Major Challenges and Limitations of Big Data Analytics

2

Bo Cao and Jim Reilly

Mental disorders have been considered as the top burden among global health problems, contributing about 32.4% years lived with disability (YLDs) and a cost of 2.5 trillion US dollars including both the direct and indirect costs (Vigo et al. 2016; Whiteford et al. 2013; Trautmann et al. 2016). The economic cost from mental disorders is expected to double by 2030. Because mental disorders usually appear early in the life, they may become a life-time burden for the patients and the caregivers. With the increasing number of patients in mental disorders and a growing aging population, the life burden and economic cost of mental disorders will be more than those of cardiovascular disease, common infections and cancer. However, unlike other physical diseases, we still highly rely on symptoms and do not have objective markers to make diagnosis of mental disorders. Once patients are diagnosed with mental disorders, we respond with a trial-and-error procedure to treat them. We seem to lack a good way to know the best treatment for a patient in advance and to provide optimal personalized treatment. These two major issues are pressing grand challenges to psychiatrists and researchers in the field of mental disorders.

The emerging field of "big data" in psychiatry opens a promising path to precise diagnosis and treatment of mental disorders. Over years of debating and hard work, researchers have come to an agreement that mental disorders are complicated and one disorder is probably not caused by a single change in the genes or neurons. However, by using high-dimensional data, such as genome-wide transcription and

B. Cao (✉)
Department of Psychiatry, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB, Canada
e-mail: cloudbocao@gmail.com

J. Reilly
Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada

I. C. Passos et al., *Personalized Psychiatry*,
https://doi.org/10.1007/978-3-030-03553-2_2

brain images, and integrating information from different modalities, we may be able to development methods of precise diagnosis and treatment prediction of mental disorders. Because the dimension of the data available is so high, a large number of observations are required correspondingly to develop and validate any model or method based on the data, which lead to a big volume of data with high dimensions and high instances. With the help of big data, it becomes possible to implement technics like data mining and machine learning to establish data-driven diagnoses and treatment strategies of mental disorders. Along with the opportunities brought by the big data in psychiatry are some unprecedented challenges.

In this chapter, we will name some challenges we are facing in the field of big data analytics in psychiatry. We hope to address and overcome these challenges with the joint force of researchers in related fields and alleviate the burden of mental disorders.

2.1 Challenges in Data Standardization

The data and knowledge shared should be scalable, expandable, transferrable and sustainable. This means that by increasing the volume of the data, we should achieve better performance of methods developed on the data and higher confidence of the outcomes, and we should be able the transfer the methods developed on one population to other populations and on the current generation to the future generations. One of the major challenges of big data analytics in psychiatry is that data collected globally is not always combinable due to the lack of standardization across regional centers and hospitals. Standardization can be considered as common features or measurements shared between datasets in the raw format in a strict sense. The measurements, or what data to collect, are usually determined upon an agreement across clinicians and researchers from different regions and disciplines. Standardization can also be considered as major shared information between datasets in a general sense. Even though the datasets may look different, the same features could be extracted after preprocessing. Lack of standardization is usually due to a disagreement among data collection parties, and makes it difficult to generalize the analysis based on one dataset to other datasets, or transfer the knowledge learned by the machine from one to another.

The first level of lack of standardization is from the diagnosis criteria. Although many researchers aim to move away from symptom-based diagnosis and achieve an objective diagnosis system based on biological markers, we still need to reply on the current diagnosis system to establish research samples. However, discrepancies in major diagnosis criteria across the world still exist. For example, bipolar disorder in children and adolescents is still diagnosed differently in Europe and U.S., resulting a much lower prevalence of bipolar disorder in Europe than U.S. (Soutullo and Chang 2005), and it is a debating topic whether bipolar disorder progresses or has severity stages (Berk et al. 2007; Passos et al. 2016; Cao et al. 2016; Kapczinski et al. 2016). Discrepancies of this kind will make it difficult to integrate data from different regions, as the data from patients with a certain label in one region may actually

represent different populations in another region. It may also make it difficult the apply models developed with data from one region to those from another region, when these regions have different diagnosis criteria.

The changes of major diagnosis criteria over the years may also expose challenges in the consistency of the methods developed with data based on these criteria. For example, the data collected from patients with autism spectrum disorder (ASD) based on the fifth edition of Diagnostic and Statistical Manual of Mental Disorders (DSM; DSM-5) may include patients that were labeled as another disorder according to the fourth edition of DSM (DSM-IV). Patients that were considered to have obsessive-compulsive disorder (OCD) or posttraumatic stress disorder (PTSD) according to DSM-5 might share the same biological signatures of patients with anxiety disorders diagnosed according to DSM-IV (American Psychiatric Association 2013a, b). These changes of criteria are sometimes due to disagreement among the clinicians and researchers, but with good intention to provide better mental health services and to reflect recent progress in the research in mental disorders. The changes of criteria will be always a challenge for big data analytics in psychiatry, as it will be hard to keep tracks of findings based on different versions of the criteria. However, as more data are generated, shared and utilized, we believe that the criteria based on the biological markers will eventually emerge and converge with the criteria based on symptoms.

The second level of lack of standardization is from the different variables or modalities collected from regional data. Researchers have already realized the value of multi-modality data in psychiatry, which usually provide a more thorough understanding of mental disorder mechanisms and a better performance of computational models in making classifications and predictions of diagnosis and treatment responses compare to data of single modality. However, it is not always possible to collect all the crucial modalities. For example, magnetic resonance imaging (MRI) can provide non-invasive measurement of brain structure and functions in-vivo, and is a powerful tool for psychiatric research especially when combined with genetic measurements (Stein et al. 2011). However, a MRI scanner is luxury equipment for many hospitals in the developing countries, and many research projects may have to drop the MRI component due to the shortage of financial support even when the patient resource is sufficient. Some scanning procedures may require dedicated expertise, such as MR spectroscopy, advanced diffusion tensor imaging and scanning very young children or patients under states involving excessive head movements (Cao et al. 2017a), which may also become challenges for hospitals and research centers without corresponding supports.

Different variables, assessments and outcome indicators may also be used in electronic health records (EHR) and health information (HI) across regions and nations. It is quite common that even with the same diagnosis criteria, clinicians and health service providers from different regions or countries may have different interpretations of the criteria and different ways to record cases. They may also add their own insight or adapt a general procedure to meet the need of local populations. All these variations of recording the patient information may lead to various measurements that are unique to certain data collection, which will cause

difficulty when a method developed on one dataset is being transferred or applied to another dataset. The EHR and HI are emerging technology in mental health, and each country is still trying to implement them efficiently according to its own medical, privacy, political and financial environments. However, it is important for researchers and policy makers to realize the necessity to facilitate a communicable and compatible health record system for the future global effort in mental health research.

