# Brain Extraction from Normal and Pathological Images: A Joint PCA/Image-Reconstruction Approach

Xu Han[a,*], Roland Kwitt[b], Stephen Aylward[c], Bjoern Menze[d], Alexander Asturias[e], Paul Vespa[f], John Van Horn[e], Marc Niethammer[a]

[a]*Department of Computer Science, UNC Chapel Hill, USA*
[b]*Department of Computer Science, University of Salzburg, Austria*
[c]*Kitware Inc., USA*
[d]*Department of Computer Science, Technical University of Munich*
[e]*USC Institute of Neuroimaging and Informatics*
[f]*David Geffen School of Medicine, UCLA Medical Center*

**Abstract**

Brain extraction from 3D medical images is a common pre-processing step. A variety of approaches exist, but they are frequently only designed to perform brain extraction from images *without* strong pathologies. Extracting the brain from images exhibiting strong pathologies, for example, the presence of a brain tumor or of a traumatic brain injury (TBI), is challenging. In such cases, tissue appearance may substantially deviate from normal tissue appearance and hence violates algorithmic assumptions for standard approaches to brain extraction; consequently, the brain may not be correctly extracted.

This paper proposes a brain extraction approach which can explicitly account for pathologies by jointly modeling normal tissue appearance and pathologies. Specifically, our model uses a three-part image decomposition: (1) normal tissue appearance is captured by a statistical appearance model (via principal component analysis (PCA)), (2) pathologies are captured via a total variation term, and (3) the skull and surrounding tissue is captured by a sparsity term. Due to its convexity, the resulting decomposition model allows for efficient optimization. Decomposition and image registration steps are alternated to allow statistical modeling of normal tissue appearance in a fixed atlas coordinate system. As a beneficial side effect, the decomposition model allows for the identification of potentially pathological areas and the reconstruction of a quasi-normal image in atlas space.

We demonstrate the effectiveness of our approach on four datasets: the publicly available IBSR and LPBA40 datasets which show normal image appearance, the BRATS dataset containing imaging with brain tumors, and a dataset containing clinical TBI images. We compare the performance with other popular brain extraction models: ROBEX, BET, BSE and a recently proposed deep learning approach. Our model performs better than these competing approaches on all four datasets. Hence, our approach is an effective method for brain extraction for a wide variety of images with high-quality brain extraction results.

*Keywords:* Brain Extraction, Image Registration, PCA, Total-Variation, Pathology

## 1. Introduction

Brain extraction[1] from volumetric magnetic resonance (MR) or computed tomography images is a common pre-processing step in neuroimaging as it allows to spatially focus further analyses on the areas of interest. The most straightforward approach to brain extraction is by manual expert delineation. Unfortunately, such expert segmentations are time consuming and very labor intensive and therefore not suitable for large-scale imaging studies. Moreover, brain extraction is complicated by differences in image acquisitions and the presence of tumors

*Corresponding author
*Email address:* xhs400@cs.unc.edu (Xu Han)

[1]We avoid the commonly used term skull stripping. We are typically interested in removing more than the skull from an image and are instead interested only in retaining the parts of an image corresponding to the brain.

and other pathologies that add to inter-expert segmentation variations.

Many methods have been proposed to replace manual delineation by automatic brain extraction. In this paper, we focus on and compare with the following four widely-used or recently published brain extraction methods, which cover a wide range of existing approaches:

- *Brain Extraction Tool (BET):* BET [1] is part of FSL (FMRIB Software Library) and is a widely used method for brain extraction. BET first finds a rough threshold based on the image intensity histogram, which is then used to estimate the center-of-gravity (COG) of the brain. Subsequently, BET extracts the brain boundary via a surface evolution approach, starting from a sphere centered at the estimated COG.

- *Brain Surface Extraction (BSE):* BSE [2] is part of BrainSuite[2]. BSE uses a sequence of low-level operations to isolate and classify brain tissue within T1-weighted MR images. Specifically, BSE uses a combination of diffusion filtering, edge detection and morphological operations to segment the brain.

- *Robust Learning-based Brain Extraction System (ROBEX):* ROBEX [3] is another widely used method which uses a random forest classifier as the discriminative model to detect the boundary between the brain and surrounding tissue. It then uses an active shape model to obtain a plausible result.

- *Deep Brain Extraction:* We additionally compare against a recently proposed deep learning approach for brain extraction [4] which uses a 3D convolutional neural network (CNN) trained on normal images and images with mild pathologies. Specifically, it is trained on IBSR v2.0[3], LPBA40 [5] and OASIS [6] datasets. We use this model as is without additional fine-tuning for other datasets.

