Received: 26 July 2017 | Revised: 15 December 2017 | Accepted: 9 January 2018

DOI: 10.1002/widm.1248

WILEY WIREs DATA MINING AND KNOWLEDGE DISCOVERY

OVERVIEW

Machine learning for bioinformatics and neuroimaging

Angela Serra[†] | Paola Galdi[†] | Roberto Tagliaferri

NeuRoNeLab, DISA-MIS, University of Salerno, Salerno, Italy

Correspondence
Roberto Tagliaferri, NeuRoNeLab, DISA-MIS, University of Salerno, Via Giovanni Paolo II, 84084 Fisciano, Salerno, Italy.
Email: rtagliaferri@unisa.it

Machine Learning (ML) is a well-known paradigm that refers to the ability of systems to learn a specific task from the data and aims to develop computer algorithms that improve with experience. It involves computational methodologies to address complex real-world problems and promises to enable computers to assist humans in the analysis of large, complex data sets. ML approaches have been widely applied to biomedical fields and a great body of research is devoted to this topic. The purpose of this article is to present the state-of-the art in ML applications to bioinformatics and neuroimaging and motivate research in new trend-setting directions. We show how ML techniques such as clustering, classification, embedding techniques and network-based approaches can be successfully employed to tackle various problems such as gene expression clustering, patient classification, brain networks analysis, and identification of biomarkers. We also present a short description of deep learning and multiview learning methodologies applied in these contexts. We discuss some representative methods to provide inspiring examples to illustrate how ML can be used to address these problems and how biomedical data can be characterized through ML. Challenges to be addressed and directions for future research are presented and an extensive bibliography is included.

This article is categorized under:
Application Areas > Health Care
Technologies > Computational Intelligence
Fundamental Concepts of Data and Knowledge > Motivation and Emergence of Data Mining
Fundamental Concepts of Data and Knowledge > Key Design Issues in Data Mining

KEYWORDS

bioinformatics, classification, clustering, deep learning, dimensionality reduction, feature selection, machine learning, multi-view learning, networks, neuroimaging

1 | INTRODUCTION

In the last decade, the amount of biomedical data gathered by the bioinformatics and neuroscience communities has grown exponentially. This large amount of data, coming in different forms such as genomes, gene expression data, gene or PPI networks in bioinformatics, or many modalities of structural and functional imaging in neuroscience, led to the need for efficient and effective computational tools to analyze and interpret these data.

[†] These authors contributed equally.

Bioinformatics is an interdisciplinary field in which new computational methods are developed to analyze biological data and to make biological discoveries (Luscombe, Greenbaum, Gerstein, et al., 2001). In genetics and genomics, bioinformatics tools were applied to the processes of sequencing and annotating genomes (Craig Venter et al., 2001). Many text mining applications were developed to analyze biological literature and organize biological data into easy to query ontologies (Cohen & Hersh, 2005). Bioinformatics also plays a central role in the understanding of gene and protein expression and regulation (Ansel, Rosenzweig, Zisman, Melamed, & Gesundheit, 2016; Hunt, 2005). At a systems biology level, it helps catalogue the biological pathways and analyze the interaction networks underlying specific biological mechanisms (Ideker, Galitski, & Hood, 2001; Kitano, 2002). Since omics data (i.e., gene expression, microRNA expression [miRNA], copy number variation [CNV], single nucleotide polymorphism [SNP] and PPI) are high dimensional and can be characterized by different temporal resolutions, advanced methodologies for their analysis and interpretation are required.

In neuroscience, neuroimaging techniques, such as computerized tomography (CT), positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and diffusion tensor imaging (DTI), are used to study brains in vivo and to understand the inner workings of the nervous system. One of the main research goal of this field is to analyze the human brain network, often referred to as human connectome, in order to understand its anatomical and functional organization. Gaining such an understanding is fundamental to support early diagnoses of neurological disorders and to improve treatments of these pathologies. Indeed, connectome data analysis has led to the discovery of biomarkers associated to several neurological conditions (Nir et al., 2012; Rudie et al., 2013; Sun, Chen, Collinson, Bezerianos, & Sim, 2015). Neuroimaging data are complex and high dimensional and come in a wide range of spatial and temporal resolutions; for these reasons, advanced analysis techniques are necessary to describe data derived from each imaging method.

Machine learning (ML) deals with the analysis and development of models and algorithms able to learn from data in order to perform predictive analysis. Here, learning is intended as the capability of ML models to adapt their configurations to optimally explain the observed data. Usually, the objective is to extract knowledge from a data collection in order to make predictions about new, previously unseen, patterns. ML can be roughly and nonexhaustively divided into supervised (Kotsiantis, Zaharakis, & Pintelas, 2007), unsupervised (Wong, Li, & Zhang, 2016), and semisupervised (Hajighorbani, Reza Hashemi, Minaei-Bidgoli, & Safari, 2016; Zhu, 2005) learning. In supervised learning, data are associated with an outcome variable of interest, and the objective is to infer a model that relates the variables of interest to the observations. On the other hand, when there are no outcome variables associated to data and the point of the analysis is to explore how data are geometrically and statistically organized, this setting is defined as unsupervised learning. In semisupervised learning both labeled and unlabeled data are used. Semisupervised learning can be applied both in combination with unsupervised and supervised learning, for example, in clustering tasks a few labeled samples can be used to define the initial structure of the clusters. In classification tasks, semisupervised learning can be used to improve the classifier performances: first labeled data can be used to train the classifier and obtain the labels of the unlabeled data. Then the new labeled data can be used to retrain the classifier on a higher number of samples.

ML techniques have been widely applied to solve these problems. In fact, a large body of research is devoted to biological and neuroimaging data mining, involving many tasks such as classification (Rathore, Habes, Iftikhar, Shacklett, & Davatzikos, 2017; Yang, Yang, Zhou, & Zomaya, 2010), clustering (Rui & Wunsch, 2010), network analysis (Greicius, Krasnow, Reiss, & Menon, 2003; Kitano, 2002), and dimensionality reduction (Saeys, Inza, & Larrañaga, 2007).

The objective of this work is to review the ML state-of-the art approaches and their application in the fields of bioinformatics and neuroimaging. We discuss some representative methods to provide inspiring examples and to motivate research in new trend-setting directions. We believe that this review will provide valuable insights and serve as a starting point for prospective investigators.

The article is organized as follows: in Section 2, a discussion about dimensionality reduction and feature selection in the biomedical field is provided, with particular focus on the differences between univariate and multivariate approaches with examples of applications for feature selection in biomarker prioritization, single nucleotide polymorphism analysis, detection of functional brain networks, and low-dimensional embedding of fMRI data. In Section 3, clustering techniques and their application are discussed. In particular, examples of clustering for the identification of co-expressed genes and for patient subtyping are reported. Moreover, applications of clustering for brain parcellation and feature extraction in fMRI data are discussed. In Section 4, examples of the most common supervised classification methods are described with their application in drug repositioning, mass spectrometry-based proteomics classification, patient classification from neuroimaging data, and multimodal parcellation of human cerebral cortex. In Section 5, network-based approaches in system biology and functional network modeling with fMRI data are discussed. In Section 6, a discussion on the differences between deep learning and classical shallow learning methods is reported. Examples of application of CNNs in RNA in situ hybridization (ISH), DNA- and RNA-binding protein and prediction of clinical neurodevelopmental outcomes from structural brain network are reported, together with an example of use of deep auto-encoders (AEs) for the diagnostic of the Alzheimer's disease. In Section 7, concluding remarks are provided. Additionally, three boxed sections were added to provide further information on advanced methods for biomedical data analysis: the first box describes the basic concepts of multiview learning; the second one presents some adaptations of standard methods taking into account the special spatial and temporal characteristics of bioinformatics and neuroimaging domains; and the third box introduces two of the most popular deep learning models.

2 | DIMENSIONALITY REDUCTION AND FEATURE SELECTION

The high dimensionality of biomedical data often requires the application of a preprocessing step aimed to decrease the size of the data set before performing further analyses. There are two main approaches to achieve this, that are dimensionality reduction techniques and feature selection. Dimensionality reduction methods involve the transformation of a high-dimensional data set into a simpler representation that still preserves most of the relevant information contained in the data. The most commonly applied method is principal component analysis (PCA), that consists in a linear transformation that projects the original data into a new space where the variable with the highest variance is projected into the first axis; the variable with the second highest variance is projected into the second axis, and so on. The reduction is obtained considering only the principal components, that is, a subset of the variables that account for most of the variability in the data. One limitation of PCA is that it is based on the assumption that data follow a Gaussian distribution, therefore it is unable to represent data distributed over more complex manifolds. While PCA is based on orthogonal transformation to obtain linearly uncorrelated features, Independent Component Analysis (ICA) (Hyvärinen & Oja, 2000), works by identifying statistically independent components in data. Other approaches are based on factor analysis, projection pursuit, regression, and topologically continuous maps (Carreira-Perpinán, 1997; Fodor, 2002). The main drawback of dimensionality reduction techniques is that inevitably some information is lost in the process, and this might hinder interpretability, especially in the case of noninvertible projections that do not allow to go back to the initial representation. When the problem at hand requires to preserve original features, feature selection methods may be preferable. The goal of feature selection is to express high-dimensional data with a low number of features to reveal significant underlying information. It is mainly used as a preprocessing step for other computational methodologies. Three different approaches are proposed in the literature: the univariate or multivariate filter methods and the multivariate wrapper and embedded methods (Saeys et al., 2007). In the filter methods, the features are ranked according to a predefined criterion, then a percentage of the top-ranked features is retained and the others are discarded. This approach is independent by the classifier. In the univariate approaches, each feature (e.g., a gene or a voxel) is evaluated and ranked individually. Examples of univariate filters are the *t*-test, the χ^2 test, or the Wilcoxon rank sum (Garcia-Chimeno, Garcia-Zapirain, Gomez-Beldarrain, Fernandez-Ruanova, & Garcia-Monco, 2017; Shuke, 2017). These approaches are fast and scalable, since the computational complexity is linear in the number of features, but they ignore the dependencies between features. To tackle this problem, multivariate filtering methodologies were proposed that evaluate the separation capability of groups of features taken together. These approaches are slower and less scalable than the univariate methodologies (the number of possible subsets grows exponentially with the number of features), and they are still independent from the classifier. Examples of multivariate feature ranking approaches are Markov blanket filters (Wang et al., 2017; Zhu, Ong, & Dash, 2007) and correlation and fast-correlation based methods (Heider, Genze, & Neumann, 2017). While filter techniques identify the best features independently from the model selection step, wrapper methods combine the model selection step with the feature subset search. In fact, the goodness of each group of features is evaluated by training and testing a specific classification model. In this way, the feature selection procedure is strongly related to the selected classifier and, compared to filtering methods, these are more computational expensive and have a higher risk of overfitting, but they can in general achieve higher accuracy, since they try to build the best possible model given the available data. Examples of wrapper selection approaches are greedy forward selection or backward elimination strategies (Hannah Immanuel & Jacob, 2017; Polat, Mehr, & Cetin, 2017). The embedded techniques for feature selection methods search for the optimal subset of features inside the classifier during its construction. This means that the search is performed in the combined space of feature subsets and hypotheses. Like the wrapper approaches, the embedded techniques are strongly linked to the classifier and then specific to the learning algorithm. Compared to the wrapper methods, they are less computationally intensive. Examples of these applications are the use of Random Forest (RF) internal measures, such as mean decrease accuracy and Gini index, or the feature selection based on support vector machine (SVM) weights. Both univariate and multivariate approaches have the common goal of finding the smallest set of features useful to correctly classify objects. Accuracy and stability are the two main requirements for feature selection methodologies. Most of the effort in the past has been spent in finding methods with high accuracy in order to increase the predictive power of the selected features.

2.1 | Feature selection for biomarker prioritization

Thinking about the problem of finding the biomarker for a disease, stability rises up as an important property of the algorithm because it should find out the same set of features across different runs. Fortino, Kinaret, Fyhrquist, Alenius, and Greco (2014)) proposed a wrapper feature selection method that combines fuzzy logic and RFs and is able to guarantee good

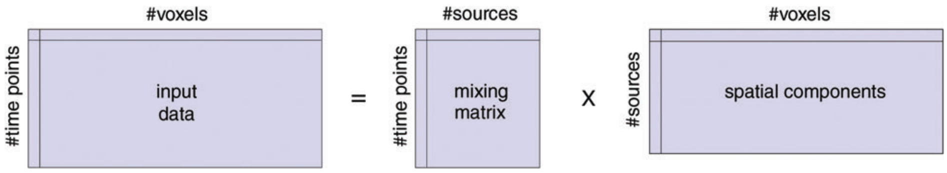


FIGURE 1 Independent component analysis on functional magnetic resonance imaging data: the input matrix consists of time series associated to brain voxels; the mixing matrix contains, for each time point, the relative contribution of each independent component to the global signal; the components matrix indicates the contribution of the single voxels to the components

performance and high stability. The algorithm consists of three parts: in the first step, called discretization, the gene expression data are transformed in fuzzy patterns (FPs). FPs provide information about the most relevant features of each category; in the second step, prior knowledge about the FPs is used to train a RF and classify data; in the third step, selected features are ranked by a permutation variable importance measure. The method was tested on several gene expression multiclass data sets and compared with other two RF-based feature selection methods: varSelRF (Díaz-Uriarte & De Andres, 2006) and Borda (Kursa, Rudnicki, et al., 2010). *F*- and *G*-scores were evaluated on 30 iterations to estimate accuracy. These measures are particularly appropriate for multiclass unbalanced problems (Yu et al., 2013). During the iterations, the significantly consistent features were selected. The final stability metric was defined as the ratio between the number of consistent features and the total number of selected features. The results show that their system has similar or better results compared to the other methods proposed in the literature.

2.2 | SNP analysis

SNPs are mutations of a single nucleotide position due to evolution and passed through hereditarily. They account for most of the genetic differences across individuals and are the basis of many disease–gene association studies. Since the number of SNPs in the genome is high (about 7 millions Kruglyak & Nickerson, 2001) it is important to identify the subset of relevant SNPs that is sufficiently informative to differentiate patients affected by different pathologies. Several computational methods for SNPs selection have been proposed based on different criteria. For example Charlon et al. (2016) proposed a PCA-base algorithm to identify the SNPs that better characterize systemic autoimmune diseases. By applying such transformation, the authors of the above paper were able to retain SNPs with the largest absolute projection values (i.e., loadings or contributions) to the 100 first principal components. For each principal component, they fitted a Gaussian mixture model with two mixture components. For each SNP, the probability to be assigned to each Gaussian model is computed, from which a classification uncertainty can be derived. Only the strong contributors have null uncertainty, therefore SNPs with a null classification uncertainty are selected. With this approach, the authors were able to identify SNPs with good classification capability and also to reduce the effect of the ancestry information that is one of the main source of genetic variations between individuals that often hide other variations that cause the diseases.

2.3 | ICA for detecting functional brain networks

The high-dimensional nature of functional neuroimaging data requires the application of methods to reduce the intrinsic complexity of functional data analysis. ICA is often applied to fMRI data in order to extract spatially independent sources of the signal and detect noise components (De Martino et al., 2007; Formisano et al., 2002; McKeown et al., 1997; McKeown, Hansen, & Sejnowsk, 2003).

Given a matrix of time series associated to brain voxels, ICA determines a number of independent sources whose combined contributions produce the fMRI signal (see Figure 1). We can distinguish between spatial and temporal ICA depending on whether we require the spatial patterns or the time courses to be independent, but the most commonly applied is the spatial ICA, and this is also due to the fact that there are many more voxels than time points. One of the most relevant applications of ICA in this context is the identification of resting state networks, that is, functionally connected regions of the brain that are detected when the brain is not involved in any particular task. Six main networks can be distinguished: the default mode network, the visual network, the fronto-parietal network, the sensori-motor network, the auditory network, and the self-referential network (Tedeschi & Esposito, 2012). These networks then become the input of further investigation, as is the case of the default mode network, that is the network that is active when the brain is awake and at rest, and several studies have found out that there are some difference in the activation of this network between healthy and nonhealthy subjects (Broyd et al., 2009).

2.4 | Low-dimensional embedding of fMRI data

In 2006, Shen, Wang, Liu, and Hu (2010)) presented a ML approach for the discriminative analysis of resting-state functional connectivity patterns of schizophrenia, that combines in a single framework feature selection technique, low-dimensional embedding and self-organizing clustering. In the proposed method, resting-state functional networks are treated as points distributed in a high-dimensional feature space, and the goal is to detect spatiotemporal patterns associated with schizophrenic symptoms that are assumed to lie on a low-dimensional manifold embedded in the feature space. In a first step of the analysis, starting from functional connectivity networks extracted from each subject, a correlation coefficient method is applied to extract the most discriminative features. Then, a dimensionality reduction step follows, which relies on locally linear embedding (Roweis & Saul, 2000), a nonlinear manifold learning technique. Finally, K-means clustering is applied in the low-dimensional space to identify two groups (patients vs. healthy controls) and the two resulting clusters are labeled as to maximize the classification rate. This approach attained good classification performance and generalization ability and is a prominent example of the potential ability of ML to support diagnosis and treatment evaluation of schizophrenia.

3 | CLUSTERING

Cluster analysis in an unsupervised learning technique that is able to identify structures from data without any previous knowledge on their distribution. It is an exploratory technique used to group a set of objects in such a way that patterns in the same group (called cluster) are more similar to each other than to those in other clusters. The notion of similarity degree is central to all the goals of cluster analysis. In fact, the clustering results strongly depend on the adopted similarity measure. The choice of the best similarity measure can only come from considerations about the subject matter.

So far, a large number of different clustering algorithms have been proposed. Here we review some of the most famous and most applied algorithms in the biomedical field and we report some examples of applications.

3.1 | Partitioning clustering

The aim of partitioning clustering is to obtain a partition of data where each point belongs to a unique cluster. Hartigan and Wong (1979) is one of the best-known clustering algorithm. Formally, let $X = \{x_1, \ldots, x_n\}$ be a set of N points in a multidimensional space and let K be an integer value, the K-means algorithm seeks to find a set of K vectors μ_k that minimize the within cluster sum of squares $WCSS = \sum_{h=1} k \sum_{x_i \in C_h} d(x_i, \mu_h)$ where C_h is the hth cluster and μ_h is the corresponding centroid. K-means works well in many practical problems especially when the resulting clusters are compact and with hyperspherical shape and it runs in approximately linear time. On the other side, its main drawbacks are that it can be easily trapped in local minima during the optimization process and it is sensitive to the starting initialization of the centroids. Indeed, it is a good practice to set a random seed before its execution in order ensure the reproducibility of the experiments. Moreover, since K-means uses the means as centroids of the clusters, it is sensitive to noise. Another issue with K-means is that it requires to fix the number of clusters, even though this is usually unknown in the data and has to be estimated using clustering analysis. This is a problem not only for K-means but also for many other partitioning clustering methods, even though some of them can estimate the most appropriate number of groups during their execution. Many efforts have been devoted to overcome the limitations of K-means clustering (Xu & Wunsch II, 2009), for example, the Partitioning around medoids algorithm (Kaufman & Rousseeuw, 1987) uses the median centrally located object in the cluster as medoids, in order to overcome the effect of the outliers on the cluster prototypes.

3.2 | Hierarchical clustering

Hierarchical clustering algorithms (Theodoridis & Koutroumbas, 2008) are the most commonly used methods to identify data structures in bioinformatics. They result in a hierarchical tree (called dendrogram) that represents a nested set of partitions. Cutting a dendrogram at a particular level produces a partition into K disjoint clusters. Depending on the methodology used to build the hierarchy, the clustering results may vary. For example, the single linkage merges clusters that have the nearest distance between two pairs of patterns, one for each cluster. It tends to generate clusters with elongated chain structures. It is effective for well-separated clusters. On the other hand, the complete linkage merges clusters based on the furthest distance between pairs of points. It is effective for small and compact clusters. The centroid linkage computes the squared Euclidean distance between cluster centroids. It assumes that data can be represented in the Euclidean space. Hierarchical clustering has been widely applied in bioinformatics and neuroimaging applications. For example, Sørlie et al. (2001) used this method to identify the gene patterns that distinguish breast carcinoma tumor subclasses. Examples of applications of hierarchical clustering in gene expression analysis are discussed in Section 3.1 and in Hand and Heard (2005). Hierarchical clustering was also applied in neuroimaging field to measure connectivity in fMRI resting-state data (Cordes, Haughton, Carew, Arfanakis, & Maravilla, 2002) and to build brain atlas (Blumensath et al., 2013).