The third level of lack of standardization is from varies of protocols in data collection. Although some datasets shared the same variables, they may show quite significant difference in the same variables due to different protocols of data collections, storage and preprocessing. For example, in a large multi-center neuroimaging dataset, the study site is one of the most significant contributors to the variance even in some of the basic measurements like cortical and subcortical region volumes (Panta et al. 2016). The effect of the study sites may be attributed to several sources, such as different brands of scanners, scanning sequences and parameters, preprocessing pipelines and even different instructions for the patients. Since it is not possible to use the same scanner and technicians to perform all the data collection, one strategy could be using common phantoms across study sites and follow the protocol in a well-established large-sample study, such as the human connectome project (http://www.humanconnectomeproject.org). Another strategy is to include a well-represented sample of healthy subjects that serves as the reference when the measurements of current dataset are compared to other public datasets (Cao et al. 2017b). The difference in the measurements between the healthy subjects in different datasets could be used to calibrate the corresponding measurements for all the subjects including patients and healthy subjects, so that different patient populations from different datasets can be compared directly (Fig. 2.1).

Another challenge in data standardization is the fast evolving technics in biology, imaging and computational analysis. We are in such a fast pace in the development of new technologies in biology and the ways that we can measure the genes, neurons, brain anatomy, networks and functions are evolving every day. New standard measurements that were not possible or affordable are being introduced more frequently than ever. Thus, it is a great challenge for us to think ahead when new data collection is planned. It is also important to keep updating and correcting knowledge derived from data collected previously. A result no matter how intuitive at the time of publication could be found less accurate when a new method is developed. For example, the segmentation of hippocampal subfields were found to be less accurate in an older version of method compared to the new version (Andreasen et al. 2011; Cao et al. 2017c, 2018), and findings using the previous version of method need to be updated and interpreted with caution (Van Leemput et al. 2009; Haukvik et al. 2015). For another example, researchers have generally believed that there is no lymphatic system in our brain, until very recently some study confirmed that our brain actually has a lymphatic system to circulate immune cells and wastes using advanced MRI imaging technics (Absinta et al. 2017). This will not only change the textbooks about the lymphatic system, but will also bring

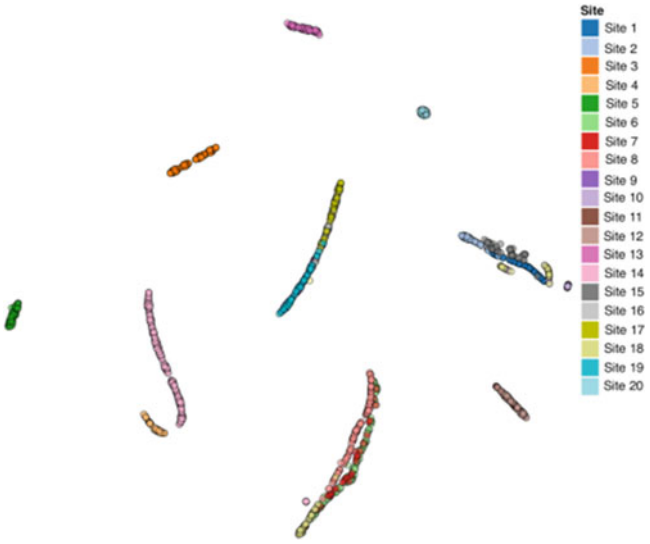

Fig. 2.1 Effect from study sites in a large sample multi-center neuroimaging study. Adapted from Panta et al. (2016)

new possible measurements about brain immunometabolism in mental disorders involving altered immune activities like neural or glial inflammation.

Although it is convenient to have the exact same measurements in datasets collected across regions for the purpose of implementing many machine learning algorithms and analyses, the advance of computational algorithms may provide more tolerance of less standardized data. Traditional methods, such as support vector machines (SVM) and regularized linear regressions have made substantial progress in big data analytics in psychiatry. However, they may require relatively strict standardization across the datasets when a model using them needs to be generalized and transferred from one dataset to another dataset. New progress in deep learning networks may relieve some of the restrictions in the variables collected in different datasets because methods like deep learning may involve an integrated feature learning process that does not need the raw data to be in the exact form from different datasets (Rajkomar et al. 2018). New computational algorithms may help to automatically "standardize" features from different variables in different datasets, and make it easy to transfer models across datasets.

The challenges due to lack of standardization could be partly overcome with good strategic planning and collaboration between developed and developing regions. The data and methods shared in the research community have made substantial contribution to the progress of mental disorder research and brain research in general. A transparent ecology to share the lessons learned during the data collection and sharing, and an open environment to facilitate the agreement on the variables

and protocols in patient evaluation and data collection will advance the progress in big data analytics in psychiatry.

2.2 Challenges in Machine Learning in Psychiatry

2.2.1 Overview of Machine Learning in Psychiatry

The machine learning (ML) paradigm is the new frontier in brain health research. The brain is far too complicated an organism to enable modeling by classical means, a process which would typically involve the use of mathematical and physical constructs or laws to predict brain *behaviour* in some way. However, our understanding of the brain is currently at such an underdeveloped state that we as humans know of no encompassing set of physical and mathematical laws that can adequately describe brain behaviour over a wide range of circumstances. In fact, the concept of humans trying to understand their own brains is a conundrum, well expressed by Emerson Pugh in the early 1930s: "*If the human brain were so simple that we could understand it, we would be so simple that we couldn't.*"

Fortunately however, the machine learning paradigm allows us to circumvent this difficulty, at least in part. Machine learning can be used to construct a rudimentary model that can predict behaviour of a complex system in a limited sense. The machine learning model compares measurements describing a system under test with previous measurements of similar systems whose behaviour has been observed and is therefore known. Because the machine learning method can then predict behavior of the complex system, it in essence constructs a rudimentary model of the system itself.

We now give a simple example of how a machine learning model can be developed that could train a "Man from Mars" to distinguish whether a particular human specimen is male or female. In this problem, there are two *classes*; male and female. We must first have available a set of N humans and their corresponding male/female class labels. Since the Man from Mars has very little prior knowledge about distinguishing male humans from female humans, he assembles a large group of measurements (*features*) from each human sample. This list of features (referred to as the *candidate features*) are only his guesses of which measurements might be discriminative between the classes. Let us say the candidate features he chooses in this case are hair length, number of teeth, skin colour, voice pitch, and weight. These candidate features are fed into a *feature selection* algorithm (to be described later) that identifies only those features which are discriminative between the classes. We observe that skin colour, number of teeth, and to some extent weight, have little bearing in determining gender. So the feature selection algorithm selects hair length and voice pitch from the list of candidate features. (We prefer only two features so we can plot in 2 dimensions). We can interpret these features as axes in a Cartesian coordinate space (called the *feature space*), and then plot the corresponding hair length and voice pitch values for each of our N human samples as a point in this feature space, as shown in Fig. 2.2. We see that the points representing the male