In addition to these methods, many other approaches have been proposed. For example, Segonne et al. [7] proposed a hybrid approach which combines watershed segmentation with a deformable surface model. Watershed segmentation is used to obtain an initial estimate of the brain region which is then refined via a surface evolution process. Another recently proposed method is the Brain Extraction Based on non-local Segmentation Technique (BEaST) [8] approach. BEaST is inspired by patch-based segmentation techniques. In particular, it identifies brain patches by assessing candidate patches based on their sum-of-squared-difference (SSD) distance to known brain patches. However, as mentioned by the authors, tumors and lesions may create problems for BEaST. 3dSkullStrip is part of the AFNI (Analysis of Functional Neuro Images) package [9]. It is a modified version of BET. In contrast to BET, it uses image data inside and outside the brain during the surface evolution to avoid segmenting the eyes and the ventricles. Lastly, Multi-Atlas Skull-Stripping (MASS) [10] is an another approach for brain segmentation which has shown excellent performance on normal (IBSR, LBPA40) and close to normal (OASIS) image datasets. One of its main disadvantages is its runtime.

Even though all these brain extraction methods exist and are regularly used, a number of challenges for automatic brain extraction remain:

- Many methods show varying performances on different datasets due to differences in image acquisition (e.g., slightly different sequences or differing voxel sizes). Hence, a method which can reliably extract the brain from images acquired with a variety of different imaging protocols would be desirable.

- Most methods only work for images which appear normal or show very minor pathologies. Strong pathologies, however, may induce strong brain deformations or strong localized changes in image appearance, which can impact brain extraction. For example, for methods based on registration, the accuracy of brain extraction will depend on the accuracy of the registration, which can be severely affected in the presence of pathologies. Hence, a brain extraction method which works reliably even in the presence of pathologies (such as brain tumors or traumatic brain injuries) would be desirable.

Inspired by the low-rank + sparse (LRS) image registration framework proposed by Liu et

[2] `http://brainsuite.org`

[3] Available at `https://www.nitrc.org/projects/ibsr`. This is a different dataset than the IBSR dataset that we use in this paper.

al. [11] and our prior work on image registration in the presence of pathologies [12], we propose a brain extraction approach which can tolerate image pathologies (by explicitly modeling them) while retaining excellent brain extraction performance in the absence of pathologies. Specifically, in our registration approach for pathological images [12], we decompose an image into a pathological and a quasi-normal part. The quasi-normal part is designed to be close to the appearance space of normal images (as captured by an appearance model via PCA of a set of normal images in atlas space). This quasi-normal image is then used for the registration to atlas space, thereby largely avoiding registration issues caused by the presence of pathologies. The decomposition and registration steps are repeated to convergence.

Our *proposed* brain extraction approach makes use of a similar alternating decomposition and registration strategy. However, the decomposition splits an image into *three* different parts: (i) a quasi-normal part which is close to the PCA-space of a set of normal brains (for example, the brains extracted in the OASIS dataset), (ii) a total variation (TV) part which captures pathologies inside the brain, and (iii) a sparse part which captures regions outside the brain (including the skull, for example). The TV and the sparse parts are locally weighted so that the TV part of the decomposition only captures pathologies *inside the brain* and the sparse part captures any non-brain regions *outside the brain*. These weightings and the quasi-normal appearance are effectively captured in atlas space and the image will automatically be atlas-aligned.

### *1.1. Contributions*

The contributions of our work are as follows:

- *(Robust) brain extraction:* Our method can reliably extract the brain from a wide variety of images. We achieve state-of-the-art results on images with normal appearance, slight, and strong pathologies. Hence our method is a generic brain extraction approach.
- *Pathology identification:* Our method captures pathologies via a total variation term in the decomposition model.
- *Quasi-normal estimation:* Our model allows the reconstruction of a quasi-normal image, which has the appearance of a corresponding pathology-free or pathology-reduced image. This quasi-normal image also allows for accurate registrations to, e.g., a normal atlas.
- *Extensive validation:* We extensively validate our approach on four different datasets, two of which exhibit strong pathologies. We demonstrate that our method achieves state-of-the-art results on all these datasets using a *single* fixed parameter setting.
- *Open source:* Our approach is available as open-source software.

### *1.2. Organization*

The remainder of the paper is organized as follows. Section 2 introduces the datasets that we use and discusses our proposed model, including the pre-processing, the decomposition and registration, and the post-processing procedures. Section 3 presents experimental results on 3D MRI datasets demonstrating that our method consistently performs better than BET, BSE, ROBEX, and the deep learning approach for all four datasets. Section 4 concludes the paper with a discussion and an outlook on possible future work.

## 2. Materials and Methods

### *2.1. Datasets*

We use the ICBM 152 non-linear atlas (2009a) [13] as our normal control atlas. ICBM 152 is a 1x1x1 mm template with 197×233×189 voxels, obtained from T1-weighted MRIs. Importantly, it also includes the brain mask. As the ICBM 152 atlas image itself contains the skull, we can obtain a *brain-only* atlas simply by applying the provided brain mask.

We use five different datasets for our experiments. Specifically, we use one (OASIS, see below) of the datasets to build our PCA model and the remaining four to test our brain extraction approach.

