Glossary

Accuracy: Proportion of correct predictions made by a classifier.

Activation function: Function (typically nonlinear) that is applied to the weighted sum of the inputs of a neuron in a neural network.

Algorithm: Instructions to follow in order to solve a problem.

Artificial neural network (ANN): Neural networks composed of layers of connected neurons (sometimes referred to as nodes or units) loosely inspired by biological neural networks.

Autoencoder: Unsupervised neural network that aims to copy its input to its output by compressing the input into a latent representation (encoder) and then reconstructing the output from this representation (decoder).

Balanced accuracy: Accuracy obtained when averaging the accuracy for the different classes. Balanced accuracy differs from accuracy in case of unbalanced datasets, where the number of observations in one category is greater than the number of observations in the other category; in contrast, in the context of fully balanced datasets, balanced accuracy and accuracy are the same.

Backpropagation: Short for "backward propagation of errors"—a method for adjusting each weight in a deep neural network based on how much it contributes to the overall error.

Bias: Error that arises from erroneous assumptions in the machine learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

Bias—variance trade-off: Property of machine learning models where lower bias-related error is associated with higher variance-related error and vice versa.

Big Data: Umbrella term used for large volumes of often heterogeneous datasets that cannot be handled by traditional data processing techniques.

Categorical variable: Variable with a discrete set of possible values (e.g., diagnostic label, gender).

Centroid: Center of a cluster as determined by a k-means algorithm. For instance, if k is 4, then the k-means algorithm will produce 4 centroids.

Class: Group to which an observation (e.g., participant) can belong (e.g., patient or control group).

Classification: Machine learning task in which the target is a categorical variable that can be binary (e.g., patients vs. controls) or multiclass (e.g., patients vs. controls vs. unaffected siblings).

Classification threshold: Probability value that defines the criterion used to assign a predicted probability to either class (e.g., assign a participant to the patients' group, if the predicted probability given by the classifier is higher than 0.5; otherwise assign the participant to the control group).

Classifier: Machine learning algorithm that uses labeled examples to create a model that can be used for classification.

Class imbalance: Problem that arises when the classes being predicted contain a very different number of observations.

Cluster: Group of observations (e.g., participants) that have some resemblance according to a similarity metric.

Clustering: Problem of assigning observations (e.g., participants) to one or more clusters.

Clustering algorithm: Subclass of unsupervised algorithms used to discover inherent groupings in the data. Examples include k-means and hierarchical clustering.

Coefficient of determination: Goodness-of-fit measure for linear regression models that indicates the proportion of the variance in the dependent variable that is explained by the predictors. Also known as the R-squared.

Confounder: Extraneous variable thought to be associated with both the features and target variable that may bias the results.

Confusion matrix: Summary table that describes the performance of a classification model by grouping predictions into four categories: true positives (patients correctly predicted as patients), true negatives (controls correctly predicted as controls), false positives (controls incorrectly predicted as patients), and false negatives (patients incorrectly predicted as controls). Confusion matrices can be used to calculate different performance metrics, such as sensitivity or specificity.

Continuous variables: Variables with a continuous range of possible values (e.g., symptoms severity scale, age).

Convergence: State reached during iterative training in which training loss and validation loss do not improve after a certain number of iterations.

Convolutional layer: Main layer in convolutional neural networks responsible for applying one or more convolutional filters and output one or more feature maps via convolutional operations between the incoming data and the convolutional filters.

Convolutional neural network: Variation of the neural networks, typically used for the analysis of images, usually made up of convolutional, pooling, and fully connected layers. Also known as ConvNet.

Convolutional filter: Square-shaped grid of discrete numbers that stores a pattern used in the convolutional operation; it is informally referred to as feature detector.

