CHAPTER

16

Multimodal integration

Sandra Vieira[1], *Walter Hugo Lopez Pinaya*[1,2], *Rafael Garcia-Dias*[1], *Andrea Mechelli*[1]

[1] **Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom;** [2] **Centre of Mathematics, Computation, and Cognition, Universidade Federal do ABC, Santo André, São Paulo, Brazil**

16.1 Introduction

The last decade has seen a stepping up in the quest for machine learning–based translational tools to assist clinical decision-making in psychiatry and neurology. Many studies have now been conducted in several brain disorders, resulting in a plethora of encouraging findings (Woo, Chang, Lindquist, & Wager, 2017). However, despite this rapid progress, the deployment of machine learning–based tools in clinical practice will require higher performances than those reported in the existing literature. A limitation of such literature is that most studies have focused on a single data modality, for example, clinical (Mechelli et al., 2017), genetic (Trakadis, Sardaar, Chen, Fulginiti, & Krishnan, 2019), neuropsychological (Wu et al., 2016), or neuroimaging (Arbabshirani, Plis, Sui, & Calhoun, 2017; Wolfers, Buitelaar, Beckmann, Franke, & Marquand, 2015; Woo et al., 2017) data. As different types of data convey complementary information, their combination may allow a richer representation of brain disorders (Wolfers et al., 2015) and better predictive models (Schumann et al., 2014).

The vast majority of machine learning studies of brain disorders have used a single neuroimaging technique. Each neuroimaging technique can only capture a specific type of alteration and therefore can only allow for a partial representation of the underlying neurobiological mechanism (Liu et al. 2015). Fig. 16.1 shows a comparison among the strengths of different neuroimaging techniques according to the metrics described in Liu et al. (2015). In contrast, we know that brain disorders are associated

Machine Learning
https://doi.org/10.1016/B978-0-12-815739-8.00016-X

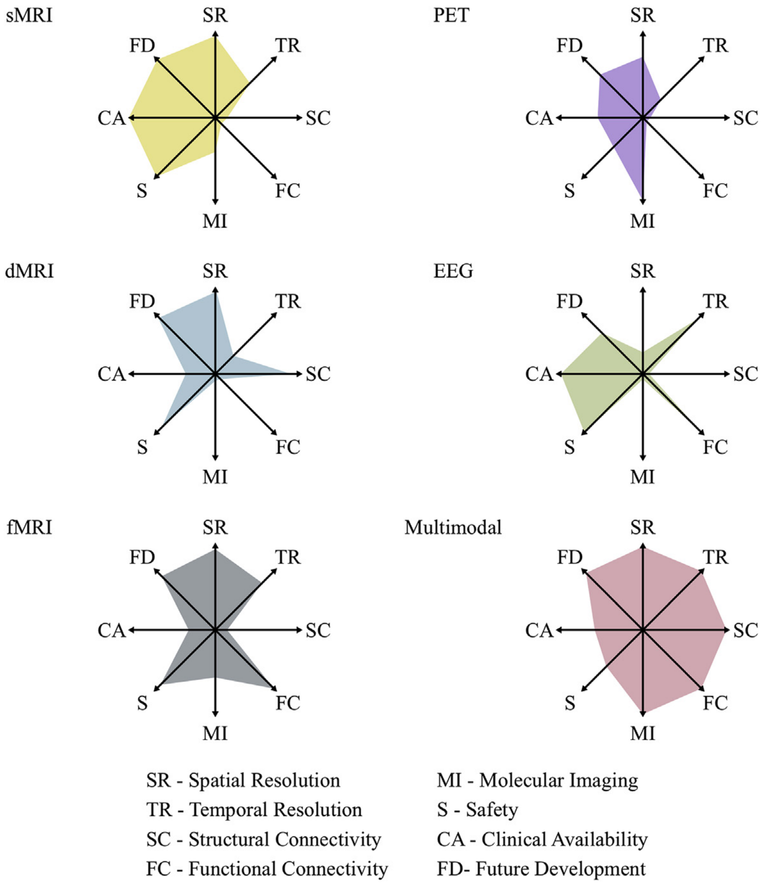

FIGURE 16.1 Graphical representation of the strengths of different neuroimaging techniques including sMRI, diffusion MRI (dMRI), functional MRI (fMRI), positron emission tomography (PET), electroencephalogram, and multimodal neuroimaging. *Adapted from Liu, S., Cai, W., Liu, S., Zhang, F., Fulham, M., Feng, D., et al. (2015). Multimodal neuroimaging computing: A review of the applications in neuropsychiatric disorders.* Brain Informatics, 2*(3), 167–180. https://doi.org/10.1007/s40708-015-0019-x.*

with multiple types of neurobiological alterations. For example, Alzheimer's disease (AD) presents with structural atrophy (deToledo-Morrell et al., 2004; Jack et al., 2005), decreased blood perfusion (Ramírez et al., 2013), and reduced glucose metabolism (Garibotto et al., 2017). Multimodal integration allows us to leverage on the strengths of the different neuroimaging techniques to build a more complete characterization of disease mechanisms. For example, by combining structural magnetic resonance imaging (sMRI) and electroencephalogram (EEG), we can develop a neurobiological model of the disease with both spatial and temporal high resolution.

Given the strong rationale for using multimodal data, one is left wondering why there are not more multimodal machine learning studies being published. After all, current neuroimaging studies tend to use comprehensive assessment protocols that include imaging, clinical, sociodemographic, cognitive, or omics data. It turns out that the implementation of this kind of statistical analysis is not trivial. Many analytical strategies typically require a high level of technical expertise with regard to the handling of the data, the implementation of the machine learning algorithm, or, more likely, both (Calhoun & Sui, 2016). In addition, there is no gold standard approach to combine multimodal data, and the available techniques are difficult to categorize into a straightforward taxonomy (Dahne et al., 2015).

However, a useful approach for categorizing multimodality techniques is based on the stage in which they are implemented along the machine learning pipeline (Noble, 2004). According to this approach, we can have *early*, *intermediate*, or *late* integration (Fig. 16.2). In early integration methods, the focus is on the data itself. Here, data can be simply concatenated or, as we will see later, subjected to a series of complex transformations to enable the extraction of multimodal latent features, which are then used as input features in a machine learning model. In intermediate data integration, different modalities are integrated as part of the model's learning process. Finally, in late integration methods, a separate machine learning model is trained for each modality and then the predictions of the different models are combined to make a final decision. Importantly, because early and late methods are applied before and after the machine learning algorithm, respectively, they can, in principle, be used in combination with any machine learning algorithm; on the other hand, intermediate methods integrate multimodal data as part of the machine learning algorithm itself and therefore are specific to the different types of machine learning algorithms.