3.3 | Mixture Models

In the clustering mixture model (McLachlan & Basford, 1988), each group in the population is assumed to be represented by a different probability distribution. The population is modeled by a finite mixture distribution of the form $p(x) = \sum_{i=1}^{K} \pi_i p(\mathbf{x}, \theta_i)$, where π_i are the mixing proportions, $\left(\sum_{i=1}^{K} \pi_i = 1\right)$ and $p(\mathbf{x},\ \theta_i)$ is an n-dimensional probability function depending on the parameter vector θ_i. The solution depends on three sets of parameters to be estimated: the values of π_i, the components of the vectors θ_i and the number of clusters in the population (K). A popular method falling in this category is the Expectation-maximization (EM) algorithm. For example, in bioinformatics, it has been used to perform molecular characterization (Lin, 2016), gene co-expression clustering (Xianxue, Guoxian, & Wang, 2017), discovering molecular pathways in PPI networks (Segal, Wang, & Koller, 2003), and so on. In neuroinformatics it has been applied to cluster and quantify fiber tracts data (Mahnaz Maddah, Grimson, Warfield, & Wells, 2008) and for time series clustering (Rani & Sikka, 2012).

3.4 | Density-based clustering

Density-based clustering supposes that the clusters are represented by dense regions of points in the data space, separated by regions of lower density. The most famous density-based clustering algorithm is the DBSCAN, proposed by Ester et al. (1996). It is based on the density-reachability model that connects points within a certain distance. It depends on two parameters: the (ϵ) distance threshold and the minimum number of objects to form a cluster *minPts.* The algorithm first finds the neighbors of each point that are at a distance less than ϵ, and identifies the core points with more than *minPts* neighbors. Then, it finds the connected components of the core points on the neighbor graph, ignoring all the noncore points. Finally, it assigns each noncore point to a nearby cluster if the cluster is an $\epsilon -$ neighbor, otherwise it assigns the point to noise. DBSCAN has many advantages compared to partitive clustering: it does not require to specify the number of clusters to be retrieved, it can find clusters of different shapes (not only Gaussian), and it is robust to outliers. Indeed, it is widely applied both in the bioinformatic and neuroscience fields (Galdi et al., 2017; Grubbs et al., 2017; Pennacchietti et al., 2017; Poole, Leinonen, Shmulevich, Knijnenburg, & Bernard, 2017; Tench, Tanasescu, Constantinescu, Auer, & Cottam, 2017). On the other side, the quality of the DBSCAN strongly depends on the similarity measure and it is not able to cluster data sets with large differences in densities, since the $minPts - \epsilon$ combination cannot be chosen appropriately for all clusters.

3.5 | Spectral clustering

The spectral clustering algorithms make use of the spectrum (eigenvalues and eigenvectors) of the similarity matrix to perform dimensionality reduction and then cluster the objects in a lower-dimensional space. Indeed, starting from a similarity matrix A, where $A_{i,j}$ represents the similarity between the samples i and j, the strategy is to compute the associated Laplacian matrix and then to apply the clustering method only to its relevant eigenvectors. A common algorithm is the one proposed by Meila and Shi (2001) where the clustering is performed on the eigenvectors associated to the highest eigenvalues of the random walk normalized Laplacian matrix $P = D^{-1}\ A$. Spectral clustering counts many applications in bioinformatic domains (Higham, Kalna, & Kibble, 2007) such as the construction of libraries of protein fragments (Elhefnawy, Li, Wang, & Li, 2017), multiview clustering of patients subtyping (Hobbs et al., 2017; Zhang, Xiaohua, & Jiang, 2017), study of DNA methylation (Sheffield et al., 2017) and so on. In the neuroimaging field, it has been recently applied to identify biomarkers for autism spectrum disease from resting state imaging (Abraham et al., 2017). Moreover, it has been used to identify time varying networks for functional magnetic resonance imaging data (Cribben & Yi, 2017). One disadvantage of these methods is their computational complexity. For this reason, some optimized implementations have been proposed in the literature, such as fast approximations (Yan, Huang, & Jordan, 2009) or parallel versions (Song, Chen, Bai, Lin, & Chang, 2008) of the spectral clustering methodology.

3.6 | Affinity propagation

The affinity propagation method is based on the concept of message passing between data points (Frey & Dueck, 2007). It takes as input pairwise similarities between data points and finds out the most representative items of the data set to build clusters around them. It operates by simultaneously considering all data points as candidate exemplars and exchanging real-valued messages between data points until a good set of exemplars and clusters emerges. The message passing procedure is based upon two kinds of messages. The responsibility, sent from data point i to the candidate exemplar point k, that reflects the accumulated evidence for how well-suited point k is to serve as the exemplar for point i, taking into account other potential exemplars for point i. The availability, sent from the candidate exemplar point k to point i that reflects the accumulated evidence about how appropriate it would be for point i to choose point k as its exemplar, taking into account the support from other points for which point k could be an exemplar. Affinity propagation method does not require to select the number of clusters and compared to others clustering algorithm it usually gives in output more clusters with uneven cluster size. However, even if the number of clusters is not required as input, affinity propagation requires to set a parameter (preferences) for each point: points with larger values of preferences are more likely to be chosen as exemplars. The number of exemplars, that is, of clusters, is influenced by the input preference values. If not differently specified, these values are initialized as the median of the input similarities. This method has been widely applied to solve computational biology and neuroimaging tasks (Fonseca et al., 2017; Hong et al., 2017; Meng et al., 2017; Salman, Du, & Calhoun, 2017).

3.7 | Fuzzy clustering

In contrast with partitioning clustering, in fuzzy cluster methodologies data points can belong to more than a single cluster. Indeed, a membership score is assigned to each data point for each cluster. Thus, points on the edge of a cluster, with lower membership grades, may belong in the cluster to a lesser degree than points in the center of cluster. The most common fuzzy clustering method is the fuzzy c-means, that works quite similarly to K-means, apart from the addition of membership values in the objective function. Fuzzy clustering has been widely applied in gene expression and gene coexpression network clustering (Alok, Saha, & Ekbal, 2008; Jiang, Li, Min, Qi, & Rao, 2017; Wang & Chen, 2017). It has been also applied in neuroimaging for tumor segmentation tasks (Manocha, Bhasme, Gupta, Panigrahi, & Gandhi, 2017), lesion detection (Kinani et al., 2017), and automatic brain parcellation (Vercelli et al., 2016).

3.8 | Biclustering

Biclustering was first proposed by Cheng and Church (2000), and derives its name from the fact that clustering is performed simultaneously on both the features (genes or voxels) and the samples in the experiment. This methodology stems from the assumption that in a biological system only a subset of features is involved with a specific biological process, which becomes active only under some experimental conditions. In this case, the inclusion of all features in sample clustering or all samples in features clustering not only increases the computational burden, but could impair the clustering performance due to the effect of these unrelated features or samples, which are treated as noise. The complexity of the biclustering problem has been shown to be NP-complete (Cheng & Church, 2000), hence many heuristics have been proposed to solve the problem based on different principles such as divide et conquer (block clustering, Cheng & Church, 2000) greedy iterative search, (FLOC, Wang, Wang, Yang, & Yu, 2002; xMOTIF, Segal, Taskar, Gasch, Friedman, & Koller, 2001; OPSM, Lazzeroni & Owen, 2002) exhaustive bicluster enumeration [SAMBA, Divina & Aguilar-Ruiz, 2006). It has been widely applied both in bioinformatics (Greene, Lin, Wang, Ye, & Wittenberg, 2017; Lu & Liu, 2017; Rengeswaran, Mathaiyan, & Kandasamy, 2017; Wang et al., 2017) and neuroscience (Gupta et al., 2017).

3.9 | Subspace methods

Since biomedical data are high-dimensional their analysis usually arises some problems such as data visualization. Moreover, as the number of dimensions grows the concept of distance becomes less precise and the complete enumeration of all subspaces becomes intractable. Furthermore, given a large number of attributes, it is likely that some attributes are correlated; hence, clusters might exist in arbitrarily oriented affine subspaces. Subspace clustering is an extension of traditional clustering that localizes the search for relevant dimensions allowing to find clusters that exist in multiple, possibly overlapping subspaces. The main computational issue with subspace clustering is that giving a d dimensional space, the number of possible subspaces in which clustering can be performed is 2^d. Hence, some heuristic algorithms have been developed that use the downward-closure property to build higher-dimensional subspaces by combining only lower-dimensional ones that already contain clusters. Examples of these algorithms are the CLIQUE (Agrawal, Gehrke, Gunopulos, & Raghavan, 2005) and SUBCLU (Kailing, Kriegel, & Kröger, 2004). SUBCLUE was proven to be more effective than CLIQUE in the identification of groups of co-expressed genes (Kailing et al., 2004).

3.10 | Projective methods

Projective clustering is a class of methods in which the input consists of high-dimensional data, and the goal is to discover those subsets of the input items that are strongly correlated in subspaces of the original space (Xianxue et al., 2017). Each subset of correlated points, together with its associated subspace, defines a projective cluster. Thus, although all cluster points are close to each other when projected on the associated subspace, they may be spread out in the full-dimensional space. This makes projective clustering algorithms particularly useful when mining or indexing data sets for which full-dimensional clustering is inadequate (as is the case for most high-dimensional data sets). Moreover, such algorithms compute projective clusters that exist in different subspaces, making them more general than global dimensionality-reduction techniques. Recently, ensemble of clustering projective methods were used to cluster cancer gene expression data sets obtaining more stable and robust to noise solutions (Xianxue et al., 2017).

3.11 | Consensus clustering

Different clustering algorithms give different solutions and also the same algorithm can have different behaviors depending on its input parameters. A class of methods called consensus clustering has been proposed to find a unique solution across the different clustering results on the same data set coming from different clustering methods or from different runs of the same algorithm (Vega-Pons & Ruiz-Shulcloper, 2011). One of the most commonly encountered consensus clustering approaches is that proposed by Monti, Tamayo, Mesirov, and Golub (2003) that generates multiple perturbed versions of the original data by computing random subsamples of the input data. Then, a consensus (or co-association) matrix $M \in R^{n \times n}$ (where n is the number of data points) is built, where each entry $M(i, j)$ is the count of how many times items i and j were assigned to the same cluster across different partitions, normalized by the number of times that the two objects were present in the same subsample. The final clustering can be obtained using the consensus matrix as a similarity matrix to be given as input to a hierarchical clustering algorithm. The main advantage of consensus clustering is that it gives stable and reliable solutions. However, one drawback is that diversity alone cannot guarantee the quality of the results, especially if we try to merge poor or incompatible clusterings. Moreover, compared to the use of one single clustering technique, consensus clustering is computationally more expensive. Consensus clustering has been applied in bioinformatics in patient subtyping (Ringner & Staaf, 2017; Wang et al., 2017), drug repositioning (Sadacca et al., 2017), and protein binding (Flock et al., 2017). In neuroscience, it has been applied to find consensus between clusterings of brain connectivity matrices (Liu, Abu-Jamous, Brattico, & Nandi, 2017; Rasero et al., 2017) and as a step of feature extraction in a preprocessing pipeline to identify stable groups of voxels in resting state fMRI data (Galdi et al., 2017).

3.12 | Multiview clustering

In recent years, multiple experiments have been made available for the same sets of samples. For example, in bioinformatics, analyses can be based on multiple experiments investigating different facets of the same phenomena, such as gene expression, miRNA expression, PPI, genome wide association and so on, in order to capture information regarding different aspects of biological systems. In the same way, neuroscience data analysis can benefit from different imaging modalities that allow to study different features of the nervous system (e.g., structural vs. functional organization). Compared to the limited perspective offered by single-view analyses, the integration of multiple views can provide a deeper understanding of the underlying principles governing complex systems. To obtain a better understanding of these complex phenomena, by integrating different views of the data, many multiview clustering algorithms have been proposed (see Section 6.4 for details on multiview learning techniques). Some examples are the methods based on matrix factorization that integrate clustering solutions obtained on each single view (Zong, Zhang, Zhao, Yu, & Zhao, 2017). Other approaches use modifications of the classical K-means clustering algorithm (Chen, Xiaofei, Huang, & Ye, 2013; Xu, Han, Nie, & Li, 2017). Other methods work on the integrative analysis of networks built on each view by using an iterative optimization analysis based on the local neighborhood and then applying spectral clustering on the final integrated matrix (Bo et al., 2014).

3.13 | Clustering evaluation

Even though a wide variety of clustering algorithms exists and many efforts have been performed to solve all the problems related to clustering, the main difficulty related to these methodologies concerns the number of clusters (K). Indeed, in several approaches it must be specified before estimating the remaining parameters, but in real problems it is usually unknown. Moreover, depending on the specific choice of the preprocessing method, the distance measure, the cluster algorithm and other parameters, different runs of the clustering procedure will produce different results. Therefore, it is very important to validate the relevance of the clusters (Handl, Knowles, & Kell, 2005). Thus, many metrics for clustering validation have been proposed in the literature, that measures cluster properties or compare the results with prior knowledge. For example, the Davies-Bouldin index (Davies & Bouldin, 1979), the Dunn index (Dunn, 1973), or the silhouette index (Bandyopadhyay & Saha, 2008) can be used to estimate the optimal value of K. On the other side, the Adjusted Rand index (Steinley, 2004), the Jaccard index (Real & Vargas, 1996), the Fowlkes and Mallows index (Fowlkes & Mallows, 1983), the F-measure or the Normalized Mutual Information (NMI) index can be used to measure the concordance between clustering results and known sample groupings.

3.14 | Clustering for the identification of coexpressed genes

The most common example of clustering in bioinformatics is its use in grouping genes in expression data. Gene expression is the process through which the information coded in the genes is converted into functional structures operating in the cell. It provides the evidence that a gene has been activated (Luscombe et al., 2001). This activation is measured, for example, in microarray or in next-generation sequencing (NGS) experiments. In microarray essays, the expression value for thousands of genes in a set of samples is obtained (Quackenbush, 2001), while in NGS methodology a scanning of the whole genome is performed, allowing to investigate known transcripts and to explore new ones (Wang, Gerstein, & Snyder, 2009). From this kind of data, information related to which are the coexpressed genes in different samples can be extracted by using clustering techniques. This is an example of clustering application where genes with similar expression level across all samples are grouped into the same cluster.

A rich literature related of classical clustering algorithm (i.e., K-means, Roweis & Saul, 2000; self-organizing maps, SOM, Kohonen, 2012; hierarchical clustering Theodoridis & Koutroumbas, 2008; and EM, McLachlan & Basford, 1988) adapted or directly applied to gene expression data has been produced. Moreover, new algorithms have been developed specifically for gene expression data such as CLICK (Sharan & Shamir, 2000), CAST (Ben-Dor, Shamir, & Yakhini, 1999), and DHC (Jiang, Pei, & Zhang, 2003). Jiang, Tang, and Zhang (2004) wrote a comprehensive and critical overview of these methods and their application to gene expression data.

Starting from the work of Eisen, Spellman, Brown, and Botstein (1998), hierarchical clustering has been widely applied in gene expression clustering. It does not require a predefined number of clusters to be selected. Since it computes a complete hierarchy of data, represented as a dendrogram, it is useful for visualization purposes (Alizadeh et al., 2000). A flat partition in clusters can be determined afterward by cutting the dendrogram at a specific height. The choice of the height can be arbitrary and the results can change based on its value. Another approach, the Pvclust (Suzuki & Shimodaira, 2006), has been proposed to solve this problem some years ago. It is a variant of the classical hierarchical clustering that is able to assess the uncertainty in the analysis. For each cluster it evaluates a *p* value that indicates how strong the cluster is supported by data. Pvclust is a freely available R package and it has been widely applied in many bioinformatics applications (Borg et al., 2017; Ellebedy et al., 2016; Esnault et al., 2017).

Also partitive clustering, such as K-means, has been widely applied (Galdi, Napolitano, & Tagliaferri, 2014; Pal, Ray, & Ganivada, 2017; Rennert et al., 2016). The main advantage of K-means is that it is simple and fast (Jiang et al., 2004). But, unlike the hierarchical clustering, K-means requires to specify the number of clusters. This is one of the main drawbacks, since the number of gene clusters is usually unknown in advance. To identify the optimal number of clusters, K-means is usually run for different values of *k* and the clustering results are then compared (Pham, Dimov, & Nguyen, 2005).

Each clustering algorithm, applied on different data sets, can have different performance and there is no absolute winner among the algorithms described in this section. For example, if the data set contains few outliers and the number of clusters is known, K-means or SOM can outperform other approaches. On the other hand, for a gene expression data set with a high level of noise and no prior knowledge on the number of clusters, CAST or CLICK may be a better choice.

Once clusters are obtained, independently from the algorithm applied, they need to be validated: this is the process through which the quality and reliability of clusters are assessed. The clustering validation can be done in terms of homogeneity, that is, when objects in the same clusters are closer to each other than those in different clusters (Sharan & Shamir, 2000). For example, a manner to perform this task is to assess the cluster coherence, by testing the robustness of a clustering result with that obtained with the addition of noise. Furthermore, gene expression clustering can be validated from a biological point of view. For example, (Tavazoie, Hughes, Campbell, Cho, and Church (1999) created a mapping from the genes in each resulting cluster into 199 known functional categories. For each cluster, the *p* values were calculated to measure the functional category enrichment.

3.15 | Clustering for the identification of patient subtypes

Many diseases—for example, cancer, neuro-psychiatric and autoimmune disorders—are difficult to treat because of the remarkable degree of variation among affected individuals (Saria & Goldenberg, 2015). Precision medicine (Hood & Friend, 2011) tries to solve this problem by individualizing the practice of medicine. It considers individual variability in genes, lifestyle, and environment with the goal of predicting disease progression and transitions between disease stages, and targeting the most appropriate medical treatments (Mirnezami, Nicholson, & Darzi, 2012).

A central role in precision medicine is played by patient subtyping, that is the task of identifying subpopulations of similar patients that can lead to more accurate diagnostic and treatment strategies. Identifying disease subtypes can help not only the science of medicine, but also the practice. In fact, from a clinical point of view, refining the prognosis for similar individuals can reduce the uncertainty in the expected outcome of a treatment on each individual.

BOX 1

MULTIVIEW LEARNING

Multiview learning is the branch of ML concerning with the analysis of multimodal data, that is, a set of items represented by different sets of features extracted from multiple data sources. The fast spread of this learning technique is motivated by the continuing increase of real applications based on multiview data. For example, both in bioinformatics and in neuroimaging, multiple experiments can be available for a set of samples (e.g., they can consist of images, signals, or text related to the same patients). Depending on the nature of data and on the statistical problem to address, the integration of heterogeneous data can be performed at different levels: early, intermediate, and late (see Figure 2). Early integration consists, in fact, in concatenating all the variables from the multiple views to obtain a single feature space, but without changing the nature or general format of data. Intermediate integration transforms each data view in a common feature space. For example, in classification problems every view can be transformed into a similarity matrix (or kernel) and these matrices can then be combined to obtain more accurate results. Finally, in the late integration approach, a distinct analysis work-flow is carried out separately for each view and only the results are integrated.

A number of data integration approaches for patient subgroup discovery were recently proposed, based on supervised classification, unsupervised clustering or biclustering (Liu, Dong, & Liu, 2016; Planey & Gevaert, 2016; Roger Higdon et al., 2015; Taskesen et al., 2016). To improve the model accuracy for patient stratification, other omics data types can be used, such as miRNA expression, methylation or copy number alterations, in addition to gene expression. For example, somatic copy number alterations provide good biomarkers for cancer subtype classification (Nielsen et al., 2008). Data integration approaches to efficiently identify subtypes among existing samples have recently gained attention. The main idea is to identify groups of samples that share relevant molecular characteristics.

