with data sets that include a large number of predictor variables or more determinate outcome variables, as they use support vectors that are based on points near the margins and favor a hyperplane that best separates these data points with the widest possible margin. They can also better separate data sets with complex boundaries by modeling nonlinear solutions with well-established RBF kernels (Mernard, 2002; Pochet & Suykens, 2006).

Linear discriminant analysis. LDA can be used to address similar classification problems as SVM but relies on a somewhat different approach. Its mathematical model, the *linear discriminant function*, is also a composite of predictor variables and estimated coefficients, but it transforms observations into a new dimension, which corresponds to the highest *eigenvalue*, in such a way that the distance between group means, or *centroids*, is maximized. Graphically, this discriminant function passes through the centroids of the two groups and is used to predict the probability of membership in the two classes, based on the entire training set (see Figure 5). It is also based on the more restrictive assumption that the predictor variables are normally distributed. Thus, LDA tends to perform well with predictor variables that are normally distributed and generate probabilistic rather than determinative predictions about group membership and when modeling class differences based on group means rather than observations at the boundary is preferable. However, LDA is sensitive to outliers in the training data, performs more poorly with observations at the margin, and produces biased estimates when assumptions of normality are violated; in contrast, SVMs are more flexible, as they rely on fewer assumptions, and they perform better when a large number of variables generate nearly certain predictions about class membership and when predicting group membership of marginal cases (Pohar, Blas, & Turk, 2004).

In addition to these guidelines based on theoretical properties and statistical assumptions, other empirical findings from planned comparisons and controlled simulations have informed our understanding of performance differences among different classification methods and the conditions under which each approach is most successful (see Table 1). For example, De Smet et al. (2006) compared SVM and LR-based classification to predict depth of infiltration in endometrial carcinoma patients based on transvaginal sonography, and found that the SVM classifier performed significantly better than the LR classifier in prospective predictions, as measured by area under the curve (AUC) at 77% and 66%, respectively, although the training set performances did not differ significantly between the two approaches, which suggests that overfitting problems substantially affected LR performance. Salazar, Velez, and Salazar (2012) analyzed performance of SVM and LR with simulated data from six types of statistical distributions as well as real microarray data in predicting disease status of several types of

Table 1. General guidelines for selecting among classification methods.

	Support vector machine (SVM)	Logistic regression (LR)	Linear discriminant analysis (LDA)
Modeling technique	Directly models decision boundary (discriminant)	Directly models decision boundary (discriminant)	Models class distributions (generative)
Separation technique	Hyperplane with maximum distance between marginal cases of each class	Logistic function of probability of membership in each class	Linear function between centroids of each class
Linearity	Linear or nonlinear (with kernel)	Nonlinear logistic function and linear decision boundary	Linear
Appropriate predictor variables	Large number of determinant predictor variables	Small number of probabilistic predictor variables	Normally distributed, probabilistic predictor variables
Appropriate sample size	Large	Large	Small
Computational time	Fast	Slow	Slow

Note. Comparison of three widely used classification methods.
SVM: support vector machine; LR: logistic regression; LDA: linear discriminant analysis.

cancer patients, and found that SVMs performed equal to or better than LR, as measured by misclassification rate, with most statistical distributions and mixed data sets. In contrast, Chen et al. (2009) found mixed results in the comparison of SVM and LR in diagnosing malignant versus benign tumors using a database of breast ultrasound volumes. In this study, LR performed better than SVM, particularly at local regions of the ROC, when using 3D power Doppler imaging; however, SVMs performed better than LR and completed training and diagnosis at faster rates when using texture analysis.

Other investigators have examined differences in performance between *discriminant classifiers* (e.g., SVMs and LR), which directly model the decision boundary between classes—that is, *conditional probability distribution*, P (y|x)—and then use this model to predict class membership of each new observation, and *generative classifiers* (e.g., LDA), which begin by modeling individual class distributions—that is, *joint probability distribution*, P (x,y)—and then use this model to select the class with the higher probability for each new observation. These studies suggest that discriminative classifiers tend to perform better than generative classifiers, presumably, in part, because they involve fewer assumptions, but that generative classifiers tend to reach their asymptotic error rate more quickly during training and hence perform better than discriminative classifiers in especially small data sets (Ng & Jordan, 2001).

SVMs and Behavioral Studies of Emotion

Over the past few decades, researchers have begun to use machine learning tools such as SVMs in the behavioral study of human emotion. These tools and the resulting algorithms have been used largely to distinguish among various emotions on the basis of facial expressions, speech/prosody, and physiological data, and have been applied to help solve problems in affective computing to improve interaction between human users and machine interfaces as well as in health monitoring settings to improve the early detection of disease states or processes.