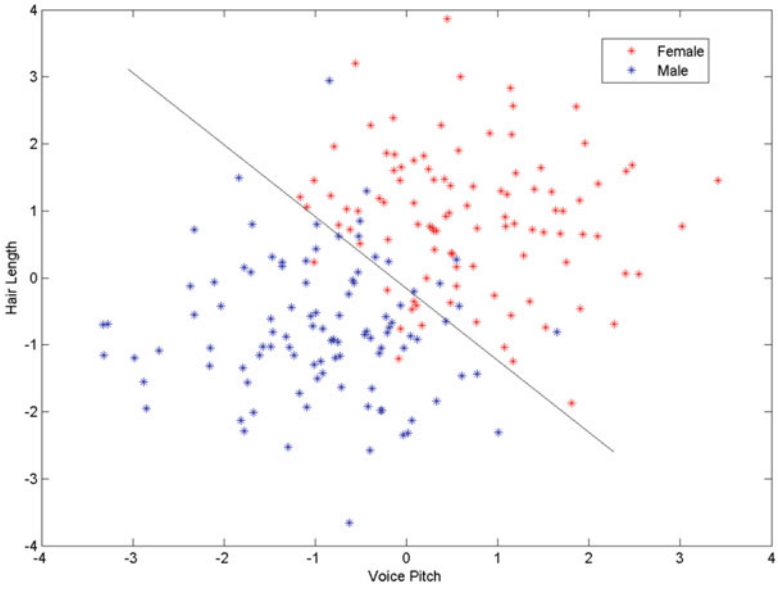

Fig. 2.2 Feature space for the "Man from Mars" example

and female samples tend to cluster into two distinct regions in the sample space—females in the upper right, and males in the lower left.

We then design a *classifier*, which in this simple case is a straight line that separates the two classes as cleanly as possible. Now that our Man from Mars has his rudimentary model constructed, he can determine the gender of a previously unseen human by measuring their hair length and voice pitch and plot the corresponding point in the feature space. The gender is determined by which side of the line the point falls on.

Let the number of selected features be M. The M features collected from each of the available N humans may be assembled into N vectors $\boldsymbol{x}_n, n = 1, \ldots, N$, each of which is of dimension $(M \times 1)$. Let us denote the corresponding (binary) class label for each human (sample) as y_n. Then the set $(\boldsymbol{x}_n, y_n), n = 1, \ldots, N$ is called the *training set.*

Our Man may wish to determine the accuracy of his rudimentary machine learning model. He may accomplish this using a *validation procedure*, which is an essential part of the machine learning process.

In Fig. 2.2 we see that some samples from each class fall on the wrong side of the boundary. This is because in this case there are some men with long hair and high voices and women with short hair and low voices. Misclassification is unavoidable in most machine learning problems; however we wish to minimize this effect by choosing the best possible combination of features and the best possible classification rule.

Thus we see there are three major components of a machine learning modelling process; these are feature selection, classification and validation. We discuss each of these components more thoroughly in the sequel, with a view to how each of the respective algorithms behave in applications related to psychiatry and neuroscience.

2.2.2 Feature Selection, Classification, and Validation Algorithms

2.2.2.1 The Feature Selection Process

In typical applications in psychiatry and neuroscience, and in many medical applications in general, the number of candidate features tends to be large but the number of available training samples is few. This scenario is difficult for the machine learning paradigm, since according to Bellman's "curse of dimensionality"(Bellman and Dreyfus 1962), the number of training samples required to maintain classification performance at a specified level grows exponentially with the number of features used by the classifier. So to maintain satisfactory levels of classification accuracy, especially in the presence of few training samples, we require the number of features adopted by the machine learning model to be as small as possible. As we have seen previously, this is accomplished using a feature selection process.

Feature selection methods, in the general sense, identify features which have a high level of statistical dependency with the class label. This means the values of selected features change significantly with class. Another interpretation of feature selection is in the *data compression*, or *dimensionality reduction* context. That is, a feature selection process identifies features which preserve the underlying characteristics of the data with as high fidelity as possible using as few features as possible.

One of the issues worthy of consideration in feature selection is that it is necessary to examine the relevance of *groups* of features rather than just features individually. An example is shown in Fig. 2.3 where it is seen that each feature individually is not discriminative; however, when considered jointly the two classes separate cleanly. Thus, an ideal feature selection algorithm must examine all possible combinations of all available N candidate features for relevance. This is a problem with combinatorial complexity and so is computationally intractable. We must therefore resort to a suboptimal approach for selecting features if we are to circumvent these computational difficulties. In practice, all practical feature selection approaches are suboptimal in some sense.

Feature selection is an intensively studied topic and accordingly there are a very large number of feature selection algorithms available in the literature. An extensive list of modern feature selection methods is provided in Armanfard et al. (2017). A feature selection method that has proven to be very effective in applications related to brain research is the minimum redundancy maximum relevance (mRMR) method (Peng et al. 2005). The mRMR method uses *mutual information* as a measure of statistical dependency. It is an iterative greedy approach where in each iteration a single feature is chosen which has the maximum mutual information with the class labels (relevance) but minimum mutual information (redundancy) with the set of

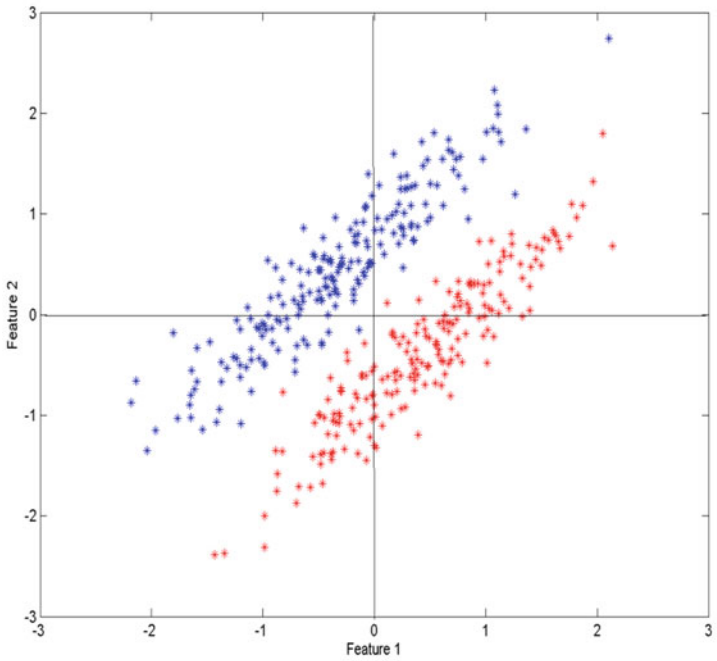

Fig. 2.3 A feature space in 2 dimensions, where neither feature is discriminative on its own, yet jointly they are highly discriminative

features chosen in previous iterations. C code for the mRMR method is available on line at http://home.penglab.com/p_publication.html.