**OASIS.** We use images from the Open Access Series of Imaging Studies (OASIS)[4] [6] to build the PCA model for our brain extraction approach. The OASIS cross-sectional MRI dataset consists of 416

[4]The OASIS data is available online at `http://www.oasis-brains.org`.

sagittal T1-weighted MRI scans from subjects between 18 and 96 years of age. In this data corpus, 100 of the subjects over 60 years old have been diagnosed with very mild to mild Alzheimer's disease (AD). The original scans were obtained with in-plane resolution $1 \times 1$ mm ($256 \times 256$), slice thickness = 1.25 mm and slice number = 128. For each subject, a gain-field corrected atlas-registered image and its corresponding masked image in which all non-brain voxels have been assigned an intensity of zero are available. Each image is resampled to $1 \times 1 \times 1$ mm isotropic voxels and is of size $176 \times 208 \times 176$.

We randomly pick 100 images and their brain masks to build our PCA model of the brain. Specifically, we register the brain-masked images to the brain-masked ICBM atlas using a B-spline registration. We use `NiftyReg` [14] to perform the B-spline registration with normalized cross-correlation (NCC) as similarity measure. To normalize image intensities, we apply an affine transform to the image intensities of the warped images so that the 1st percentile is mapped to 0.01 and 99th percentile is mapped to 0.99 and then crop the image intensities to be within $[0, 1]$. We then perform PCA on the now registered and normalized images and retain the top 50 PCA modes for our statistical appearance model. This is similar to an active appearance model [15].

We *evaluate* our approach on four datasets, which all provide brain masks. Although, in our study, we focus on T1-weighted images only, our model can be applied to other modalities as long as the PCA model is also built from data acquired by the same modality. The datasets we use for validation are described below.

**IBSR.** The Internet Brain Segmentation Repository (IBSR)[5] contains MR images from 20 healthy subjects of age $29.1 \pm 4.8$ years including their manual brain segmentations, provided by the Center for Morphometric Analysis at Massachusetts General Hospital. All coronal 3D T1-weighted spoiled gradient echo MRI scans were acquired using two different MR systems: ten scans (4 males and 6 females) were performed on a 1.5T Siemens Magnetom MR system (with in-plane resolution of $1 \times 1$ mm and slice thickness of 3.1 mm); another ten scans (6 males and 4 females) were acquired from a 1.5T General Electric Signa MR system (with in-plane resolution of $1 \times 1$ mm and slice thickness of 3 mm). Segmentations of the brain images into white matter, grey matter and cerebrospinal fluid (CSF) are provided. While, in principle, the union of the segmentations of white matter, grey matter and CSF should represent the desired brain mask, this is not exactly the case (see Fig. 1). To alleviate this issue for each segmentation, we use morphological closing to fill in remaining gaps and holes inside the brain mask and, in particular, to disconnect the background inside the brain mask from the surrounding image background. The structuring element we use for closing is a ball with a radius of 1 voxel using an 18-connected neighborhood[6]. We then find the connected component for the background and consider its complement the brain mask. Fig. 1 shows the pre-processing result after these refinement steps, compared to the original IBSR segmentation (i.e., the union of whiter matter, grey matter, and the CSF).

(a) (b) (c)

**Figure 1:** Example coronal slice of (a) an IBSR MR brain image, (b) the corresponding original IBSR brain segmentation (i.e., union of white matter, grey matter and CSF) and (c) the refined brain segmentation result.

**LPBA40.** The LONI Probabilistic Brain Atlas (LPBA40) dataset of the Laboratory of Neuro Imaging (LONI) [5] consists of 40 normal human brain volumes. LPBA40 contains images of 20 males and 20 females of age $29.20 \pm 6.30$ years. Coronal T1-weighted images with slice thickness 1.5 mm were acquired using a 1.5T GE system. Images for 38 of the subjects have in-plane resolution of $0.86 \times 0.86$ mm; the images for the remaining two subjects have a resolution of $0.78 \times 0.78$ mm. A manually segmented brain mask is available for each image.

**BRATS:** We use twenty representative image volumes of low and high grade glioma patients from

[5]The IBSR data is freely available online at `https://www.nitrc.org/projects/ibsr`.

[6]The 18-voxel connectivity is also used for other morphological operations in this paper

the Brain Tumor Segmentation (BRATS 2016) dataset [16] that include cases with large tumors, deformations, or resection cavities. We do not use the BRATS images available as part of the BRATS challenge as these have already been pre-processed (i.e., brain-extracted and co-registered). Instead, we obtain a subset of twenty of the originally acquired images. The BRATS dataset is challenging as the images were acquired with different clinical protocols and various different scanners from multiple ($n = 19$) institutions. Furthermore, the BRATS images have comparatively low resolution and some of them contain as few as 25 axial slices (with slice thickness as large as 7mm). The in-plane resolutions vary from 0.47×0.47 mm to 0.94×0.94 mm with image grid sizes between 256×256 and 512×512 pixels. We manually segment the brain in these images to obtain an accurate brain mask for validation.