For example, SNF (Bo et al., 2014) is an intermediate integration network fusion methodology able to integrate multiple genomic data (e.g., mRNA expression, DNA methylation, and miRNA expression data) to identify relevant patient subtypes (see Box 1 for details on multiview learning). The method first constructs a patient similarity network for each view. Then, it iteratively updates the network with the information coming from other networks in order to make them more similar at each step. At the end, this iterative process converges to a final fused network. The authors tested the method combining mRNA expression, miRNA expression, and DNA methylation from five cancer data sets. They showed that the similarity networks of each view have different characteristics related to patient similarity while the fused network gives a clearer picture of the patient clusters. They compared the proposed methodology with iClust (Shen et al., 2012) and the clustering on concatenated views. Results were evaluated with the silhouette score for clustering coherence, Cox log-rank test p value for survival analysis for each subtype and the running time of the algorithms. SNF outperformed single view data analysis and it was able to identify cancer subtypes validated by survival analysis.

Serra et al. (2015) proposed a late integration methodology (see Box 1) for identifying patient subtypes in cancer data sets called MVDA. The approach consists of four main steps (see Figure 3): the first is the prototype extraction, where the features are clustered in order to reduce the data dimension; the second is prototype ranking, where the prototypes are ranked based on their class separability scores; the third is a single-view clustering step on each view; the last one is the integration of the single-view clustering results with a matrix factorization approach. The approach was validated on six different cancer multiview data sets downloaded from The Cancer Genome Atlas (TCGA), the Memorial Sloan-Kettering Cancer Center, and from NCBI GEO. Views in the data sets include gene expression, miRNA expression, RNASeq, miRNASeq, protein expression, CNV, and clinical data. The goodness of clustering was evaluated with two criteria: clustering purity with respect to class labels and NMI between clustering assignment and class labels. MVDA was compared with classical single-view clustering algorithms like k-means, Partition around medoids and Hierarchical clustering with Ward's criterion. Clusterings coming from the integration of multiple views reached a better performance compared to the single-view results. Moreover, the method was compared with TW-k-means and SNF: MVDA performed similarly or better than the other two integration methodologies.

3.16 | Clustering for brain parcellation

The most common application of clustering to neuroimaging data is brain parcellation: the brain is divided into a certain number of nonoverlapping regions (called parcels) according to a given criterion of intraregion homogeneity. Different imaging modalities provide different neurobiological information that can be used to define brain regions, for example, gyrosulcal anatomy from structural MRI data; anatomical connectivity from diffusion imaging or functional connectivity from resting-state functional magnetic resonance images. Brain parcellations are usually adopted because the native resolution of brain images is too high compared to the structures of interest (Thirion, Varoquaux, Dohmatob, & Poline, 2014); for

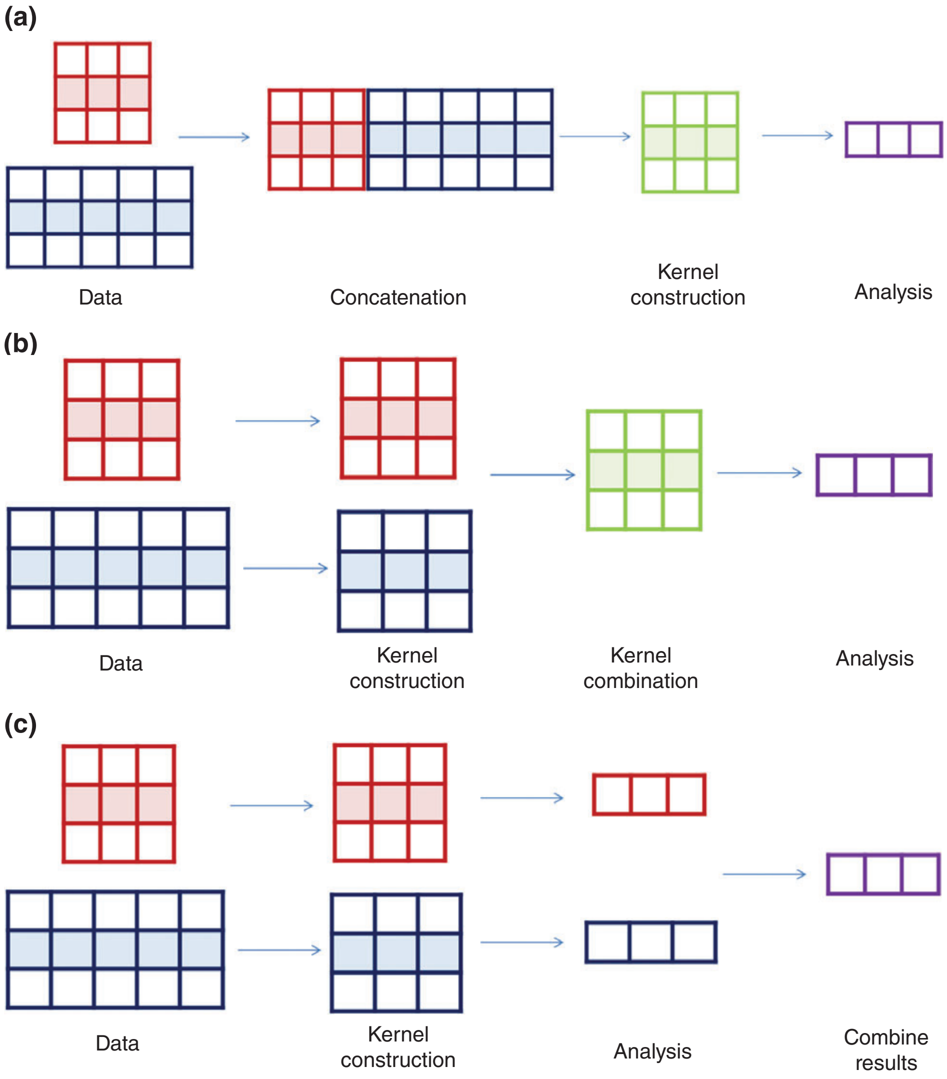
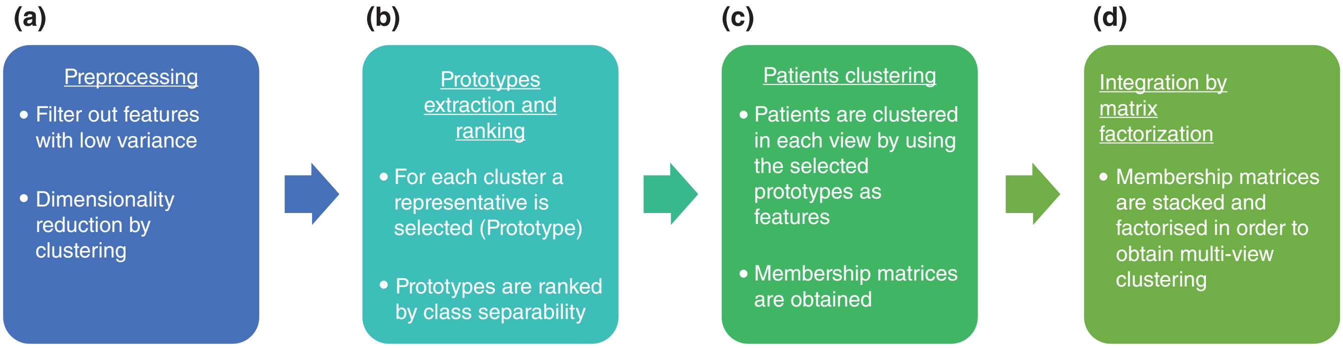


FIGURE 2 Data integration stages as proposed by Pavlidis, Weston, Cai, and Grundy (2001). They proposed a Support Vector Machine (SVM) kernel function in order to integrate microarray data. In early integration methodologies, SVMs are trained with a kernel obtained from the concatenation of all the views in the data set (a). In intermediate integration, first a kernel is obtained for each view, and then the combined kernel is used to train the SVM (b). In the late integration methodology a single SVM is trained on a single kernel for each view and then the final results are combined (c)

example, in functional imaging, a task-related activation may span across multiple voxels.[1] Moreover, working at the level of brain regions instead of single voxels is especially useful in cohort studies where the goal is obtaining averaged signals in a group of subjects. In this context, the application of clustering can decrease significantly data dimensionality allowing for the application of more sophisticated analysis techniques.

The advantage of data-driven parcellations obtained with clustering over anatomically defined parcellations is that they can provide a better model for the signal, while a predefined atlas might not fit data well, since even when data are projected in a standard space, there is still a nonnegligible intersubject variability.


(a)

Preprocessing

- Filter out features with low variance
- Dimensionality reduction by clustering

(b)

Prototypes extraction and ranking

- For each cluster a representative is selected (Prototype)
- Prototypes are ranked by class separability

(c)

Patients clustering

- Patients are clustered in each view by using the selected prototypes as features
- Membership matrices are obtained

(d)

Integration by matrix factorization

- Membership matrices are stacked and factorised in order to obtain multi-view clustering

FIGURE 3 MVDA integration methodology. The approach is composed of four steps. As a first step, the features are clustered and replaced by their prototypes, in order to reduce the input dimension (part a of the figure). Second, the prototypes are ranked by the patient class separability and the most significant ones are selected (part b of the figure). Third, the patients are clustered and the membership matrices are obtained (part c of the figure). Fourth, a late integration approach is used to integrate clustering results (part d of the figure)

Commonly adopted clustering methods for brain parcellation are mixture models (Golland, Golland, & Malach, 2007; Lashkari et al., 2012; Tucholka et al., 2008), K-means clustering (Flandin et al., 2002; Kahnt, Chang, Park, Heinzle, & Haynes, 2012; Thomas Yeo et al., 2011), hierarchical clustering (Eickhoff et al., 2011; Michel et al., 2012; Orban et al., 2014), and spectral clustering (Chen et al., 2013; Thirion et al., 2006). A spatial constraint is often added to the above models to obtain spatially contiguous regions (Cameron Craddock, Andrew James, Holtzheimer, Xiaoping, & Mayberg, 2012).

The choice of the number of clusters is tightly connected to the phenomena under investigation, but general rules to assess the quality of a parcellation can be derived considering the clustering stability and reproducibility. In functional parcellation, empirical studies have shown that the minimum number of parcels needed to guarantee reproducibility across subject is about 200 (Shen, Papademetris, & Todd Constable, 2010; Thirion et al., 2014).

3.17 | Clustering-based feature extraction in fMRI data analysis

Functional neuroimaging data describe brain activity in some form, and consist in volumetric images of the brain acquired over time (see Box 2). In fMRI data analysis, clustering can be applied to raw time series data in order to detect regions that show similar activation patterns (Goutte, Toft, Rostrup, Nielsen, & Hansen, 1999). However, the low signal-to-noise ratio that characterizes this type of data, together with the increasing spatial and temporal resolution of available data sets, make this approach impractical.

An alternative method is to use clustering to identify structures in data, after the raw time series have been preprocessed (Thirion & Faugeras, 2004). An example in this sense is the application of clustering techniques on spatial maps derived from ICA (see Section 2.3) on resting-state fMRI data (Galdi et al., 2017). In this approach, voxels are clustered using Pearson's correlation coefficient as a similarity measure. To obtain a more stable set of clusters, clustering solutions are enhanced through consensus techniques, that is, different partitions are combined into a final clustering in order to improve the quality of individual data clustering. At the end of this process, a representative feature (e.g., the mean) is extracted from each cluster, reducing the data dimensionality to the number of clusters that is chosen in order to guarantee sufficiently big groups of voxels that can be easily mapped into the anatomical brain regions (Figure 4). The extracted features are then used to train a model to perform classification tasks and consequently identifying the regions of the brain which are relevant to discriminate among different classes of subjects.

BOX 2

SPATIAL AND TEMPORAL CONSTRAINTS IN BIOMEDICAL DATA

Biomedical data such as gene expression data and fMRI data have intrinsic spatial and temporal characteristics. For example, in fMRI data analysis the primary goal is to understand the neural activity over time, the functional connectivity between neurons and regions and their relationship with stimuli. fMRI data may be viewed as a multivariate time series formed by a single time series for each voxel. Many efforts have been made to incorporate temporal and spatial correlations in the statistical analysis of time series since classical approaches are unable to take into account the spatial and temporal dependencies in the data. For example, several approaches have been proposed to compute brain parcellation by identifying groups of voxels that are spatially contiguous and functionally similar (Blumensath et al., 2013; Cameron Craddock et al., 2012). Moreover, classical clustering algorithms such as K-means and hierarchical clustering have been applied with ad-hoc similarity measures for time series such as the cross-correlation function (Goutte et al., 1999). In the same way as fMRI data, also omic data can be measured over time in order to characterize the dynamics of biological processes. The most common experiments are performed in gene expression data analysis where the expression for the same gene is evaluated at different time points. Even if classical ML methods have been applied to the analysis of gene expression time series, they assume independence between the input features and do not take into account the intrinsic temporal correlation in the data (Bar-Joseph, Gitter, & Simon, 2012). Hence, specific algorithms have been proposed such as the clustering algorithm CAGED (Ramoni, Sebastiani, & Kohane, 2002) that groups genes based on their trajectories, or Hidden Markov models that group genes with respect to their transcriptional trends (Schliep, Schönhuth, & Steinhoff, 2003). Time series analysis also finds applications in clinical tasks where the change in gene expression values over time is used to monitor patients' responses to disease treatments (Huang et al., 2011). Using temporal information to predict the outcome on the basis of dynamic expressions instead of static expressions was proved to obtain better performances (Lin, Kaminski, & Bar-Joseph, 2008).

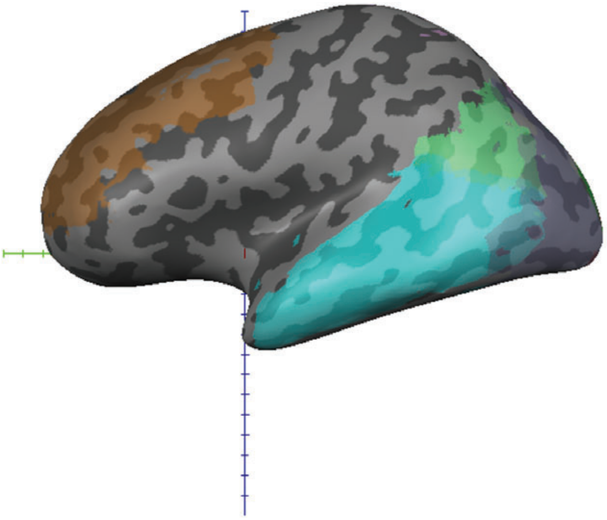


FIGURE 4 One of the advantage of clustering-based feature selection techniques is that the anatomical information associated to features is preserved, therefore at the end of the analysis region of interest can be easily visualized

4 | CLASSIFICATION

In ML, classification is part of the so-called supervised learning methodologies. Supervised learning is defined as the problem of estimating a functional relation between the features x_i of each sample of the data set and the outcome variable y_i:

$$y_i \approx f(x_i)$$

Depending on the type of the outcome variable, supervised learning divides into two subcategories: if the outcome variable is categorical, that is, can only assume a finite set of discrete values, the problem is referred to as classification; when the outcome variable can assume values in a continuous range (e.g., the level of glucose in blood), the problem is referred to as regression. Here, the focus will be only on the description of classification models. We talk of binary classification when the outcome variable can only assume two values, for example, -1 (healthy control) or 1 (Parkinson's disease), and of multi-class classification when there are more than two possible outcomes, for example, in patient subtyping. One-class classification is a special case where a model is trained to recognize a single class of objects (Khan & Madden, 2009), as in the identification of genes related to diseases (Yang, Li, Mei, Kwoh, & Ng, 2012).

Predictive models can belong to two families, namely parametric and nonparametric models. Parametric models assume that the function to be estimated belongs to a family described by a finite set of parameters denoted by $\boldsymbol{w}$. In this case, learning corresponds to the estimation of parameters such that the estimated function optimally describes the data. Nonparametric models, on the other hand, do not limit the relation between the input features of the data set and the outcome variable to a particular function family. The complexity of nonparametric models is automatically tuned during the data training step. In the following we review some of the most popular models.

4.1 | Support vector machine

Linear SVM is one of the most used among nonparametric models. It assumes that the outcome variable is linearly related to the corresponding input features

$$y \approx f_{\boldsymbol{w}}(\boldsymbol{x}_i) = \boldsymbol{w}^T \boldsymbol{x}_i.$$

The dependence of f on the vector $\boldsymbol{w}$ is made explicit, where $\boldsymbol{w}$ corresponds to the parameters to be learned. The assumption of linearity implies that, geometrically, a hyper-plane can separate the observations of a class from the observation of the other class. Specifically, a hyper-plane is uniquely defined by an orthogonal vector corresponding to the parameters $\boldsymbol{w}$ to be estimated. If the samples are actually separable by a hyper-plane, without errors, then there must exist at least a choice of $\boldsymbol{w}$ such that for each class $\hat{y} = f_{\boldsymbol{w}}(\boldsymbol{x}_i)$ it will equal the corresponding outcome variable y_i for each sample of the data set. Among the possible solutions, the SVM model finds the separating hyperplane that maximizes the margin (see Figure 5), defined as the smallest distance between any sample and the separating hyperplane, by solving a quadratic programming problem (Cortes & Vapnik, 1995).

SVMs can be used also as a nonlinear nonparametric classifier. In fact, instead of computing a linear function f, data can be transformed by applying a nonlinear kernel function. Kernel functions enable the classifier to operate in a high-dimensional, implicit feature space obtained by computing the inner products between the images of all pairs of data in the feature space. This approach is called the "kernel trick" (Aizerman, 1964).

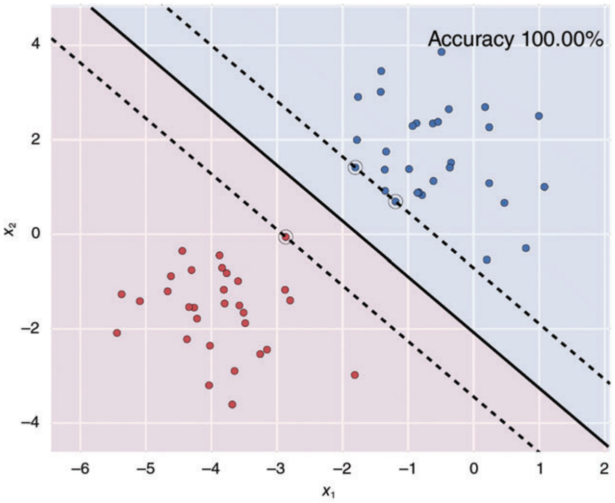


FIGURE 5 Linearly separable problem solved by a linear support vector machine. The area enclosed by the dashed lines corresponds to the margin. The circled samples of both classes represent the support vectors and define the decision boundary shown by the solid line

Prediction for unlabelled inputs, that is, those not in the training set, is obtained with the application of a similarity kernel function k, between the unlabeled input $\mathbf{x}^{'}$ and each of the training inputs $\mathbf{x}_i$.

Originally, SVMs were proposed to solve binary classification problems. In many practical applications, especially in the biomedical area, the number of classes is generally higher than two. These multiclass problems are usually solved by decomposing them into binary subproblems and building SVM classifiers which distinguish between one class versus the rest (one-vs.-all approach) or between each pair of classes (one-vs.-one approach) (Duan & Sathiya Keerthi, 2005; Hsu & Lin, 2002). In the one-versus-all approach, a new instance is classified by a winner-takes-all strategy, meaning that the classifier with the highest probability determines the class. In the one-versus-one approach, the classification is performed by a max-wins voting strategy, meaning that every classifier assigns the instance to one of the two classes, then the number of assignments for each class is counted, and the sample is assigned to the class with most votes. Since SVMs can deal very well with high dimensional data, they have been widely applied in both neuroscience and bioinformatics. Further readings about their applications can be found in the following surveys (Byvatov & Schneider, 2003; Zhou, Wang, Liu, Ogunbona, & Shen, 2014). The main disadvantages of SVM are its computational demands and the fact that it is susceptible to overfitting, depending on the adopted kernel (Hsu, Chang, Lin, et al., 2003).