Often in feature selection problems, the scale of the candidate features can vary over many orders of magnitude. This extensive range of values can pose difficulties for the feature selection and classification algorithms. This issue may be conveniently resolved by normalizing the values of each feature using e.g. their z-score. That is, all values x_{mn} of the mth feature are replaced with the value $x'_{mn} = \frac{x_{mn}-u_m}{\sigma_m}, n = 1, \ldots, N$, where μ_m and σ_m are the mean and standard deviation respectively of the mth feature evaluated over the N available samples from the training set.

2.2.2.2 The Classification Process

The features are selected so that the samples from each class in the training set separate (i.e. cluster) as well as possible into two (in the binary case) distinct regions in the feature space, each region corresponding to a class. In a typical machine learning scenario, the two classes seldom separate cleanly; there is usually some overlap between the clusters representing each of the classes. The classifier may be described as a mathematical rule that maps a prescribed (i.e. test) point in the feature space into a class, in some optimal fashion that minimizes the occurrence

of a classification error. That is, the classifier determines the most likely cluster that a test point belongs to. Note that points which fall into an overlap region between clusters may not classify correctly.

There are many types of classifier. The support vector machine (SVM) (Haykin 2009; Hastie et al. 2009) is a well-established classification method that has been shown to behave well in psychiatric applications, with a built in SVM function available in later versions of Matlab and Tensor Flow. The basic version of the SVM classifier formulates a hyperplane that separates the two classes so that the *margin* is maximized. The margin is the distance from the closest points in each class to the hyperplane. These closest points are referred to as *support vectors*; hence the name of the classifier.

Classification is a very mature topic and consequently there are many types of classification methods, in addition to the SVM, that are available in the literature. Examples include K Nearest Neighbor (KNN), the Linear Discriminant Analyzer (LDA), the naïve Bayes classifier, decision trees, etc. These are all described in Hastie et al. (2009). There is also the well-known multi-layer perceptron as described in Rumelhart (1986) and Haykin (2009).

Decision trees are specifically useful in the present context since they form the basis of more sophisticated classifiers which we discuss later in this section. There are several tree-based training methods that are discussed in Hastie et al. (2009) and Bishop (2006). A characteristic of the decision tree is that it produces unbiased outputs with high variance; hence, they are not useful as is for classification.

Classifiers, as well as many feature selection algorithms, usually have at least one associated parameter whose value must be tuned to produce optimal classification performance in a given scenario. For example, the SVM classifier incorporates a user-defined parameter that controls the tradeoff between increasing the margin size and ensuring that the training sample feature vectors $\mathbf{x}_n$ lie on the correct side of the margin. Another example is the parameter K (number of nearest neighbours) in the KNN classifier. Details on how to select a suitable value for these parameters are described in Sect. 13.2.2.3.

Classification in the Nonlinearly Separable Case: In Figs. 2.2 and 2.3, we have shown simple cases where the class clusters separate linearly. While this is the easiest case to deal with from the theoretical perspective, in practice the boundaries between the classes are seldom linear, as shown in the example on the left in Fig. 2.4. In this case, it can be seen that if a linear boundary is used to separate the feature space on the left, significant classification error will result.

Fortunately, under certain conditions, various forms of classifier like the SVM can be easily adopted to the nonlinear boundary case using the so-called *kernel trick* e.g., Bishop (2006). The kernel trick is applicable if the only numerical operations performed by the classifier are inner products. The kernel trick in effect maps the original data in the original Cartesian space through a nonlinear transformation Φ into a higher-dimensional space where ideally, the data separate linearly, as shown on the right in Fig. 2.4. The interesting feature of the kernel trick however is that the nonlinear transformation is not performed explicitly. Instead, it may be induced simply by replacing each inner product operation of the form $\boldsymbol{x}^T\boldsymbol{z}$ involved in the

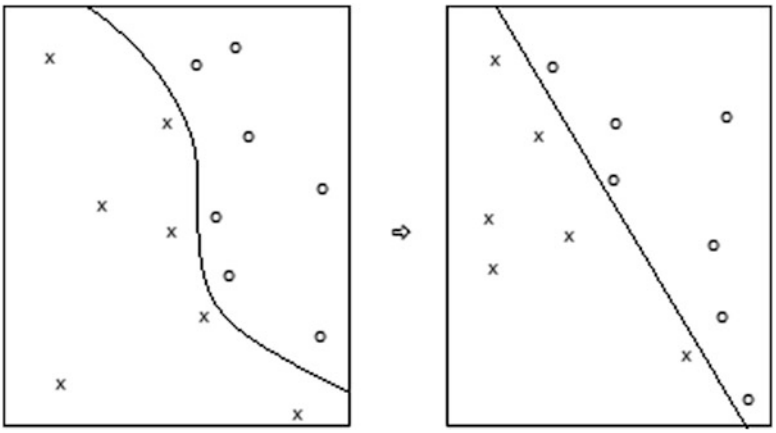

Fig. 2.4 Transformation of a nonlinear feature space (left) in to a linear separable space (right)

implementation of the classifier algorithm with a kernel function $k(\boldsymbol{x}, \boldsymbol{z})$,where $\boldsymbol{x}$ and $\boldsymbol{z}$ are feature vectors in the present case.

Kernel functions can be interpreted as similarity measures; the larger the value of the function, the more similar are the vector arguments **x** and **z**. They must obey the property that its associated Gram matrix be positive definite. Examples of valid kernel functions are the Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\gamma ||\boldsymbol{x} - \boldsymbol{z}||_2^2\right)$, where γ is a real-valued user-defined parameter, and the polynomial kernel $k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T \boldsymbol{z} + c)^d$, where c and d are also real-valued user-defined parameters. The respective parameters are adjusted so that the boundary in the transformed space is as linear as possible. More details on all aspects of the kernelization process are available in Müller et al. (2001) and Bishop (2006).

Machine Learning Methods Specifically Recommended for Use in Brain Research The first such approach which has proven useful in brain studies is the mRMR feature selection scheme in conjunction with an SVM classifier (Khodayari-Rostamabad et al. 2010, 2013; Ravan et al. 2011, 2012; Colic et al. 2017). For example, in Khodayari-Rostamabad et al. (2013) this approach was used to predict response of patients with major depressive disorder to treatment with an SSRI.

Adaboost Another approach uses *boosting* (Bishop 2006) where the idea is to aggregate many "weak" classifiers (learners) into one that is very "strong". The *Adaboost* algorithm (Schapire 2003) is a well-known example of such a method. This method uses multiple instances of weak learners. For training, each weak classifier weighs each sample of the training set differently, with more weight being placed on the samples which the classifiers get wrong. The Adaboost algorithm combines the feature selection and classification roles and typically uses decision trees as the weak learner. It forms its final output decision on a majority vote amongst the weak learners. The Adaboost algorithm has the desirable property that, provided the individual weak learners give better than chance accuracy, then the

probability of error of the aggregate classifier decays exponentially as the number of learners increases (Schapire 2003).