**TBI.** Finally, we use our own Traumatic Brain Injury (TBI) dataset which contains 8 TBI images as well as manual brain segmentations. These images have been resampled to 1×1×1 mm isotropic voxel size with image size between $192 \times 228 \times 170$ and $256 \times 256 \times 176$. Segmentations are available for healthy brain, hemorrhage, edema and necrosis. To generate the brain masks, we always use the union of healthy tissue and necrosis. We also include hemorrhage and edema if they are contained within healthy brain tissue.

Fig. 2 shows example images from each dataset to illustrate image variability. IBSR and LPBA40 contain images from normal subjects and include large portions of the neck; BRATS has very low out-of-plane resolution; and the TBI dataset contains large pathologies and abnormal skulls.

### *2.2. Review of related models*

As mentioned previously, brain extraction is challenging because it requires the identification of all non-brain tissue which can be highly variable (cf. Fig. 2). Our brain extraction approach is based on image alignment to an atlas space where a brain mask is available. However, this requires a reliable registration approach which can tolerate variable image appearance as well as pathologies (i.e., brain tumors, traumatic brain injuries, or general head injuries resulting in skull deformations and fractures). In both cases, no one-to-one mapping between image and atlas space may be available and a direct application of standard image similarity measures for image registration may be inappropriate.

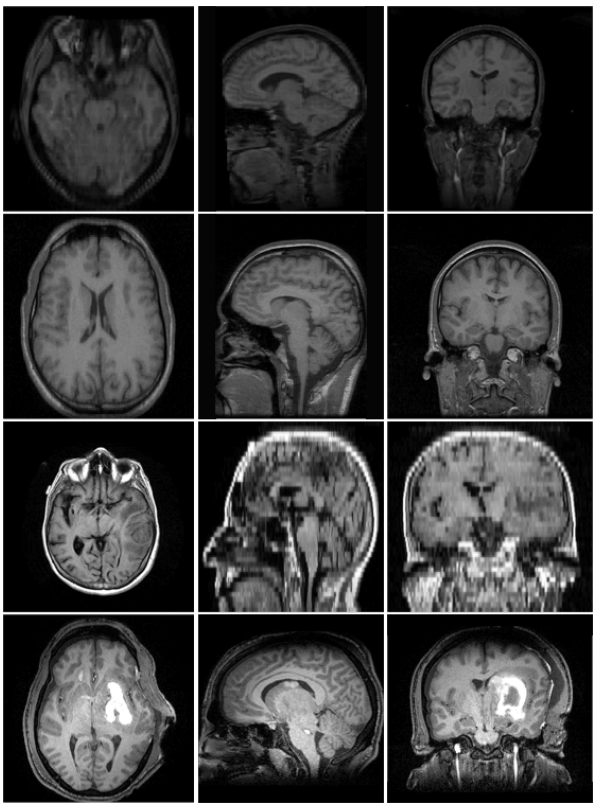


**Figure 2:** Illustration of image appearance variability on a selection of images from each (evaluation) database. From *top* to *bottom*: IBSR, LPBA40, BRATS and TBI.

A variety of approaches have been proposed to address the registration of pathological images. For example, cost function masking [17] and geometric metamorphosis [18] exclude the pathological regions when measuring image similarities. However, these approaches require prior segmentations of the pathologies, which can be non-trivial and/or labor intensive. A conceptually different approach is to learn the normal image appearance from population data and to estimate a quasi-normal image from a pathological image. Then, the quasi-normal image can be used for registration [19]. The low-rank + sparse (LRS) image registration framework, proposed by Liu et al. [11], follows this idea by iteratively registering the low-rank components from the input images to the atlas, and then re-computes the low-rank components. After convergence, the image is well-aligned with the atlas.

Our proposed brain extraction model builds upon our previous PCA-based approach for pathological

image registration [12] which, in turn, builds upon and removes many shortcomings of the low-rank + sparse approach of Liu et al. [11]. We therefore briefly review the low-rank + sparse technique in Sec. 2.2.1 and the PCA approach for pathological image registration in Sec. 2.2.2. We discuss our proposed model for brain extraction in Sec. 2.3.

#### *2.2.1. Low-Rank + Sparse (LRS)*

The standard LRS decomposition requires minimization of the energy

$$E(L,S) = \|L\|_* + \lambda\|S\|_1, \quad \text{s.t.} \quad D = L + S \ , \ (1)$$

where $D$ is a given data matrix, $\|\cdot\|_*$ is the nuclear norm (i.e., a convex approximation for the matrix rank), and $\lambda > 0$ weighs the contribution of the sparse part, $S$, in relation to the low-rank part $L$. In imaging applications, $D$ contains all the (vectorized) images: each image is represented as a column of $D$. The low-rank term captures common information across columns. The sparse term, on the other hand, captures uncommon/unusual information. As Eq. (1) is convex, minimization results in a global minimum.