Facial Expression Analysis

Machine learning tools such as SVMs have been widely used to automatically identify facial expressions from standardized image databases as well as real-world images; these efforts frequently achieve over 90% classification accuracy, with chance agreement ranging from 14% to 33% (Pantic & Rothkrantz, 2000). The process of training a classifier to recognize emotional expressions generally involves three steps: face detection, data extraction, and expression classification. Investigators have developed increasingly sophisticated tools to improve performance and generalizability in each of these areas. These classifiers typically involve either a single binary decision or multiple binary decisions that can be linked in hierarchical fashion to select among several different types of basic emotions. For example, An, Yang, and Bhanu (2015) developed a smile-detection classifier based on extreme learning machine methods, and successfully identified naturalistic facial expressions (smile vs. no-smile) from several large databases containing over 7,500 images with 93.1% classification accuracy (chance agreement $\leqslant$ 57%). Their approach incorporated automatic face detection, feature extraction, and facial registration without any manual labeling or key-point detection, which is an important step toward fully automatic, real-time facial expression recognition.

Other investigators have developed SVM-based algorithms that classify sample faces into multiple categories. For example, Susskind, Littlewort, Bartlett, Movellan, and Anderson (2007) used six independent classifiers, each of which distinguished between an acted neutral face and basic emotion (i.e., anger, disgust, fear, happiness, sadness, and surprise). Sample images were given a standardized rating from each of the six classifiers and the highest rating was selected, in winner-takes-all fashion. This approach resulted in a mean classification accuracy of 79.2% (chance agreement = 17%), compared to human performance at 89.2% accuracy, and was highest for happiness, sadness, and surprise (100%) followed by anger and disgust (75%) and fear (25%). Song, Han, and Hong (2010) developed an online learning approach that adapted to new, acted facial samples following an initial training phase and used support vector pursuit learning in order to reduce the number of training data

stored and minimize computational requirements, which allowed their classifier to successfully operate in real time. Their algorithm successfully classified new faces into five emotional expressions (anger, happiness, neutral, sadness, and surprise) with 92.7% classification accuracy (chance agreement = 20%), based on automatic extraction of 12 distance measurements between the eyes, eyebrows, and lips.

Finally, Tan et al. (2016) classified induced facial expressions based on facial electromyography data taken from the corrugator supercilii (i.e., "frowning muscle") and zygomaticus major (i.e., "smiling muscle") of adult participants as they viewed the presented images. This algorithm classified images into five distinct categories from the circumplex model of emotion (Larsen & Diener, 1992): neutral valence and low arousal, positive valence and high arousal, positive valence and low arousal, negative valence and high arousal, and negative valence and low arousal. The classifier achieved accuracy scores of 75.69% to 100.00% (chance agreement = 20%); it also showed significant differences in performance between younger and older, but not between female and male, participants.

Speech Emotion Recognition Systems

Researchers have also used SVM-based methods in behavioral studies of human emotion to examine emotional expressions in speech/prosody. Most of these speech emotion recognition (SER) systems are based on prosodic features such as pitch, energy, and speaking rate extracted from audio recordings of professional actors instructed to express particular emotions. They also frequently involve distinguishing among multiple emotions or between challenging binary pairs of emotions while still maintaining classification accuracies. For example, Harimi, AhmadyFard, Shahzadi, and Yaghmaie (2015) developed a SER system to analyze both prosodic and spectral features from a database composed of 535 samples provided by 10 professional actors. They noted that while most SER systems successfully discriminate between emotions on the basis of arousal (e.g., anger vs. sadness), they are far less successful in distinguishing between emotions on the basis of valence (e.g., anger vs. joy), which produces the majority of errors in many SER systems. Harimi et al. addressed this problem by using a linear SVM and nonlinear features to classify angry versus joyful voices, the most challenging binary comparison, with 99.1% (chance agreement = 40–60%) and 98.85% accuracy (chance agreement = 31–69%) for female and male voices, respectively. Harimi et al. also extended this approach to multiemotional problems consisting of seven different emotions (anger, boredom, disgust, fear, joy, neutral, sadness) and achieved classification accuracies of 94.58% (chance agreement = 5–20%). Other investigators have also extended binary SVMs to multiemotional classification problems using one-versus-one and one-versus-rest classifiers or hierarchical approaches designed to minimize the number of required features; these researchers have achieved classification accuracies up to 94.7% (chance agreement ⩽ 20%) for five-class problems (Hassan & Damper, 2012; Lee, Mower, Busso, Lee, & Narayanan, 2011).

