FEATURE EXTRACTION FRAMEWORK

Data mining is to be based on a number of brain structure volume features, which are automatically extracted from patient MRI scans[1].

How it works

Single NIfTI volumes of the brain are first partitioned into three classes: grey matter, white matter and background. This procedure also incorporates an approximate image alignment step and a correction for image intensity non-uniformities. This procedure is done using the *Segment*[2] tool from within the *SPM12*[3] software, which runs within the MATLAB[4] programming language.

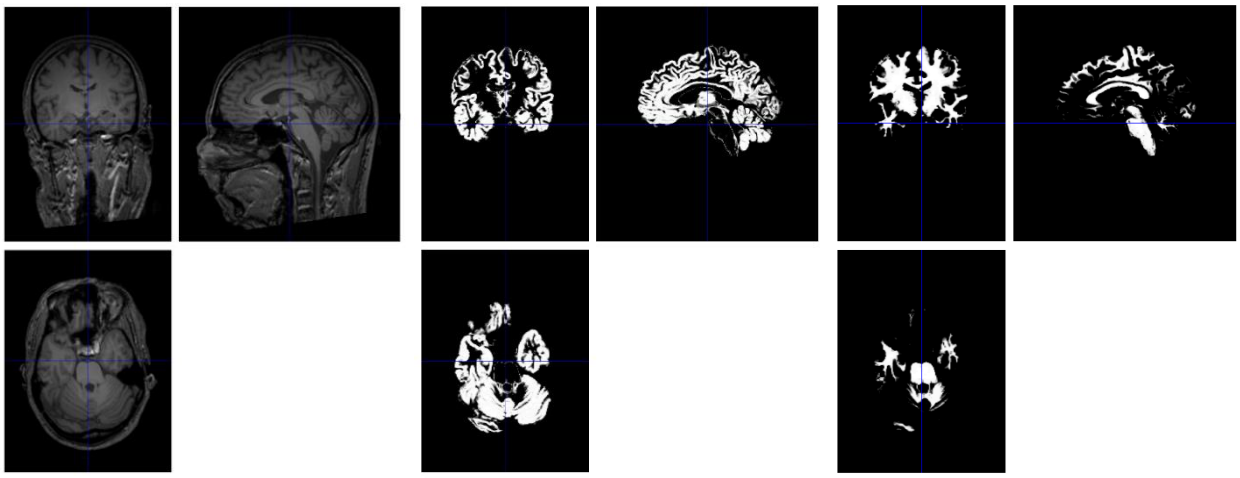

Figure 1. An original T1-weighted MRI scan (left), along with automatically extracted grey (middle) and white matter (right) tissue maps. The tissue maps encode the probability of each tissue type (given the model and data).

Tissue atlases, precomputed from the training data (see later), are then spatially registered with the extracted grey and white matter maps, using the *Shoot*[5] tool from *SPM12*. The warps estimated from this registration step are then used to project other pre-computed image data in to alignment with the original scans (and their grey and white matter maps).

The rules of probability[6] are then used to combine the various images to give a probabilistic label map for each brain structure. These probabilities are summed for each structure, to give probabilistic volume estimates. These estimates serve as features for data mining. Optionally, the method also allows maximum probability label maps to be saved.

---

[1] http://xkcd.com/1425/

[2] Ashburner J, Friston KJ. Unified segmentation. Neuroimage. 2005 Jul 1;26(3):839-51.

[3] http://www.fil.ion.ucl.ac.uk/spm/software/spm12/

[4] http://uk.mathworks.com/products/matlab/

[5] Ashburner, John, and Karl J. Friston. "Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation." NeuroImage 55.3 (2011): 954-967.

[6] For example, for labelling the hippocampus, we have: P(structure=hippocampus) = P(structure=hippocampus | tissue=grey) × P(tissue=grey) + P(structure=hippocampus | tissue=white) × P(tissue=white) + P(structure=hippocampus | tissue=other) × P(tissue=other).