In this chapter, we introduce some of the most popular methods for multimodal data integration in machine learning studies of brain disorders, as well as some innovative and promising approaches. We first describe an emerging and promising approach known as *data fusion*. This can be considered an early data integration approach, as it typically involves the application of the methods to the data themselves; the resulting features convey joint information about the different modalities, which can then be used as input data for the machine learning models. Next, we describe intermediate data integration. Here, we discuss kernel-based data integration approaches in the context of Support Vector Machine (SVM)—the most popular machine learning technique in the investigation of brain disorders. Kernel-based data integration methods are arguably some of the most prevalent strategies for combining multimodal data in the machine learning literature. However, although useful,

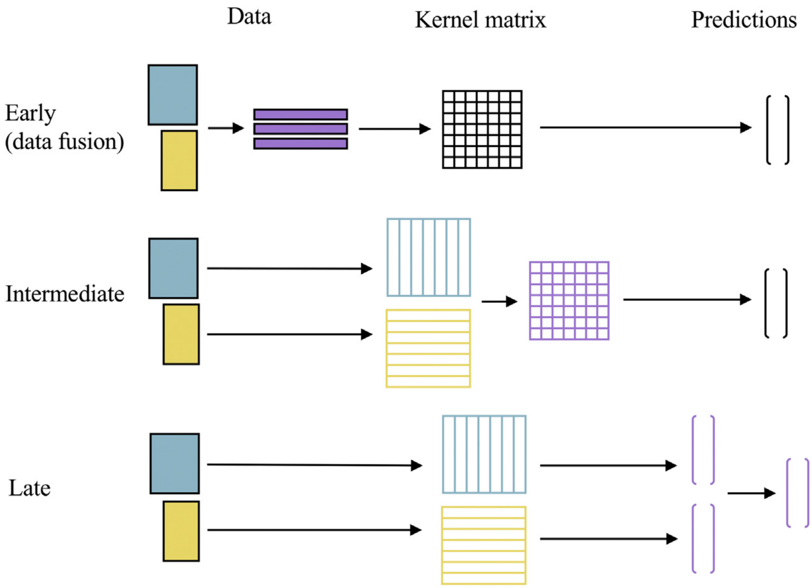

FIGURE 16.2 Illustration of three types of multimodal analysis according to where multimodal data integration takes place along the machine learning pipeline (for simplicity the illustration depicts the combination of two modalities in the context of Support Vector Machine [SVM]). The stage in which multimodal integration takes place is shown in purple (gray in print version). Early methods are implemented straight onto the data, via concatenation of the different modalities or data fusion. The latter aims to find latent features that best represent the interactions between the different modalities; features are then put through an SVM in which a single kernel matrix is computed. In intermediate methods, the different modalities are trained using separate kernels which are subsequently integrated as part of the training process of the SVM. Finally, late methods build an SVM model for each modality and combine the predictions of the different models to make a final decision of class membership. *Adapted from Noble, W. S. (2004). Support vector machine applications in computational biology. In B. Scholkopf. K. Tsuda. J. P. Vert (Eds.),* Kernel methods in computational biology *(pp. 71–92). Cambridge, MA: MIT Press. Retrieved from www.gs.washington.edu/noble/name-change.*

these methods do not fully exploit interactions between modalities. For example, they might not be able to detect a change in functional magnetic resonance imaging (fMRI) data that is contingent to a change in sMRI data. To overcome this limitation, more recent machine learning studies are using deep learning, which is capable of learning complex interactions between the different modalities. Finally, we discuss late data integration methods. These are probably the simplest and most straightforward to implement and have been widely used in the literature.

Given that most multimodal machine learning studies have used neuroimaging data, our description of the above methods focuses on the analysis of multimodal neuroimaging data. However, all methods covered in this chapter can be generalized to other types of data (e.g., neuropsychological scores). In the last part of this chapter, we present exemplar cases of multimodal machine learning studies from the brain disorders literature.

16.2 Early multimodal data integration: data fusion

Data fusion leverages on the interactions between modalities to extract meaningful features that convey joint information stored across multiple sources of data. The features resulting from this process can then be used for different purposes, including machine learning. Therefore, data fusion takes place early in the machine learning pipeline as it is implemented directly on the data.

Calhoun and Sui (2016) provide a useful framework for data fusion methods in the context of neuroimaging data. Of special importance is the distinction between *asymmetric* and *symmetric* techniques. Asymmetric data fusion refers to the use of one modality to constrain another. For example, EEG data can be used as regressors in the analysis of fMRI data to extract voxels that are associated with EEG data. On the other hand, symmetric data fusion treats multiple modalities equally, taking full advantage of the joint information across multiple types of data. Symmetric data fusion methods can be further divided into hypothesis- and data-driven models. While hypothesis-driven models incorporate specific knowledge about the problem, for example, a specific influence of one brain region upon another (e.g., structural equation modeling), they may miss important relationships that are not included in the hypotheses. On the other hand, data-driven models do not require a priori hypotheses and thus are useful for exploring the data. For this reason, symmetric data-driven fusion is often described as the most powerful way of dealing with multimodal data (see Calhoun and Sui (2016) and Sui, Adali, Yu, Chen, and Calhoun (2012) for a review).

In this section, we provide an overview of some of the most commonly used symmetric data-driven techniques: joint independent component analysis (jICA), multimodal canonical correlation analysis (mCCA), and the combination of the two (jICA + mCCA).

16.2.1 jICA

jICA is an extension of the popular independent component analysis (ICA). Briefly, ICA is a technique for revealing hidden factors that underlie a set of observable data. ICA has been widely used to solve blind source separation problems (Fig. 16.3); these include, for example, the problem of deriving brain waves recorded using multiple sensors and the problem of removing interfering radio signals reaching a mobile phone.

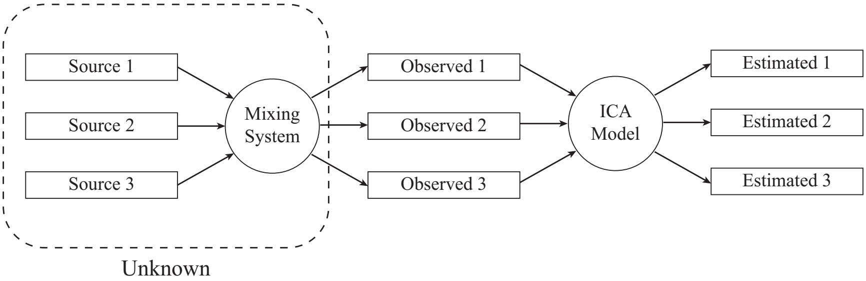

FIGURE 16.3 Diagram of a blind source separation problem. The main components of this diagram are the sources (e.g., the speakers in the cocktail problem; here referred to as Source 1, 2, and 3); the mixing system which determines the linear combination between the sources; the observable data (e.g., the data captured by the microphones; here referred to as Observed 1, 2, and 3); the ICA model that tries to disentangle the sources; and the estimated individual sources (e.g., the clean voice of each person recovered via ICA; here referred to as Estimated 1, 2, and 3). When applying ICA, it is assumed that the observed variables are linear combinations of the speakers' speech signals and that the sources are independent and non-Gaussian.

Among the multiple blind source separation problems that can be addressed using ICA, perhaps the most famous one is the so-called "cocktail party" problem (Bell & Sejnowsky, 1995). In this problem, we are at a cocktail party with many people speaking simultaneously and multiple microphones in the room, each one recording a mixture of the voices. Here, the aim is to recover the clean voice of each person. The ICA algorithm can help us achieve this aim by enabling the separation of the recorded mixed signals into individual sources.