Convolutional operation: Element-wise multiplication of some input data and a convolution filter of the same rank (e.g., 2D matrix and 2D filter) typically of a smaller size (5x5 matrix and 3x3 filter), followed by the sum of all the resulting values; the resulting single value indicates the presence or absence of the pattern conveyed in the convolutional filter. The operation starts by applying the convolutional filter to the top-left corner of the incoming data and the process is repeated again after sliding the convolutional filter to the right; this repetition takes place from right to left and from top to bottom until the whole incoming data are processed. The resulting set of single values is referred to as feature map.

Cross-validation (CV): Resampling procedure used to evaluate how well a machine learning model will generalize to new data. It involves dividing the data into k nonoverlapping subsets that are used for training and testing the model; if k equals 10, for example, it then becomes 10-fold CV; if k equals the number of observations, then it is usually referred to as leave-one-out CV.

Curse of dimensionality: Common problem in brain disorders research that arises when the number of features (i.e. dimensions) is substantially larger than the number of observations (e.g., the number of voxels is much higher than the number of participants in a typical neuroimaging study). In this case, reducing the number of features can be beneficial to alleviate computational demands, remove redundant or irrelevant information, and reduce the risk of overfitting. Also known as the as the "small n, large p" problem.

Data fusion: Family of methods for multimodal data analysis, which aims to extract joint information between modalities.

Discrimination map: Summary of the feature weights; in neuroimaging, this is typically displayed with a brain image showing the weights for each brain region or each voxel.

Decision trees: Nonparametric supervised learning algorithm that can be used for classification or regression problems. It works by splitting observations (e.g., participants) into branches based on simple binary decision rules inferred from the features.

Deep learning: Class of representation learning methods capable of learning increasingly abstract features through consecutive nonlinear transformations in a layer-wise fashion.

Deep neural network: Deep learning network comprised of several fully connected layers. Also referred to as multilayer perceptron, multilayer network, fully connected neural network, and similar variations.

Dimensionality: Number of features in a dataset. For example, a preprocessed structural brain scan may have up to 150,000 features, each one corresponding to one voxel.

Dimensionality reduction: Transforming a high-dimensionality dataset into a dataset with lower dimensionality while ensuring minimal loss of information.

Double dipping: In the context of machine learning, this term refers to the use of the same data for feature engineering and model selection. For example, selecting features based on the results of a group comparison conducted using the whole sample and then using only the significant features as input for a machine learning algorithm.

Drop-out: Regularization technique for deep neural networks in which a random number of neurons and their respective incoming and outgoing connections is temporarily removed from the network during training.

Early-stopping: Regularization technique that stops model training when training loss ceases to decrease and when the validation loss starts to increase.

Ensemble learning algorithm: Method that combines multiple classifiers to create a better-performing model.

Euclidean distance: Distance metric that calculates the shortest straight line between two points. Formally, it is defined as the square root of the sum of the squares of the differences between the corresponding coordinates of the points.

Example: Term used to refer to a vector of values corresponding to the features (e.g., extracted regional volumes) for one participant. Other terms often used with the same meaning include "instance," "sample," and "observation."

Epoch: Number of times an iterative algorithm runs through the entire dataset.

Evaluation metric: Metric used to measure the quality of a model (e.g., accuracy, sensitivity, root of mean squared error). Also known as performance metric.

Feature: Single variable used as input for the machine learning algorithm. For example, one voxel in a brain scan. Also known as independent variable in classical statistics.

Feature set: Set of features used as input data for a machine learning algorithm. For example, all voxels in a brain scan.

Feature engineering: General term for data transformations that aim to create informative feature sets before they are inputted into a machine learning algorithm (e.g., feature selection, normalization).

Feature extraction: Process of extracting features from the data (e.g., using FreeSurfer to extract cortical thickness from neuroanatomical scans).

Feature map: Output of the convolutional operation from one convolutional feature.

Feature selection: Process of selecting relevant features from a dataset for a particular task.

Feature space: Collection of all values from all features used as input data for a machine learning algorithm.

Feature vector: Vector containing the data of an observation with multiple features (in tabular data, a feature vector corresponds to a row).