Physiological Signal Analysis

Researchers have also used SVMs to distinguish among various emotions on the basis of physiological data. For example, Verma and Tiwary (2014) used a combination of modeling, clustering, and classification approaches to distinguish among 13 emotional states in 32 subjects from the Database for Emotion Analysis Using Physiological Signals, in which subjects' physiology was recorded as they rated 40 one-minute excerpts of music videos (Koelstra et al., 2012). Statistical analyses relied on a multimodal fusion approach that combined central (electroencephalography [EEG]) as well as several peripheral (Galvanic skin response, electromyography, electroculogram, blood volume pressure, respiration pattern, and skin temperature) measures of physiological changes. These data were used first to validate an a priori three-dimensional model of emotion consisting of valence, arousal, and dominance dimensions, from which a cluster analysis then produced the following five-cluster solution: (1) happiness/joy/fun/excitement; (2) love/cheerfulness/pleasure; (3) anger/hate; (4) sadness; and (5) melancholy/sentimentality. Verma and Tiwary (2014) then used these data to train four independent classifiers that distinguished among emotional states and found that SVM outperformed other methods, with classification accuracies ranging from 77.96% (love; fun) to 80.28% (cheerful) for 13 different emotions (chance agreement ⩽ 13%).

SVMs and Neural Studies of Emotion

Human neuroimaging, including structural modalities such as MRI and functional modalities such as fMRI and EEG, has become an important tool in exploring the neural structures and brain networks involved in emotional processing. Because neuroimaging scans assessments typically generate large amounts of multidimensional data that can be used in empirically driven ways to improve prediction, SVM-based methods are being used increasingly for both exploratory analyses and data reduction purposes, particularly in decoding studies (Naselaris, Kay, Nishimoto, & Gallant, 2011) that predict emotional states from selected neural features.

Decoding in Emotional Face Paradigms

In addition to classifying emotional face images, SVM-based methods and multivoxel pattern analysis have also been used to analyze neural data obtained while participants view emotional face images. Building on methods and findings from vision neuroscience (e.g., Kanwisher, McDermott, & Chun, 1997; Kay, Naselaris, Prenger, & Gallant, 2008; Nishimoto et al., 2011), Zhang et al. (2016) successfully classified facial expression images into four emotions (neutral, fearful, angry, and happy) by decoding fMRI activation signals taken from voxels distributed across face-selective regions of interest (ROIs). These face-selective ROIs, which were empirically defined for each subject by localizer scans, included the fusiform face area (FFA), amygdala, superior temporal sulcus, and anterior inferior

temporal cortex. Furthermore, Zhang et al. found that these four regions exhibited important functional differences: the FFA and anterior inferior temporal cortex successfully discriminated among different participants on the basis of facial identity; the superior temporal sulcus performed better than other regions in distinguishing between neutral versus emotional (i.e., fearful/angry/happy) faces; and the amygdala performed better than other regions at classifying fearful versus nonfearful faces. Other studies have used similar methods to decode emotional expressions from static images (Harry, Williams, Davis, & Kim, 2013) or dynamic videos (Said, Moore, Engell, & Haxby, 2010), based on fMRI activation data taken from other brain regions, including the FFA, frontal operculum, and early visual cortex as well as EEG data taken from right occipital areas (Hidalgo-Munoz et al., 2013).

Decoding in Other Emotional Paradigms

Skerry and Saxe (2014) used SVM-based methods to successfully classify emotions as positive versus negative valence based on neural data taken from the medial prefrontal cortex as participants viewed facial expression images or animated videos in which a character's emotion could only be identified from the situation. Although both classifiers were independently trained using one stimulus set (e.g., facial expressions), they were able to successfully classify stimuli from the other set (e.g., animated situations), demonstrating sensitivity to emotional valence that generalizes across different stimulus types, including perceived versus inferred emotions as well as static images versus dynamic animations. Moreover, a classifier trained on animated situations also successfully discriminated between trials in which participants received rewards in the form of monetary gains (i.e., positive valence) versus punishments in the form of monetary losses (i.e., negative valence), based on neural features from a particular subregion of the medial prefrontal cortex, suggesting that neural representations in some areas also generalize across attributed versus experienced emotions.