The challenge in this problem is that the sources and the mixing system are unknown. To solve this, the ICA algorithm makes a few assumptions about the sources and the mixing system. It assumes that the sources are statistically independent (i.e., observing the value of one component does not provide any information about the value of the other) and nonnormally distributed. In addition, the ICA algorithm makes the further assumption that the mixing system is linear. In the end, the ICA algorithm will create an unmixing matrix that estimates values from the sources, where these estimates are usually called *independent components* (ICs). The resulting ICs are assumed to capture the main latent properties of the data; for example, in the case of fMRI data, the different components are thought to reflect distinct functional networks during a certain task. There are alternative versions of ICA; however, a technical description of these versions is beyond the scope of this book.

Now, considering a multimodal dataset with two types of data, we could simply apply this same procedure twice, one for each modality, to capture the main latent properties. However, this would result in unrelated unmixing matrices and different ICs for each modality.

An alternative option would be to simply concatenate the multimodal data and extract the ICs; in this case, the resulting latent component will be based on the characteristics of both modalities, and therefore this can be considered an example of data fusion. This is the idea of the jICA algorithm (Fig. 16.4).

16.2.2 Multimodal canonical correlation analysis

Canonical correlation analysis (CCA) is a multivariate statistical model that captures the linear interrelationships between two groups of features. Specifically, CCA seeks to reexpress two groups of features (where we have its values along many individuals) as multiple pairs of new transformed features (Fig. 16.5A). A new feature (called *canonical variate*) is associated with the original group of features via a set of weights, which are defined by a canonical vector (Fig. 16.5B). One canonical variate is formed for each group of features; this means that, in a new pair of canonical variates, the canonical variate for the group of features number 1 has one canonical vector defining the weights, and the canonical variate for the group of features number 2 has a different canonical vector. The challenge is how to define the weights of the canonical vectors. In CCA, these weights are chosen in a way that maximizes the correlation between the two canonical variates, where the resulting correlation coefficient is called the canonical correlation coefficient (CCC). The CCC measures the strength of the relationship between the two canonical variates (Fig. 16.5C). A detailed description of the technical algorithm used to maximize the CCC is outside the scope of this chapter. Using CCA, we can create several pairs of canonical covariates where the different canonical

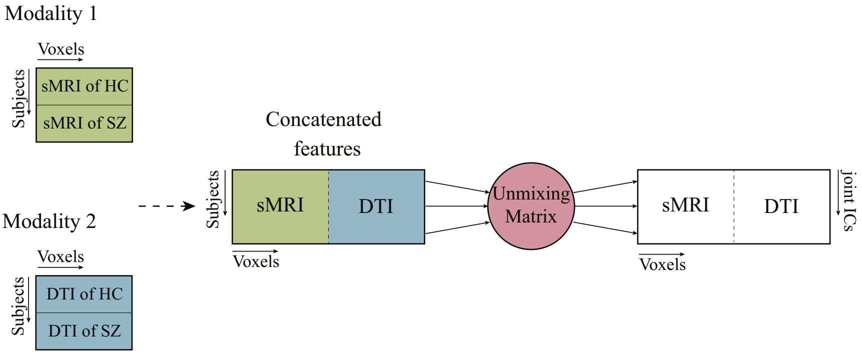

FIGURE 16.4 The joint ICA concatenates the two modalities (e.g., sMRI and DTI data from healthy control (HC) and patients with schizophrenia (SZ)). Then, it finds a common unmixing matrix from which joint ICs are extracted. The resulting independent component will contain joint information from both modalities.

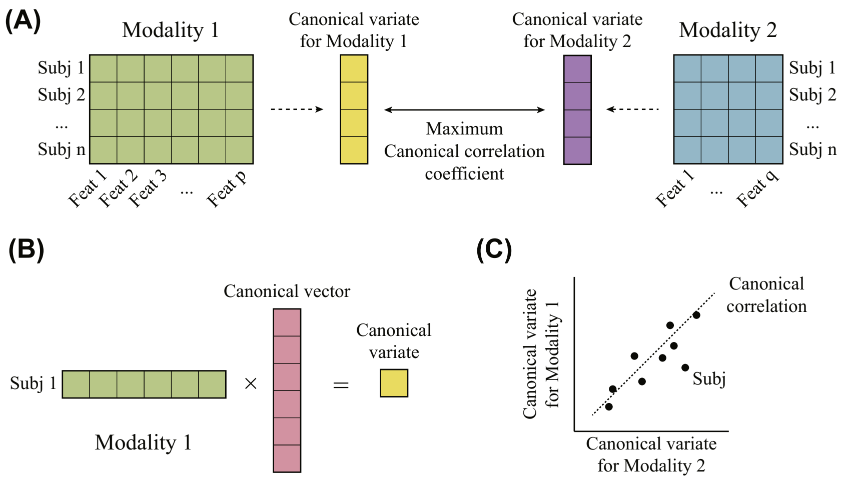
(A)
Modality 1
Canonical variate for Modality 1
Canonical variate for Modality 2
Modality 2
Subj 1
Subj 2
...
Subj n
Maximum Canonical correlation coefficient
Feat 1
Feat 2
Feat 3
...
Feat p
Feat 1
...
Feat q
(B)
Canonical vector
Canonical variate
Subj 1
×
=
Modality 1
(C)
Canonical correlation
Canonical variate for Modality 1
Subj
Canonical variate for Modality 2

FIGURE 16.5 Canonical correlation analysis (CCA). (A) Two groups of features (Feat), one with p and the other with q number of features, can be expressed as canonical variates. Each pair of canonical variates are defined in a way to maximize their correlation (indicated by the canonical correlation coefficient). (B) Each value on the canonical variate is computed by the weighted sum of the features by the canonical vector. (C) Scatter plot showing each subject of the dataset represented by the canonical variates and the correlation between the covariates. *Adapted from Wang, H. T., Smallwood, J., Mourao-Miranda, J., Xia, C. H., Satterthwaite, T. D., Bassett, D. S., et al. (2018). Finding the needle in high-dimensional haystack: A tutorial on canonical correlation analysis. Retrieved from http://arxiv.org/abs/1812.02598.*

vectors are orthogonal (independent) from each other, thereby representing different relationships found among the groups of features.

CCA can be easily applied to the analysis of multimodal data (mCCA; Fig. 16.6) to link multiple measures together and identify multivariate patterns between two data modalities (Correa, Li, Adalı, & Calhoun, 2008). For example, CCA allows a data matrix comprised of brain measurements (e.g., brain regional volumes) to be analyzed with respect to a second data matrix comprised of behavioral measurements (e.g., neuropsychological scores). Here, mCCA decomposes each modality into canonical variates that have the maximum CCC across the two modalities. Compared to jICA that constrains two groups of features to have one single unmixing matrix, mCCA is a more flexible technique allowing for the estimation of a distinct mixing matrix for each modality.