Fine-tuning: Term typically used in the context of deep learning to refer to the process of adjusting the neural network's parameters to the task at hand, after being initialized with prelearned weights.

Fully connected layer: Layer in artificial neural networks in which each neuron is connected to the neurons in the previous layer. Also known as dense layers.

Gaussian distribution: Data that follow a normal distribution.

Generative model: Class of machine learning algorithms that model (i.e. generate) the distribution of individual classes and use this information to determine the probability that a new observation was created from the same mechanism that created the data from each class; it contrasts with discriminative models which learn decision boundaries between classes based on past data (e.g., Support Vector Machine, logistic regression, linear regression).

Gradient Descendent: Optimization algorithm that finds the minimum of a function by iteratively moving in the direction of steepest descent; it is the most commonly used optimization method to train deep learning networks.

Hold-out method: Simplest model evaluation method in which the data are split into training and test set once. When using the hold-out method, a common split is using 80% of data for training and the remaining 20% of the data for testing.

Hyperparameter: Property of a machine learning algorithm that needs to be defined prior to or during the learning process either by the user or by some automated method. Note that hyperparameters are different from the model's parameters, which are determined during the training of the algorithm. For example, in deep learning, the number of layers is a hyperparameter and the weights are a parameter.

Hyperparameter tuning: Adjustment of the hyperparameters of a model to improve its performance.

Hyperplane: Boundary that separates a space into two subspaces. For example, a line is a hyperplane in a 2D feature space, and a plane is a hyperplane in a 3D feature space. It is most commonly used in the context of Support Vector Machine, where a hyperplane is the boundary separating a high-dimensional space.

In-sample: Term referring to the data used to train the model; similar to training set.

Independent component analysis: Widely used blind source separation technique that attempts to find statistically latent independent components in the data.

Instance: Term used to refer to a vector of values corresponding to the features (e.g., extracted regional volumes) for one participant. Other terms often used with the same meaning include "sample," "example," and "observation."

Interpretability: Degree to which a model can be readily understood and explained. For example, in deep learning networks, it can be difficult to decipher the complex computations taking place across the several layers; conversely, linear regression models are more transparent and more interpretable.

Kernel: Similarity function for pairs of data points; used by several machine learning algorithms such as Support Vector Machine.

Kernel trick: Use of kernels to transform the data into a feature space which allows the classes to be separated more easily.

K-fold cross-validation: Type of cross-validation in which the data are randomly divided into k roughly equally groups; at each iteration of the cross-validation, all data apart from one group are used for training and the remaining group is used for testing. This process is repeated k times, each time with a different group used for testing. The number k gives name to the method, e.g., in 10-fold CV, the data are split into 10 groups.

Label: Target to be predicted by a classifier.

Leave-one-out cross-validation: Type of cross-validation in which all data apart from one subject are used for training and the remaining subject is used for testing. This process is repeated until each subject has been used for testing.

Linear regression: Regression algorithm that aims to estimate the linear relationship between features and a continuous target.

Logistic regression: Classification algorithm that assigns observations to a categorical target; this is achieved by using the logistic sigmoid function to derive a probability value that can then be mapped to two or more discrete classes.

Loss: Measure of how much the model's predictions deviate from the true targets; the lower the loss, the better the model. The value is determined by the so-called loss function.

Loss function: Method used to measure the loss in a machine learning algorithm; for example, Support Vector Machine uses the hinge loss function. Also known as cost function.

Machine learning: Area of artificial intelligence that is concerned with identifying patterns from data and uses these patterns to make predictions about unseen data.

Machine learning algorithm: Series of instructions used to generate a machine learning model. Examples include logistic regression and Support Vector Machine.

Mean absolute error: Performance metric used in regression models; it is the average of the absolute differences between predicted and actual values.

Mean squared error: Performance metric used in regression models; it is the sum of the square of the difference between the predicted and target values, divided by the number of observations.

Min-max scaling: Type of scaling/normalization that transforms features by scaling each feature to a given range.