Random Forest An additional (related) concept is *bagging*, which is short form for "aggregate bootstrapping". A widely used classification algorithm in this respect is the *random forest* (RF) classifier (Hastie et al. 2009; Breiman and Spector 1992). Like Adaboost, the RF classifier uses a multiplicity of decision trees, and again the final output decision is based on a majority vote over the individual decision trees. Unlike Adaboost, the input to each decision tree for training is a resampled (with replacement) version of the complete training set, and the feature inputs at each node are also randomly chosen. The RF classifier has the advantage that, unlike the other forms of classifier we have discussed so far, it is insensitive to the overfitting phenomenon, to be discussed later. It too combines the feature selection and classification processes. The RF classifier has been successfully used e.g. in detecting onset of epileptic ictal periods (Colic et al. 2017).

The Localized Feature Selection Method (LFS) Conventional forms of feature selection methods are global; i.e., they assign a single set of features that attempts to characterize the entire sample (i.e. training sample) space. In contrast, the LFS method (Armanfard et al. 2016a, 2017) allows the choice of selected features to vary across the sample space, thus adapting to variations such as nonlinearities, discontinuities or nonstationarities that may appear across different regions of the sample space. Each training sample is treated as a representative point for its surrounding region and as such is assigned its own distinct set of features. These (local) feature sets are determined by solving a straightforward optimization procedure in the form of a linear program. The LFS method, unlike deep learning methods, is therefore very easy to train. The LFS method is well suited to the "data poor" case where the number of candidate features far exceeds the number of available training samples, and is also immune to the overfitting problem (to be discussed). The LFS method has proven to be successful in predicting emergence in coma patients (Armanfard et al. 2016b).

There are also deep learning methods (Le Roux and Bengio 2008) that are currently a very active area of research. In particular, autoencoders (Le 2015) have the desirable characteristic of being able to automatically generate features directly from the data. Deep learning methods have been very successful in many applications; however, generally they require large, noise-free training sets. In many applications in psychiatry and related fields training data is very hard to come by, and so deep learning methods have not proven very successful for the applications at hand. It is for this reason they are not considered further in this chapter.

2.2.2.3 Validation and Measurement of Performance

Validation is a very important component of the machine learning model. It is used in conjunction with the available training set to determine classification accuracy of a proposed machine learning implementation. It is also instrumental in tuning

the parameters that are associated with the feature identification, classification or kernelization procedures.

Before we discuss validation per se, we consider two different forms of error associated with the machine learning model. These are *training error* and *generalization* error. Training error is the classification error using the training set itself. An example is shown in Fig. 2.2. In this case, because of the overlap between the classes in the training set, we see that a linear boundary (as determined e.g. by an SVM) cannot separate the two classes without error. Generalization error on the other hand arises if a new sample which is not contained in the training set is incorrectly classified. The validation process estimates the generalization error of the respective machine learning model based on the training error.

The usual form of validation is *cross-validation*, where the available training set is split into two parts—the larger which is referred to as a *training set*, and the other the *test set*. The machine learning model is built using only the data in the training set. The performance of the resulting model is then evaluated by feeding the test set samples into the classifier and comparing the classification results with the corresponding labels provided by the test set.

The most common method is k-fold cross validation. Here the entire training set is partitioned into k contiguous groups, with each group containing the integer closest to N/k samples. The procedure iterates k times, where in each iteration one group is held out for testing and the remaining groups are used for training. Each group is left out once. The fold error is the average error rate over the samples in the group, and the overall error rate is the average of the group error. Leave-one-out cross validation (LOOCV) is a form of k-fold cross validation, except there are $k = N$ folds (i.e. there is only one sample in each test group). LOOCV works well in the small N case, but often is computationally expensive because the entire modelling process must be repeated N times. The method can be susceptible to high variance in the generalization error estimate. A third form of validation is bootstrapping. It is similar to k-fold cross validation except that in each fold the training and test groups are chosen randomly with replacement.

As discussed in Hastie et al. (2009), the cross-validation estimate of the generalization error is subject to both bias and variance. Bias happens because the number of training samples available in each fold may be inadequate to train the underlying model accurately. This results in the cross-validation procedure overestimating the generalization error. There is also a variance associated with the cross-validation estimate, since it is obtained by averaging over a finite number of samples. As k decreases, the variance increases but the bias decreases. Breiman and Spector (1992) and Kohavi (1995) suggest that a value of $k = 5$ or 10 gives a reasonable compromise between these two counter-acting effects.

A cross validation procedure can also be used to tune the parameters of the machine learning model. For example, if we are using a k-fold process for performance evaluation, the data in the training set in each fold is subjected to a second, inner cross-validation loop. In each fold of the inner loop, the data is again split into a "tuning" set and a test set. The inner loop is repeated several times using

different values of the parameter, and the value giving the best performance is then selected for that fold of the outer loop.

A very important consideration in cross validation is that the training and test sets be kept completely separate. If a data sample is included in the training set and then afterwards is again used for testing, then performance is biased upwards, because the machine learning model has been specifically trained to avoid errors over *all* samples in the training data.

As an example of the machine learning process in psychiatry applications, we now briefly describe a study (Khodayari-Rostamabad et al. 2013) which used machine learning to predict response to SSRI treatment for major depressive disorder, based on analysis of the EEG. The training set consisted of EEG measurements from 22 patients who were diagnosed with MDD and whose response to the treatment was recorded after several months of treatment. The set of candidate features consisted of power spectral density measurements at many frequency values from all electrodes, and spectral coherence values from all pairs of electrodes over the same set of frequency values. The study used 20 electrodes and 50 frequency values, which resulted in over 10,000 candidate features. The mRMR feature selection algorithm was used to reduce this set down to 10 or fewer features which have the most relevance with the recorded response to the treatment. An SVM classifier was used and the estimated correct classification rate was approximately 85%. This study therefore provides a good indication that machine learning methods can adequately predict response using EEG analysis.

2.2.3 Further Considerations in the Development of a Machine Learning Model

2.2.3.1 The Over/Underfitting Problem

Consider the situation shown in Fig. 2.2 where a linear boundary does not cleanly separate the training samples into their respective classes. The temptation in this case may be to build a classifier that that can generate a more flexible boundary that works its way around the misfit points and so places the misplaced samples on the correct side of the boundary. This increased flexibility can be achieved by introducing additional parameters into the classifier model. In this case, the classifier can be trained so that the training error reduces to zero. Let us assume that the underlying true but unknown boundary corresponding to the physical process that generates the data is in fact linear. Then new data points placed where the flexible boundary has been diverted may not classify properly, and so the generalization error degrades in this case. This phenomenon is called *overfitting* and is a result of the machine learning model over-adapting to the training set (i.e. the boundary is allowed to become too "wiggly").

Another form of overfitting occurs when the dimension of the feature space becomes too large in proportion to the number of training samples. For example, in the linearly separable case, an n-dimensional hyperplane can always separate any arbitrary class configuration of $n + 1$ data points. So as the number of features

increases, the classifier has more freedom to fit the training data, which implies the training error decreases, but at the cost of increased generalization error. It is fortunate that a properly executed cross-validation procedure will detect the presence of overfitting.