16.2.3 Multimodal CCA + joint ICA

The combination of the mCCA and jICA takes advantage of the benefits of each method. The jICA algorithm allows us to estimate distinct independent components; however, it creates one single unmixing matrix for all modalities. On the other hand, mCCA outputs a group of canonical

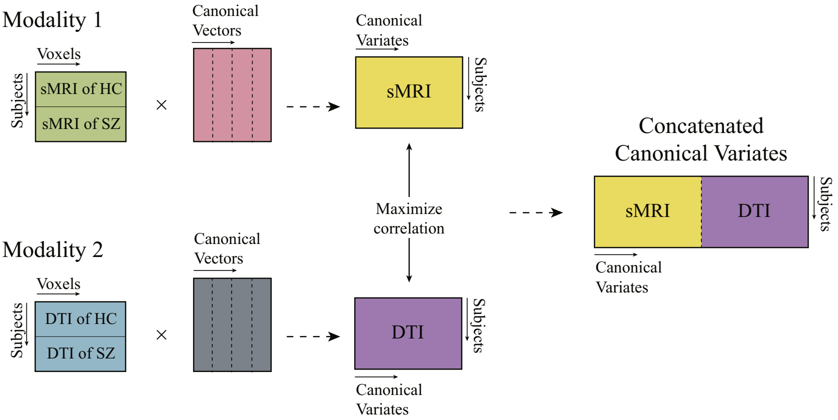

FIGURE 16.6 Schematic of mCCA, a multivariate tool that can simultaneously consider two variable sets from different modalities to uncover essential hidden associations. mCCA estimates a mixing matrix for each modality such that they are maximally correlated.

vectors for each modality, but the resulting canonical variates may not differ from each other sufficiently in some cases. There is evidence that this combination tends to outperform each individual method without increasing the computational cost (Sui et al., 2011). In brief, mCCA first links the different modalities by finding the associated covariates with the maximal correlation between them. The final canonical covariates are then concatenated and jICA extracts the joint independent components (Sui et al., 2012) (Fig. 16.7).

16.3 Intermediate multimodal integration: kernel-based methods and deep learning

16.3.1 Kernel-based methods

Here, we describe two examples of intermediate data integration methods: unweighted sum of kernels and multi-kernel learning. Both these approaches are specific to kernel-based machine learning algorithms; therefore, we refer to them as "kernel-based" methods and explain them in the context of SVM. This algorithm is described in detail in Chapter 6. In its simplest form, SVM uses the dot product (linear kernel function) that calculates similarities in the input data and stores this information in the kernel matrix. This creates a new feature space where the SVM model can use a linear decision boundary to separate the classes.

Now, to integrate multiple modalities during the training process of an SVM and generate a single output, we can build a kernel matrix for each

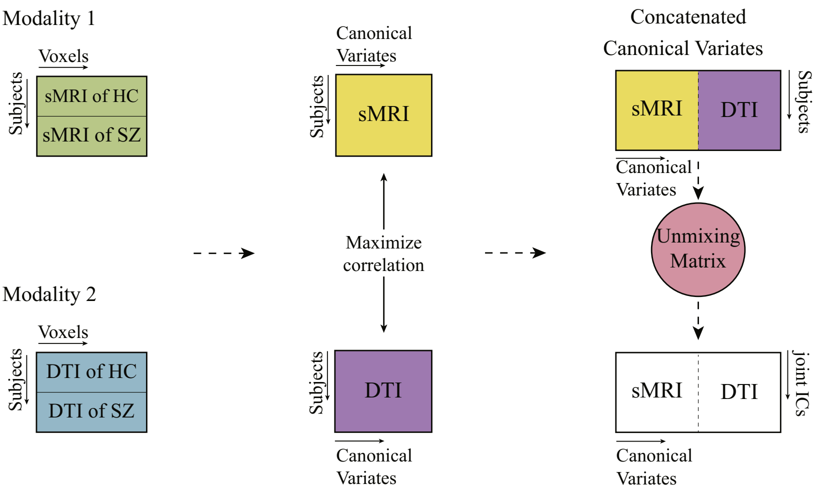

FIGURE 16.7 Multimodal CCA + joint ICA. The two methods are combined by using mCCA to estimate the independent components, and the using these as input for jICA.

modality and find a linear combination of the resulting kernel matrix to create a final decision boundary. There are at least two different ways of doing this: (1) unweighted simple sum of kernels and (2) multi-kernel learning (Fig. 16.8).

16.3.1.1 *Unweighted simple sum of kernels*

If we think of a linear kernel as a way of representing the similarity between data points within a certain modality, we can integrate different types of data simply by building a kernel for each modality and then adding them together to create a single kernel matrix (K') that represents all modalities. An SVM is then trained and tested using this new integrated kernel (K'). Importantly, different modalities may have different numbers of features and may be scaled differently. To correct for this, each kernel must be normalized before the different kernels are added together. Note that, because the kernels are simply summed, the data from the different modalities have equal weightings in terms of their contribution to the decision function.

16.3.1.2 *Multi-kernel learning*

Unlike the previous method where the kernels were simply summed with equal weightings, the multi-kernel learning (MKL) approach aims to automatically find the optimal way of combining the individual kernels. This is achieved by estimating the "best kernel" from a linear combination of "base" kernels, one from each modality. For more details on MKL, see Chapter 8.

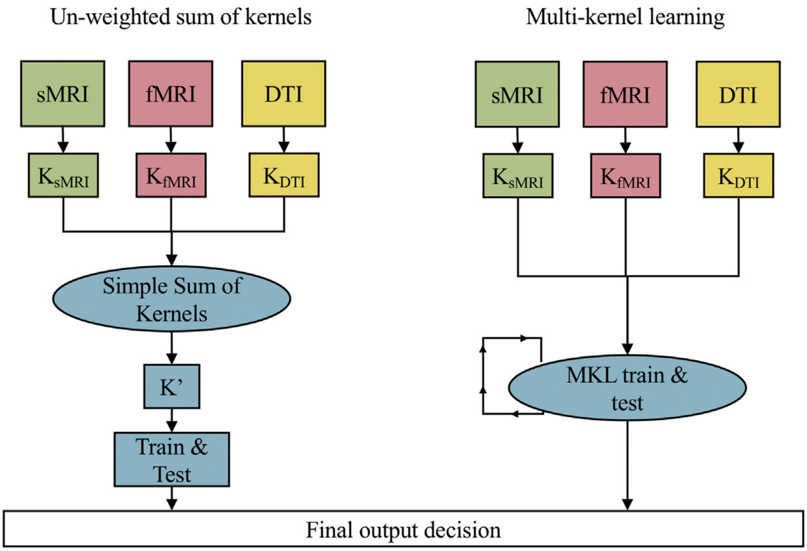

FIGURE 16.8 Flowchart depicting the processing pipeline for un-weighted sum of kernels and multi-kernel learning. MRI, magnetic resonance imaging; sMRI, structural MRI; DTI, diffusion tensor imaging; fMRI, functional MRI; $K_{sMRI}/K_{fMRI}/K_{DRI}$, kernel matrix for sMRI/fMRI/DTI data; K′, integrated kernel matrix. *Adapted from Pettersson-Yeo, W., Benetti, S., Marquand, A. F., Joules, R., Catani, M., Williams, S. C. R., et al. (2014). An empirical comparison of different approaches for combining multimodal neuroimaging data with support vector machine.* Frontiers in Neuroscience, 8*, 189.*

16.3.2 Deep learning

Although kernel-based methods are useful, they do not fully exploit interactions between modalities. To overcome this limitation, more recent machine learning studies are using deep learning—a promising approach that is gaining considerable attention after outperforming traditional machine learning models in multiple problem domains (Durstewitz, Koppe, & Meyer-Lindenberg, 2019) (see Chapters 9—11). Part of a wider family of representation learning, deep learning seeks to find highly abstract representations from the data through consecutive layers of nonlinear transformations (LeCun, Bengio, & Hinton, 2015). Several studies have now shown the potential of such approach in neuroimaging (Durstewitz et al., 2019; Vieira, Pinaya, & Mechelli, 2017). The ability to automatically discover high-level representations may be particularly useful in brain disorders because different sources of information such as genetics and neuroimaging data are unlikely to have a simple linear correspondence (Plis et al., 2014). In addition, because we often lack hypotheses about how different modalities may be related to each other, data-driven methods such as deep learning may particular useful (Durstewitz et al., 2019).