In practice, applying the LRS model requires forming the matrix $D$ from all the images. $D$ is of size $m \times n$, where $m$ is the number of voxels, and $n$ is the number of images. For 3D images, $m \gg n$ (typically). Assuming all images are spatially well-aligned, $L$ captures the quasi-normal appearance of the images whereas $S$ contains pathologies which are not shared across the images. Of course, in practice, the objective is image alignment and hence the images in $D$ cannot be assumed to be aligned a-priori. Hence, Liu et al. [11] alternate LRS decomposition steps with image registration steps. Here the registrations are between all the low-rank images (which are assumed to be approximately pathology-free) and an atlas image. This approach is effective in practice, but can be computationally costly, may require large amounts of memory, and has the tendency to lose fine image detail in the quasi-normal image reconstructions, $L$. In detail, the matrix $D$ has a large number of rows for typical 3D images, hence it can be costly to store. Furthermore, optimizing the LRS decomposition involves a singular value decomposition (SVD) at each iteration with a complexity of $\mathcal{O}(min\{mn^2, m^2n\})$ [20] for an $m \times n$ matrix. While large datasets are beneficial to capturing data variation, the quadratic complexity renders LRS computationally challenging in these situations.

However, it is possible to overcome many of these shortcomings while staying close to the initial motivation of the original LRS approach. The following Section 2.2.2 discusses how this can be accomplished.

#### *2.2.2. Joint PCA-TV model*

To avoid the memory and computational issues of the low-rank + sparse decomposition discussed above, we previously proposed a joint PCA/Image-Reconstruction model [12] for improved and more efficient registration of images with pathologies. In this model, we have a collection of normal images and register all the normal images to the atlas *once*, using a standard image similarity measure. These normal images do not need to be re-registered during the iterative approach. We mimic the low-rank part of the LRS by a PCA decomposition of the atlas-aligned normal images from which we obtain the PCA basis and the mean image. Let us consider the case when we are now given a *single* pathological image $I$. Let $\hat{I}$ denote the pathological image after subtracting the mean image $M$ and $B$ the PCA basis matrix. $\hat{L}$ and $T$ are images of the same size as $I$[7]. Specifically, we minimize

$$E(T, \hat{L}, \boldsymbol{\alpha}) = \frac{1}{2}\|\hat{L} - B\boldsymbol{\alpha}\|_2^2 + \gamma\|\nabla T\|_{2,1}, \quad \text{s.t. } \hat{I} = \hat{L} + T \qquad (2)$$

where $\|\nabla T\|_{2,1} = \sum_i \|\nabla T_i\|_2$ and $i$ denotes spatial location. This model is similar to the Rudin-Osher-Fatemi (ROF) image denoising model [21]. It results in a total variation (TV) term, $T$, which captures the parts of $\hat{I}$ that are (i) relatively large, (ii) spatially contiguous, and (iii) cannot be explained by the PCA basis, e.g., pathological regions. The quasi-low-rank part $\hat{L}$ remains close to the PCA space but retains fine image detail. The quasi-normal image $L$ can then be reconstructed as $L = M + \hat{L}$. We refer to this model as our joint PCA-TV model.

As in the LRS approach, we can register the quasi-normal image $L$ to atlas space and alternate decomposition and registration steps. However, in contrast to the LRS model, the PCA-TV model registers only *one* image ($L$) in each registration step and consequently requires less time and memory to compute. Furthermore, the reconstructed

[7] Images are vectorized for computational purposes; but the spatial gradient $\nabla$ denotes the gradient in the spatial domain.

quasi-normal image, $L$, retains fine image detail as pathologies are captured via the total variation term in the PCA-TV model.

### *2.3. Proposed brain extraction approach*

The following sections describe how our proposed brain extraction approach builds upon the principles of the PCA-TV model (Section 2.3.1), and discusses image pre-processing (Section 2.3.2), the overall registration framework (Section 2.3.3), and post-processing steps (Section 2.3.4).

#### *2.3.1. Joint PCA-Sparse-TV model*

The PCA-TV model captures the pathological information well, but it does not model non-brain regions (such as the skull) appropriately. The skull is, for example, usually a thin, shell-shape structure and other non-brain tissue may be irregularly shaped with various intensities. The only commonality is that all these structures surround the brain. Specifically, if a test image is aligned to the atlas well, these non-brain tissues should *all* be located outside the atlas' brain mask. Hence, we reject these non-brain regions via a spatially distributed sparse term. We penalize sparsity heavily inside the brain and relatively little on the outside of the brain. This has the effect that it is very cheap to assign voxels outside the brain to the sparse term; hence, these are implicitly declared as brain outliers. Of course, if we would already have a reliable brain mask we would not need to go through any modeling. Instead, we assume that our initial affine registration provides a good *initial alignment* of the image, but that it will be inaccurate at the boundaries. We therefore add a constant penalty close to the boundary of the atlas brain mask. Specifically, we create two masks: a two-voxel-eroded brain mask, which we are confident is within the brain and a one-voxel-dilated brain mask, which we are confident includes the entire brain. We then obtain the following model:

$$\begin{aligned} E(S,T,\hat{L},\boldsymbol{\alpha}) = \frac{1}{2}\|\hat{L} - B\boldsymbol{\alpha}\|_2^2 + \gamma\|\nabla T\|_{2,1} \\ + \|\boldsymbol{\Lambda} \odot S\|_1, \\ \text{s.t. } \hat{I} = \hat{L} + S + T \end{aligned} \tag{3}$$

where $\boldsymbol{\Lambda} = \Lambda(\boldsymbol{x}) \geq 0$ is a spatially varying weight

$$\Lambda(\boldsymbol{x}) = \begin{cases} \infty, & \boldsymbol{x} \in \text{Eroded Mask (inside)} \\ \lambda, & \boldsymbol{x} \in \text{Dilated Mask and} \\ & \boldsymbol{x} \notin \text{Eroded Mask (at boundary)} \\ 0, & \boldsymbol{x} \notin \text{Dilated Mask (outside)} \end{cases} \tag{4}$$

with $\boldsymbol{x}$ denoting the spatial location. Further, in Eq. (3), $\odot$ indicates an element-wise product and $\gamma \geq 0$ weighs the total variation term.

We refer to this model as our joint PCA-Sparse-TV model. It decomposes the image into three parts. Similar to the PCA-TV model, the quasi-low-rank part $\hat{L}$ remains close to the PCA space and the TV term, $T$, captures pathological regions. Here, the PCA basis is generated from normal images that have been already brain-extracted. Therefore $\hat{L}$ only contains the brain tissue. Different from the previous model, we add a spatially distributed sparse term, $S$, which captures tissue outside the brain, e.g., the skull. In effect, since $\Lambda$ is very large inside the eroded mask, none of the image inside the eroded mask will be assigned to the sparse part. Conversely, all of the image outside the dilated mask will be assigned to the sparse part. We then integrate this PCA-Sparse-TV model into the low-rank registration framework. This includes three parts: pre-processing, iterative registration and decomposition, and post-processing as we will discuss in the following.

#### *2.3.2. Pre-processing*

Fig. 3 shows a flowchart of our pre-processing approach as discussed in the following paragraphs.

**Intensity normalization.** Given a test image from which we want to extract the brain, we first affinely transform the image intensities to standardize the intensity range to $[0, 1000]$. The goal of this step is to remove negative and small intensity values ($< 1$), which is required as an input to subsequent bias field correction which log tranforms image intensities for processing. Specifically, we first compute the 1st and the 99th percentile of the voxel intensities. We then affinely transform the image intensities of the entire image such that the intensity of the 1st percentile is mapped to 100 and of the 99th percentile to 900. As this may result in intensities smaller than zero or larger than 1000 for the extreme ends of the intensity distribution, we crop the intensities to be within $[0, 1000]$.

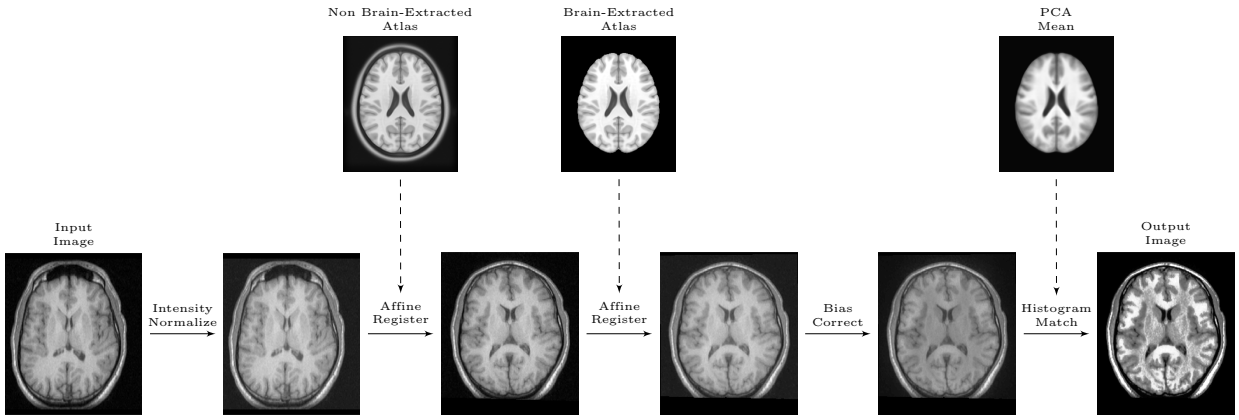



**Figure 3:** Preprocessing flow chart: Input image is the original image. Eventually, the output image will be fed into the registration/decomposition framework.