Decoding in Real-Time fMRI

Other investigators working with brain–computer interfaces have developed online SVMs capable of decoding brain states in real time and providing immediate neurofeedback to participants (LaConte, 2011). For example, Sitaram et al. (2011) trained an online classifier to distinguish among happiness, sadness, and disgust in real time using fMRI activation signals from the whole brain as well as from a priori ROIs in healthy individuals while they recalled emotionally salient events from their personal lives. Hollmann et al. (2011) developed an online classifier to decode whole-brain fMRI activation data taken from participants during social interaction while playing the ultimatum game, a commonly used paradigm in game theory in which a participant decides to either accept or reject a monetary proposal from another player. This classifier distinguished, with a predictive accuracy of about 70% (chance agreement $\leqslant$ 56.8%), between motivational states leading players to either accept or reject these offers before they communicated their decisions.

SVMs and the Neural Studies of Affective Disorders

Disorders of emotion, including major depressive disorder (MDD) and bipolar disorder (BD), are among the most burdensome psychiatric illnesses, both for the individual and for society (Merikangas et al., 2007; Whiteford et al., 2013). MDD is characterized by profound changes in affect, including low mood and loss of pleasure (anhedonia) in addition to symptoms related to motivation, sleep, appetite, attention, and psychomotor processes. Individuals with BD experience depressive episodes as well as at least one episode of mania, a state characterized by high levels of arousal and other behavioral abnormalities.

SVMs have been used to explicate relations in high-dimensional data to help improve the characterization and treatment of these disorders. Notably, SVMs allow researchers to categorize individual participants into classes that are difficult to define explicitly, in contrast to traditional tools that are typically limited to drawing group-level inferences; this feature of SVMs is particularly important in efforts to develop translational neuroscience tools that will help inform clinical decisions about specific patients. Studies of SVMs in affective disorders have focused on identifying patterns of brain activity that can be used to corroborate psychiatric diagnoses as well as predict treatment outcome and prognosis (Kipli, Kouzani, & Williams, 2013; Orrù, Pettersson-Yeo, Marquand, Sartori, & Mechelli, 2012).

Identification of Major Depressive Disorder

The most common application of SVMs in the study of affective disorders is to diagnose and identify individuals with psychiatric disorders. A large and growing literature suggests that MDD can be identified using SVMs and neuroimaging features (for reviews, see Kipli et al., 2013; Orrù et al., 2012). These studies have used either functional (Fu et al., 2008; Hahn et al., 2011; Marquand, Mourao-Miranda, Brammer, Cleare, & Fu, 2008; Nouretdinov et al., 2011; Patel et al., 2015; Zeng et al., 2012) or structural (Costafreda, Chu, Ashburner, & Fu, 2009; Gong et al., 2011; Mwangi, Ebmeier, Matthews, & Douglas Steele, 2012; Nouretdinov et al., 2011; Patel et al., 2015; Sacchet, Prasad, Foland-Ross, Thompson, & Gotlib, 2015) features derived from MRI. In a recent meta-analysis, Kambeitz et al. (2016) analyzed 33 primary studies that collectively included 912 patients with MDD and 894 healthy controls, and found classification accuracies of 77% sensitivity and 78% specificity; they also conducted separate analyses for several neuroimaging modalities and found that each modality achieved somewhat different classification accuracies: structural MRI (70% sensitivity, 71% specificity), diffusion tensor imaging (88% sensitivity, 92% specificity), task-based fMRI (75% sensitivity, 77% specificity), and resting-state fMRI (85% sensitivity, 83% specificity). To date, Zeng et al. (2012), who assessed resting-state fMRI

functional connectivity in 24 individuals with MDD and 29 healthy controls, have reported the highest classification accuracy of 100% in patients and 89.7% in healthy controls; in this study, connectivity of the amygdala exhibited the strongest discriminative power. In order to confirm the generalizability of these findings, however, it is critical that researchers conduct large-cohort, multisite studies that compare the accuracy of models trained on one site and tested on data from other sites.

Identification of Bipolar Disorder

Several studies have used SVMs to differentiate individuals with BD from healthy controls. Costafreda et al. (2011) assessed fMRI features derived from a verbal fluency task in 32 patients with BD and 40 healthy controls and achieved rates of 56% sensitivity and 89% specificity. Schnack et al. (2014) used gray matter density features from 66 individuals with BD and 66 healthy controls to achieve 55% sensitivity and 63% specificity. Redlich et al. (2014) analyzed similar gray matter density features from 29 individuals with BD and 29 healthy controls and achieved 75.9% and 65.5% accuracies at two different within-sites trials.