Deep learning is able to extract high-level cross-information from multiple modalities and output individual-level predictions all in the same model (Fig. 16.9). This is an important departure from the data integration methods described earlier in this chapter, which allowed one to (1) extract meaningful cross-information from multiple modalities and subsequently use this as input features for a machine learning algorithm (early data integration—data fusion); (2) input the different modalities separately and integrate them during the learning phase of the machine learning algorithm (traditional intermediate methods such as the kernel-based methods); or (3) input the different modalities separately and integrate them during the testing phase (late data integration). In addition, contrary to previous methods that only rely on the linear associations between modalities, deep learning is also capable of finding nonlinear associations in the data.

An important limitation of deep learning is that it tends to require large amounts of data. A related limitation is that it can be hard to predict how much data are enough to leverage this approach. However, with the rapid expansion of multisite consortia and data sharing initiatives, these limitations are likely to become less of an obstacle in the future. It is expected that many more multimodal studies based on deep learning will emerge in the next few years.

16.4 Late multimodal integration: ensemble methods

Ensemble methods can be used to integrate multimodal data by using the output of each individually trained classifier to form an ensemble

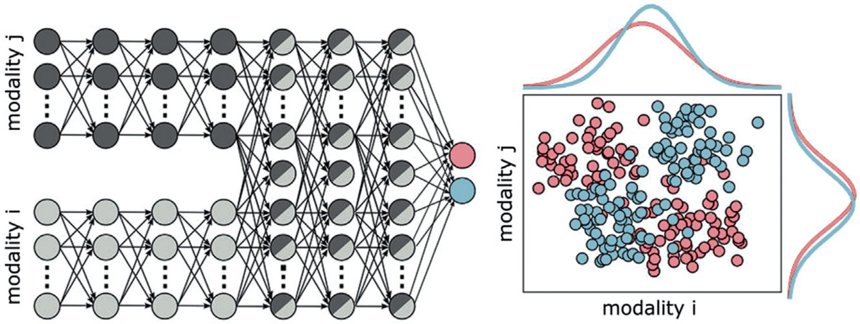

FIGURE 16.9 Illustration of a generic multimodal deep learning network. Left: lower layers represent modality-specific properties while deeper layers may learn complex combinations of features from the different modalities. Right: a nonlinear combination of data from two modalities may allow for easier separation of two disorders compared to data from a single modality. *Adapted from Durstewitz, D., Koppe, G., Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry.* Molecular Psychiatry, 1. *https://doi.org/10.1038/s41380-019-0365-9; Calhoun, V. D., Sui, J. (2016). Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness.* Biological Psychiatry. Cognitive Neuroscience and Neuroimaging, 1*(3), 230–244. https://doi.org/10.1016/j.bpsc.2015.12.005.*

decision. There are several possible approaches for multimodal integration with ensemble methods. Here, we focus on two voting methods: soft voting and majority voting. As these methods only use the predictions of the classifiers, they can be applied to any machine learning algorithm (Fig. 16.10).

16.4.1 Soft voting

The late integration of different modalities can be implemented by training a machine learning model with each single modality, resulting in N models for N modalities. For example, let us imagine a situation in which we have to predict the category of an individual (e.g., patient vs. healthy control) for whom we have multimodal data. In this case, each type of data will be used to train a model, and each model will predict the probability of this individual belonging to each of the two categories. Using soft voting, we can combine these predictions by calculating the mean probability over the two categories; we can then select the category with the highest probability as the final prediction. For example, if the predictive probability of an individual being a patient based on their sMRI, fMRI and neuropsychological data are 0.3, 0.6, and 0.8,

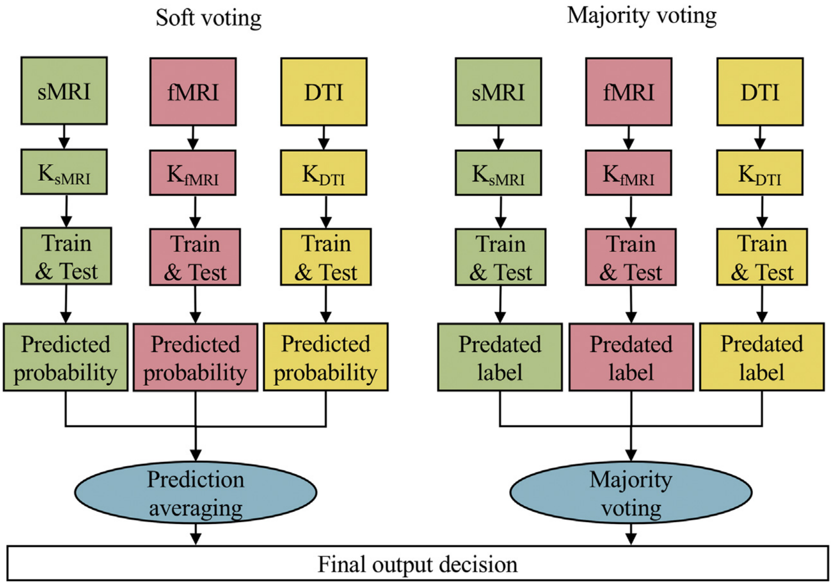

FIGURE 16.10 Two voting methods for late data integration: soft voting and majority voting. *Adapted from Pettersson-Yeo, W., Benetti, S., Marquand, A. F., Joules, R., Catani, M., Williams, S. C. R., et al. (2014). An empirical comparison of different approaches for combining multimodal neuroimaging data with support vector machine.* Frontiers in Neuroscience, 8, *189.*

respectively; the soft voting method averages these values to 0.57. This means that the probability of this individual being a patient is higher than the probability of them being a healthy control ($P = 1 - 0.57 = 0.43$). We can therefore conclude that, based on this ensemble of models, the individual in question belongs to the patient class.

16.4.2 Majority voting

Similar to soft voting, majority voting also uses the output of each base classifier. However, in this case, it is the predicted class labels (e.g., 0 or 1) that are considered, rather than the predicted probabilities (e.g., 0.5, 0.7, 0.2). The final class label is the class with the largest number of predictions among the base classifiers. In the example above, in which the predictive probability of being a patient was 0.3, 0.6, and 0.8 for sMRI, fMRI, and neuropsychological data, respectively, the equivalent binary predicted probabilities would be “not patient,” “patient,” and “patient”; because two of the three predictions are “patient,” the final predicted class label using the majority voting method would be “patient.” Had there been four instead of three modalities to combine, with predicted labels “not patient”, “not patient,” “patient,” and “patient,” it would not have been possible to determine a final predicted class label. Therefore, in cases where there is an even number of data types to be combined, there is a risk of tied decisions; in such cases, the final predicted class label is determined by breaking the ties using any arbitrary heuristic provided, which is chosen a priori.