Underfitting occurs when the model is not flexible enough to fit the data. This could happen for example when the number of selected features is too small to adequately separate the training set. An example of the underfitting problem is as follows. Suppose we have a data set which separates cleanly with three features. In this case a classifier algorithm such as SVM would specify a boundary plane in the corresponding 3 dimensional space to separate the classes. Suppose now that we discarded one of the features used only two of the three features. Then all the data would be projected onto the remaining 2 dimensional plane and the two classes may overlap with each other, thus reducing performance.

If the number of features is too large, we have overfitting, and if too small, we have underfitting. So how to choose a good value? One valid method is to repetitively train a machine learning model for an increasing number of features (starting e.g. at 1) and test each model using a cross-validation procedure. We should see the error decrease initially as the number of features increases, because underfitting becomes less of an issue. But then as the number increases further, the error will bottom out to a plateau, and then begin to increase, due to overfitting. The best number of features to use may be taken as that corresponding to the minimum error.

2.2.3.2 Missing Data

In many applications, particularly in medicine, the feature vector associated with a specific data sample may not contain all the values or measurements of the specified selected features. When data is collected during studies, missing data may result from patient incompliance, patient drop-out, measurements being too inconvenient or expensive to acquire, etc. The problem is that many machine learning algorithms will not execute properly when some data from the feature vectors are missing. Thus some value for the missing features must be supplied in order for the algorithm to run properly on a computer. The problem is that an improperly substituted value for a missing value may adversely impact the accuracy of the machine learning model. So what value do we supply that will minimize this impact?"

There are many approaches available to address this question. One is simply to delete any incomplete samples. However, in doing so, we are throwing away useful data, and so this is an undesirable option. Other approaches therefore attempt to estimate suitable values for the missing features, based on the available remaining data. The process of filling in missing data is generally referred to as *imputation*. There are many forms of imputation, many of which are well discussed in García-Laencina et al. (2010). The basic idea behind imputation is that the statistical dependencies that may exist between the different feature values in a training sample are exploited to estimate the missing value. The difficulty with this approach is that in some cases, e.g., the mRMR method, the features are specifically selected so

that the statistical dependencies between feature values is minimized. Thus in some cases imputation is an ineffective method.

In cases where there is significant correlation between feature values in a data sample, we can use ordinary regression to impute the missing data. Another approach is to use more sophisticated model-building statistical methods such as the EM algorithm. Yet another approach is to use a second-level machine learning approach to estimate the missing data in the primary problem. These methods are all discussed well in García-Laencina et al. (2010).

Perhaps the most sensible approach to handle the missing data case is to use feature selection and classification methods that can be adapted to tolerate missing data. Two such methods are the random forest (RF) and the localized feature selection (LFS) approach. When some features in the training set are missing, the training procedure for both algorithms is easily modified to accommodate this case. However, when testing data contains missing values, both models may have to be partially re-trained so that missing features in the test data are excluded. This can be expensive from the computational perspective, but the data imputation process involves a significant computational cost as well. At this point it is not known how the performance of the RF or LFS approaches to handling missing data compare to that of imputation methods. However, in the case where there is little statistical dependency between the selected feature values, the LFS and RF methods will almost surely perform better than methods using imputation.

2.2.3.3 Imbalanced Data

The data imbalance problem occurs when the training set consists of many more samples of one class than another. These are referred to as the majority vs. minority classes, respectively. For example, if a research study involves testing the human population at large for psychiatric illness, we are likely to find far more healthy subjects than ill patients. Thus the training set becomes imbalanced. Imbalanced data sets become a problem in the machine learning context, since the model is hindered in learning the distributive properties of the minority class. For example, in a case where the split between the majority vs. minority classes in the training set is 90% vs. 10%, the model need only output a majority class decision in all cases and overall, it would be correct 90% of the time. However, in this case the minority class would be misclassified 100% of the time. As a further example, studies (Woods et al. 1993) have been performed where machine learning was used to detect cancer from a mammography data set. The data set contained a 40:1 imbalance in favour of the noncancerous class. The results showed accuracy rates of close to 100% for the noncancerous case, and only approximately 10% for the cancerous class. Thus a large proportion of cancerous cases would be incorrectly classified as noncancerous. This case has more severe consequences than incorrectly diagnosing a noncancerous patient. This example illustrates that in the imbalanced data case, it is necessary to consider more refined performance metrics, such as receiver operating characteristic (ROC) curves and others that can weigh errors from the different classes in different degrees (He and Garcia 2009).

The negative consequences of the imbalanced data case become more severe when the class distributions in the feature space become more complex. This could happen e.g., if the distribution of one or both of the classes devolves into multiple clusters, or a single cluster of complex shape, instead of the ideal case where each class is represented by a single well-defined cluster. The situation is particularly severe in the high-dimensional case with few training data, since then there are not enough samples for the model to learn the characteristics of the minority class.

There are effective methods that have been developed to mitigate the imbalanced data problem. One such method that has shown a great deal of success in many applications is the synthetic minority oversampling technique (SMOTE) (Chawla et al. 2002). It balances the dataset by sampling (generating) synthetic minority class samples, and discarding some majority class samples, if necessary. The synthetic minority class samples are generated by selecting a specific minority class training sample at random. Artificial samples are generated by placing a new sample on a straight line between the minority sample under consideration and one of its K nearest neighbours of the same class. This sampling process can be repeated many times to generate as many synthetic minority class samples as desired. This method preserves the characteristics of the minority class data and has been demonstrated to work well in many situations. There are several variations on the basic method, as discussed in He and Garcia (2009). The SMOTE algorithm is included in the Tensorflow package.

The SMOTE method and its variants use sampling techniques to augment minority class samples. Another approach at handling the imbalanced data case are *cost-sensitive* methods, which effectively place more weight on minority class errors than on majority class errors during the training process. In many cases this can be achieved simply by trading off an increase in majority class error for an improvement in minority class performance. The Adaboost and LFS algorithms in particular are easily adapted to incorporate this tradeoff. In the Adaboost case, it is only necessary to modify the formulation of the distribution function over the training samples; with LFS, the tradeoff can be implemented simply by varying the parameter γ (Armanfard et al. 2016a, 2017). The literature on this topic is extensive; there is an abundant reference list in He and Garcia (2009).

2.3 Challenges from Data to Knowledge

Traditional ways of research in psychiatry tend to be reductionism and hypothesis driven, which is proved to be effective to investigate single-factor mechanism at the group-level. This approach is still the golden standard when it comes to establish the causality between a factor and the outcome, because we usually could only manipulate one or limited number of factors in experimental or clinical setups. When many factors, including genetic, physiological and behavioral factors, and their interactions need to be considered at the same time, it is usually not efficient, if not impossible, to use the reductionism approach to investigate one by one of the many possible factor combinations (Williams and Auwerx 2015). The new

big data approach could take into account all the factors without many priori assumptions, which will lead to effective outcome prediction at the individual level and new hypotheses that have been ignored previously. This approach will provide translational applications in personalized psychiatry, as the knowledge or algorithms learned from existing data could be applied on new cases. It will also provide insights of important factors and their links in mental disorders, which can then be investigated using a hypothesis-driven approach. Thus, the traditional approach and the novel big data approach are complementary to each other in future research of mental disorders.