**Atlas registration.** Next, we first align the intensity-normalized input image to the non brain-extracted atlas. Then, we affinely register the result from the first step to the brain-extracted atlas, but this time using a one-voxel-dilated brain mask in atlas space; this step has the effect of ignoring parts of the image which are not close to the brain in the registration and it gives us a better alignment in the brain region. For both steps we use `reg_aladin` of `NiftyReg` [22] disabling symmetric registration (`-noSym`). The first registration initializes the transformation using the center of gravity (CoG) of the image.

**Bias field correction.** Next, we use N4ITK [23], a variant of the popular non-parametric non-uniform intensity normalization (N3) algorithm [24], to perform bias field correction. As the image has been affinely aligned to the atlas in the previous step, we use our two-voxel-eroded brain mask as the region for bias field estimation. Specifically, we use the `N4BiasFieldCorrection` function in `SimpleITK` [25], with its default settings.

**Histogram matching:** The final step of the preprocessing is histogram matching. We match the histograms of the bias corrected image with the histogram of the mean image of the population data only within the two-voxel-eroded brain mask. This histogram matched image is then the starting point for our brain extraction algorithm.

*2.3.3. Registration framework*

Similar to the PCA-TV model, we alternate between *image decomposition* steps using the PCA-Sparse-TV model and *registration to the brain-extracted atlas*. We use a total of six iterations in our framework. In the first iteration ($k = 1$), the images are in the original space. We decompose the input image $I_1 = I$, into the quasi-normal ($L_1 = \hat{L}_1 + M$), sparse ($S_1$), and total variation ($T_1$) images by minimizing the energy from Eq. (3). We then obtain a pathology-free or pathology-reduced image, $R_1$, by adding the sparse and the quasi-normal images of the decomposition: $R_1 = L_1 + S_1$.

For the next two iterations ($k = \{2, 3\}$), we first find the affine transform $\Phi_k$ by affinely registering the pathology-reduced images from the previous iteration, $R_{k-1}$ (i.e., $R_{k-1} = L_{k-1} + S_{k-1}$), to the brain-extracted atlas. We use the one-voxel-dilated brain mask for cost-function masking which allows the registration to focus only on the brain tissue. We then apply the transform $\Phi_k$ to transform the previous input images to atlas space and obtain new input images, $I_k$, (i.e., $I_k = I_{k-1} \circ \Phi_k$). We minimize Eq. (3) again to obtain new decomposition results $(L_k, S_k, T_k)$. These decomposition/affine-registration steps are repeated two times, which is empirically determined to be sufficient for convergence. These affine registration steps result in a substantially improved alignment in comparison to the initial affine registration by itself.

The last three iterations ($k = \{4, 5, 6\}$) repeat the same process, but are different in the following aspects: (i) we now use a b-spline registration instead of the affine registration; (ii) we use the pathology-reduced image and cost function masking only for the first B-spline registration step, as we did in the previous affine steps. For the remaining two steps, we use the quasi-normal images $L_{k:k=\{5,6\}}$ as the

moving images and we do not use the mask during the registrations. The use of the mask is no longer necessary as registrations are now performed using the quasi-normal image; (iii) we use the non-greedy registration strategy of the original low-rank + sparse framework [26], in which we deform the quasi-normal image back to the image space of the third iteration (after the affine steps) in order to avoid accumulating deformation errors.

These steps further refine the alignment, in particular, close to the boundary of the brain mask. After the last iteration, the image is well-aligned to the atlas and we have all the transforms from the original image space to atlas space. As a side effect, the algorithm also results in a quasi-normal reconstruction of the image, $L_6$, an estimate of the pathology, $T_6$, and an image of the non-brain tissue $S_6$, all in atlas space.

The *overall algorithm* steps are listed below:

- 0) Pre-processing steps: prepare the input image $I_1$, as described in Section 2.3.2.
- 1) First step: decompose input image $I_1$ into $\{L_1, S_1, T_1\}$, and obtain the pathology-free image $R_1$.
- 2) Affine steps ($k = \{2, 3\}$): (i) find the *affine* transform $\Phi_k$, that maps the $R_{k-1}$ to the brain-extracted atlas with cost function masking; (ii) apply the transform $\Phi_k$ to update the input images: $I_k = I_{k-1} \circ \Phi_k$; (iii) decompose the input image $I_k$ into $\{L_k, S_k, T_k\}$ and obtain pathology-free image $R_k$.
- 3) B-spline step ($k = 4$): (i) find the *b-spline* transform $\Phi_4$, that maps $R_3$ to the brain-extracted atlas with cost function masking; (ii) apply the transform $\Phi_4$ to update the input images: $I_4 = I_3 \circ \Phi_4$; (iii) decompose the input image $I_4$ into $\{L_4, S_4, T_4\}$.
- 4) B-spline steps ($k = \{5, 6\}$): (i) find the *b-spline* transform $\Phi_k$, that maps $L_{k-1} \circ (\Phi_{k-1})^{-1}$ to the brain-extracted atlas; (ii) apply the transform $\Phi_k$ to update the input image: $I_k = I_{k-1} \circ \Phi_k$; (iii) decompose the input image $I_k$ into $\{L_k, S_k, T_k\}$.
- 5) Post-processing steps: generate the resulting brain-extracted image and the brain mask in the original image space (see below).