16.5 Application to brain disorders

Although a few multimodal machine learning studies have been conducted, these are not yet commonplace in the brain disorders literature. For example, in three recent metaanalyses of machine learning neuroimaging studies in depression, schizophrenia, and bipolar disorder, there is almost no mention of multimodal studies (Kambeitz et al., 2015, 2017; Librenza-Garcia et al., 2017). Perhaps this is not surprising given the recency of the field. Indeed, there are still many areas for improvement in single-modality studies. For example, while the vast majority of studies have used neuroimaging data, several other types of data are yet to be fully explored, such as clinical, behavioral, neuropsychological, or omics data; most studies so far are limited to a small range of disorders (AD, schizophrenia, bipolar, and depression), and current findings from single-modality neuroimaging studies need to be replicated in larger and more diverse samples. Multimodal data integration can also be technically demanding, especially when it comes to more advanced methods such as data fusion and deep learning. Nevertheless, there is a growing body of evidence showing the potential of combining multimodal data and machine learning for building powerful predictive models. In this section, we provide some exemplar studies from the literature with respect to three main groups of methods covered in this chapter: kernel-based and ensemble methods, data fusion, and deep learning.

16.5.1 Multimodal studies with kernel-based and ensemble methods in psychosis

Kernel-based and ensemble methods are among the most common strategies for integrating multimodal data in machine learning. Cabral et al. (2016) used neuroanatomical and brain function data to classify patients with schizophrenia and healthy controls. A separate regularized logistic regression model was first trained for each modality. The final prediction for each participant was then estimated by taking the mean of the two trained models′ decision scores, i.e., soft voting. The ensemble model was able to classify patients and controls with an accuracy of 75%, outperforming both sMRI and fMRI classifiers by 4% and 5%, respectively. In one of the few attempts so far to combine genetic and neuroimaging data to classify patients with schizophrenia and healthy controls, Yang, Liu, Sui, Pearlson, and Calhoun (2010) trained separate SVM models for single-nucleotide polymorphism (SNP) and fMRI data. The predictions of the two models were then combined via majority voting (i.e., hard voting) to make a final decision. The combination of genetic and fMRI data achieved better accuracy than either type of data alone, suggesting that the two modalities capture different but partially complementary aspects of the illness (Yang et al., 2010).

In an attempt to empirically evaluate different approaches, Pettersson-Yeo et al. (2014) used several data integration methods to classify individuals at ultrahigh risk of psychosis, first episode psychosis, and healthy controls based on sMRI, fMRI, and diffusion tensor imaging (DTI). Four methods were compared in terms of their ability to enhance the best single-modality classification accuracy: (1) an unweighted sum of kernels, (2) MKL, (3) soft voting, and (4) hard voting. Results showed that, while the integration of multiple modalities can enhance classification accuracy by up to 13%, such integration does not always lead to improved performance; this depends on the chosen method and specific diagnostic comparison under consideration. Importantly, it was also found that simple methods (e.g., soft voting) may perform better than more complex alternatives (e.g., MKL).

Overall, these studies provide encouraging evidence that combining different modalities could improve the accuracy of classification compared to single-modality approaches. However, as highlighted in Pettersson-Yeo et al. (2014), most studies have only used neuroimaging data. The integration of more diverse types of data (e.g., genetic, cognitive, clinical, and neuroimaging), which are likely to convey less overlapping information, could potentially provide a more complete picture of brain disorders and further improve the accuracy of classification.

16.5.2 Data fusion in schizophrenia and bipolar disorder

Data fusion methods can be used independently from machine learning methods. For example, the earliest studies using data fusion, dating back to the early 2000s (e.g., Calhoun, Adali, Pearlson, & Kiehl, 2006), did not involve the use of machine learning. In one of these studies, Calhoun et al. (2006) used jICA to fuse fMRI and EEG data from healthy controls. Later the same team expanded this work to investigate patients with schizophrenia, bipolar disorder, and healthy controls. This involved using mCCA + jICA to fuse fMRI and DTI (Sui et al., 2011), sMRI, DTI, and fMRI (Sui et al., 2013), as well as using parallel ICA to fuse genetic and imaging data (Meda et al., 2010; Pearlson, Liu, & Calhoun, 2015; Rashid et al., 2019).

More recently, studies have started to use data fusion methods to generate features that can be used as input for machine learning models. In one such studies, Sui et al. (2018) aimed to predict cognition scores in patients with schizophrenia. To achieve this, sMRI, dMRI, and fMRI data were first analyzed with mCCA with reference (mCCAR; an extension of the mCCA) plus jICA to extract latent joint independent components across the three modalities. The extracted components included areas that are part of the central executive network, the salience network, and the default mode network, in line with the triple network model of major psychopathology proposed by Menon (2011). The same components were then used to predict cognitive scores using multiple linear regression. The association between predicted and true cognitive scores achieved a moderate correlation of 0.5. Notably, when the same procedure for data fusion was applied to two validation cohorts and the resulting components were used as predictors in the regression model trained in the discovery cohort, there was a correlation of 0.2 and 0.4 between the predicted and true cognitive scores for each one of the validation cohorts. Overall, this study provides seminal evidence that data fusion, combined with even a simple model such as linear regression, can provide valuable insight into structural and functional alterations in schizophrenia.

16.5.3 Multimodal neuroimaging studies with deep learning in Alzheimer's disease

The flexibility of deep learning networks means that architectures can be easily extended to process multimodal data. Therefore, although the application of deep learning in brain disorders research is recent, there have already been a number of studies combining different neuroimaging modalities. Unsurprisingly, the vast majority of these studies used the widely popular ADNI dataset to classify AD and mild cognitive impairment (MCI) from controls (Vieira et al., 2017). Most studies have used some variation or combination of deep neural networks (DNNs), convolutional neural networks (CNNs), or autoencoders (AE), which are explained in more detail in Chapters 9—11.

As an example, Vu, Yang, Nguyen, Oh, and Kim (2017) used a 3D CNN model to combine sMRI and PET data. As depicted in Fig. 16.11, features were extracted from each modality and then combined in later hidden layers. The initial layer was pretrained with an AE to learn basic properties of the data regardless of diagnostic label. The CNN was then initialized with these pretrained weights and the whole network was fined-tuned via gradient descent. This model classified AD with an accuracy of 91%. In another recent study, Zhou, Thung, Zhu, and Shen (2019) attempted to combine imaging and genetic data using a similar approach, although in this case DNNs were employed in a three-stage scheme (Fig. 16.12). In stage 1, the initial layers learned latent representations for each modality independently. In stage 2, the next group of layers learned the joint latent feature representations for each pair of modality combination (sMRI and PET, sMRI and SNP, PET and SNP) using the high-level features learned from stage 1. Finally, in stage 3, the network predicted the diagnostic labels by fusing the learned joint latent feature representations from the previous stage. The resulting model was able to distinguish between AD and healthy controls with 92% accuracy and between MCI and healthy controls with 75% accuracy. In addition, when the same model was used in two multiclass classifications, it achieved reasonable performance: (1) 64% accuracy for the classification of AD, MCI, and healthy controls; (2) 55% accuracy for the classification of four groups including individuals with "progressive" MCI who developed AD at follow-up, individuals with "stable" MCI who did not develop AD at follow-up, individuals with AD and healthy controls.