4.2 | Linear discriminant analysis

Linear discriminant analysis (LDA) is a parametric multivariate statistical model that assumes predictor variables to be normally distributed with equal covariance matrices. As such, this method is not indicated when the normality assumption is not fulfilled and in presence of outliers. In practice, means and covariances are estimated on the training set by maximum likelihood or maximum a posteriori estimate. LDA can be used both as a linear classifier and to determine which variables most contribute to discriminate among classes. It works by projecting data on a one-dimensional axis to find a linear combinations of predictor variables that maximizes the ratios of between-group to within-group sums of squares (McLachlan, 2004). In neuroscience, LDA has been recently applied to electroencephalography (EEG) data to discriminate between healthy elderly controls and patients affected by Alzheimer's disease (Neto, Biessmann, Aurlien, Nordby, & Eichele, 2016). Adaptations of LDA have been proposed for classification of gene expression data (Guo, Hastie, & Tibshirani, 2006; Huang, Quan, He, & Zhou, 2009; Li, Wang, Wang, Xue, & Wong, 2017; Pan & Zhang, 2017).

4.3 | Logistic regression

Logistic regression is a linear regression model for categorical dependent variables that makes no assumptions on the shape of the data distribution. The goal of logistic regression is to determine the combination of variables with the greatest probability of predicting the expected outcome. It uses maximum likelihood estimation to compute probabilities using a logistic function to model the relationship between dependent and independent variables (Hosmer Jr, Lemeshow, & Sturdivant, 2013). This approach has the advantage of providing probabilities associated to outcomes, but it tends to have a high bias. To avoid overfitting and improve generalization ability, logistic regression is often used in combination with regularization: a penalty term is added to the objective function to optimize in order to penalize too complex models that tend to memorize training data. Due to the flexibility and the ease of interpretation of this model, it is widely applied in classification tasks in bioinformatics (Ayers & Cordell, 2010; Bazzoli & Lambert-Lacroix, 2016; Choi et al., 2017; Jostins & McVean, 2016; Wu, Chen, Hastie, Sobel, & Lange, 2009) and neuroscience (Alkan, Koklukaya, & Subasi, 2005; Ryali, Supekar, Abrams, & Menon, 2010; Zhang, Hu, Ma, & Xu, 2015), often as an alternative to LDA since it does not require the normality assumption (Pohar, Blas, & Turk, 2004).

4.4 | Decision trees

Tree-based models are a family of nonparametric methods based on the principle of *divide et impera*. Decision trees partition the whole space of input features into nonoverlapping rectangular regions $R_1, R_2, \cdots, R_M$ aligned with the axes, then assign a simple model to each region. The Classification And Regression Trees (CART) method (Breiman, Friedman, Olshen, & Stone, 1984) assigns a constant value c_m to each partition R_m and produces the prediction function

$$f(\boldsymbol{x}_i) = \sum_{m=1}^{M} c_m I(\boldsymbol{x}_i \in R_m)$$

where $I(q)$ equals 1 when the condition q is true, and 0 otherwise. If the data demand it, nonlinear dependencies between the input features and the outcome variable can be estimated.

The learning algorithm of the CART model is a greedy procedure that splits the feature space into smaller and smaller regions minimizing some impurity criterion. Starting with the whole data set, for each splitting variable j and split point s, a partition into two new regions is produced, and the reduction in the impurity criterion obtained by the split is evaluated. Then, the split that produces the highest reduction in the impurity is retained. Each region is recursively split into smaller and purer regions until a termination criterion is met, such as the maximum level of complexity for the model, or a minimum number of observations in a region to be split. Common purity criteria for classification problems are the Gini index (Loh, 2011) and the cross-entropy index (Chen, Kar, & Ralescu, 2012; Foody, 1995). The resulting recursive partitioning scheme can be represented by a binary tree as in Figure 6.

Predictions are obtained by traversing the nodes of the tree from the root to a leaf. When a leaf is reached, the predicted outcome variable is the most frequent class in that region. Intelligible rules can be easily derived from a fitted decision tree. Each path from the root to a leaf node can be translated into a chain of conditions that result in a classification of the observations. This makes decision trees valuable particularly in medical sciences, where the inspection of decision rules can bring insights about the experimental question.

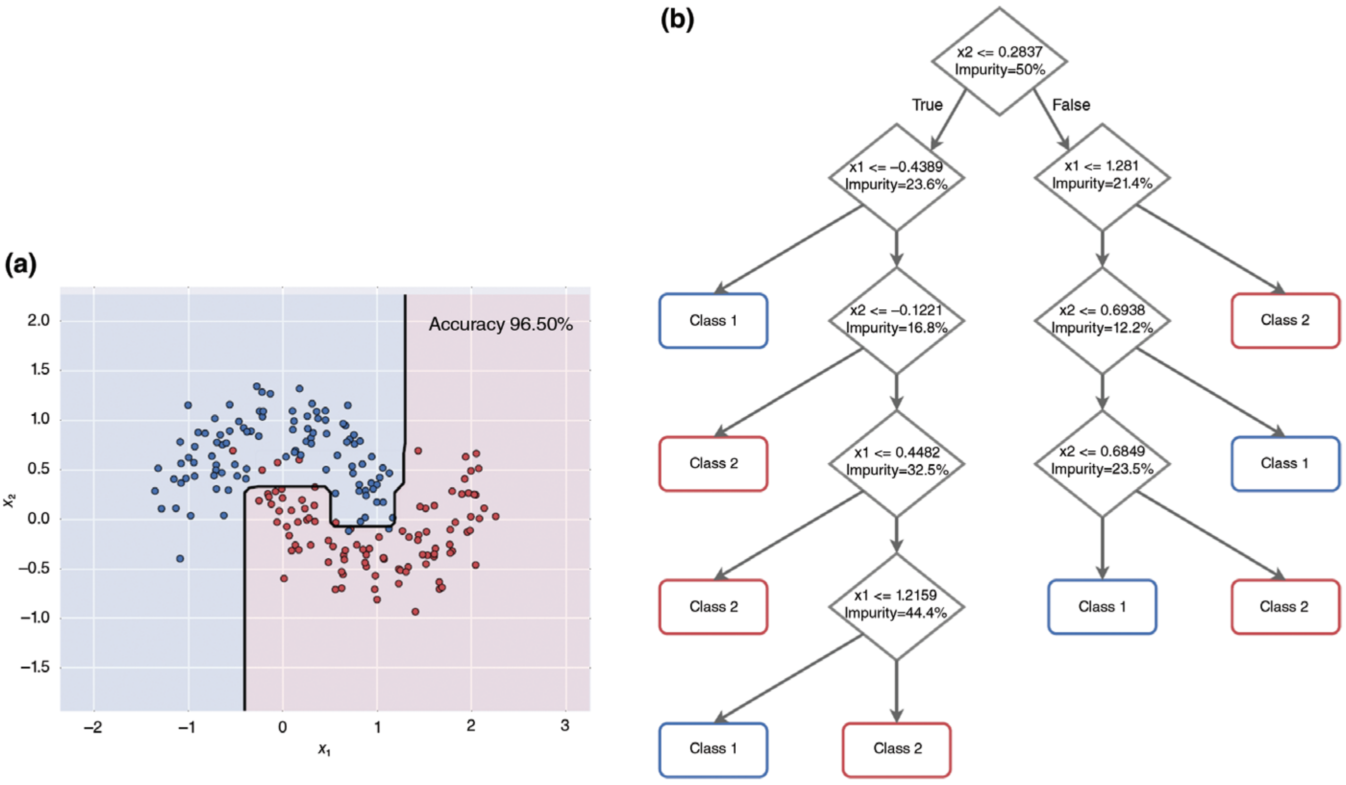


FIGURE 6 (a) A nonlinearly separable data set. The flexibility of the decision tree improves classification performance. (b) The binary tree derived from the trained classifier. Predictions are computed by visiting the tree and performing the tests included inside the diamond-shaped nodes. When a rectangular node is reached, the predicted label corresponds to the reported class

Decision trees are more flexible compared to the linear SVM, but their higher sensitivity to the data makes them more unstable. Small perturbations in the data set can produce different models and variable predictions, especially along the decision boundary. This trade-off must be taken into account when modeling a data set. Usually, model selection is a trial and error procedure, where different models fitted to the data set are compared in order to choose the best model by balancing flexibility and stability. Decision trees have many applications in bioinformatics (Che, Liu, Rasheed, & Tao, 2011). They have also been applied in patient classification from neuroimaging data (Libero, DeRamus, Lahti, Deshpande, & Kana, 2015; Mudali, Teune, Renken, Leenders, & Roerdink, 2015).

4.5 | Random forest

A RF is an ensemble method based on bagging (Breiman, 2001). A large set of independent classifiers (decision trees) are aggregated to produce a more accurate classification with respect to each single model. The higher is the number of independent trees used to train the RF, the smaller is the variance of prediction. The independence of the predictors is ensured by training each tree on a bootstrapped training data set and randomly selecting a subset of features, each time a split of the data set is estimated. This makes RF able to deal with multiple features which may be correlated. Usually, each tree of the forest is trained on a bootstrapped data set of roughly two thirds of the observations of the original data set. The remaining portion is then used to estimate the generalization performance of the tree. The aggregation of these estimates is called Out of Bag and measures the prediction ensemble error. One main advantage of using RF classifier is that the features can be ranked based on their average measure of improvement in the purity criterion, that can be considered as an index of feature relevance for classification. A disadvantage of the RF is its sensitivity to the input parameters (Huang & Boutros, 2016). RFs have been widely applied both in bioinformatics (Qi, 2012) and in neuroimaging (Sarica, Cerasa, & Quattrone, 2017).

4.6 | Bayesian classifiers

This class of models includes all classifiers that use Bayes' theorem on conditional probabilities to predict class membership minimizing the probability of misclassification. Given a data set X and a set of class labels $Y = \{1, 2, \ldots, K\}$, we denote $P(x)$ the probability $P(X = x)$, $P(c)$ the probability $P(Y = c)$ and $P(c|x)$ the conditional probability $P(Y = c|X = c)$. A Bayesian classifier is a function C defined as follows:

$$C(x) = \underset{c \in \{1,2,\ldots,K\}}{\operatorname{argmax}} P(c \mid x)$$

Assuming that the information about the classes (prior probabilities $P(c)$) and the distribution of data in classes $P(x|c)$ are known, Bayesian classifiers compute the posterior probability $P(c|x)$ using the Bayes' theorem:

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)},$$

where $P(x) = \sum_{c \in \{1,2,\ldots,K\}} P(x \mid c)P(c)$. Naive Bayesian classifiers further assume that features values are independent given the class. This conditional independence assumption simplifies the computation of $P(x|c)$, also known as the likelihood term, that is derived with a maximum likelihood estimation. The advantage of this method is that class memberships are easy and fast to compute; however, it is not guaranteed to perform well when the assumptions are not met. These models have many applications in bioinformatics, such as RNA sequences classification (Knight, Ivanov, & Dougherty, 2014; Wang, Garrity, Tiedje, & Cole, 2007), prediction of PPI sites (Murakami & Mizuguchi, 2010), and mass spectrometry data analysis (Webb-Robertson, Metz, Waters, Zhang, & Rewers, 2017). It has also been adopted in MRI-based diagnostic of pathological brains (Zhou et al., 2015).

4.7 | K-nearest neighbors (KNN)

This non-parametric model is based on a very simple concept: the outcome of a prediction for a test instance depends on the labels of its KNN in the feature space, which are usually determined using Euclidean distance, and class membership is assigned with a majority voting scheme. No prior knowledge is needed about the distribution of data, which is only locally approximated. The choice of k depends on the data, but in general it should be large enough to reduce the effect of noise on classification and small enough to avoid including instances of other classes in the neighborhood. In presence of unbalanced classes the risk is to incorrectly assign memberships to the over-sampled class; this problem can be mitigated by assigning weights to neighbors which are proportional to the inverse of the distance with the test instance, so that nearer objects contribute more. This model has been applied in proteomics to study protein localization (Alex Xijie & Moses, 2016; Chou & Shen, 2006; Horton et al., 2007; Qiao, Yan, & Li, 2017; Wang, Li, Zhang, & Wang, 2016) and to predict protein types

(Shen & Chou, 2005). For its ability to model the local structure of data, it has also been used in brain tissue segmentation (Ghayoor, Paulsen, Kim, & Johnson, 2016; Havaei, Jodoin, & Larochelle, 2014; Vrooman et al., 2007). Even if the KNN method is simple to implement and naturally handles multiclass classification, the main drawback is that the test step is quite slow when exploring a large search space to find nearest neighbors, while other algorithms, such as SVM or LDA, once trained, can classify test instances in constant time.

4.8 | Rule induction

Rule-based classifiers are a class of models that, starting from a set of observations, derive rules that identify subgroups of objects that are characterized by some properties of interest. The induced rules have the form: IF *condition* THEN *class*, where the condition is a set of attribute-value pairs derived from the features describing training instances. For this reason, they differ from traditional methods, like, for example, discriminant analysis approaches, which build black-box models that do not allow to infer relationships among the variables of interest and produce a single function that is applied invariably to any new instance. There are three classes of rule induction methods: (a) separate-and-conquer (or covering) strategies, which search for a rule that explains patterns of the training data that form a subgroup, then recursively learn more rules until every object is assigned to a class; (b) divide-and-conquer strategies are those used by decision trees, that generate one rule for each path from the root to a leaf; (c) exhaustive search strategies explore all the rules that predict the class label, that can later be filtered using a minimum quality criterion. Separate-and-conquer methods are usually preferred to the other strategies because they can handle overlapping rules (contrary to divide-and-conquer methods) and do not generate redundant rules (as in exhaustive rule search) (Pham, Clemente, Satou, & Ho, 2005). Let us consider the application of rule induction for the identification of miRNA regulatory modules (Tran, Satou, & Bao Ho, 2008). Given a set of miRNA and a set of their target genes, the goal is to find subsets of miRNA and corresponding target genes (regulatory modules). In this case, the attributes are expression profiles of miRNAs and mRNAs. For a given gene, Pearson's correlation coefficient is used to compute similarities between pairs of genes and a threshold is used to divide the gene set into two classes of similar and dissimilar genes. A separate-and-conquer rule induction strategy is then applied to produce a set of miRNA-mRNA regulatory rules. Finally, only significant rules, containing miRNAs with highly correlated expression profiles, are retained. Other applications of rule induction in bioinformatics include the description of gene sets (Gruca & Sikora, 2017) and disease subtyping (Cangelosi et al., 2013; Huang, Huang, Lee, & Weng, 2015).

4.9 | Meta-classification

Meta-classifiers are a class of models that do not perform classification using training data directly, but use meta-information derived from other learning models. The goal is to build a final model able to obtain a better generalization, to achieve higher accuracy and to obtain more stable results (Lemke, Budka, & Gabrys, 2015; Vilalta & Drissi, 2002). Several approaches exist that fall into this category. Bagging strategies consist in subsampling the original data set multiple times and training a weak classifier (a simple model able to predict above chance) on each subsample; the final predictions are computed integrating the results of base classifiers, typically with a majority voting scheme. A very popular example is RF, an ensemble of decision tree classifiers. Boosting approaches, like AdaBoost, iteratively update the training data increasing the weight of misclassified instances to focus on objects that are hard to classify and produce ever more accurate models. Stacked generalization works by training multiple base learners on the training data and then building one or more high-level classifiers that are trained using the predictions of low-level classifiers as features. Meta-classifiers are useful when little prior knowledge is known about data distribution or the functional relationship between predictors and outcome, since the only task for the investigator is to train very simple models (e.g., decision trees) with very general assumptions. Moreover, these models tend to be less sensitive to overfitting than simple classifiers. One drawback is that meta-classifiers do not provide interpretable coefficients (Westreich, Lessler, & Funk, 2010). Applications in biomedical settings include cancer classification (Bhanot, Alexe, Venkataraghavan, & Levine, 2006; Fung & Ng, 2004), lesion classification (Hong, Bernhardt, Gill, Bernasconi, & Bernasconi, 2017; Tsirogiannis, Frossyniotis, Nikita, & Stafylopatis, 2004) and membrane protein type prediction (Wang, Yang, & Chou, 2006).

4.10 | Model selection

In the first step of a model building process, the set of collected observations is split into two data sets, defined as the training set and the test set. This is required since training and evaluating a model performance on the same data yields overoptimistic results. A typical rule of thumb for splitting train and test data is 80%:20%. Each selected model is fitted on the training set. To compare the performances of the fitted models, they are validated on the hold-out observations of the test set, that have not been used during training but for which the correct outcome variable is known. The general idea of supervised learning is that the more the training set is representative of the unknown population of observations, the more reliable a model with high performances on the test set is. There are different metrics to evaluate classification results. The most common is the accuracy, which corresponds to the count of test observations correctly classified, divided by the size of the test set. In binary classification, accuracy can be defined in terms of true/false positives and true/false negatives, that corresponds to items correctly/incorrectly assigned to the positive and negative classes, respectively; for example, the positive class might indicate patients and the negative class could represent control subjects. The accuracy is then the number of true positives and true negatives over the total number of items. Counting only the number of true positives (negatives) over the total number of items classified as positive (negative) gives the specificity of the classifier for the positive (negative) class. The sensitivity for a classifier for a given class is defined as the number of correctly classified instances over the total number of items belonging to that class.[2] A high specificity corresponds to a low number of false positives, while a high sensitivity implies a low number of false negatives. According to the problem at hand, the trade-off between specificity and sensitivity might have different implications: considering the example of a classifier used to diagnose patients over healthy controls, a false positive translates into erroneously diagnose a healthy subject as sick, and a false negative indicates a failure in the diagnosis of a patient. The receiver operating characteristic (ROC) curve is a useful tool for model selection that plots the sensitivity as a function of the false-positive rate: a classifier with maximum sensitivity and no false positives would be a point with coordinates (0,1) in the ROC space, with a corresponding area under the curve (AUC) equal to 1. Another measure used to find a balance between sensitivity and specificity is the F-score, defined as the harmonic mean of the two scores. When the models to be trained have also hyper-parameters to be optimized, like the C parameter[3] of the SVMs, an additional amount of observations is further held out from the training set, usually 20%, to form the validation set. Different configurations of the hyper-parameters of the same model are fitted on the training set. Then, models are compared by evaluating the performances on the validation set. The best-performing model on the validation set, is then compared to other models as usual on the test set. The necessity of a validation set is dictated by the fact that when looking for the best values of hyper-parameters, over-fitting (Hawkins, 2004) must be taken into account. A model is said to overfit when it starts memorizing, rather than generalizing from the training set, but the performances on the hold-out data are poor. Since the choice of the hyper-parameters can be seen as actual learning, different configurations of hyper-parameters cannot be evaluated on the test set, otherwise overconfident performances will be obtained. Depending on the size of train, validation and test sets, more or less variability is associated to the estimated parameters (more training data, more stable parameters) or to the performances (more test data, more stable performance assessment). As is the case in neuroimaging experiments, the number of observations is rarely above a hundred. When data are not sufficient to obtain stable estimates of parameters and performances, a smarter way of using the available data is cross-validation (Kohavi et al., 1995). Cross-validation is an alternative method to perform hyper-parameter validation, which consists in averaging several performance evaluations corresponding to different splits of the data. The training set is divided into k chunks or folds of approximately the same size, usually 5 or 10 based on the size of the training set. Each fold, in turn, will be a validation set, on which every model with a choice of hyper-parameters, trained on the remaining $k - 1$ folds, will be evaluated. The performances of each fold are averaged and the best-performing configuration of hyper-parameters will be chosen. The model with best cross-validated configuration of hyper-parameters, is then trained on the whole training set, to obtain a more stable estimate of the parameters. The model obtained in this way can be then compared to different models evaluating the performances on the test set as in the above case.