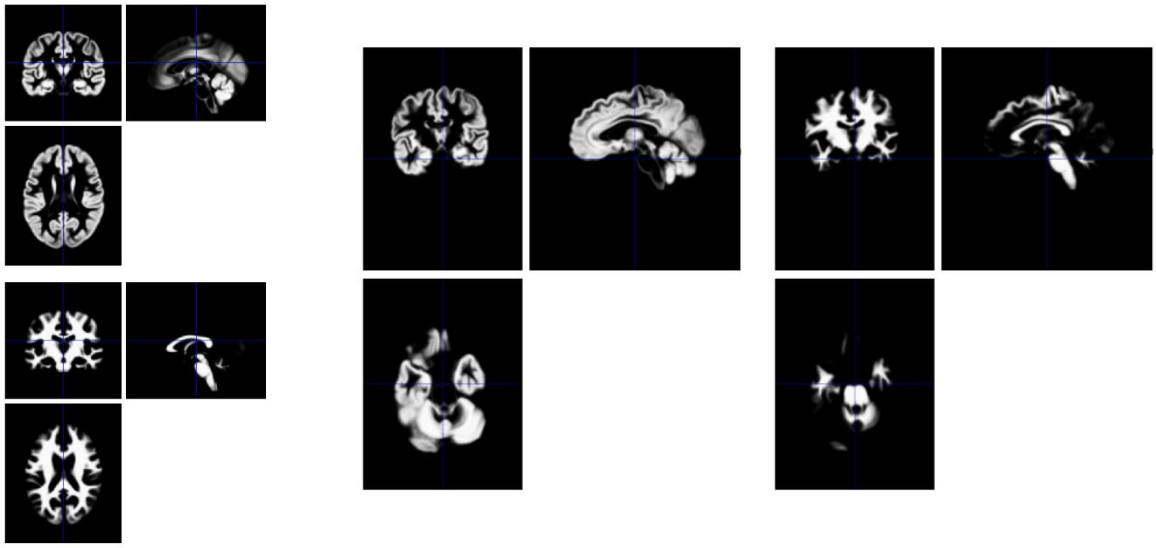

Figure 2. Grey and white matter from the original tissue atlases (left), along with registered versions (middle and right).

Motivation for approach

Needs to be fully automated and robust

The datasets are too large for manual interventions to be involved. Also, privacy concerns prohibit visual inspection of the image data.

Small number of features

The format of the current database framework (comma separated value text files, interrogation via SQL and privacy preserving requirements) precludes the use of many modern machine learning approaches suitable for images. For this reason, the Mbytes of information in each patient's image is reduced to a small number of features, which are more suitable for privacy-preserving data mining. This allows sufficient statistics, such as mean vectors and covariance matrices accumulated from many patients' data, to be passed easily from each hospital site.

Features interpretable by clinicians

The proposed rule-based data mining strategies emphasizes interpretability over accuracy. For this reason, the image features that enter into the data mining ought to be interpretable by clinicians. The features chosen are the volumes of various brain structures. This requires a robust and fully automated way to segment (label) brain scans into several different regions. Alternative feature types could have included cortical thickness of different regions, but these can usually not be obtained in a fully automated way. Some manual intervention is often required.

Needs to be relatively fast

The feature extraction procedure is to be applied to several thousand scans, so computational speed is important. The privacy-preserving nature of the work means that feature extraction can only be done at each hospital site, which precludes the use of the HBP supercomputing facilities. The current state-of-the-art in automated brain labelling are very computationally expensive, relying on estimating the detailed

nonlinear registration of multiple images from the training set with each patient's image. Each of these registration steps may take several hours[7]. Instead, the approach used in SP8 involves a single image alignment between the patient's image and a pre-defined template derived from averaged data[8] from multiple training images. The approach for propagating the labels is also much simpler (and faster) than the state-of-the-art methods.