Taken collectively, the existing literature suggests that multimodal deep learning tends to perform better than single modality studies as well as multimodal studies with traditional "shallow" methods (Durstewitz et al., 2019; Vieira et al., 2017). However, most studies have used moderate sample sizes, which can lead to overfitting in the "data-hungry" deep

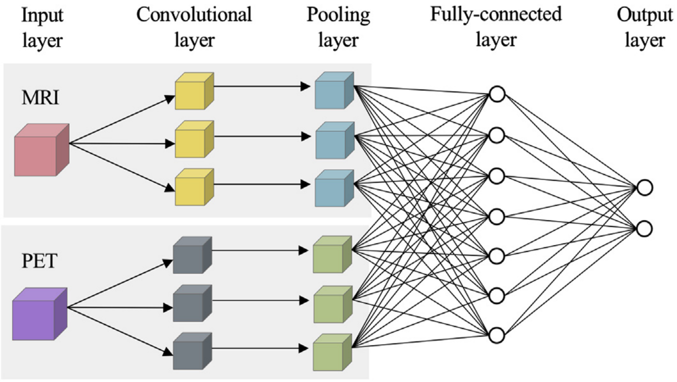

FIGURE 16.11 CNN model pretrained with AE to integrate sMRI and PET. *Adapted from Vu, T. D., Yang, H.J., Nguyen, V. Q., Oh, A.R., Kim, M.S. (2017). Multimodal learning using convolution neural network and Sparse Autoencoder.* In 2017 IEEE international conference on big data and smart computing (BigComp) *(pp. 309–312). IEEE. https://doi.org/10.1109/BIGCOMP.2017.7881683.*

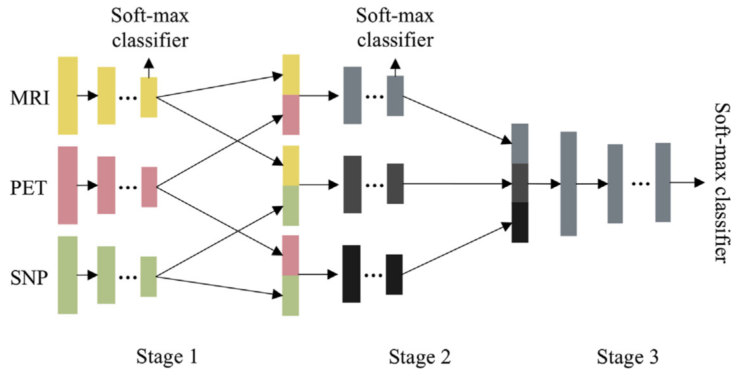

FIGURE 16.12 DNN model that combines sMRI, PET and genetic data. *Adapted from Zhou, Thung, Zhu, and Shen (2019).*

learning models. As larger samples become increasingly available, multimodal deep learning studies are likely to continue to grow in the next few years.

16.6 Conclusion

Multimodal machine learning is a recent yet fast-growing topic in brain disorders research. The aim is to capitalize on the complementary nature of different modalities to build better prediction models. In this chapter, we introduced three classes of data integration methods according to the stage in which they are implemented along the machine learning pipeline. Most studies have used either an intermediate or late data integration method, typically MKL or some form of ensemble approach. However, more recently, data fusion has been gaining increasing attention as a powerful way of uncovering latent cross-information between modalities. The use of data fusion in combination with machine learning is very recent and, so far, only a limited number of studies have been conducted. Initial evidence has shown promise, and this is likely to be an expanding area in the next coming years. Finally, deep learning is also gaining considerable interest as a possible avenue for multimodal prediction modeling. Its ability to extract high-level multimodal features may be particularly useful to uncover relations in data that are likely to be associated in a complex manner, such as imaging and genetics. Although these models tend to require large amounts of data, the increasingly easy access to large sample size will likely propel significant advances in multimodal deep learning models. Additionally, an increasing number of studies are also collecting a wide range of measurements, such as imaging, genetic, and neuropsychological data. The efficient use of this information is a challenging task but one that promises significant gains in the pursuit of comprehensive models of brain disorders.

16.7 Key points

- Multimodal machine learning leverages on the complementary nature of different types of data to build better prediction models.
- Multimodal methods can be categorized into early, intermediate, and late integration methods, depending on what stage in which the data integration takes place along the machine learning pipeline.
- Early data integration methods involve transforming the data before putting them through a machine learning algorithm.
- Data fusion is a promising early data integration method that involves the extraction of cross-modal information to be used as input features for machine learning.
- Intermediate data integration methods combine data during the training phase; examples include sum of kernels, multi-kernel learning and deep learning.
- Deep learning is a promising approach capable of both extracting high-level features from multimodal data and using them for individual-level predictions.
- Late data integration methods use the outputs of different models, each trained on a different modality, to make a final decision; ensemble methods are a classic example.

References

deToledo-Morrell, L., Stoub, T., Bulgakova, M., Wilson, R., Bennett, D., Leurgans, S., et al. (2004). MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. *Neurobiology of Aging, 25*(9), 1197–1203. https://doi.org/10.1016/J.NEUROBIOLAGING.2003.12.007.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage, 145*, 137–165. https://doi.org/10.1016/J.NEUROIMAGE.2016.02.079.

Bell, A. J., & Sejnowsky, T. J. (1995). An information maximisation approach to blind separation and blind deconvolution. *Neural Computation, 7*(1), l.

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

Calhoun, V. D., Adali, T., Pearlson, G. D., & Kiehl, K. A. (2006). Neuronal chronometry of target detection: Fusion of hemodynamic and event-related potential data. *NeuroImage, 30*(2), 544–553. https://doi.org/10.1016/J.NEUROIMAGE.2005.08.060.

Calhoun, V. D., & Sui, J. (2016). Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging, 1*(3), 230–244. https://doi.org/10.1016/j.bpsc.2015.12.005.

Correa, N. M., Li, Y.-O., Adalı, T., & Calhoun, V. D. (2008). Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia. *IEEE Journal of Selected Topics in Signal Processing, 2*(6), 998–1007. https://doi.org/10.1109/JSTSP.2008.2008265.

Dahne, S., Bieszmann, F., Samek, W., Haufe, S., Goltz, D., Gundlach, C., et al. (2015). Multivariate machine learning methods for fusing multimodal functional neuroimaging data. *Proceedings of the IEEE, 103*(9), 1507–1530. https://doi.org/10.1109/JPROC.2015.2425807.

Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry, 1*. https://doi.org/10.1038/s41380-019-0365-9.

Garibotto, V., Herholz, K., Boccardi, M., Picco, A., Varrone, A., Nordberg, A., et al. (2017). Clinical validity of brain fluorodeoxyglucose positron emission tomography as a biomarker for Alzheimer's disease in the context of a structured 5-phase development framework. *Neurobiology of Aging, 52*, 183–195. https://doi.org/10.1016/J.NEUROBIOLAGING.2016.03.033.

Jack, C. R., Petersen, R. C., Xu, Y. C., O'Brien, P. C., Smith, G. E., Ivnik, R. J., et al. (2005). Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology, 52*(7), 1397–1403. https://doi.org/10.1212/01.wnl.0000180958.22678.91.