Differentiating Major Depressive Disorder From Bipolar Disorder

Affective disorders are often misdiagnosed, which can lead to considerable negative outcomes for patients, including prolonged suffering from delays in providing effective treatment (Singh & Rajput, 2006). Machine learning is promising in improving the differentiation among different types of affective disorders. Several studies have used SVMs to distinguish MDD from BD. In the first of these studies, Redlich, et al. used voxel-based morphometry (VBM) to differentiate individuals diagnosed with MDD from those with BD across two sites. The features were selected based on neural models of emotion regulation and included volumetric measurements from prefrontal cortex, amygdala, thalamus, striatum, and hippocampus. Classification accuracy was 75.9% and 65.5% when training and testing were conducted within-sites (using cross-validation), and 63.8% and 69.0% when training was conducted between-sites. Similarly, Sacchet, Livermore, Iglesias, Glover, & Gotlib (2015) used subcortical structures to differentiate 57 individuals with MDD from 40 individuals with BD and achieved classification rates ranging from 56.0% to 62.9% (chance agreement = 41–59%). In addition, Grotegard et al. (2014) used fMRI activation during presentation of several types of emotional faces to distinguish between 22 individuals with MDD and 22 individuals with BD and achieved a range of classification rates from 56.8% to 79.6% (chance agreement $\leqslant$ 52%), with the highest results for the contrast of sad versus happy faces. These classification accuracies, which are somewhat lower when distinguishing between individuals with MDD and BD compared to accuracies obtained when distinguishing between a single disorder and healthy controls, suggest greater difficulty in distinguishing between different psychiatric disorders compared to a given psychiatric disorder and healthy controls, presumably due to similarity in many neural features among psychiatric disorders.

Treatment Prediction in Affective Disorders

Investigators have used SVMs to predict the outcomes of interventions in affective disorders. These studies typically compare depressed individuals who responded to a particular treatment (treatment-responsive) to those who did not (treatment-resistant). In the first of these studies, Costafreda, Ashburner, & Fu (2009) were able to correctly predict clinical remission following administration of fluoxetine with 88.9% sensitivity and 88.9% specificity in a cohort of 18 depressed individuals using features derived from VBM. The same group used fMRI features from an implicit sad-face viewing task and was able to predict treatment response to cognitive behavioral therapy in 16 individuals with MDD with a sensitivity of 71% and specificity of 86% (Costafreda, Khanna, Mourao-Miranda, & Fu, 2009). Using VBM and SVMs, Gong et al. (2011) reported 65–70% accuracy in differentiating 46 depressed individuals who did and did not respond to a variety of antidepressant treatments (50% chance agreement) using gray and white matter features. Similarly, Liu et al. (2012) used VBM to differentiate 35 individuals with MDD who were either antidepressant treatment-resistant or treatment-responsive and achieved 82.9% accuracy (chance agreement $\leqslant$ 52%) with both white and grey matter features. More recently, Patel et al. (2015) used multimodal functional and structural imaging features to distinguish between 33 antidepressant treatment-responsive and treatment-resistant individuals with late-life depression. The most accurate classification of treatment response achieved 88.89% sensitivity and 90% specificity and used structural and functional connectivity features combined with age, gender, and education. While these initial findings are promising, it will be important in future research to investigate large-scale multitreatment studies in order to identify optimal treatments on an individual-by-individual basis.

SVMs and the Prediction of the Onset of Affective Disorders

Only one study, conducted by Foland-Ross et al. (2015), has used SVMs to predict the onset of affective disorder in currently healthy individuals. In that study, structural scans were collected from 33 never-disordered adolescents, who were assessed regularly for the onset of depression over a period of 5 years. Using SVMs, baseline cortical thicknesses from regions implicated in MDD and emotion regulation were used to predict the onset of depression with 69% sensitivity and 70% specificity. Feature weights indicated that right medial orbitofrontal, right precentral, left anterior cingulate, and bilateral insula were most important in constructing the SVMs. This study provides promising evidence that SVMs can be used to predict the onset of depression and, perhaps, that they can therefore be useful in prevention programs. Future research should examine the biological mechanisms underlying these trajectories and explicitly assess any clinical gains that are associated with using a machine-learning approach over traditional clinical procedures.

Limitations and Future Directions

SVMs have proven to be useful analytic tools for data reduction and classification in the scientific study of human emotion. They have demonstrated relatively high levels of accuracy in discriminating among various emotions on the basis of facial expressions, speech/prosody, and physiological signals, as well as promising results in classifying individuals diagnosed with mood disorders and predicting responses to standardized treatments.