Needs to be flexible for different MRI contrasts

MRI scans come in many different types, with a variety of image "contrasts". In some types of scans, the grey matter in the brain appears darker than the white matter, whereas in other types, the grey matter appears brighter. There are also many subtle variations due to artifacts from different scanners etc. State-of-the-art brain labelling methods usually rely on having manually labelled training scans with exactly the same properties as the patient scans to be segmented. Within a hospital environment, there are hundreds of variations in the types of scans acquired. It is not feasible to obtain labelled training data for all these varieties - simply because manually labelling scans is very time consuming and tedious, yet requires a lot of expertise. Instead of basing the labelling procedure on the raw scans, the adopted approach first segments the images into a different tissue classes, and bases the labelling procedure on these. Many automated methods are available for this tissue classification procedure, and they can usually adapt to a variety of MRI contrasts. The tissue classification method chosen is widely used, well supported and very familiar to several SP8 team members[9]. It is also under constant development, and can be extended to suit the needs of SP8.

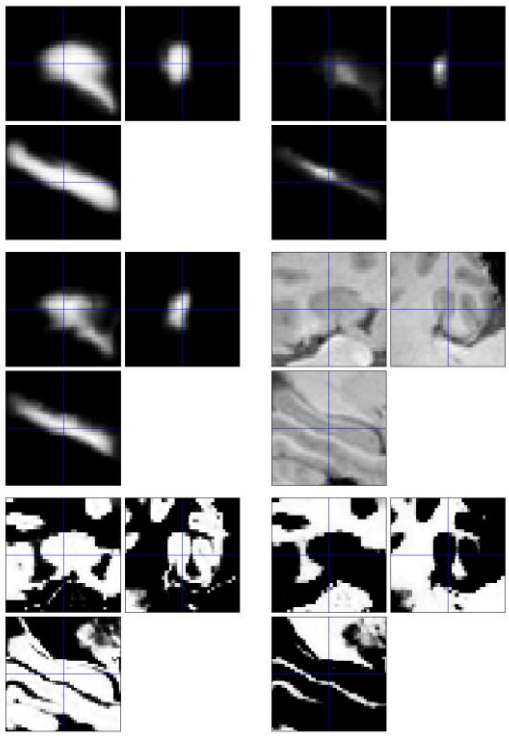

Figure 3. Detail of the label propagation procedure, in the region around the right hippocampus. The top row shows a map of the probability of each voxel being part of the hippocampus, given that it has been labelled as grey matter (left) or white matter (right). The second row shows the probability of hippocampus given that a voxel has not been labelled as grey or white matter, followed by a detail from the original scan. The bottom row shows the grey and white matter tissue probability maps.

Needs to be relatively accurate

Many label propagation approaches use the ANTs software[10] for image alignment, as it performed best in an extensive comparison[11]. In a separate evaluation, the image registration adopted for this work[12] outperformed all methods from that comparison. It also runs an order of magnitude faster than ANTs. Label fusion can be made much more accurate for scans with the same properties as the training data. Some information that could be useful for label fusion is inevitably lost through partitioning the scans into

---

[7] https://masi.vuse.vanderbilt.edu/workshop2012/index.php/Workshop_Proceedings

[8] Ashburner J, Friston KJ. Computing average shaped tissue probability templates. Neuroimage. 2009 Apr 1;45(2):333-41.

[9] Ashburner J, Friston KJ. Unified segmentation. Neuroimage. 2005 Jul 1;26(3):839-51.

[10] http://stnava.github.io/ANTs/

[11] Klein, Arno, et al. "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration." Neuroimage 46.3 (2009): 786-802.

[12] Ashburner, John, and Karl J. Friston. "Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation." NeuroImage 55.3 (2011): 954-967.

different tissue types. Even so, by using a principled probabilistic fusion method, the computed volumes are still reasonably accurate.

Scope of data

The range of image types suitable for accurate labelling is restricted. For reliable results, scans should be high resolution (1.5 mm isotropic or better), and it should be possible to visually distinguish between grey and white matter in the brain. The brain should also be visually distinct from surrounding tissues. A relatively small amount of intensity non-uniformity artifact should not impact the method. The volumetric T1-weighted scans collected for research purposes and clinical scans with similar properties can usually be easily labelled. Images of lower resolution can still be labelled, but the accuracies of the volume estimates are much lower. Artifacts that are not modelled, such as motion artifact, also degrade the accuracy of the features.