4.11 | Drug repositioning

Drug repositioning is the process by which known drugs and compounds are used to treat new indications (i.e., a different disease than that for which the drugs were placed on the market) (Sleigh & Barton, 2010). The main advantage of drug repositioning over traditional drug development is that the risk of failure of adverse toxicology is reduced because the known drugs have already passed a number of toxicity tests. Classical methods for drug repositioning rely on the response of the cell (at the level of the genes) after treatment, or on disease-to-drug relationships, merging several information levels (Dudley et al., 2011; Gottlieb, Stein, Ruppin, & Sharan, 2011; Sanseau et al., 2012). However, these approaches encountered some obstacles such as the noisy structure of the gene expression and the few amount of genomic data related to many diseases. Multiview biological data and their integration can significantly increase the ability of the scientific community to reposition existing drugs. Usually, these approaches use machine-learning or network theory algorithms to integrate and analyze multiple layers of information such as the similarity of the drugs based on how similar are their chemical structures, or on how close are their targets within the PPI network, and on how correlated are the gene expression patterns after treatment. For example, Napolitano et al. (2013), for each drug, integrated three different omics views: genome-wide gene expression measures, chemical structure, and drugs targets. They applied a kernel-based late integration approach (see Box 1) where for each view they constructed a distance matrix and then they combined these matrices by creating a mean kernel. The first similarity matrix is the correlation between the gene expression patterns; the second depends on how similar are the drugs with respect to their chemical structure and the last one is the distance matrix between drug targets in their PPI network. The combined matrix was used to train the multiclass SVM classifier in order to predict therapeutic classes. Their results show a high accuracy in the classification task that allows for the repositioning of systematically misclassified drugs.

4.12 | Classification of mass spectrometry-based proteomics

High-throughput techniques, such as microarray, are commonly used to measure gene expression (Allison, Cui, Page, & Sabripour, 2006). In conjunction with a mass spectrometer they can be used to measure protein abundance (Aebersold & Mann, 2003), providing a genome-wide transcription or translation monitoring. These experiments often result in the evaluation of a high number of features and a limited number of samples, that is, an ill-posed problem also related to the commonly known "curse-of-dimensionality" problem (Somorjai, Dolenko, & Baumgartner, 2003). Then, the selection of the most relevant features with such a small number of samples is the key issue in microarray or mass spectrometry-based proteomics classification problem. Ensemble methods, such as RF, offer the advantage of their ability in dealing with data of small sample size and high dimensionality, such as those generated by microarray and mass-spectrometry-based proteomics studies. Geurts et al. (2005) applied the RF classifier to identify biomarkers for two different inflammatory diseases (rheumatoid arthritis [RA] and inflammatory bowel diseases [IBD]) using two mass spectrometry-based data sets. In both cases, they performed a binary classification task by classifying RA and IBD patients versus healthy subjects. The study showed that, compared with other classifiers, RFs, on average, give the lowest error rate with the smallest variance. RFs properties were also investigated in other studies. For example, Lee, Lee, Park, and Song (2005) compared different ensemble methods (bagging and boosting) with RFs, with the same experimental settings, and found that RFs were the model with best performances. In another study (Díaz-Uriarte & De Andres, 2006), RFs were compared with linear classifiers (such as SVM with feature selection) on 10 different microarray data sets, leading to the conclusions that RFs are able to preserve predictive accuracy even though yielding smaller feature sets. Izmirlian (2004) demonstrated other advantages of RFs, such as robustness to noise and computational speed in classifying proteomics data.

4.13 | Patient classification from neuroimaging data

The SVM model is a popular choice as a classifier for neuroimaging data, probably due to its ability to cope with high-dimensional data. For example, LaConte, Strother, Cherkassky, Anderson, and Xiaoping (2005) used SVM for temporal classification of block design fMRI data. In this context, given all brain voxels at a specific time point, the classification task consists in assigning the experimental design value for that time. Therefore, each data point represents the fMRI image at a certain time. Following the SVM algorithm, the input vectors are mapped to a high-dimensional feature space and the model attempts to find a linear decision boundary in this space. In this work, the authors consider different methods to compute summary maps from SVMs, such as the direct visualization of the SVM training weight vector or the feature space weighting in function of the distance from the margin. The experimental results suggested that linear kernels are preferable in this application and that SVMs seem to be robust against different preprocessing strategies, although the most evident effect is observed with the detrending of input time series.

Another popular classifier is RF (Sarica et al., 2017). Fratello et al. (2017) propose a multiview approach (see Box 1) for the classification of subjects affected by neurodegenerative diseases, specifically Amyotrophic Lateral Sclerosis and Parkinson's Disease. This model combines a clustering-based feature extraction methodology with an ensemble classifier based on RF and fuse the information coming from two imaging modalities: fMRI (functional connectivity) and DTI (structural connectivity). Clustering is applied on each of the views to obtain a compressed representation of the input data: voxels are agglomerated and each brain parcel is represented by its median. These agglomerated features are then used to train the multiview classifier in two alternative ways: (a) training two separate RFs (one for each view) and then merging the outputs with a majority voting scheme (each Decision Tree expresses a vote); (b) the two views are integrated into the learning phase and a single RF is trained. Both models exhibited ensemble classification accuracies significantly above chance. This work is an example of how ML techniques can be adapted to leverage on the complementary information derived from multiview data.

4.14 | Multimodal parcellation of human cerebral cortex

In 2016, Glasser et al. (2016) proposed a multiview late integration approach (see Box 1) for brain parcellation that combines the information coming from several imaging modalities: architectural measures of relative cortical myelin content and cortical thickness from structural images; pattern of activation from task fMRI, functional connectivity from resting state fMRI. After a set of potential areal borders is determined (considering sharp transitions in two or more of the above cortical properties), a machine-learning classifier is trained to identify parcels in individual subjects using multimodal areal fingerprints, that is, a pattern of features derived from the different data views. With this approach, 180 regions per hemisphere were identified. The great advantage introduced by the adoption of an automated ML method is that, once the classifier is trained, the parcellation can be applied to new unseen subjects, without redefining the borders from scratch.

5 | NETWORK-BASED APPROACHES

Complex network theory has an important role in a wide range of disciplines (Albert & Barabási, 2002; Dorogovtsev & Mendes, 2013), ranging from engineering, social sciences (Reka, Jeong, & Barabasi, 1999), communications, systems biology (Barabasi & Oltvai, 2004; Mitra, Carvunis, Ramesh, & Ideker, 2013) and neuroscience (Bassett & Sporns, 2017; Ed & Sporns, 2009; Fallani, Richiardi, Chavez, & Achard, 2014; Wang, Zuo, & He, 2010). A network (or equivalently, a graph) is a mathematical abstraction that represents a set of objects, called nodes, and their relationships, called edges. The concept of network is cross disciplinary and it is independent from the kind of objects and relations that it represents. Formally, a graph G is defined as the pair $G = (V, E)$, where $V = v_1, \dots, v_n$ is the finite set of objects representing the nodes of the graph, and $E = e_1, \dots, e_m$ is the finite set of objects representing the edges. Each edge in E is a connection between a pair of nodes (x, y) in V. If there is a relevant sorting order in the pair (x, y) then the graph G will be said to be oriented (or directed) and x will be said the source of the edge and y the destination. Conversely, if there is no relevant order, the graph G will be said to be unoriented (or undirected). In terms of information flow into the network, in an oriented graph, the information can transit only from x to y. On the contrary, in an undirected graph, the information can flow in both ways. Moreover, nodes and edges can have attributes that identify specific properties of the objects and their interactions represented in the graph. For example, nodes can have labels representing their name, size, color, etc., while usually edges can have a numeric weight that represents the connection strength between the two end nodes (in this case the graph is said to be weighted). In a visual representation, the nodes of a graph are usually denoted as circles and the edges are denoted as arrows going from the source to the destination. Generally, undirected edges are represented as lines without arrows. Networks are a suitable tool to model complex entities and their interactions. There are many problems that can be solved using these structures. For example, they allow to infer information about the global structure of the connections (network topology); to identify groups of entities which have homogeneous characteristics (communities); to calculate similarity between entities based on the number of paths that join them together. Many networks models have been applied to study interactions between biological phenotypes or to study connectivity in the brain.

5.1 | Network analysis in systems biology

Complex biological systems can be represented and analyzed as computable networks. There are different kinds of biological networks under study in the field of systems biology, the most common ones being: PPI networks, gene regulatory networks, gene co-expression networks. Other examples of networks in systems biology are related to the study of the interactions between phenotypic entities.

In PPI networks, proteins are nodes and their interactions are edges (Rivas & Fontanillo, 2010). PPI networks mainly represent information about how different proteins coordinate to operate with other proteins to perform biological processes within the cell (Pellegrini, Haynor, & Johnson, 2004). For many of the proteins their complete sequence is already known, but their molecular function needs to be fully determined. This prediction can be performed by comparing their interactions with other biomolecules.

Gene regulatory networks are directed graphs (see Figure 7) that represent a collection of molecular regulators (DNA, RNA, protein) that interact with each other and with other substances in the cell, such as transcription factors (Carninci et al., 2005), to govern the gene expression levels of mRNA and proteins (Hecker, Lambeck, Toepfer, Van Someren, & Guthke, 2009). They are often studied to identify gene motifs, that are small sets of recurring regulation patterns and constitute the basic building blocks of transcription networks (Alon, 2007).

Gene coexpression networks are undirected graphs (see Figure 7) where the nodes are the genes and there is an edge between a pair of genes if they are significantly co-expressed in the samples (Stuart, Segal, Koller, & Kim, 2003).

Network models are also used to study interactions between different kinds of phenotypic entities. Many studies related to the interactions between genes and diseases have been performed, analyzing complex networks where the diseases and the genes represent the nodes and the edges between them represent their interaction strength (Piñero et al., 2016; Zickenrott, Angarica, Upadhyaya, & Del Sol, 2016). For example, DisGeNET (Piñero et al., 2016) is a network-based exploratory platform developed to understand the underlying mechanisms of complex diseases. In fact, many efforts have been made to identify connections between genes and diseases (Botstein & Risch, 2003; Kann, 2010), but there are increasing evidences that most diseases arise due to complex interactions among environmental risk factors and multiple genetic variants (Hirschhorn & Daly, 2005).

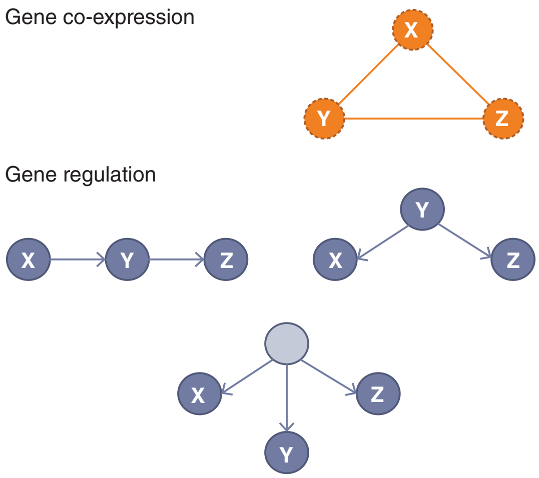


FIGURE 7 An example of gene co-expression network (orange undirected graph) and gene regulation network (violet directed graph)

Graphs can efficiently represent complex phenomena and they can be rapidly analyzed with ad hoc algorithms that consider the topological relatedness of their constituents. The main hypothesis is that patterns of similarity between sets of phenotypes could be used as an indication of biological association. In previous studies, network-based models were used to perform drug repositioning tasks (Iorio et al., 2010) and to construct network of interactions between diseases based on their symptoms (Zhou, Menche, Barabási, & Sharma, 2014). Both the works used network substructures to make inferences between the entities.

For example Iorio et al. (2010)), starting from transcriptomic data related to drug treatments on human cells, constructed a network of interactions between drugs to characterize their mode of action. Each drug was represented by the ranked list of genes sorted by their differential expression values with respect to their controls. Then, the similarities between each couple of drug, that represents the edges of the network, were computed by using the Inverse Total Enrichment Score that is based on the Kolmogorov–Smirnov test. Then they scanned the network in search of communities, to identify groups of drugs with similar effects. Moreover, to reposition a new drug, the distances between its molecular alteration pattern and those of the drugs in the communities were calculated. Then the drugs were predicted to have the same behavior of those in the closest community.

Zhou et al. (2014) constructed a human disease network based on symptom similarity. They parsed thousands of research articles in PubMed (Wheeler et al., 2007) related to diseases, computed the term frequency of each symptom (Mesh term (Lowe & Octo Barnett, 1994)) associated to each disease, and then used a bipartite network projection method to compute similarities between diseases based on how many symptoms they shared. Then, they created a network of diseases interactions based on how many genes or proteins were shared by each couple of diseases. Then, in order to identify similarities between diseases they used global measures coming from network theory to characterize couples of diseases or disease communities. For example, they used the Dijkstra's algorithm (Cormen, 2009) to compute the shortest path between diseases; then, using this information as a dissimilarity measure, they clustered diseases with the complete linkage algorithm.

5.2 | Functional network modeling with fMRI data

Observing how activations of brain regions vary over time by means of functional imaging enables the investigation of functional connectivity, that is, the existence of functional relations between functional regions of the brain (Smith et al., 2013). Network modeling is the natural choice as a tool to investigate these phenomena.

Once a parcellation has been applied to the brain, each brain region can be modeled as a node while edges represent a functional connection (see Figure 8). The simplest and most common method to measure connectivity is computing

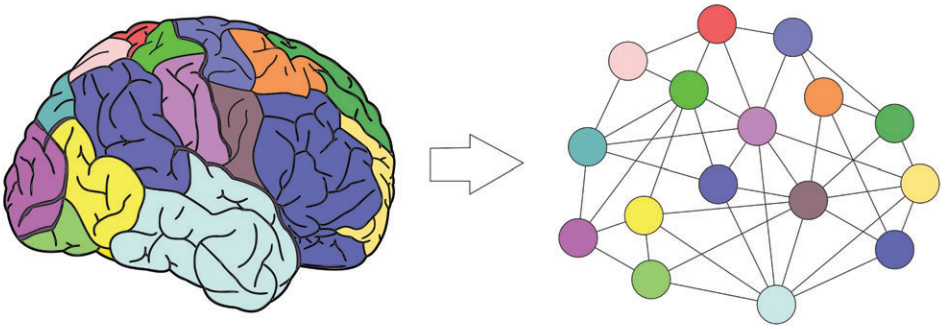


FIGURE 8 Network modeling applied to the investigation of brain connectivity: nodes represent brain regions and an edge between two regions indicates that there is a functional connection between the measured brain activity in the two regions

Pearson's correlation coefficients between the time series associated to pairs of brain regions, ignoring causal relationships. Alternatively, partial correlation can be used to estimate direct connection, but still ignoring directionality. Methods that try to model directionality involve the study of temporal lag between pair of time series, conditional dependencies, and nongaussianities in the time series (Hyvärinen & Smith, 2013; Smith et al., 2011).

Many graph theory concepts have been applied to the analysis of functional networks, including clustering coefficients and motifs, network modularity, node centrality, small worldness, and general network efficiency measures (Ed & Sporns, 2009). One of the limitations of the network-based approaches is that their efficacy is strongly affected by the accuracy of the network model, from its resolution to the thresholds chosen to select a subset of relevant edges. Another concern is related to the level of abstraction provided by a network model: a change in the value of a graph-theory-based measure such as communication path length might be difficult to interpret as an effective change in brain connectivity (Smith et al., 2013). However, several studies have found a connection between network properties and pathological conditions such as Alzheimer's disease (Stam, Jones, Nolte, Breakspear, & Scheltens, 2006; Supekar, Menon, Rubin, Musen, & Greicius, 2008) and Schizophrenia (Liu et al., 2008; Rubinov et al., 2009).

Bayesian network (BN) models have been applied in connectivity analysis (Bielza & Larrañaga, 2014; Ide, Zhang, & Li Chiang-shan, 2014; Rajapakse & Zhou, 2007). BNs are directed acyclic graphs used to model dependencies between variables in the presence of uncertainty (Koller & Friedman, 2009). They can be static or dynamic depending on whether the modeled network is assumed to be evolving over time, which is a natural assumption when modeling fMRI time series. BNs can be used to infer associations between nodes and to test competing hypotheses for patterns of connectivity, and also to study alterations in connectivity in clinical population (Wu et al., 2011). Two limitations of this approach are that, when learning the optimal network structure from data, the model search space increases exponentially with the number of brain parcels included in the model, and that cyclic relationships cannot be modeled. Markov networks (or Markov Random Field) are undirected graphs (that might be cyclic) that model random variables having a Markov property that have also been used in functional connectivity estimation (Liu, Awate, Anderson, & Thomas Fletcher, 2014). A main drawback of these models is the assumption that each voxel belongs to a single functional network, while several brain regions (e.g., the precuneus) are known to participate in a high number of integrated networks.

6 | DEEP LEARNING APPLICATIONS IN BIOMEDICINE

Deep learning is a large class of ML models that are composed of multiple processing layers able to represent data with a high level of abstraction (LeCun, Bengio, & Hinton, 2015). The main differences between traditional shallow learning (i.e., neural network with one or two hidden layer, or SVM) and deep learning is that the former does not deal with raw data and requires a feature extraction step to be performed before the learning process (LeCun et al., 2015). On the other hand, deep neural networks (DNNs) can act as feature detector units, where each layer extract increasingly more abstract and sophisticated features from the original raw data input.

Since the term deep learning refers to a wide class of techniques, one of the major challenges in deep learning applications is the selection of the most adequate model (i.e., CNNs, deep belief networks, stacked AEs (SAE) or restricted Boltzmann machines) for the current task. Bengio and Goodfellow (Bengio, Goodfellow, & Courville, 2015) classified the different DNNs models into three categories: networks for supervised learning, that are designed to provide discriminative power in classification problems; networks for unsupervised learning, that are designed to identify high-order data correlation; hybrid or semisupervised networks which aim to classify data using the outputs of an unsupervised model, in order to speed up the learning process.

One of the main problem of DNNs is that they are very complex models, due to their high number of hyper-parameters. These parameters depend on different variables, such as: architectural aspects (such as the number of layers or the transfer functions), the optimization type (e.g., learning rates and momentum values) or the type of regularization (e.g., dropout probabilities). This, linked to the nonlinearity of the model, makes DNNs optimization a challenging and time-consuming task. See this overview (Schmidhuber, 2015) for more details on DNNs models and their optimization.

Nevertheless, DNNs have already been successfully applied in a wide range of biomedical applications (Angermueller, Pärnamaa, Parts, & Stegle, 2016; Mamoshina, Vieira, Putin, & Zhavoronkov, 2016; Plis et al., 2014; Vieira, Pinaya, & Mechelli, 2017). In fact, the same techniques that have showed to perform well in image and voice analysis can be applied, with some adjustments, to medical imaging and biological data.

6.1 | CNNs application for RNA ISH

RNA ISH is a techniques able to localize and visualize the gene expression in a group of cells, in a specific tissue or in a whole organisms (Bayani & Squire, 2004). This method is helpful to illustrate changes in expression patterns during development. It was used to construct the Allen Developing Mouse Brain Atlas that contains expression maps for more than 2000 genes in different developing stages of the brain (Thompson et al., 2014). This operation was usually performed by hand, but then, with the application of CNNs (see Box 3), Zeng, Li, Mukkamala, Ye, and Ji (2015) were able to perform an automatic annotation of the gene expression patterns. DNNs models require a large number of labeled images to be trained. One way to overcome this limitation is the use of a transfer learning approach, where the network is trained on one data set and then used as feature extractor on other data sets (Donahue et al., 2014; Gupta, Ayhan, & Maida, 2013; Oquab, Bottou, Laptev, & Sivic, 2014; Razavian, Azizpour, Sullivan, & Carlsson, 2014; Zeiler & Fergus, 2014). Zeng et al. applied transfer learning from natural images to ISH images. They used the OverFeat (Sermanet et al., 2013) model, trained on the ImageNet data set, and then generalized and used as feature extractor on ISH images. The results show that using convolutional models as feature extractors it is possible to achieve better performances on the tasks of annotating gene expression patterns at multiple levels of brain structures. They achieved an overall average AUC of 0.894 ± 0.014, as compared with 0.820 ± 0.046 yielded by the bag-of-words approach.