Kambeitz, J., Cabral, C., Sacchet, M. D., Gotlib, I. H., Zahn, R., Serpa, M. H., et al. (2017). Detecting neuroimaging biomarkers for depression: A meta-analysis of multivariate pattern recognition studies. *Biological Psychiatry, 82*(5), 330–338. https://doi.org/10.1016/J.BIOPSYCH.2016.10.028.

Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., et al. (2015). Detecting neuroimaging biomarkers for schizophrenia: A meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology, 40*(7), 1742–1751. https://doi.org/10.1038/npp.2015.22.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539.

Librenza-Garcia, D., Kotzian, B. J., Yang, J., Mwangi, B., Cao, B., Pereira Lima, L. N., et al. (2017). The impact of machine learning techniques in the study of bipolar disorder: A systematic review. *Neuroscience and Biobehavioral Reviews, 80*, 538–554. https://doi.org/10.1016/J.NEUBIOREV.2017.07.004.

Liu, S., Cai, W., Liu, S., Zhang, F., Fulham, M., Feng, D., et al. (2015). Multimodal neuroimaging computing: A review of the applications in neuropsychiatric disorders. *Brain Informatics, 2*(3), 167–180. https://doi.org/10.1007/s40708-015-0019-x.

Mechelli, A., Lin, A., Wood, S., McGorry, P., Amminger, P., Tognin, S., et al. (2017). Using clinical information to make individualized prognostic predictions in people at ultra high risk for psychosis. *Schizophrenia Research, 184*, 32–38. https://doi.org/10.1016/j.schres.2016.11.047.

Meda, S. A., Jagannathan, K., Gelernter, J., Calhoun, V. D., Liu, J., Stevens, M. C., et al. (2010). A pilot multivariate parallel ICA study to investigate differential linkage between neural networks and genetic profiles in schizophrenia. *NeuroImage, 53*(3), 1007–1015. https://doi.org/10.1016/J.NEUROIMAGE.2009.11.052.

Menon, V. (2011). Large-scale brain networks and psychopathology: A unifying triple network model. *Trends in Cognitive Sciences, 15*(10), 483–506. https://doi.org/10.1016/j.tics.2011.08.003.

Noble, W. S. (2004). Support vector machine applications in computational biology. In B. Scholkopf, K. Tsuda, & J. P. Vert (Eds.), *Kernel methods in computational biology* (pp. 71–92). Cambridge, MA: MIT Press. Retrieved from www.gs.washington.edu/noble/name-change.

Pearlson, G. D., Liu, J., & Calhoun, V. D. (2015). An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders. *Frontiers in Genetics, 6*, 276. https://doi.org/10.3389/fgene.2015.00276.

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

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E.a., Bockholt, H. J., Long, J. D., et al. (2014). Deep learning for neuroimaging: A validation study. *Frontiers in Neuroscience, 8*(August), 1–11. https://doi.org/10.3389/fnins.2014.00229.

Ramírez, J., Górriz, J. M., Salas-Gonzalez, D., Romero, A., López, M., Álvarez, I., et al. (2013). Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features. *Information Sciences, 237*, 59–72. https://doi.org/10.1016/J.INS.2009.05.012.

Rashid, B., Chen, J., Rashid, I., Damaraju, E., Liu, J., Miller, R., et al. (2019). A framework for linking resting-state chronnectome/genome features in schizophrenia: A pilot study. *NeuroImage, 184*, 843–854. https://doi.org/10.1016/J.NEUROIMAGE.2018.10.004.

Schumann, G., Binder, E. B., Holte, A., de Kloet, E. R., Oedegaard, K. J., Robbins, T. W., et al. (2014). Stratified medicine for mental disorders. *European Neuropsychopharmacology, 24*(1), 5–50. https://doi.org/10.1016/J.EURONEURO.2013.09.010.

Sui, J., Adali, T., Yu, Q., Chen, J., & Calhoun, V. D. (2012). A review of multivariate methods for multimodal fusion of brain imaging data. *Journal of Neuroscience Methods, 204*(1), 68–81. https://doi.org/10.1016/J.JNEUMETH.2011.10.031.

Sui, J., He, H., Pearlson, G. D., Adali, T., Kiehl, K. A., Yu, Q., et al. (2013). Three-way (N-way) fusion of brain imaging data based on mCCA + jICA and its application to discriminating schizophrenia. *NeuroImage, 66*, 119–132. https://doi.org/10.1016/J.NEUROIMAGE.2012.10.051.

Sui, J., Pearlson, G., Caprihan, A., Adali, T., Kiehl, K. A., Liu, J., et al. (2011). Discriminating schizophrenia and bipolar disorder by fusing fMRI and DTI in a multimodal CCA+ joint ICA model. *NeuroImage, 57*(3), 839–855. https://doi.org/10.1016/J.NEUROIMAGE.2011.05.055.

Sui, J., Qi, S., van Erp, T. G. M., Bustillo, J., Jiang, R., Lin, D., et al. (2018). Multimodal neuromarkers in schizophrenia via cognition-guided MRI fusion. *Nature Communications, 9*(1), 3028. https://doi.org/10.1038/s41467-018-05432-w.

Trakadis, Y. J., Sardaar, S., Chen, A., Fulginiti, V., & Krishnan, A. (2019). Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics, 180*(2), 103–112. https://doi.org/10.1002/ajmg.b.32638.

Vieira, S., Pinaya, W. H. L., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience and Biobehavioral Reviews, 74*, 58–75. https://doi.org/10.1016/J.NEUBIOREV.2017.01.002.

Vu, T. D., Yang, H.-J., Nguyen, V. Q., Oh, A.-R., & Kim, M.-S. (2017). Multimodal learning using convolution neural network and Sparse Autoencoder. In *2017 IEEE international conference on big data and smart computing (BigComp)* (pp. 309–312). IEEE. https://doi.org/10.1109/BIGCOMP.2017.7881683.

Wang, H.-T., Smallwood, J., Mourao-Miranda, J., Xia, C. H., Satterthwaite, T. D., Bassett, D. S., et al. (2018). *Finding the needle in high-dimensional haystack: A tutorial on canonical correlation analysis*. Retrieved from http://arxiv.org/abs/1812.02598.

Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., & Marquand, A. F. (2015). From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience and Biobehavioral Reviews, 57*, 328–349. https://doi.org/10.1016/J.NEUBIOREV.2015.08.001.

Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience, 20*(3), 365–377. https://doi.org/10.1038/nn.4478.

Wu, M.-J., Passos, I. C., Bauer, I. E., Lavagnino, L., Cao, B., Zunta-Soares, G. B., et al. (2016). Individualized identification of euthymic bipolar disorder using the Cambridge Neuropsychological Test Automated Battery (CANTAB) and machine learning. *Journal of Affective Disorders, 192*, 219–225. https://doi.org/10.1016/J.JAD.2015.12.053.

Yang, H., Liu, J., Sui, J., Pearlson, G., & Calhoun, V. D. (2010). A hybrid machine learning method for fusing fMRI and genetic data: Combining both improves classification of schizophrenia. *Frontiers in Human Neuroscience, 4*, 192. https://doi.org/10.3389/fnhum.2010.00192.

Zhou, T., Thung, K.-H., Zhu, X., & Shen, D. (2019). Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Human Brain Mapping, 40*(3), 1001–1016. https://doi.org/10.1002/hbm.24428.