Although SVMs have been successful in these types of classification and prediction settings, thus far they have been limited in their use as fundamental discovery tools. In this regard, only a small number of studies have moved beyond simple classification or prediction based on a limited selection of relevant features. Indeed, SVMs are designed to work with all of the specified data to achieve optimal predictive or classification accuracy, often with limited attention to selecting appropriate features and minimal regard for whether the resulting models are sufficiently parsimonious or cohesive to inform the development of scientific theories. For example, SVMs have shown promising results in classifying emotional states or clinical groups by rather indiscriminately using a large number of neural features extracted from brain scans; however, these studies have not yet built a convincing case about what types of neural features (e.g., functional activation vs. functional connectivity; insula vs. amygdala seeds) are most discriminative nor have they integrated findings into a cohesive mechanistic account of the neural basis of emotion. Thus, future work should integrate SVM analyses with feature selection techniques and traditional methods of hypothesis testing to generate a more parsimonious account of the underlying mechanisms responsible for producing distinct classes of emotions or patient groups. This is particularly pressing in clinical settings in which investigators strive not only to predict patient responses to available treatments but also to improve diagnostic systems, to inform our understanding of the neural mechanisms responsible for differences in treatment response and to generate novel targets for intervention. In addition, some investigators have now turned their attention to tool development that strives to adapt SVMs to address more complex structural differences between groups; to examine multigroup (Susskind et al., 2007), multimodal (Verma & Tiwary, 2014), and nonlinear (Harimi et al., 2015) classification problems; to generate probabilistic predictions (Hollmann et al., 2011; Sacchet, Livermore, Iglesias, Glover, & Gotlib, 2015); and to build regression-based models (Qin et al., 2015). Future studies should continue to develop such classification and group-comparison tools, and characterize their relative strengths and limitations, in order to build better tools that enable us to explore important questions that remain open in affective science.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iD

Chris H. Miller https://orcid.org/0000-0001-6581-6375

References

An, L., Yang, S., & Bhanu, B. (2015). Efficient smile detection by extreme learning machine. *Neurocomputing*, *149*(Pt. A), 354–363.

Casella, G., Fienberg, S., & Olkin, I. (2015). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Chen, S. T., Hsiao, Y., Huang, Y., Kuo, S., Tseng, H., Wu, H., & Chen, D. (2009). Comparative analysis of logistic regression, support vector machine, and artificial neural network for the differential diagnosis of benign and malignant solid breast tumors by the use of three-dimensional power Doppler imaging. *Korean Journal of Radiology*, *10*, 464–471.

Chih-Chung, C., & Chih-Jen, L. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3). https://doi.org/10.1145/1961189.1961199

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

Costafreda, S., Fu, C., Picchioni, M., Toulopoulou, T., McDonald, C., Kravariti, E., . . . McGuire, P. (2011). Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder. *BMC Psychiatry*, *11*(18), 1–10.

Costafreda, S., Khanna, A., Mourao-Miranda, J., & Fu, C. (2009). Neural correlates of sad faces predict clinical remission to cognitive behavioural therapy in depression. *NeuroReport*, *20*, 637–641.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society*, *20*, 215–242.

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.

De Smet, F., De Brabanter, J., van den Bosch, T., Pochet, N., Amant, F., van Holsbeke, C., . . . Timmerman, D. (2006). New models to predict depth of infiltration in endometrial carcinoma based on transvaginal sonography. *Ultrasound in Obstetrics & Gynecology*, *27*, 664–671.

Esbensen, K., & Geladi, P. (2010). Principles of proper validation: Use and abuse of re-sampling for validation. *Journal of Chemometrics*, *24*, 168–187.

Foland-Ross, L., Sacchet, M., Prasad, G., Gilbert, B., Thompson, P., & Gotlib, I. (2015). Cortical thickness predicts the first onset of major depression in adolescence. *International Journal of Developmental Neuroscience*, *46*, 125–131.

Fu, C., Mourao-Miranda, J., Costafreda, S., Khanna, A., Marquand, A., Williams, S., & Brammer, M. (2008). Pattern classification of sad facial processing: Toward the development of neurobiological markers in depression. *Biological Psychiatry*, *63*(7), 656–662.

Gaspar, P., Carbonell, J., & Oliveria, J. (2012). On the parameter optimization of support vector machines for binary classification. *Journal of Integrative Bioinformatics*, *9*(3). https://doi.org/10.1515/jib-2012-201

Gong, Q., Wu, Q., Scarpazza, C., Lui, S., Jia, Z., Marquand, A., . . . Mechellia, A. (2011). Prognostic prediction of therapeutic response in depression using high-field MR imaging. *NeuroImage*, *55*, 1497–1503.