6.2 | DNA- and RNA-binding protein with deep leaning CNN networks

The understanding of the sequence specificities of DNA- and RNA-bind protein is a crucial point in the development of models of regulatory biological processes and the identification of random disease variants. Usually, the sequence specificities are fully characterized using position weight matrices (PWMs), that are easy to interpret and are easy to be scanned over a genomic sequence to detect potential binding sites (Stormo, 2000). Many other traditional shallow classification models have been proposed (Kazan, Ray, Chan, Hughes, & Morris, 2010; Rohs et al., 2010; Siggers & Gordân, 2013), but they have to cope with different problems such as: (a) data produced by different technologies that come in different formats; (b) the huge amount of data to be analyzed; (c) the fact that each data acquisition has its own artifacts, biases, and noise. Alipanahi, Delong, Weirauch, and Frey (2015) adopted a deep CNN to predict sequence specificities and binding scores to cope with all these problems. Their method, called DeepBind, is able to capture binding specificities directly from raw sequence data and jointly discovers new sequence motifs and the rules needed to combine them in a predictive binding score. The training phase was performed using a set of sequences with their experimentally determined binding scores. Sequences can have varying lengths and binding scores can be real-valued measurements or binary class labels. For each sequence, the model computes a binding score in four steps: (a) a convolution stage in which it scans a set of motif detectors (each motif detector M_i, $i = 1, \ldots, k$ is a 4×4 matrix similar to PWM) across the sequence; (b) the rectification stage isolates positions with a good match by shifting the response of detector M_k by b_k and clamping all negative values to zero; (c) The pooling stage computes the maximum and average of each motif detector's rectified response across the sequence; maximizing helps to identify the presence of

BOX 3

CONVOLUTIONAL NEURAL NETWORKS (CNNs) AND AUTOENCODERS (AEs)

CNNs were originally designed for the analysis of images, to leverage the spatial structure of neighboring pixels. In fact, in traditional fully connected architectures, close and far pixels are treated in the same way, ignoring topology. These networks are inspired by the workings of the visual cortex, where simple neurons respond to motifs in a small localized region (the so called receptive field) and complex neurons respond to patterns in larger regions (Angermueller et al., 2016). An example of CNN architecture is shown in Figure 9. In a CNN layer, each hidden unit is connected to a small group of contiguous neurons from the previous layer and all the units share the same weights, so that all neurons respond to the same pattern across the image (e.g., an edge). Different layers respond to different patterns. The term convolutional refers to the convolution operation applied to each receptive field, that consists in computing the weighted sum of input neurons, and applying an activation function. Local connectivity and weight sharing drastically reduce the number of parameters compared to a fully connected network. Pooling layers are added to further compress input dimensionality by computing the average or the maximum over adjacent neurons.

AEs are feedforward, nonrecurrent neural networks with an output layer having the same number of nodes as the input layer. An example of autoencoder architecture is reported in Figure 10. In this model, backpropagation is applied to reconstruct the input data minimizing reconstruction error. In this sense, they are considered unsupervised networks, since no external information is fed into the model. The learnt weights can then be used as input features for a traditional supervised learning model. For instance, AEs can learn a low-dimensional representation from high-dimensional input data (Hinton & Salakhutdinov, 2006). In this case, the compressed representation can be extracted from a small central hidden layer. Another application is the reconstruction of input starting from partially corrupted samples.

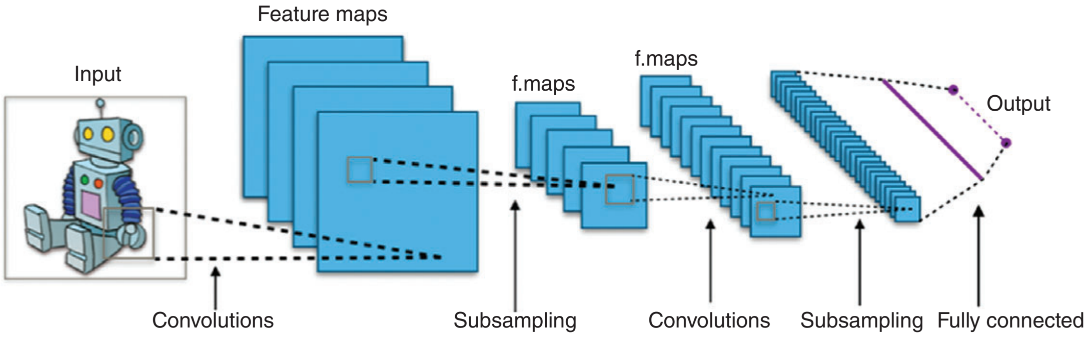


FIGURE 9 Typical CNN architecture. Image by Aphex34 (Own work) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

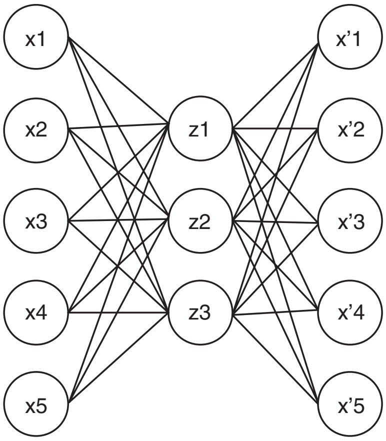


FIGURE 10 Example of autoencoder architecture with five inputs (×1, ×2, ×3, ×4, and ×5) and outputs ($x'1$, $x'2$, $x'3$, $x'4$, and $x'5$) and one hidden layer with three neurons ($z1$, $z2$, and $z3$)

longer motifs, whereas averaging helps to identify cumulative effects of short motifs, and the contribution of each is determined automatically by learning; (d) these values are fed into a nonlinear neural network with weights *W*, which combines the responses to produce a score. Though there is no single agreed upon metric for evaluating the quality of sequence specificity predictions, the model was compared across different data sets and with different evaluation metrics with respect to other state-of-the-art methodologies and it was shown to outperform all of them both using in vivo and in vitro data.

6.3 | Deep AEs for the diagnostic of the Alzheimer's disease

The ability of deep learning models to detect complex patterns can be exploited in neuro-biological studies to identify biomarkers of neurological disorders (Plis et al., 2014; Vieira et al., 2017). Recently, there has been a growing interest in the application of AEs (see Box 3) to extract abstract features from several neuroimaging modalities, often used in a multiview fashion with SAE. In Liu et al. (2015), a multilayered neural network consisting of several AEs and a soft-max layer is used for the diagnostic of the Alzheimer's disease. MR and PET imaging modalities are fused by jointly training the AEs with the concatenated MR and PET inputs. To avoid neurons that are activated only by one modality, in a pretraining phase, a series of samples is presented to the network where the inputs of one of the modalities are replaced by zeros. In this way, the first AE is trained to reconstruct the original inputs from the inputs that are mixed with hidden modalities. Then, higher layers learn to reconstruct the high-level representation from the noisy one propagated through lower layers, thus inferring the correlations between the different modalities. In Suk, Wee, Lee, and Shen (2016), AEs are used to extract hierarchical nonlinear relationships between functionally connected regions of the brain, following the idea that the functional organization of the brain is dynamic rather than static.

6.4 | CNNs for brain networks

CNNs (see Box 3) are usually applied to analyze images, since they leverage the locality of image features, and some examples of applications to brain images are present in the literature (Bengio et al., 2015; Gao & Hui, 2016; Gupta et al., 2013; Krizhevsky, Sutskever, & Hinton, 2012; Payan & Montana, 2015; Sarraf & Tofighi, 2016); however, it is possible to apply CNNs also to connectome data, taking advantage of the structural or functional organization of brain regions. Kawahara et al. (2017) proposed a CNN model (BrainNetCNN) designed for the prediction of clinical neurodevelopmental outcomes from structural brain networks of infants born preterm. Contrary to traditional image-based CNN, BrainNetCNN exploits the topology of brain networks to build convolutional filters based on edge-to-edge, edge-to-node, and node-to-graph relationships, thus avoiding the need for fully connected layers and consequently reducing the number of parameters to be learnt. The input of the CNN consists of individual 90 × 90 adjacency matrices, derived from DTI imaging and tractography: edge weights represent the number of connections between pairs of predefined anatomical regions. The model has been proven to outperform CNNs with the same number of parameters but built only on fully connected layers.

7 | CONCLUSION

ML is a fundamental ingredient for the analysis of complex systems as those studied by biomedical sciences. Firstly, the advances in high-throughput technologies for the acquisition of biomedical data have created the need of sophisticated methods able to cope with the complexity of big data. Secondly, the number and the heterogeneity of existing modalities to describe biological and neurobiological phenomena require the development of intelligent systems to integrate the information coming from several domains. For these reasons, more and more studies are moving toward multiview data integration, since working with a single modality provides only a partial picture of the problem under investigation.

The goals of multiview learning and data integration are to obtain higher precision and greater statistical power than those provided by single-view data sets. Moreover, integrative analysis can be useful in validating results from different data sets, under the assumption that if information from independent data sources is concordant, it is more likely that the results are reliable than in the case of information coming from a single source. On the other side, since the typical number of available samples in biomedical task is not enough to avoid bias and overfitting problems, these technique must be combined with feature selection or dimensionality reduction in order to avoid the curse of dimensionality. A new trend is to apply ad hoc model to each view and then combine their results.

Following a holistic approach to the study of biological systems, one of the tasks to be accomplished is to model the interactions between entities at different abstraction levels, as cells, tissues, and organisms in systems biology and neurons, synapses and functional hubs in neurosciences. With this goal in mind, the application of network-based analyses and many concepts borrowed from graph theory represent effective tools to gain useful insights about the organization and workings of these systems.

However, the challenge relies not only in modeling and analyzing the massive volume of available data, but mainly in extracting new knowledge to be exploited for advances in precision medicine, diagnostics, and treatments of pathological conditions. In this sense, the availability of ever more powerful hardware resources has made possible the adoption of deep learning models, that are a promising approach when dealing with the task of inferring relevant information from big data. One of the main disadvantages of deep learning in the applications to biomedical data is the fact that deep models require a higher number of samples to be trained, even if in some cases this problem can be solved by using transfer learning techniques and by combining deep learning models with statistical and computational learning methodologies to reduce the number of features and make available the most of the data. Moreover, there are many other potential challenges, including the interpretation of the deep learning results and the selection of an appropriate architecture and its hyper-parameters.

Despite some limitations, the results outlined in this paper are almost always very encouraging confirming the central role of ML techniques and suggesting new promising research areas, such as multiview learning and deep learning approaches, in biomedical data analysis.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

NOTES

[1] Volumetric pixels in a three-dimensional grid.

[2] The concepts of sensitivity and specificity can be trivially extended to multi-class problems.

[3] The C value of the SVM is a controlling hyper-parameter that weights the contribution of the errors to the objective function.

RELATED WIREs ARTICLES

Machine learning

REFERENCES

Abraham, A., Milham, M. P., Di Martino, A., Cameron Craddock, R., Samaras, D., Thirion, B., & Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage*, *147*, 736–745.

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, *422*(6928), 198–207.

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (2005). Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery*, *11*(1), 5–33. https://doi.org/10.1007/s10618-005-1396-1

Aizerman, M. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, *25*, 821–837.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97.

Alex Xijie, L., & Moses, A. M. (2016). An unsupervised knn method to systematically detect changes in protein localization in high-throughput microscopy images. *PLoS One*, *11*(7), e0158712.

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838.

Alizadeh, A. A., Eisen, M. B., Eric Davis, R., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, *403*(6769), 503–511.

Alkan, A., Koklukaya, E., & Subasi, A. (2005). Automatic seizure detection in eeg using logistic regression and artificial neural network. *Journal of Neuroscience Methods*, *148*(2), 167–176.

Allison, D. B., Cui, X., Page, G. P., & Sabripour, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews. Genetics*, *7*(1), 55–65.

Alok, A. K., Saha, S., & Ekbal, A. (2008). Semi-supervised clustering for gene-expression data in multiobjective optimization framework. *International Journal of Machine Learning and Cybernetics*, *8*(2), 421–439.

Alon, U. (2007). Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, *8*(6), 450–461.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878.

Ansel, A., Rosenzweig, J. P., Zisman, P. D., Melamed, M., & Gesundheit, B. (2016). Variation in gene expression in autism spectrum disorders: An extensive review of transcriptomic studies. *Frontiers in Neuroscience*, *10*, 601.

Ayers, K. L., & Cordell, H. J. (2010). Snp selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, *34*(8), 879–891.

Bandyopadhyay, S., & Saha, S. (2008). A point symmetry-based clustering technique for automatic evolution of clusters. *IEEE Transactions on Knowledge and Data Engineering*, *20*(11), 1441–1457.

Barabasi, A.-L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113.

Bar-Joseph, Z., Gitter, A., & Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, *13*(8), 552–564.

Bassett, D. S., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, *20*(3), 353–364.

Bayani, J., & Squire, J. A. (2004). Fluorescence in situ hybridization (FISH). *Current Protocols in Cell Biology*, *22.4*(Suppl 23), 1–52.

Bazzoli C., & Lambert-Lacroix, S. (2016). Classification using LS-PLS with logistic regression based on both clinical and gene expression variables. Available at https://hal.archives-ouvertes.fr/hal-01405101/

Ben-Dor, A., Shamir, R., & Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology*, *6*(3–4), 281–297.

Bengio, Y., Goodfellow, I. J., & Courville, A. (2015). Deep learning. *Nature*, *521*, 436–444.

Bhanot, G., Alexe, G., Venkataraghavan, B., & Levine, A. J. (2006). A robust meta-classification strategy for cancer detection from ms data. *Proteomics*, *6*(2), 592–604.

Bielza, C., & Larrañaga, P. (2014). Bayesian networks in neuroscience: A survey. *Frontiers in Computational Neuroscience*, *8*, 131.

Blumensath, T., Jbabdi, S., Glasser, M. F., Van Essen, D. C., Ugurbil, K., Behrens, T. E. J., & Smith, S. M. (2013). Spatially constrained hierarchical parcellation of the brain with resting-state fmri. *NeuroImage*, *76*, 313–324.

Bo, W., Mezlini, A. M., Demir, F., Fiume, M., Zhuowen, T., Brudno, M., … Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, *11*(3), 333–337.

Borg, N. J., Mitchell, M. S., Lukacs, P. M., Mack, C. M., Waits, L. P., & Krausman, P. R. (2017). Behavioral connectivity among bighorn sheep suggests potential for disease spread. *The Journal of Wildlife Management*, *81*(1), 38–45.

Botstein, D., & Risch, N. (2003). Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, *33*, 228–237.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Boca Raton, FL: CRC Press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Broyd, S. J., Demanuele, C., Debener, S., Helps, S. K., James, C. J., & Sonuga-Barke, E. J. S. (2009). Default-mode brain dysfunction in mental disorders: A systematic review. *Neuroscience & Biobehavioral Reviews*, *33*(3), 279–296.

Byvatov, E., & Schneider, G. (2003). Support vector machine applications in bioinformatics. *Applied Bioinformatics*, *2*(2), 67–77.

Cameron Craddock, R., Andrew James, G., Holtzheimer, P. E., Xiaoping, P. H., & Mayberg, H. S. (2012). A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, *33*(8), 1914–1928.

Cangelosi, D., Blengio, F., Versteeg, R., Eggert, A., Garaventa, A., Gambini, C., … Varesio, L. (2013). Logic learning machine creates explicit and stable rules stratifying neuroblastoma patients. *BMC Bioinformatics*, *14*(7), S12.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., … RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) (2005). The transcriptional landscape of the mammalian genome. *Science*, *309*(5740), 1559–1563.

Carreira-Perpinán, M. A. (1997). *A review of dimension reduction techniques* (Technical Report CS-96-09). Department of Computer Science, University of Sheffield, 9, 1–69.

Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. *Frontiers in Aging Neuroscience*, *9*, 329.

Charlon, T., Martínez-Bueno, M., Bossini-Castillo, L., David Carmona, F., Di Cara, A., Wojcik, J., … Alarcón-Riquelme, M. E. (2016). Single nucleotide polymorphism clustering in systemic autoimmune diseases. *PLoS One*, *11*(8), e0160270.

Che, D., Liu, Q., Rasheed, K., & Tao, X. (2011). Decision tree and ensemble learning algorithms with their applications in bioinformatics. In Arabnia H., & Tran QN. (Eds.), *Software tools and algorithms for biological systems* (pp. 191–199). New York: Springer.

Chen, H., Li, K., Zhu, D., Jiang, X., Yuan, Y., Lv, P., … Liu, T. (2013). Inferring group-wise consistent multimodal brain networks via multi-view spectral clustering. *IEEE Transactions on Medical Imaging*, *32*(9), 1576–1586.

Chen, X., Xiaofei, X., Huang, J. Z., & Ye, Y. (2013). Tw-k-means: Automated two-level variable weighting clustering algorithm for multiview data. *IEEE Transactions on Knowledge and Data Engineering*, *25*(4), 932–944.

Chen, X., Kar, S., & Ralescu, D. A. (2012). Cross-entropy measure of uncertain variables. *Information Sciences*, *201*, 53–60.

Cheng, Y., & Church, G. M. (2000). Biclustering of expression data. *Ismb*, *8*, 93–103.

Choi, S. H., Labadorf, A. T., Myers, R. H., Lunetta, K. L., Dupuis, J., & DeStefano, A. L. (2017). Evaluation of logistic regression models and effect of covariates for case–control study in rna-seq analysis. *BMC Bioinformatics*, *18*(1), 91.

Chou, K.-C., & Shen, H.-B. (2006). Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic k-nearest neighbor classifiers. *Journal of Proteome Research*, *5*(8), 1888–1897.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, *6*(1), 57–71.

Cordes, D., Haughton, V., Carew, J. D., Arfanakis, K., & Maravilla, K. (2002). Hierarchical clustering to measure connectivity in fmri resting-state data. *Magnetic Resonance Imaging*, *20*(4), 305–317.

Cormen, T. H. (2009). *Introduction to algorithms*. Cambridge, MA: MIT Press.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cribben, I., & Yi, Y. (2017). Estimating whole-brain dynamics by using spectral clustering. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, *66*(3), 607–627.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1*(2), 224–227.

De Martino, F., Gentile, F., Esposito, F., Balsi, M., Di Salle, F., Goebel, R., & Formisano, E. (2007). Classification of fmri independent components using ic-fingerprints and support vector machine classifiers. *NeuroImage*, *34*(1), 177–194.

Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 3.

Divina, F., & Aguilar-Ruiz, J. S. (2006). Biclustering of expression data with evolutionary computation. *IEEE Transactions on Knowledge and Data Engineering*, *18*(5), 590–602.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). *Decaf: A deep convolutional activation feature for generic visual recognition.* International Conference on Machine Learning, Beijing, China, 647–655.

Dorogovtsev, S. N., & Mendes, J. F. F. (2013). *Evolution of networks: From biological nets to the Internet and WWW*. Oxford, England: Oxford University Press.

Duan, K., & Sathiya Keerthi, S. (2005). Which is the best multiclass svm method? An empirical study. *Multiple Classifier Systems*, *3541*, 278–285.

Dudley, J. T., Sirota, M., Shenoy, M., Pai, R. K., Roedder, S., Chiang, A. P., … Butte, A. J. (2011). Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science Translational Medicine*, *3*(96), 76–96.

Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*, 32–57.

Ed, B., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186–198.

Eickhoff, S. B., Bzdok, D., Laird, A. R., Roski, C., Caspers, S., Zilles, K., & Fox, P. T. (2011). Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation. *NeuroImage*, *57*(3), 938–949.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), 14863–14868.

Elhefnawy, W., Li, M., Wang, J., & Li, Y. (2017). Construction of protein backbone fragments libraries on large protein sets using a randomized spectral clustering algorithm. In *International symposium on bioinformatics research and applications* (pp. 108–119). Cham, Switzerland: Springer.

Ellebedy, A. H., Jackson, K. J. L., Kissick, H. T., Nakaya, H. I., Davis, C. W., Roskin, K. M., … Ahmed, R. (2016). Defining antigen-specific plasmablast and memory b cell subsets in human blood after viral infection or vaccination. *Nature Immunology*, *17*(10), 1226–1234.

Esnault, C., Gualdrini, F., Horswell, S., Kelly, G., Stewart, A., East, P., … Treisman, R. (2017). Erk-induced activation of tcf family of srf cofactors initiates a chromatin modification cascade associated with transcription. *Molecular Cell*, *65*(6), 1081–1095.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Knowledge Discovery and Data Mining*, *96*, 226–231.

Fallani, F. D. V., Richiardi, J., Chavez, M., & Achard, S. (2014). Graph analysis of functional brain networks: Practical issues in translational neuroscience. *Philosophical Transactions of the Royal Society B*, *369*(1653), 20130521.

Flandin, G., Kherif, F., Pennec, X., Malandain, G., Ayache, N., & Poline, J.-B. (2002). *Improved detection sensitivity in functional mri data using a brain parcelling technique.* Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002, 467–474.