It is still crucial to transform the complex data with understandable representation in low dimensions in many cases, because we can visualize the data in 2D and 3D dimensions, static or changing over time. Visualization will help us to see high-dimensional data in an intuitive space. It will show data distributions for certain measurements and overlay measurements onto each other to show their interactions, which will help to understand the mechanisms underlying different measurements, identify the outliers and unusual cases, discover major variance contributors, select subsets of data for post-hoc analysis and so on. Although most of these tasks could also be done with proper mathematical tools directly applied at the high-dimensional data, it is challenging to make sense of the data when the dimension of the data is high and data involve multiple modalities. Moreover, visualization in low dimension is helpful for researchers to demonstrate certain concepts and convey the knowledge to the audience without professional data science training, such as some clinicians and patients. For example, a visualization method called t-distributed stochastic neighbor embedding (t-SNE) can help researchers see a large sample of high-dimensional multi-modal brain imaging data (Panta et al. 2016). We can easily see the reliable difference between images from 1.5 and 3 T scanners, and there seems to be no apparent difference in the scanning time. These observations may provide further confidence for us to combine existing images scanned at varied time of the day or to plan new scans without much concern of scanning time, while make us to be cautious about data that have been scanned or are going to be scanned with different magnetic field intensities. Big data visualization is still an emerging field and psychiatry will benefit from the development of it, yet it is also a challenging field with respect to the number of factors that need to be considered in mental health.

Big data in psychiatry armed with advanced machine learning and artificial intelligence technics will become one of the strongest tools in the research of mental disorders. However, as an interdisciplinary field, the collaboration between experts in psychiatry, neuroscience, psychology, computer science, mathematicians, and software engineers is not replaceable by the novel methods of big data analytics. The value of big data will not be appreciated by the public until it is converted to massive knowledge of mechanisms of mental disorders or translational tools that can guide the diagnosis and treatment of mental disorders. It is only when the interdisciplinary experts make joint forces together that the big data in psychiatry can reach its full potential to become beneficial knowledge and the corresponding challenges that we have discussed can be overcome (Fig. 2.5).

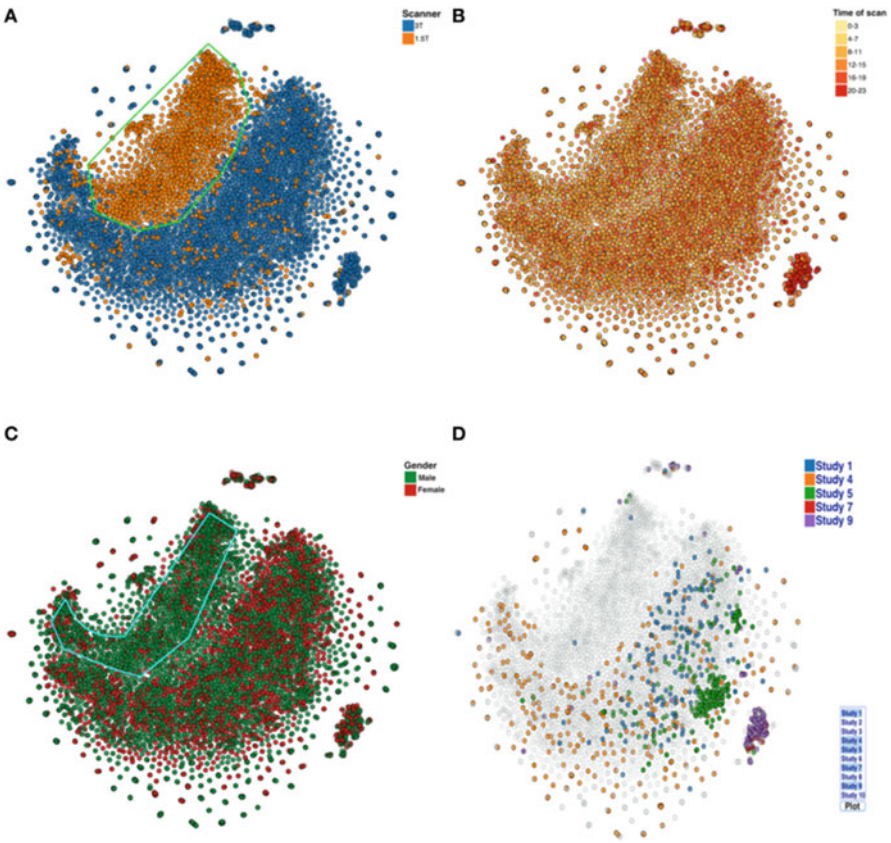

Fig. 2.5 t-SNE plots color coded by (**a**) scanner type (**b**) scan acquisition time (**c**) gender, and (**d**) studies. Adapted from Panta et al. (2016)

References

Absinta M, Ha SK, Nair G et al (2017) Human and nonhuman primate meninges harbor lymphatic vessels that can be visualized noninvasively by MRI. Elife. 6:e29738. https://doi.org/10.7554/eLife.29738.001

American Psychiatric Association (2013a) Diagnostic and statistical manual of mental disorders, 5th Edition (DSM-5). Diagnostic Stat Manual of Mental Disorder 4th Ed TR. 280. https://doi.org/10.1176/appi.books.9780890425596.744053

American Psychiatric Association (2013b) Highlights of changes from DSM-IV to DSM-5. Focus (Madison) 11(4):525–527. https://doi.org/10.1176/appi.focus.11.4.525

Andreasen NC, Nopoulos P, Magnotta V, Pierson R, Ziebell S, Ho B-C (2011) Progressive brain change in schizophrenia: a prospective longitudinal study of first-episode schizophrenia. Biol Psychiatry 70(7):672–679. https://doi.org/10.1016/j.biopsych.2011.05.017

Armanfard N, Reilly JP, Komeili M (2016a) Local feature selection for data classification. IEEE Trans Pattern Anal Mach Intell 38(6):1217–1227. https://doi.org/10.1109/TPAMI.2015.2478471

Armanfard N, Komeili M, Reilly JP, Mah R, Connolly JF (2016b) Automatic and continuous assessment of ERPs for mismatch negativity detection. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol 2016. IEEE, Piscataway, pp 969–972. https://doi.org/10.1109/EMBC.2016.7590863

Armanfard N, Reilly JP, Komeili M (2017) Logistic localized modeling of the sample space for feature selection and classification. IEEE Trans Neural Networks Learn Syst 29(5):1396–1413. https://doi.org/10.1109/TNNLS.2017.2676101

Bellman RE, Dreyfus SE (1962) Applied dynamic programming. Ann Math Stat 33(2):719–726. https://doi.org/10.1289/ehp.1002206

