Machine Learning best practices

Focus on data :

* Characteristics
* Class balance
* Number of samples
  + Data augmentation techniques can be used
  + Always visualize the sample before training
* Check the task to be performed considering all the above
* Divide data into train, validation and test groups
  + Usually 70%, 20%, 10%
  + Attention to avoid cross-reference (same subject on several groups)
  + The test data should be used only after the model is ready
* Apply normalization
  + If performing CNN fine-tuning check the normalization to be performed (the one performed on original training)
* Cross-validation
  + If the number of samples is small use cross-validation

Deep CNN:

* Check the results visualizing heat maps or gradients
* The CNN can “learn” a different task than the one you are interested (but correlated to it)

Training:

* Plan the training approach
* Reading several files during training is time-consuming (each NifTi or Dicom). Convert the files to one big file (h5) or to numpy files.
* Run the training using few data on the local computer before uploading to the cloud
  + Check the input data and the network
* If using extended 2D (sequencial slices from the same subject) check if all input slices are from the same acquisition
* Save the statistics data and the network weights
* Perform the coding in a way you can reproduce past experiments

<http://karpathy.github.io/2019/04/25/recipe/>