Flock, T., Hauser, A. S., Lund, N., Gloriam, D. E., Balaji, S., & Madan Babu, M. (2017). Selectivity determinants of gpcr–g-protein binding. *Nature*, *545*(7654), 317–322.

Fodor, I. K. (2002). *A survey of dimension reduction techniques* (Technical Report No. UCRL-ID-148494). Lawrence Livermore National Lab, Livermore, CA.

Fonseca, L., van Pul, C., Lori, N., van den Boom, R., Andriessen, P., Buijs, J., & Vilanova, A. (2017). Automatic atlas-based segmentation of brain white matter in neonates at risk for neurodevelopmental disorders. In Schultz T., Özarslan E., & Hotz I. (Eds.), *Modeling, analysis, and visualization of anisotropy* (pp. 355–372). Cham, Switzerland: Springer.

Foody, G. M. (1995). Cross-entropy for the evaluation of the accuracy of a fuzzy land cover classification with fuzzy ground data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *50*(5), 2–12.

Formisano, E., Esposito, F., Kriegeskorte, N., Tedeschi, G., Di Salle, F., & Goebel, R. (2002). Spatial independent component analysis of functional magnetic resonance imaging time-series: Characterization of the cortical components. *Neurocomputing*, *49*(1), 241–254.

Fortino, V., Kinaret, P., Fyhrquist, N., Alenius, H., & Greco, D. (2014). A robust and accurate method for feature selection and prioritization from multi-class omics data. *PLoS One*, *9*, e107801.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*(383), 553–569.

Fratello, M., Caiazzo, G., Trojsi, F., Russo, A., Tedeschi, G., Tagliaferri, R., & Esposito, F. (2017). Multi-view ensemble classification of brain connectivity images for neurodegeneration type discrimination. *Neuroinformatics*, *15*(2), 199–213.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–976.

Fung, B. Y. M., & Ng, V. T. Y. (2004). *Meta-classification of multi-type cancer gene expression data*. In *Proceedings of the 4th international conference on data mining in bioinformatics* (pp. 31–39). London, UK: Springer-Verlag.

Galdi, P., Fratello, M., Trojsi, F., Russo, A., Tedeschi, G., Tagliaferri, R., & Esposito, F. (2017). Consensus-based feature extraction in rs-fmri data analysis. *Soft Computing*, 1–11.

Galdi, P., Napolitano, F., & Tagliaferri, R. (2014). Consensus clustering in gene expression. In *International meeting on computational intelligence methods for bioinformatics and biostatistics* (pp. 57–67). Cham, Switzerland: Springer.

Gao, X. W., & Hui, R. (2016). *A deep learning based approach to classification of ct brain images.* SAI Computing Conference (SAI), 2016, London, United Kingdom, IEEE, 28–31.

Garcia-Chimeno, Y., Garcia-Zapirain, B., Gomez-Beldarrain, M., Fernandez-Ruanova, B., & Garcia-Monco, J. C. (2017). Automatic migraine classification via feature selection committee and machine learning techniques over imaging and questionnaire data. *BMC Medical Informatics and Decision Making*, *17*(1), 38.

Geurts, P., Fillet, M., De Seny, D., Meuwis, M.-A., Malaise, M., Merville, M.-P., & Wehenkel, L. (2005). Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics*, *21*(14), 3138–3145.

Ghayoor, A., Paulsen, J. S., Kim, R. E. Y., & Johnson, H. J. (2016). *Tissue classification of large-scale multi-site mr data using fuzzy k-nearest neighbor method.* SPIE Medical Imaging, International Society for Optics and Photonics, 97841V–97841V.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., … van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*, 171–178.

Golland, P., Golland, Y., & Malach, R. (2007). *Detection of spatial activation patterns as unsupervised segmentation of fmri data.* Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007, 110–118.

Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). Predict: A method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, *7*(1), 496.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F., & Hansen, L. K. (1999). On clustering fMRI time series. *NeuroImage*, *9*(3), 298–310.

Greene, J., Lin, M., Wang, J., Ye, J., & Wittenberg, G. (2017). 339-biclustering of blood gene expression data identifies patient subtypes with different biological pathologies in major depressive disorder. *Biological Psychiatry*, *81*(10), S139.

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, *100*(1), 253–258.

Grubbs, K. J., Bleich, R. M., Santa Maria, K. C., Allen, S. E., Farag, S., Team, A. B., … Bowers, A. A. (2017). Large-scale bioinformatics analysis of bacillus genomes uncovers conserved roles of natural products in bacterial physiology. *mSystems*, *2*(6), e00040–e00017.

Gruca, A., & Sikora, M. (2017). Data-and expert-driven rule induction and filtering framework for functional interpretation and description of gene sets. *Journal of Biomedical Semantics*, *8*(1), 23.

Guo, Y., Hastie, T., & Tibshirani, R. (2006). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, *8*(1), 86–100.

Gupta, A., Ayhan, M., & Maida, A. (2013). *Natural image bases to represent neuroimaging data.* International Conference on Machine Learning, 987–994.

Gupta, C. N., Castro, E., Rachkonda, S., van Erp, T. G. M., Potkin, S., Ford, J. M., et al. (2017). Biclustered independent component analysis for complex biomarker and subtype identification from structural magnetic resonance images in schizophrenia. *Frontiers in Psychiatry*, *8*, 179.

Hajighorbani, M., Reza Hashemi, S. M., Minaei-Bidgoli, B., & Safari, S. (2016). A review of some semi-supervised learning methods. *IEEE-2016, First International Conference on New Research Achievements in Electrical and Computer Engineering.*

Hand, D. J., & Heard, N. A. (2005). Finding groups in gene expression data. *BioMed Research International*, *2005*(2), 215–225.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, *21*(15), 3201–3212. https://doi.org/10.1093/bioinformatics/ bti517

Hannah Immanuel, M. S., & Jacob, S. G. (2017). Feature selection techniques for Alzheimer's disease: A review. *International Journal of Engineering Technology Science and Research*, *4*(7), 494–499.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society*, *28*, 100–108. https://doi.org/10.2307/2346830

Havaei, M., Jodoin, P.-M., & Larochelle, H. (2014). *Efficient interactive brain tumor segmentation as within-brain knn classification.* 2014 22nd International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, IEEE, 556–561.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12.

Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., & Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems*, *96*(1), 86–103.

Heider, D., Genze, N., & Neumann, U. (2017). Efs: An ensemble feature selection tool implemented as r-package and web-application. *BioData Mining*, *10*(1), 21.

Higdon, R., Earl, R. K., Stanberry, L., Hudac, C. M., Montague, E., Stewart, E., et al. (2015). The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders. *Omics: A Journal of Integrative Biology*, *19*(4), 197–208.

Higham, D. J., Kalna, G., & Kibble, M. (2007). Spectral clustering and its use in bioinformatics. *Journal of Computational and Applied Mathematics*, *204*(1), 25–37.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, *6*(2), 95–108.

Hobbs, B. D., Morrow, J. D., Celli, B. R., Bueno, R., Criner, G. J., DeMeo, D. L., Hersh, C. P., Silverman, E. K., & Cho, M. H. (2017). Chronic obstructive pulmonary disease subtyping through multiple-omics data integration. *C21. Omics in lung disease*, American Thoracic Society, A4964–A4964.

Hong, H., Yin, X., Li, F., Guan, N., Bo, X., & Luo, Z. (2017). *Predicting potential gene ontology from cellular response data.* Proceedings of the 5th International Conference on Bioinformatics and Computational Biology, ACM, 5–10.

Hong, S.-J., Bernhardt, B., Gill, R., Bernasconi, N., & Bernasconi, A. (2017). Connectome-based pattern learning predicts histology and surgical outcome of epileptogenic malformations of cortical development. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 390–397). Cham, Switzerland: Springer.

Hood, L., & Friend, S. H. (2011). Predictive, personalized, preventive, participatory (p4) cancer medicine. *Nature Reviews. Clinical Oncology*, *8*(3), 184–187.

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Hajime, H., Adams-Collier, C. J., & Nakai, K. (2007). Wolf psort: Protein localization predictor. *Nucleic Acids Research*, *35*(suppl_2), W585–W587.

Hosmer Jr., D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. *398*). Hoboken, NJ: John Wiley & Sons.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification. Available at https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415–425.

Huang, B. F. F., & Boutros, P. C. (2016). The parameter sensitivity of random forests. *BMC Bioinformatics*, *17*(1), 331.

Huang, D., Quan, Y., He, M., & Zhou, B. (2009). Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data. *Journal of Experimental & Clinical Cancer Research*, *28*(1), 149.

Huang, G.-M., Huang, K.-Y., Lee, T.-Y., & Weng, J. T.-Y. (2015). An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. *BMC Bioinformatics*, *16*(1), S5.

Huang, Y., Zaas, A. K., Rao, A., Dobigeon, N., Woolf, P. J., Veldman, T., et al. (2011). Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. *PLoS Genetics*, *7*(8), e1002234.

Hunt, I. (2005). From gene to protein: A review of new and enabling technologies for multi-parallel protein expression. *Protein Expression and Purification*, *40*(1), 1–22.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4), 411–430.

Hyvärinen, A., & Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, *14*(Jan), 111–152.

Ide, J. S., Zhang, S., & Li Chiang-shan, R. (2014). Bayesian network models in brain functional connectivity analysis. *International Journal of Approximate Reasoning*, *55*(1), 23–35.

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: Systems biology. *Annual Review of Genomics and Human Genetics*, *2*(1), 343–372.

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., … di Bernardo, D. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, *107*(33), 14621–14626.

Izmirlian, G. (2004). Application of the random forest classification algorithm to a seldi-tof proteomics study in the setting of a cancer prevention trial. *Annals of the New York Academy of Sciences*, *1020*(1), 154–174.

Jiang, D., Pei, J., & Zhang, A. (2003). *Dhc: a density-based hierarchical clustering method for time series gene expression data.* Proceedings of Third IEEE Symposium on Bioinformatics and Bioengineering, 2003, IEEE*,* 393–400.

Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *16*(11), 1370–1386.

Jiang, Z., Li, T., Min, W., Qi, Z., & Rao, Y. (2017). Fuzzy c-means clustering based on weights and gene expression programming. *Pattern Recognition Letters*, *90*, 1–7.

Jostins, L., & McVean, G. (2016). Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. *Bioinformatics*, *32*(12), 1898–1900.

Kahnt, T., Chang, L. J., Park, S. Q., Heinzle, J., & Haynes, J.-D. (2012). Connectivity-based parcellation of the human orbitofrontal cortex. *Journal of Neuroscience*, *32*(18), 6240–6250.

Kailing, K., Kriegel, H.-P., & Kröger, P. (2004). *Density-connected subspace clustering for high-dimensional data.* Proceedings of the 2004 SIAM International Conference on Data Mining, SIAM, 246–256.

Kann, M. G. (2010). Advances in translational bioinformatics: Computational approaches for the hunting of disease genes. *Briefings in Bioinformatics*, *11*(1), 96–110.

Kaufman, L., & Rousseeuw, P. (1987). *Clustering by means of medoids*. Amsterdam: North-Holland.

Kawahara, J., Brown, C. J., Miller, S. P., Booth, B. G., Chau, V., Grunau, R. E., … Hamarneh, G. (2017). Brainnetcnn: Convolutional neural networks for brain networks: Towards predicting neurodevelopment. *NeuroImage*, *146*, 1038–1049.

Kazan, H., Ray, D., Chan, E. T., Hughes, T. R., & Morris, Q. (2010). Rnacontext: A new method for learning the sequence and structure binding preferences of rna-binding proteins. *PLoS Computational Biology*, *6*(7), e1000832.

Khan, S. S., & Madden, M. G. (2009). A survey of recent trends in one class classification. In *Irish conference on artificial intelligence and cognitive science* (pp. 188–197). Berlin, Heidelberg: Springer.

Kinani, V., Marie, J., Silva, A. J. R., Funes, F. G., Vargas, D. M., Díaz, E. R., & Arellano, A. (2017). Medical imaging lesion detection based on unified gravitational fuzzy clustering. *Journal of Healthcare Engineering*, *2017*, 16.

Kitano, H. (2002). Systems biology: A brief overview. *Science*, *295*(5560), 1662–1664.

Knight, J. M., Ivanov, I., & Dougherty, E. R. (2014). Mcmc implementation of the optimal bayesian classifier for non-gaussian models: Model-based rna-seq classification. *BMC Bioinformatics*, *15*(1), 401.

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*, 1137–1145.

Kohonen, T. (2012). *Self-organization and associative memory* (Vol. *8*). Berlin, Heidelberg: Springer Science & Business Media.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT Press.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. Supervised machine learning: A review of classification techniques. *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, 2007, 3-24.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, 1097–1105.

Kruglyak, L., & Nickerson, D. A. (2001). Variation is the spice of life. *Nature Genetics*, *27*(3), 234–236.

Kursa, M. B., Rudnicki, W. R., et al. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, *36*, 1–13.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., & Xiaoping, H. (2005). Support vector machines for temporal classification of block design fMRI data. *NeuroImage*, *26*(2), 317–329.

Lashkari, D., Sridharan, R., Vul, E., Hsieh, P.-J., Kanwisher, N., & Golland, P. (2012). Search for patterns of functional specificity in the brain: A nonparametric hierarchical bayesian model for group fMRI data. *NeuroImage*, *59*(2), 1348–1368.

Lazzeroni, L., & Owen, A. (2002). Plaid models for gene expression data. *Statistica Sinica*, *12*, 61–86.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, J. W., Lee, J. B., Park, M., & Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis*, *48*(4), 869–885.

Lemke, C., Budka, M., & Gabrys, B. (2015). Metalearning: A survey of trends and technologies. *Artificial Intelligence Review*, *44*(1), 117–130.

Li, D., Wang, L., Wang, J., Xue, Z., & Wong, S. T. C. (2017). *Transductive local fisher discriminant analysis for gene expression profile-based cancer classification.* 2017 I.E. EMBS International Conference on Biomedical & Health Informatics (BHI), IEEE, 49–52.

Libero, L. E., DeRamus, T. P., Lahti, A. C., Deshpande, G., & Kana, R. K. (2015). Multimodal neuroimaging based classification of autism spectrum disorder using anatomical, neurochemical, and white matter correlates. *Cortex*, *66*, 46–59.

Lin, H.-Y. (2016). Gene discretization based on em clustering and adaptive sequential forward gene selection for molecular classification. *Applied Soft Computing*, *48*, 683–690.

Lin, T.-h., Kaminski, N., & Bar-Joseph, Z. (2008). Alignment and classification of time series gene expression in clinical studies. *Bioinformatics*, *24*(13), i147–i155.

Liu, C., Abu-Jamous, B., Brattico, E., & Nandi, A. K. (2017). Towards tunable consensus clustering for studying functional brain connectivity during affective processing. *International Journal of Neural Systems*, *27*(02), 1650042.

Liu, G., Dong, C., & Liu, L. (2016). Integrated multiple “-omics” data reveal subtypes of hepatocellular carcinoma. *PLoS One*, *11*(11), e0165457.

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., … ADNI (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer’s disease. *IEEE Transactions on Biomedical Engineering*, *62*(4), 1132–1140.

Liu, W., Awate, S. P., Anderson, J. S., & Thomas Fletcher, P. (2014). A functional network estimation method of resting-state fMRI using a hierarchical Markov random field. *NeuroImage*, *100*, 520–534.

Liu, Y., Liang, M., Zhou, Y., He, Y., Hao, Y., Song, M., … Jiang, T. (2008). Disrupted small-world networks in schizophrenia. *Brain*, *131*(4), 945–961.

Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 14–23.

Lowe, H. J., & Octo Barnett, G. (1994). Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *JAMA*, *271*(14), 1103–1108.

Lu, Y., & Liu, Y. (2017). Ensemble biclustering gene expression data based on the spectral clustering. *Neural Computing and Applications*, 1–14.

Luscombe, N. M., Greenbaum, D., Gerstein, M., et al. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine*, *40*(4), 346–358.

Mahnaz Maddah, W. E., Grimson, L., Warfield, S. K., & Wells, W. M. (2008). A unified framework for clustering and quantitative analysis of white matter fiber tracts. *Medical Image Analysis*, *12*(2), 191–202.

Mamoshina, P., Vieira, A., Putin, E., & Zhavoronkov, A. (2016). Applications of deep learning in biomedicine. *Molecular Pharmaceutics*, *13*(5), 1445–1454.

Manocha, P., Bhasme, S., Gupta, T., Panigrahi, B. K., & Gandhi, T. K. (2017). Automated tumor segmentation and brain mapping for the tumor area. arXiv preprint arXiv:1710.11121.

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T.-P., Kindermann, S. S., Bell, A. J., & Sejnowski, T. J. (1997). *Analysis of fMRI data by blind separation into independent spatial components* (Technical Report No. NHRC-REPT-97-42). Naval Health Research Center, San Diego, CA.

McKeown, M. J., Hansen, L. K., & Sejnowsk, T. J. (2003). Independent component analysis of functional MRI: What is signal and what is noise? *Current Opinion in Neurobiology*, *13*(5), 620–629.

McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition*. Hoboken, NJ: John Wiley & Sons 544 pp.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. *84*). New York: Marcel Dekker.

Meila, M., & Shi, J. (2001). Learning segmentation by random walks. In Leen T. K., Dietterich T. G., & Tresp V. (Eds.), *Advances in neural information processing systems* (pp. 873–879). Cambridge, MA: MIT Press.

Meng, J., Zhang, J., Luan, Y.-S., He, X.-Y., Li, L.-S., & Zhu, Y.-F. (2017). Parallel gene selection and dynamic ensemble pruning based on affinity propagation. *Computers in Biology and Medicine*, *87*, 8–21.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., & Thirion, B. (2012). A supervised clustering approach for fMRI-based inference of brain states. *Pattern Recognition*, *45*(6), 2041–2049.

Mirnezami, R., Nicholson, J., & Darzi, A. (2012). Preparing for precision medicine. *New England Journal of Medicine*, *366*(6), 489–491.

Mitra, K., Carvunis, A.-R., Ramesh, S. K., & Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, *14*(10), 719–732.

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, *52*(1), 91–118.

Mudali, D., Teune, L. K., Renken, R. J., Leenders, K. L., & Roerdink, J. B. T. M. (2015). Classification of parkinsonian syndromes from fdg-pet brain data using decision trees with ssm/pca features. *Computational and Mathematical Methods in Medicine*, *2015*, 1–10.

Murakami, Y., & Mizuguchi, K. (2010). Applying the naïve bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics*, *26*(15), 1841–1848.

Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M., & Greco, D. (2013). Drug repositioning: A machine-learning approach through data integration. *Journal of Cheminformatics*, *5*(1), 30.

Neto, E., Biessmann, F., Aurlien, H., Nordby, H., & Eichele, T. (2016). Regularized linear discriminant analysis of eeg features in dementia patients. *Frontiers in Aging Neuroscience*, *8*, 273.

Nielsen, K. V., Ejlertsen, B., Møller, S., Jørgensen, J. T., Knoop, A., Knudsen, H., & Mouridsen, H. T. (2008). The value of top2a gene copy number variation as a biomarker in breast cancer: Update of dbcg trial 89d. *Acta Oncologica*, *47*(4), 725–734.

Nir, T., Jahanshad, N., Jack, C. R., Weiner, M. W., Toga, A. W., & Thompson, P. M. (2012). Small world network measures predict white matter degeneration in patients with early-stage mild cognitive impairment. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), Barcelona, Spain,* IEEE, 1405–1408.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). *Learning and transferring mid-level image representations using convolutional neural networks.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 1717–1724.

Orban, P., Doyon, J., Petrides, M., Mennes, M., Hoge, R., & Bellec, P. (2014). The richness of task-evoked hemodynamic responses defines a pseudohierarchy of functionally meaningful brain networks. *Cerebral Cortex*, *25*(9), 2658–2669.

Pal, S. K., Ray, S. S., & Ganivada, A. (2017). Gene function analysis. In *Granular neural networks, pattern recognition and bioinformatics* (pp. 163–193). Cham, Switzerland: Springer.

Pan, M., & Zhang, J. (2017). Correlation-based linear discriminant classification for gene expression data. *Genetics and Molecular Research*, *16*(1), gmr16019357.