Berk M, Conus P, Lucas N et al (2007) Setting the stage: from prodrome to treatment resistance in bipolar disorder. Bipolar Disord 9(7):671–678. https://doi.org/10.1111/j.1399-5618.2007.00484.x

Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin. https://doi.org/10.1117/1.2819119

Breiman L, Spector P (1992) Submodel selection and evaluation in regression. The X-random case. Int Stat Rev 60(3):291–319. https://doi.org/10.2307/1403680

Cao B, Passos IC, Mwangi B et al (2016) Hippocampal volume and verbal memory performance in late-stage bipolar disorder. J Psychiatr Res 73:102–107. https://doi.org/10.1016/j.jpsychires.2015.12.012

Cao B, Stanley JA, Passos IC et al (2017a) Elevated choline-containing compound levels in rapid cycling bipolar disorder. Neuropsychopharmacology 42(11):2252–2258. https://doi.org/10.1038/npp.2017.39

Cao B, Mwangi B, Passos IC et al (2017b) Lifespan gyrification trajectories of human brain in healthy individuals and patients with major psychiatric disorders. Sci Rep 7(1):511. https://doi.org/10.1038/s41598-017-00582-1

Cao B, Passos IC, Mwangi B et al (2017c) Hippocampal subfield volumes in mood disorders. Mol Psychiatry 22(9):1–7. https://doi.org/10.1038/mp.2016.262

Cao B, Luo Q, Fu Y et al (2018) Predicting individual responses to the electroconvulsive therapy with hippocampal subfield volumes in major depression disorder. Sci Rep 8(1):5434. https://doi.org/10.1038/s41598-018-23685-9

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority oversampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953

Colic S, Wither RG, Lang M, Zhang L, Eubanks JH, Bardakjian BL (2017) Prediction of antiepileptic drug treatment outcomes using machine learning. J Neural Eng 14(1):016002. https://doi.org/10.1088/1741-2560/14/1/016002

García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR (2010) Pattern classification with missing data: a review. Neural Comput Appl 19(2):263–282. https://doi.org/10.1007/s00521-009-0295-6

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7

Haukvik UK, Westlye LT, Mørch-Johnsen L et al (2015) In vivo hippocampal subfield volumes in schizophrenia and bipolar disorder. Biol Psychiatry 77(6):581–588. https://doi.org/10.1016/j.biopsych.2014.06.020

Haykin S (2009) Neural networks and learning machines, vol 3. Prentice Hall, Upper Saddle River doi:978-0131471399

He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239

Kapczinski NS, Mwangi B, Cassidy RM et al (2016) Neuroprogression and illness trajectories in bipolar disorder. Expert Rev Neurother 7175:1744–8360 (Electronic):1–9. https://doi.org/10.1080/14737175.2017.1240615

Khodayari-Rostamabad A, Hasey GM, MacCrimmon DJ, Reilly JP, de Bruin H (2010) A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy. Clin Neurophysiol 121(12):1998–2006. https://doi.org/10.1016/j.clinph.2010.05.009

Khodayari-Rostamabad A, Reilly JP, Hasey GM, de Bruin H, MacCrimmon DJ (2013) A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. Clin Neurophysiol 124(10):1975–1985. https://doi.org/10.1016/j.clinph.2013.04.010

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI 14(2):1–7. https://doi.org/10.1067/mod.2000.109031

Le QV A tutorial on deep learning part 2: autoencoders, convolutional neural networks and recurrent neural networks. Google Brain. 2015:1–20

Le Roux N, Bengio Y (2008) Representational power of restricted Boltzmann machines and deep belief networks. Neural Comput 20(6):1631–1649. https://doi.org/10.1162/neco.2008.04-07-510

Müller KR, Mika S, Rätsch G, Tsuda K, Schölkopf B (2001) An introduction to kernel-based learning algorithms. IEEE Trans Neural Netw 12(2):181–201. https://doi.org/10.1109/72.914517

Panta SR, Wang R, Fries J et al (2016) A tool for interactive data visualization: application to over 10,000 brain imaging and phantom MRI data sets. Front Neuroinform 10:1–12. https://doi.org/10.3389/fninf.2016.00009

Passos IC, Mwangi B, Vieta E, Berk M, Kapczinski F (2016) Areas of controversy in neuroprogression in bipolar disorder. Acta Psychiatr Scand 134(2):91–103. https://doi.org/10.1111/acps.12581

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Rajkomar A, Oren E, Chen K et al (2018) Scalable and accurate deep learning for electronic health records. npj Digit Med 1(1):1–15. https://doi.org/10.1038/s41746-018-0029-1

Ravan M, Reilly JP, Trainor LJ, Khodayari-Rostamabad A (2011) A machine learning approach for distinguishing age of infants using auditory evoked potentials. Clin Neurophysiol 122(11):2139–2150. https://doi.org/10.1016/j.clinph.2011.04.002

Ravan M, MacCrimmon D, Hasey G, Reilly JP, Khodayari-Rostamabad A (2012) A machine learning approach using P300 responses to investigate effect of clozapine therapy. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. IEEE, Piscataway, pp 5911–5914. https://doi.org/10.1109/EMBC.2012.6347339

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533. https://doi.org/10.1038/323533a0

Schapire RE (2003) The boosting approach to machine learning: an overview. Nonlinear Estim Classif 171:149–171 doi:10.1.1.24.5565

Soutullo C, Chang K (2005) Bipolar disorder in children and adolescents: international perspective on epidemiology and phenomenology. Bipolar Disord 7(6):497–506. http://onlinelibrary.wiley.com/doi/10.1111/j.1399-5618.2005.00262.x/full

Stein JL, Hibar DP, Madsen SK et al (2011) Discovery and replication of dopamine-related gene effects on caudate volume in young and elderly populations (N1198) using genome-wide search. Mol Psychiatry 16(9):927–937. https://doi.org/10.1038/mp.2011.32

Trautmann S, Rehm J, Wittchen H (2016) The economic costs of mental disorders. EMBO Rep 17(9):1245–1249. https://doi.org/10.15252/embr.201642951

Van Leemput K, Bakkour A, Benner T et al (2009) Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. Hippocampus 19(6):549–557. https://doi.org/10.1002/hipo.20615

Vigo D, Thornicroft G, Atun R (2016) Estimating the true global burden of mental illness. Lancet Psychiatry 3(2):171–178. https://doi.org/10.1016/S2215-0366(15)00505-2

Whiteford HA, Degenhardt L, Rehm J et al (2013) Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet 382(9904):1575–1586. https://doi.org/10.1016/S0140-6736(13)61611-6

Williams EG, Auwerx J (2015) The convergence of systems and reductionist approaches in complex trait analysis. Cell 162(1):23–32. https://doi.org/10.1016/j.cell.2015.06.024

Woods KS, Doss CC, Bowyer KW, Solka JL, Priebe CE, Jr WPK (1993) Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. Int J Pattern Recognit Artif Intell 7(6):1417–1436