Grotegard, D., Stuhrmann, A., Kugel, H., Schmidt, S., Redlich, R., Zwanzger, P., . . .Dannlowski, U. (2014). Amygdala excitability to subliminally presented emotional faces distinguishes unipolar and bipolar depression: An fMRI and pattern classification study. *Human Brain Mapping*, *35*, 2995–3007.

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. *Archives of General Psychiatry*, *68*, 361–368.

Harimi, A., AhmadyFard, A., Shahzadi, A., & Yaghmaie, K. (2015). Anger or joy? Emotion recognition using nonlinear dynamics of speech. *Applied Artificial Intelligence*, *29*(7), 675–696.

Harry, B., Williams, M., Davis, C., & Kim, J. (2013). Emotional expressions evoke a differential response in the fusiform face area. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00692

Hassan, A., & Damper, R. (2012). Classification of emotional speech using 3DEC hierarchical classifier. *Speech Communication*, *54*(7), 903–916.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer.

Hidalgo-Munoz, A., Pereira, A., Lopez, M., Galvao-Carmona, A., Tome, A., Vazquez-Marrufo, M., & Santos, I. (2013). Individual EEG differences in affective valence processing in women with low and high neuroticism. *Clinical Neurophysiology*, *124*(9), 1798–1806.

Hollmann, M., Rieger, J., Baecke, S., Lutzkendorf, R., Muller, C., Adolf, D., & Bernarding, J. (2011). Predicting decisions in human social interactions using real-time fMRI and pattern classification. *PLoS ONE*, *6*(10). https://doi.org/10.1371/journal.pone.0025304

Huang, C., & Wang, C. (2006). A GA-based feature selection and parameters optimization for support vector machines. *Expert Systems with Applications*, *31*, 231–240.

Jeni, L., Cohn, J., & de la Torre, F. (2013). Facing imbalanced data recommendations for the use of performance metrics. *International Conference on Affective Computing and Intell Interact Workshops*, *2013*, 245–251.

Kambeitz, J., Cabral, C., Sacchet, M., Gotlib, I., Zahn, R., Serpa, M., . . . Koutsoulris, N. (2016). Detecting neuroimaging biomarkers for depression: A meta-analysis of multivariate pattern recognition studies. *Biological Psychiatry*, *82*(5), 330–338.

Kanwisher, N., McDermott, J., & Chun, M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *17*(11), 4302–4311.

Kay, K., Naselaris, T., Prenger, R., & Gallant, J. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Keerthi, S., & Lin, C. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*, *15*(7), 1667–1689.

Kipli, K., Kouzani, A. Z., & Williams, L. J. (2013). Towards automated detection of depression from brain structural magnetic resonance images. *Neuroradiology*, *55*, 567–584.

Koelstra, S., Mühl, C., Soleymani, M., Lee, J., Yazdani, A., Ebrahimi, T., . . . Patras, I. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, *3*(1), 18–31.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, *2*, 1137–1143.

LaConte, S. (2011). Decoding fMRI brain states in real-time. *NeuroImage*, *56*(2), 440–454.

Larsen, R., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. S. Clark (Ed.), *Emotion* (pp. 25–59). Thousand Oaks, CA: SAGE.

Lee, C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Interspeech*, *53*(9–10), 320–323.

Liu, F., Wenbin, G., Dengmiao, Y., 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). http://doi.org/10.1371/journal.pone.0040968

Mathworks. (2017). MATLAB and Statistics Toolbox (Release 2017b) [Computer software]. Natick, MA: Author.

Menard, S. W. (2002). *Applied logistic regression* (2nd ed.). Thousand Oaks, CA: SAGE.

Merikangas, K. R., Akiskal, H. S., Angst, J., Greenberg, P. E., Hirschfield, R. M. A., Petukhova, M., & Kessler, R. C. (2007). Lifetime and 12-month prevalence of bipolar spectrum disorder in the national comorbidity survey replication. *Archives of General Psychiatry*, *64*, 543–552.

Mwangi, B., Ebmeier, K. P., Matthews, K., & Steele, J. D. (2012). Multicentre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. *Brain*, *135*, 1508–1521.

Naselaris, T., Kay, K., Nishimoto, S., & Gallant, J. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410.

Ng, A., & Jordan, M. (2001). On discriminative versus generative classifiers: A comparison of logistic regression and naïve Bayes. *Advances in Neural Information Processing Systems*, *14*, 605–610.

Nishimoto, S., Vu, A., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641–1646.