Pavlidis, P., Weston, J., Cai, J., & Grundy, W. N. (2001). *Gene functional classification from heterogeneous data.* Proceedings of the fifth Annual International Conference on Computational Biology, Montreal, QC, Canada, ACM, 249–255.

Payan, A., & Montana, G. (2015). Predicting Alzheimer's disease: A neuroimaging study with 3d convolutional neural networks. arXiv preprint arXiv:1502.02506.

Pellegrini, M., Haynor, D., & Johnson, J. M. (2004). Protein interaction networks. *Expert Review of Proteomics*, *1*(2), 239–249.

Pennacchietti, F., Vascon, S., Nieus, T., Rosillo, C., Das, S., Tyagarajan, S. K., … Cella Zanacchi, F. (2017). Nanoscale molecular reorganization of the inhibitory postsynaptic density is a determinant of gabaergic synaptic potentiation. *Journal of Neuroscience*, *37*(7), 1747–1756.

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of k in k-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), 103–119.

Pham, T. H., Clemente, J. C., Satou, K., & Ho, T. B. (2005). Computational discovery of transcriptional regulatory rules. *Bioinformatics*, *21*(Suppl 2), ii101–ii107.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., … Furlong, L. I. (2016). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, *49*, gkw943.

Planey, C. R., & Gevaert, O. (2016). Coincide: A framework for discovery of patient subtypes across multiple datasets. *Genome Medicine*, *8*(1), 1.

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., … Calhoun, V. D. (2014). Deep learning for neuroimaging: A validation study. *Frontiers in Neuroscience*, *8*, 229.

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

Polat, H., Mehr, H. D., & Cetin, A. (2017). Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. *Journal of Medical Systems*, *41*(4), 55.

Poole, W., Leinonen, K., Shmulevich, I., Knijnenburg, T., & Bernard, B. (2017). Multiscale mutation clustering algorithm identifies pan-cancer mutational clusters associated with pathway-level changes in gene expression. *PLoS Computational Biology*, *13*(2), e1005347.

Qi, Y. (2012). Random forest for bioinformatics. In Zhang C., & Ma Y. (Eds.), *Ensemble machine learning* (pp. 307–323). Boston, MA: Springer.

Qiao, S., Yan, B., & Li, J. (2017). Ensemble learning for protein multiplex subcellular localization prediction based on weighted knn with different features. *Applied Intelligence*, 1–12.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, *2*(6), 418–427.

Rajapakse, J. C., & Zhou, J. (2007). Learning effective brain connectivity with dynamic Bayesian networks. *NeuroImage*, *37*(3), 749–760. https://doi.org/10.1016/j.neuroimage.2007.06.003

Ramoni, M. F., Sebastiani, P., & Kohane, I. S. (2002). Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences*, *99*(14), 9121–9126.

Rani, S., & Sikka, G. (2012). Recent techniques of clustering of time series data: A survey. *International Journal of Computer Applications*, *52*(15), 1–9.

Rasero, J., Pellicoro, M., Angelini, L., Cortes, J. M., Marinazzo, D., & Stramaglia, S. (2017). Consensus clustering approach to group brain connectivity matrices. *Network Neuroscience*, *1*, 242–253.

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., & Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, *155*, 530–548.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). *Cnn features off-the-shelf: An astounding baseline for recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 806–813.

Real, R., & Vargas, J. M. (1996). The probabilistic basis of jaccard's index of similarity. *Systematic Biology*, *45*(3), 380–385.

Reka, A., Jeong, H., & Barabasi, A.-L. (1999). Diameter of the world-wide web. *Nature*, *401*(6749), 130–131.

Rengeswaran, B., Mathaiyan, N., & Kandasamy, P. (2017). Cuckoo search with mutation for biclustering of microarray gene expression data. *International Arab Journal of Information Technology*, *14*(3), 300–306.

Rennert, R. C., Schäfer, R., Bliss, T., Januszyk, M., Sorkin, M., Achrol, A. S., … Gurtner, G. C. (2016). High-resolution microfluidic single-cell transcriptional profiling reveals clinically relevant subtypes among human stem cell populations commonly utilized in cell-based therapies. *Frontiers in Neurology*, *7*, 41.

Ringner, M., & Staaf, J. (2017). P1. 02-062 consensus of gene expression phenotypes and prognostic risk predictors in primary lung adenocarcinoma. *Journal of Thoracic Oncology*, *12*(1), S525–S526.

Rivas, J. D. L., & Fontanillo, C. (2010). Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Computational Biology*, *6*(6), e1000807.

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., & Mann, R. S. (2010). Origins of specificity in protein-dna recognition. *Annual Review of Biochemistry*, *79*, 233–269.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326.

Rubinov, M., Knock, S. A., Stam, C. J., Micheloyannis, S., Harris, A. W. F., Williams, L. M., & Breakspear, M. (2009). Small-world properties of nonlinear brain activity in schizophrenia. *Human Brain Mapping*, *30*(2), 403–416.

Rudie, J. D., Brown, J. A., Beck-Pancer, D., Hernandez, L. M., Dennis, E. L., Thompson, P. M., … Dapretto, M. (2013). Altered functional and structural brain network organization in autism. *NeuroImage: Clinical*, *2*, 79–94.

Rui, X., & Wunsch, D. C. (2010). Clustering algorithms in biomedical research: A review. *IEEE Reviews in Biomedical Engineering*, *3*, 120–154.

Ryali, S., Supekar, K., Abrams, D. A., & Menon, V. (2010). Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, *51*(2), 752–764.

Sadacca, B., Hamy-Petit, A.-S., Laurent, C., Gestraud, P., Bonsang-Kitzis, H., Pinheiro, A., … Reyal, F. (2017). New insight for pharmacogenomics studies from the transcriptional analysis of two large-scale cancer cell line panels. *Scientific Reports*, *7*(1), 15126.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Salman, M. S., Du, Y., & Calhoun, V. D. (2017). *Identifying fMRI dynamic connectivity states using affinity propagation clustering method: Application to schizophrenia*. 2017 I.E. International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 904–908.

Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Brent Richards, J., Cardon, L. R., & Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. *Nature Biotechnology*, *30*(4), 317–320.

Saria, S., & Goldenberg, A. (2015). Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems*, *30*(4), 70–75.

Sarraf, S. & Tofighi, G.. (2016). Classification of Alzheimer's disease using fMRI data and deep learning convolutional neural networks. arXiv preprint arXiv: 1603.08631.

Schliep, A., Schönhuth, A., & Steinhoff, C. (2003). Using hidden markov models to analyze gene expression time course data. *Bioinformatics*, *19*(Suppl 1), i255–i263.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Segal, E., Taskar, B., Gasch, A., Friedman, N., & Koller, D. (2001). Rich probabilistic models for gene expression. *Bioinformatics*, *17*(Suppl 1), S243–S252.

Segal, E., Wang, H., & Koller, D. (2003). Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, *19*(Suppl 1), i264–i272.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

Serra, A., Fratello, M., Fortino, V., Raiconi, G., Tagliaferri, R., & Greco, D. (2015). Mvda: A multi-view genomic data integration methodology. *BMC Bioinformatics*, *16*(1), 261.

Sharan, R., & Shamir, R. (2000). *Click: a clustering algorithm with applications to gene expression analysis*. Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, 16.

Sheffield, N. C., Pierron, G., Klughammer, J., Datlinger, P., Schönegger, A., Schuster, M., … Tomazou, E. M. (2017). Dna methylation heterogeneity defines a disease spectrum in ewing sarcoma. *Nature Medicine*, *23*(3), 386–395.

Shen, H., & Chou, K.-C. (2005). Using optimized evidence-theoretic k-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. *Biochemical and Biophysical Research Communications*, *334*(1), 288–292.

Shen, H., Wang, L., Liu, Y., & Hu, D. (2010). Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. *NeuroImage*, *49*(4), 3110–3121.

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., … Sander, C. (2012). Integrative subtype discovery in glioblastoma using icluster. *PLoS One*, *7*(4), e35236.

Shen, X., Papademetris, X., & Todd Constable, R. (2010). Graph-theory based parcellation of functional subunits in the brain from resting-state fMRI data. *NeuroImage*, *50*(3), 1027–1035.

Shuke, N. (2017). Voxel-based control database generated from clinical fdg pet data for statistical analysis of brain fdg pet: Comparison with subject-based normal database. *Journal of Nuclear Medicine*, *58*(supplement 1), 1257–1257.

Siggers, T., & Gordân, R. (2013). Protein–dna binding: Complexities and multi-protein codes. *Nucleic Acids Research*, *42*(4), 2099–2111.

Sleigh, S. H., & Barton, C. L. (2010). Repurposing strategies for therapeutics. *Pharmaceutical Medicine*, *24*(3), 151–159.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., … Woolrich, M. W. (2011). Network modelling methods for fMRI. *NeuroImage*, *54*(2), 875–891.

Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., … van Essen, D. C. (2013). Functional connectomics from resting-state fMRI. *Trends in Cognitive Sciences*, *17*(12), 666–682.

Somorjai, R. L., Dolenko, B., & Baumgartner, R. (2003). Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: Curses, caveats, cautions. *Bioinformatics*, *19*(12), 1484–1491.

Song, Y., Chen, W.-Y., Bai, H., Lin, C.-J., & Chang, E. Y. (2008). Parallel spectral clustering. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 374–389). Berlin, Heidelberg: Springer.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, *98*(19), 10869–10874.

Stam, C. J., Jones, B. F., Nolte, G., Breakspear, M., & Scheltens, P. (2006). Small-world networks and functional connectivity in Alzheimer's disease. *Cerebral Cortex*, *17*(1), 92–99.

Steinley, D. (2004). Properties of the hubert-arable adjusted rand index. *Psychological Methods*, *9*(3), 386–396.

Stormo, G. D. (2000). Dna binding sites: Representation and discovery. *Bioinformatics*, *16*(1), 16–23.

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, *302*(5643), 249–255.

Suk, H.-I., Wee, C.-Y., Lee, S.-W., & Shen, D. (2016). State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *NeuroImage*, *129*, 292–307.

Sun, Y., Chen, Y., Collinson, S. L., Bezerianos, A., & Sim, K. (2015). Reduced hemispheric asymmetry of brain anatomical networks is linked to schizophrenia: A connectome study. *Cerebral Cortex*, *27*(1), 602–615.

Supekar, K., Menon, V., Rubin, D., Musen, M., & Greicius, M. D. (2008). Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. *PLoS Computational Biology*, *4*(6), e1000100.

Suzuki, R., & Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, *22*(12), 1540–1542.

Taskesen, E., Huisman, S. M. H., Mahfouz, A., Krijthe, J. H., de Ridder, J., van de Stolpe, A., … Reinders, M. J. T. (2016). Pan-Cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Scientific Reports*, *6*, 24949.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, *22*(3), 281–285.

Tedeschi, G., & Esposito, F. (2012). Neuronal networks observed with resting state functional magnetic resonance imaging in clinical populations. In Bright P. (Ed.), *Neuroimaging-cognitive and clinical neuroscience* (pp. 109–128). Rijeka, Croatia: InTech.

Tench, C. R., Tanasescu, R., Constantinescu, C. S., Auer, D. P., & Cottam, W. J. (2017). Coordinate based random effect size meta-analysis of neuroimaging studies. *NeuroImage*, *153*, 293–306.

Theodoridis, S., & Koutroumbas, K. (2008). *Pattern recognition*. Cambridge, MA: Academic Press. ISBN:9781597492720.

Thirion, B., & Faugeras, O. (2004). Feature characterization in fMRI data: The information bottleneck approach. *Medical Image Analysis*, *8*(4), 403–419.

Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., & Poline, J.-B. (2006). Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. *Human Brain Mapping*, *27*(8), 678–693.

Thirion, B., Varoquaux, G., Dohmatob, E., & Poline, J.-B. (2014). Which fMRI clustering gives good brain parcellations? *Frontiers in Neuroscience*, *8*, 167.

Thomas Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., … Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165.

Thompson, C. L., Ng, L., Menon, V., Martinez, S., Lee, C.-K., Glattfelder, K., … Jones, A. . R. (2014). A high-resolution spatiotemporal atlas of gene expression of the developing mouse brain. *Neuron*, *83*(2), 309–323.

Tran, D. H., Satou, K., & Bao Ho, T. (2008). Finding microrna regulatory modules in human genome using rule induction. *BMC Bioinformatics*, *9*(12), S5.

Tsirogiannis, G., Frossyniotis, D., Nikita, K., & Stafylopatis, A. (2004). A meta-classifier approach for medical diagnosis. In Vouros G.A., & Panayiotopoulos T. (Eds.), *Methods and Applications of Artificial Intelligence. SETN 2004, Samos, Greece. Lecture Notes in Computer Science* (Vol. 3025, pp. 154–163). Berlin Heidelberg: Springer.

Tucholka, A., Thirion, B., Perrot, M., Pinel, P., Mangin, J.-F., & Poline, J.-B. (2008). Probabilistic anatomo-functional parcellation of the cortex: How many regions? In Metaxas D., Axel L., Fichtinger G., & Székely G. (Eds.), *International conference on medical image computing and computer-assisted intervention* (pp. 399–406). Berlin, Heidelberg: Springer.

Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, *25*(03), 337–372.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Gocayne, J. D. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351.

Vercelli, U., Diano, M., Costa, T., Nani, A., Duca, S., Geminiani, G., … Cauda, F. (2016). Node detection using high-dimensional fuzzy parcellation applied to the insular cortex. *Neural Plasticity*, *2016*, 1–8.

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 & Biobehavioral Reviews*, *74*, 58–75.

Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*, *18*(2), 77–95.

Vrooman, H. A., Cocosco, C. A., van der Lijn, F., Stokking, R., Arfan Ikram, M., Vernooij, M. W., … Niessen, W. J. (2007). Multi-spectral brain tissue segmentation using automatically trained k-nearest-neighbor classification. *NeuroImage*, *37*(1), 71–81.

Wang, A., An, N., Yang, J., Chen, G., Li, L., & Alterovitz, G. (2017). Wrapper-based gene selection with markov blanket. *Computers in Biology and Medicine*, *81*, 11–23.

Wang, C., Armasu, S. M., Kalli, K. R., Maurer, M. J., Heinzen, E. P., Keeney, G. L., … Goode, E. L. (2017). Pooled clustering of high-grade serous ovarian cancer gene expression leads to novel consensus subtypes associated with survival and surgical outcomes. *Clinical Cancer Research*, *23*(15), 4077–4085.

Wang, H., Wang, W., Yang, J., & Yu, P. S. (2002). *Clustering by pattern similarity in large data sets.* Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, ACM, 394–405.

Wang, J., Zuo, X., & He, Y. (2010). Graph-based network analysis of resting-state functional MRI. *Frontiers in Systems Neuroscience*, *4*, 16.

Wang, Q., & Chen, G. (2017). Fuzzy soft subspace clustering method for gene co-expression network analysis. *International Journal of Machine Learning and Cybernetics*, *8*(4), 1157–1165.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267.

Wang, S.-Q., Yang, J., & Chou, K.-C. (2006). Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. *Journal of Theoretical Biology*, *242*(4), 941–946.

Wang, X., Li, H., Zhang, Q., & Wang, R. (2016). Predicting subcellular localization of apoptosis proteins combining GO features of homologous proteins and distance weighted knn classifier. *BioMed Research International*, *2016*.

Wang, Y. X., Ke, L., Theusch, E., Rotter, J. I., Medina, M. W., Waterman, M. S., & Huang, H. (2017). Generalized correlation measure using count statistics for gene expression data with ordered samples. *Bioinformatics*, btx641.

Wang, Z., Gerstein, M., & Snyder, M. (2009). Rna-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63.

Webb-Robertson, B.-J. M., Metz, T. O., Waters, K. M., Zhang, Q., & Rewers, M. (2017). Bayesian posterior integration for classification of mass spectrometry data. In Datta S., & Mertens B. (Eds.), *Statistical analysis of proteomics, metabolomics, and Lipidomics data using mass spectrometry* (pp. 203–211). Cham, Switzerland: Springer.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*(8), 826–833.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., … Yaschenko, E. (2007). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *35*(Suppl 1), D5–D12.

Wong, K.-C., Li, Y., & Zhang, Z. (2016). Unsupervised learning in genome informatics. In *Unsupervised learning algorithms* (pp. 405–448). Cham, Switzerland: Springer.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, *25*(6), 714–721.

Wu, X., Li, R., Fleisher, A. S., Reiman, E. M., Guan, X., Zhang, Y., … Yao, L. (2011). Altered default mode network connectivity in Alzheimer's disease—A resting functional MRI and bayesian network study. *Human Brain Mapping*, *32*(11), 1868–1881.

Xianxue, Y., Guoxian, Y., & Wang, J. (2017). Clustering cancer gene expression data by projective clustering ensemble. *PLoS One*, *12*(2), e0171429.

Xu, J., Han, J., Nie, F., & Li, X. (2017). Re-weighted discriminatively embedded *k*-means for multi-view clustering. *IEEE Transactions on Image Processing*, *26*(6), 3016–3027.

Xu, R., & Wunsch II, D. (2009). *Clustering*. Hoboken, NJ: IEEE.

Yan, D., Huang, L., & Jordan, M. I. (2009). *Fast approximate spectral clustering*. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, ACM, 907–916.

Yang, P., Li, X.-L., Mei, J.-P., Kwoh, C.-K., & Ng, S.-K. (2012). Positive-unlabeled learning for disease gene identification. *Bioinformatics*, *28*(20), 2640–2647.

Yang, P., Yang, Y. H., Zhou, B. B., & Zomaya, A. Y. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, *5*(4), 296–308.

Yu, H., Hong, S., Yang, X., Ni, J., Dan, Y., & Qin, B. (2013). Recognition of multiple imbalanced cancer types based on dna microarray data using ensemble classifiers. *BioMed Research International*, *2013*, 1–13.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In Fleet D., Pajdla T., Schiele B., & Tuytelaars T. (Eds.), *European conference on computer vision* (pp. 818–833). Cham, Switzerland: Springer.

Zeng, T., Li, R., Mukkamala, R., Ye, J., & Ji, S. (2015). Deep convolutional neural networks for annotating gene expression patterns in the mouse brain. *BMC Bioinformatics*, *16*(1), 147.

Zhang, X., Hu, B., Ma, X., & Xu, L. (2015). Resting-state whole-brain functional connectivity networks for mci classification using l2-regularized logistic regression. *IEEE Transactions on Nanobioscience*, *14*(2), 237–247.

Zhang, Y., Xiaohua, H., & Jiang, X. (2017). Multi-view clustering of microbiome samples by robust similarity network fusion and spectral clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *14*(2), 264–271.

Zhou, L., Wang, L., Liu, L., Ogunbona, P., & Shen, D. (2014). Support vector machines for neuroimage analysis: Interpretation from discrimination. In *Support Vector Machines Applications* (pp. 191–220). Cham, Switzerland: Springer.

Zhou, X., Wang, S., Xu, W., Ji, G., Phillips, P., Sun, P., & Zhang, Y. (2015). Detection of pathological brain in MRI scanning based on wavelet-entropy and naive bayes classifier. In Ortuño F., & Rojas I. (Eds.), *International conference on bioinformatics and biomedical engineering* (pp. 201–209). Cham, Switzerland: Springer.

Zhou, X. Z., Menche, J., Barabási, A.-L., & Sharma, A. (2014). Human symptoms–disease network. *Nature Communications*, *5*, 4212.

Zhu, X. (2005). *Semi-supervised learning literature survey*. Madison, WI: University of Wisconsin.

Zhu, Z., Ong, Y.-S., & Dash, M. (2007). Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, *40*(11), 3236–3248.

Zickenrott, S., Angarica, V. E., Upadhyaya, B. B., & Del Sol, A. (2016). Prediction of disease–gene–drug relationships following a differential network analysis. *Cell Death & Disease*, *7*(1), e2040.

Zong, L., Zhang, X., Zhao, L., Yu, H., & Zhao, Q. (2017). Multi-view clustering via multi-manifold regularized non-negative matrix factorization. *Neural Networks*, *88*, 74–89.

How to cite this article: Serra A, Galdi P, Tagliaferri R. Machine learning for bioinformatics and neuroimaging. *WIREs Data Mining Knowl Discov*. 2018;e1248. https://doi.org/10.1002/widm.1248