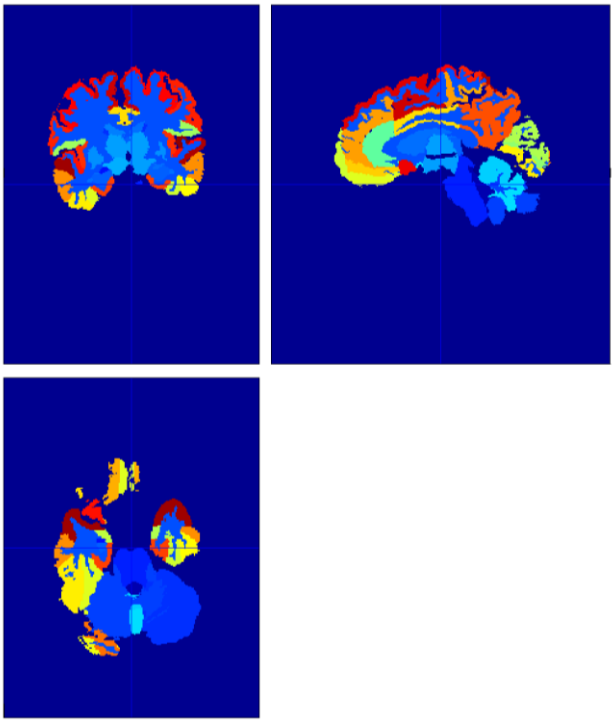

Figure 4. An automatically labelled image, showing the most probable structure labels (according to the approach).

Scans must be of brains. If the algorithm is presented with scans of other organs, it will not identify them as such and give meaningless results. Because the procedure begins with an affine registration, using a local optimization procedure, it can be susceptible to failure if the centre of the brain is too far from the centre of the magnetic field of the scanner. Providing the anterior commissure is within about 5 cm from the centre of the magnetic field, then a good solution is invariably obtained. Scan data must be presented as NIfTI format[13], after having been converted from DICOM[14] in a way that preserves the orientation and positional information from the headers (correctly using the *Image Orientation Patient*, *Image Position Patient* and *Pixel Spacing* tags). Data must be volumetric with good coverage of the brain. Brain structures that fall outside the field of view will be assigned volumes of zero. Brain structures that lie at the edge of the field of view will be assigned volumes based on how much of the structure is included within the scan.

The approach works on a single volumetric patient scan. It does not yet have the ability to combine the multiple scans typically acquired from a patient during a scanning session.

[13] http://nifti.nimh.nih.gov/nifti-1/

[14] http://dicom.nema.org/standard.html

Training data

The training labels were derived from the *MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling*[15]. These data were released under the Creative Commons Attribution-NonCommercial[16] (CC BY-NC) with no end date. Users should credit the T1-weighted MRI scans as originating from the OASIS project [17] and the labeled data as "provided by *Neuromorphometrics, Inc.* (http://Neuromorphometrics.com/) under academic subscription". These references should be included in all workshop and final publications. The brain labelling protocol[18] was developed by Jason Tourville and Ruth Carpenter at Neuromorphometrics, following consultation with the neuroscience community. Brain labels were provided for the scans of 30 subjects from the OASIS dataset (20 female).