Nouretdinov, I., Costafreda, S. G., Gammerman, A., Chervonenkis, A. l., Vovk, V., Vapnik, V., & Fu, C. H. Y. (2011). Machine learning classification with confidence: Application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. *NeuroImage*, *15*, 809–813.

Ojala, M., & Garriga, G. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, *11*, 1833–1863.

Orru, 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 & Biobehavioral Reviews*, *36*, 1140–1152.

Pantic, M., & Rothkrantz, L. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1424–1445.

Patel, M., Andreescu, C., Price, J., Edelman, K., Reynolds, C., & Aizenstein, H. (2015). Machine learning approaches for integrating clinical and imaging features in LLD classification and response prediction. *International Journal of Geriatric Psychiatry*, *30*, 1056–1067.

Pedregosa, F., Varoquax, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pochet, L., & Suykens, J. (2006). Support vector machines versus logistic regression: Improving prospective performance in clinical decision-making. *Ultrasound in Obstetrics & Gynecology*, *27*, 607–608.

Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: A simulation study. *Metodoloski zvezki*, *1*, 143–161.

Qin, J., Shen, H., Zeng, L., Jiang, W., Liu, L., & Hu, D. (2015). Predicting clinical responses in major depression using intrinsic functional connectivity. *NeuroReport*, *26*(12), 675–680.

Rakotomamonjy, A. (2004). *Support vector machines and area under ROC curve* (Technical report PSI-INSA de Rouen). https://api.semanticscholar.org/CorpusID:1648794

Redlich, R., Almeida, J., Grotegerd, D., Opel, N., Kugel, H., Heindel, W., . . . Dannlowski, U. (2014). Brain morphometric biomarkers distinguishing unipolar and bipolar depression: A voxel-based morphometry-pattern classification approach. *JAMA Psychiatry*, *71*, 1222–1230.

Sacchet, M., Livermore, E., Iglesias, J., Glover, G., & Gotlib, I. (2015). Subcortical volumes differentiate major depressive disorder, bipolar disorder, and remitted major depressive disorder. *Journal of Psychiatric Research*, *68*, 91–98.

Sacchet, M., Prasad, G., Foland-Ross, L., Thompson, P., & Gotlib, I. (2015). Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory. *Frontiers in Psychiatry*, *6*. https://doi.org/10.3389/fpsyt.2015.00021

Said, C., Moore, C., Engell, A., & Haxby, J. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision*, *10*, 1–12.

Salazar, D., Velez, J., & Salazar, J. (2012). Comparison between SVM and logistic regression: Which one is better to discriminate? *Revista Colombiana de Estadística*, *35*, 223–237.

Schnack, H., Nieuwenhuis, M., van Haren, N., Abramovic, L., Scheewe, T., Brouwer, R., . . . Kahn, R. (2014). Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder, and healthy subjects. *NeuroImage*, *84*, 299–306.

Singh, T., & Rajput, M. (2006). Misdiagnosis of bipolar disorder. *Psychiatry (Edgmont)*, *3*, 57–63.

Sitaram, R., Lee, S., Ruiz, S., Rana, M., Veit, R., & Birbaumer, N. (2011). Real-time support vector classification and feedback of multiple emotional brain states. *NeuroImage*, *56*(2), 753–765.

Skerry, A. E., & Saxe, R. (2014). A common neural code for perceived and inferred emotion. *Journal of Neuroscience*, *34*(48), 15997–16008.

Song, K., Han, M., & Hong, J. (2010). Online learning design of an image-based facial expression recognition system. *Intelligent Service Robotics*, *3*(3), 151–162.

Susskind, J., Littlewort, G., Bartlett, M., Movellan, J., & Anderson, A. (2007). Human and computer recognition of facial expressions of emotion. *Neuropsychologia*, *45*(1), 152–162.

Tan, J., Andrade, J., Li, H., Walter, S., Hrabal, D., Rukavina, S., . . . Traue, H. (2016). Recognition of intensive valence and arousal affective states via facial electromyographic activity in young and senior adults. *PLoS ONE*, *11*(1), 1–15. https://doi.org10.1371/journal.pone.0146691

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*, 91–98.

Verma, G., & Tiwary, U. (2014). Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, *102*, 162–172.

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., . . . Vos, T. (2013). Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010. *Lancet*, *382*, 1575–1586.

Zeng, 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*, 1498–1507.

Zhang, H., Japee, S., Nolan, R., Chu, C., Liu, N., & Ungerleider, L. (2016). Face-selective regions differ in their ability to classify facial expressions. *NeuroImage*, *130*, 77–90.