Misclassification: Incorrect prediction made by a classifier (e.g., predicting a participant as a control when they are actually a patient, or predicting a participant as a patient when they are actually a control).

Model: In the context of machine learning, this term refers to a data structure that stores a representation of a dataset (weights and biases); models are learned from training an algorithm on a dataset.

Model selection: Process of choosing between different machine learning models; this involves choosing between different possible algorithms (e.g., Support Vector Machine, logistic regression, etc), hyperparameters (e.g., value of C for Support Vector Machine), sets of features (e.g., regional brain volumes or voxel-level volumes or both), and even feature engineering approaches (e.g., different feature scaling methods).

Multiclass classification: Classification task with more than two classes.

Multikernel learning: Supervised machine learning algorithm used for both classification and regression that learns an optimal combination of a predefined set of kernels as part of the training process.

Multimodal: Term referring to more than one type (modality) of data. It can refer to the dataset; for example, a dataset comprising genetic and neuroimaging measures. It can also refer to the analysis; for example, a machine learning analysis that combines multiple modalities of data.

Nested cross-validation: Type of cross-validation that creates an additional inner cross-validation loop, by diving the training set into training set and validation set; typically used for hyperparameter tuning.

Neuron: Building block of artificial neural networks also known as node or unit; it outputs a single value after applying an activation function (nonlinear transformation) to a weighted sum of multiple input values.

Noise: Irrelevant information or randomness in the data.

Normalization: Commonly used practice in feature engineering to address the difference in units and range across features. Examples include min-max scaling and standardization (z-score normalization). Also referred to as scaling.

Observation: Term used to refer to a vector of values corresponding to the features (e.g., extracted regional volumes) for one participant. Other terms often used with the same meaning include "instance," "example,"and "sample."

One-hot encoding: Recoding of a categorical variable into several binary individual variables, one for each level of the original variable. For example, the variable gender (male/female) would be transformed into the variables male (yes/no) and female (yes/no).

Optimization: Process of finding the optimal parameters for a machine learning algorithm with the aim of minimizing the loss function.

Out-of-sample: Term referring to the data used to test the model; similar to test set.

Overfitting: Issue occurring when a machine learning algorithm learns the training data so closely (incorporating details and noise specific to the data) that it performs very well in the training data but poorly in new data.

Parameter: Internal variable of a machine learning algorithm whose values are learned from the data to minimize the loss function.

Performance metric: Metric used to measure the quality of a model (e.g., accuracy, sensitivity, root of mean squared error). Also known as evaluation metric.

Permutation testing: Method typically used to assess the statistically significance of a performance metric.

Pretraining: Term used in the context of deep learning to refer to the unsupervised training of a network before it is trained for a particular supervised task. For example, in a face recognition task, a network can be pretrained to learn basic properties of the data regardless of what image they convey; the resulting weights are then used to initialize the same network that is now instructed to identify faces in a process known as fine-tuning.

Principal component analysis: Unsupervised method for reducing data dimensionality by finding the best linear combination of latent orthogonal (independent) components while maximizing explained variability in the data.

Pooling layer: Layer typically used in convolutional neural networks to reduce the size of the feature maps outputted by the convolutional layer.

R-squared: Goodness-of-fit measure for linear regression models that indicates the proportion of the variance in the dependent variable that is explained by the predictors. Also known as the coefficient of determination.

Random forests: Supervised learning method for classification or regression that builds an ensemble of decision trees.

Receiver operating characteristic (ROC) curve: Evaluation metric used for classification problems; corresponds to the curve that plots the true positive rate (i.e., examples that were incorrectly predicted as positive) against the false positive rate (i.e., examples that were correctly predicted as positive).

Reinforcement learning: Class of machine learning models that are trained to maximize some notion of reward via iterative trial and error.

Regression: Machine learning task in which the target is a continuous variable (e.g., symptom severity scores).

Regularization: Group of strategies that penalize model complexity to prevent overfitting.