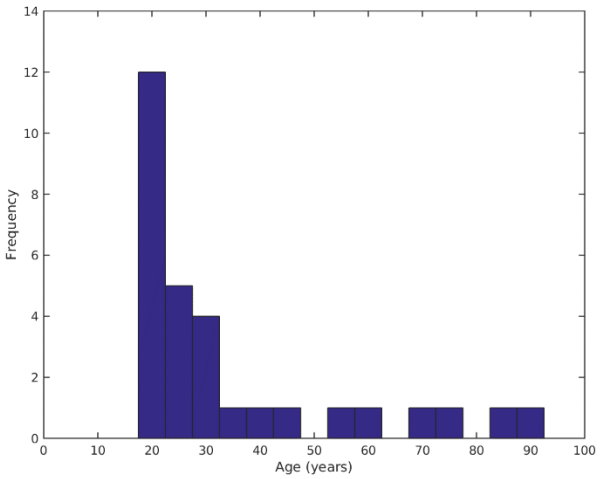

Figure 5. The distribution of ages of subjects in the training set, shown as a histogram.

Each subject had a T1-weighted MRI scan of isotropic 1 mm resolution, which had been anonymized by removing the face before being released. The dimensions of each volume ranged between 256×256×264 and 256×256×264.

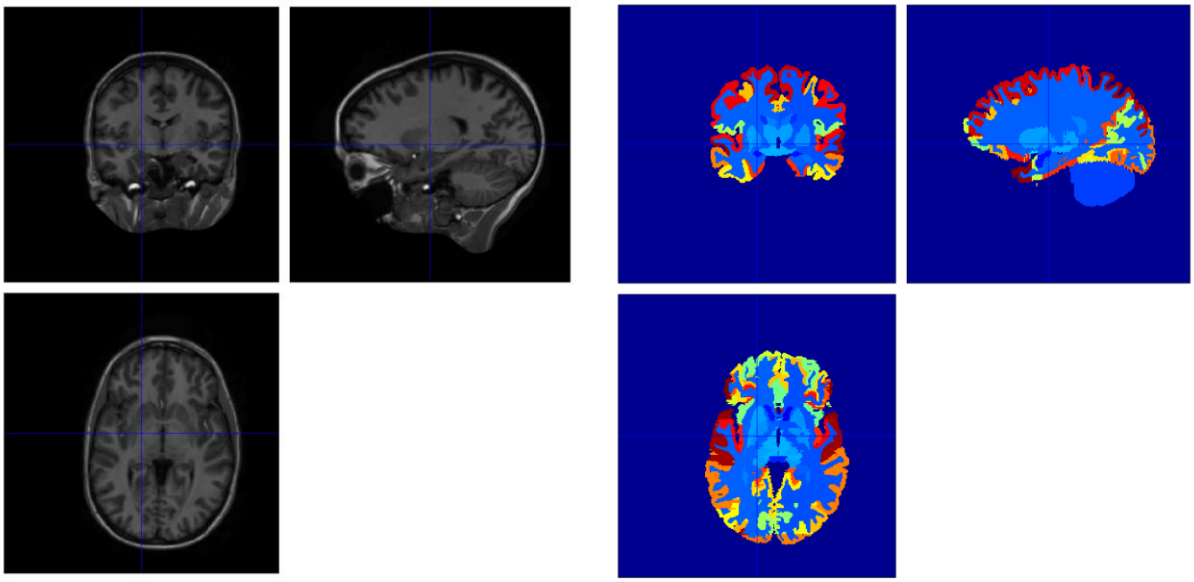

Figure 6. One of the training scans, with the MRI (left) and manually defined labels (right).

[15] https://masi.vuse.vanderbilt.edu/workshop2012/index.php/Challenge_Details
[16] https://creativecommons.org/licenses/by-nc/3.0/
[17] http://www.oasis-brains.org/
[18] http://braincolor.mindboggle.info/docs/BrainCOLOR_cortical_parcellation_protocol.pdf

Labels are coded as defined in this table:

4 3rd Ventricle	136 Right lateral orbital gyrus
11 4th Ventricle	137 Left lateral orbital gyrus
23 Right Accumbens Area	138 Right middle cingulate gyrus
30 Left Accumbens Area	139 Left middle cingulate gyrus
31 Right Amygdala	140 Right medial frontal cortex
32 Left Amygdala	141 Left medial frontal cortex
35 Brain Stem	142 Right middle frontal gyrus
36 Right Caudate	143 Left middle frontal gyrus
37 Left Caudate	144 Right middle occipital gyrus
38 Right Cerebellum Exterior	145 Left middle occipital gyrus
39 Left Cerebellum Exterior	146 Right medial orbital gyrus
40 Right Cerebellum White Matter	147 Left medial orbital gyrus
41 Left Cerebellum White Matter	148 Right postcentral gyrus medial segment
44 Right Cerebral White Matter	149 Left postcentral gyrus medial segment
45 Left Cerebral White Matter	150 Right precentral gyrus medial segment
46 CSF	151 Left precentral gyrus medial segment
47 Right Hippocampus	152 Right superior frontal gyrus medial segment
48 Left Hippocampus	153 Left superior frontal gyrus medial segment
49 Right Inf Lat Vent	154 Right middle temporal gyrus
50 Left Inf Lat Vent	155 Left middle temporal gyrus
51 Right Lateral Ventricle	156 Right occipital pole
52 Left Lateral Ventricle	157 Left occipital pole
55 Right Pallidum	160 Right occipital fusiform gyrus
56 Left Pallidum	161 Left occipital fusiform gyrus
57 Right Putamen	162 Right opercular part of the inferior frontal gyrus
58 Left Putamen	163 Left opercular part of the inferior frontal gyrus
59 Right Thalamus Proper	164 Right orbital part of the inferior frontal gyrus
60 Left Thalamus Proper	165 Left orbital part of the inferior frontal gyrus
61 Right Ventral DC	166 Right posterior cingulate gyrus
62 Left Ventral DC	167 Left posterior cingulate gyrus
63 Right vessel	168 Right precuneus
64 Left vessel	169 Left precuneus
69 Optic Chiasm	170 Right parahippocampal gyrus
71 Cerebellar Vermal Lobules I-V	171 Left parahippocampal gyrus
72 Cerebellar Vermal Lobules VI-VII	172 Right posterior insula
73 Cerebellar Vermal Lobules VIII-X	173 Left posterior insula
75 Left Basal Forebrain	174 Right parietal operculum
76 Right Basal Forebrain	175 Left parietal operculum
100 Right anterior cingulate gyrus	176 Right postcentral gyrus
101 Left anterior cingulate gyrus	177 Left postcentral gyrus
102 Right anterior insula	178 Right posterior orbital gyrus
103 Left anterior insula	179 Left posterior orbital gyrus
104 Right anterior orbital gyrus	180 Right planum polare
105 Left anterior orbital gyrus	181 Left planum polare
106 Right angular gyrus	182 Right precentral gyrus
107 Left angular gyrus	183 Left precentral gyrus
108 Right calcarine cortex	184 Right planum temporale
109 Left calcarine cortex	185 Left planum temporale
112 Right central operculum	186 Right subcallosal area
113 Left central operculum	187 Left subcallosal area
114 Right cuneus	

115 Left cuneus	190 Right superior frontal gyrus
116 Right entorhinal area	191 Left superior frontal gyrus
117 Left entorhinal area	192 Right supplementary motor cortex
118 Right frontal operculum	193 Left supplementary motor cortex
119 Left frontal operculum	194 Right supramarginal gyrus
120 Right frontal pole	195 Left supramarginal gyrus
121 Left frontal pole	196 Right superior occipital gyrus
122 Right fusiform gyrus	197 Left superior occipital gyrus
123 Left fusiform gyrus	198 Right superior parietal lobule
124 Right gyrus rectus	199 Left superior parietal lobule
125 Left gyrus rectus	200 Right superior temporal gyrus
128 Right inferior occipital gyrus	201 Left superior temporal gyrus
129 Left inferior occipital gyrus	202 Right temporal pole
132 Right inferior temporal gyrus	203 Left temporal pole
133 Left inferior temporal gyrus	204 Right triangular part of the inferior frontal gyrus
134 Right lingual gyrus	205 Left triangular part of the inferior frontal gyrus
135 Left lingual gyrus	206 Right transverse temporal gyrus
	207 Left transverse temporal gyrus