ReLU: Short for rectified linear unit; a popular nonlinear function used as the activation function in deep learning.

Representation learning: Class of machine learning methods that aim to automatically discover the best representation of the in/out data; this is in contrast with traditional machine learning algorithms in which features are manually engineered before they are used as input features.

Residual: Difference between the observed value of the dependent variable and the predicted value.

Root of mean squared error (RMSE): Performance metric used in regression; it is the square root of the mean squared error.

Sample: Term used to refer to a vector of values corresponding to the features (e.g., extracted regional volumes) for one participant. Other terms often used with the same meaning include "instance," "example," and "observation."

Scaling: Commonly used practice in feature engineering to remove the difference in units and range across features. Examples include min-max scaling and standardization (z-score normalization). Also referred to as normalization.

Semisupervised learning: Class of machine learning models that learns from both labeled and unlabeled data.

Sensitivity: Metric that measures the model's performance at classifying positive observations correctly, i.e., proportion of patients classified as such. Also known as true positive rate or recall.

Silhouette scores: Method to measure how close each point in a cluster is to the points in its neighboring clusters. It is often used to estimate the optimal value for k in k-means clustering.

Softmax: Function typically used in the last layer of a deep neural network that provides the probabilities for each possible class in a classification problem, either binary or multiclass. These probabilities must add up to 1.

Sparsity: Data containing predominately zeros. For example, a vector containing a single 1 value and a million 0 values would be defined as sparse.

Specificity: Measures the model's performance at classifying negative observations correctly, i.e., proportion of healthy controls classified as such. Also known as true negative rate.

Standardization: Type of scaling/normalization that transforms features, so that they resemble a normal distribution. Also known as z-score normalization.

Stratified cross-validation: Type of cross-validation typically used in classification problems to ensure each class is equally represented in each fold.

Supervised learning: Class of machine learning that learns the relationship between a set of observations and known targets. Most machine learning algorithms fall in this category; examples include linear regression and Support Vector Machine.

Support Vector Machine: Classification algorithm that seeks to maximize the margin between two classes by making use of kernels to map input data to a higher dimensional space.

Target: Variable to be predicted by the machine learning model. Also known as dependent variable in classical statistics.

Test set: Portion of the data used to test the trained machine learning model, i.e., data in which the trained machine learning model is used to make predictions, which are then used to calculate the model's final performance.

Train(ing) set: Portion of the data used to fit the machine learning model, i.e., data from which the machine learning model's parameters are inferred.

Transfer learning: Machine learning technique where the learned features from a model trained on a certain task are used for a related task. For example, train a model in a large publicly available neuroimaging dataset with healthy controls and then fine-tune the learned features in another model that classifies patients and controls using similar data.

True negative rate: Measures the model's performance at classifying negative observations correctly, i.e., proportion of healthy controls classified as such. Also known as specificity.

True positive rate or recall: Metric that measures the model's performance at classifying positive observations correctly, i.e., proportion of patients classified as such. Also known as sensitivity.

Underfitting: When a machine learning algorithm can neither capture the underlying trend of the data nor generalize to new data. Contrary to overfitting, the model does not fit the data well enough and therefore will perform poorly both on the training and test sets.

Unsupervised learning: A subclass of machine learning that aims to find hidden (or latent) structures in the data. Examples include clustering and dimensionality reduction.

Validation set: Portion of data used during model training to examine on how well the current parameters generalize beyond the training set.

Voxel: Unit of graphic information that defines a point in a three-dimensional space. In neuroimaging, this space corresponds to a brain volume acquired through one of the available neuroimaging techniques (e.g., magnetic resonance imaging); each voxel represents the average signal measured at a given location within the brain volume. A typical brain volume contains thousands of voxels.

Weight: Measure of the contribution of a given feature to a model's prediction; a weight of 0 indicates that the corresponding feature does not contribute to the prediction.

Z-score normalization: Type of scaling that transforms features, so that they resemble a normal distribution. Also known as standardization.