#### *2.3.4. Post-processing*

Post-processing consists of applying to the atlas mask the inverse transforms of the affine registrations in the pre-processing step and the inverse transforms of the registrations generated in the framework described in section 2.3.3. The warped-back atlas mask is the brain mask for the original image. To extract the brain in the original image space, we simply apply the brain mask on the original input image. All subsequent validations are performed in the original image space.

## 3. Experimental results

### *3.1. Experimental setup*

We evaluate our method on all four evaluation datasets. For comparison, we also assess the performance of BET, BSE, ROBEX and CNN[8] on these datasets. We use BET v2.1 as part of FSL 5.0, BSE v.17a from BrainSuite and ROBEX v1.2. We solve our PCA model via a primal-dual hybrid gradient method [27]. In addition, we implement the decomposition on the GPU and run it on an NVIDIA Titan X GPU [28] [29].

### *3.2. Evaluation Measures*

We evaluate the brain extraction approaches using the measures listed below.

*Dice coefficient.* Given two sets X and Y (containing the spatial voxel positions of a segmentation), the Dice coefficient $D(X, Y)$ is defined as

$$D(X, Y) = \frac{2|X \cup Y|}{|X| + |Y|}, \tag{5}$$

where $X \cup Y$ denotes set union between $X$ and $Y$ and $|X|$ denotes the cardinality of set $X$.

*Average, maximum and 95% surface distance.* We also measure the symmetric surface distances between the automatic brain segmentation and the gold-standard brain segmentation. This is defined as follows: the distance of a point $x$ to a set of points (or set of points of a triangulated surface $S_A$) is defined as

$$d(x, S_A) = \min_{y \in S_A} d(x, y), \tag{6}$$

where $d(x, y)$ is the Euclidean distance between the point $x$ and $y$. The average symmetric surface dis-

[8] `https://github.com/GUR9000/Deep_MRI_brain_extraction`

tances between two surfaces $S_A$ and $S_B$ is then defined as

$$ASD(S_A, S_B) = \frac{1}{|S_A| + |S_B|} \times \left( \sum_{x \in S_A} d(x, S_B) + \sum_{y \in S_B} d(y, S_A) \right), \quad (7)$$

where $|S_A|$ denotes the cardinality of $S_A$ [30] (i.e., number of elements if represented as a set or surface area if represented in the continuum). To assess behavior at the extremes, we also report the maximum symmetric surface distance as well as the 95th percentile symmetric surface distance, which is less prone to outliers. These are defined in analogy, i.e., by computing all distances from surface $S_A$ to $S_B$ and vice versa followed by the computation of the maximum and the 95th percentile of these distances.

*Sensitivity and specificity.* We also measure sensitivity (i.e., true positive (TP) rate) and specificity (i.e., true negative (TN) rate). Here TP denotes the brain voxels which are correctly labeled as brain; TN denotes the non-brain voxels correctly labeled as such. Furthermore, the false negatives (FN) are the brain voxels incorrectly labeled as non-brain and the false positives (FP) are the non-brain voxels which are incorrectly labeled as brain. Let $V$ be the set of all voxels of an image, and $X$ and $Y$ the automatic brain segmentation and gold-standard brain segmentation, respectively. The sensitivity and specificity are then defined as follows [31] :

$$sensitivity = \frac{TP}{TP + FN} = \frac{|X \cap Y|}{|Y|} \quad (8)$$

$$specificity = \frac{TN}{TN + FP} = \frac{|V| - |X \cup Y|}{|V| - |Y|} \quad (9)$$

### *3.3. Datasets of normal images: IBSR/LPBA40*

**IBSR results:** Figure 4 shows the box-plots summarizing the results for the IBSR dataset. Overall, ROBEX, BSE, BET and our model perform well on this dataset, with a median dice coefficient above 0.95. CNN does not perform satisfactorily, with low Dice scores, low sensitivity, large distance errors, and overall high variance. Our PCA model outperforms all other methods with respect to Dice scores (median close to 0.97) and distance measures. BSE also works well on most cases, but it shows larger variability and exhibits two outliers which represent failure cases. ROBEX and BET show the highest sensitivity, but reduced specificity. Conversely, our PCA model, BSE, and CNN have high specificity but reduced sensitivity (the CNN model dramatically so). Table 3 (top) shows the means and standard deviations for the test results on this dataset. Our PCA model achieves the highest mean Dice overlap score (at 0.97) with the smallest standard deviation. ROBEX and BET show slightly reduced Dice overlap measures (mean around 0.95). BSE and CNN show the lowest performance. Our PCA model also performs best for the surface distance measures; it has the lowest mean average distance, 95% distance and the lowest maximum surface distance. Overall our PCA model performs best.

In addition, we perform a one-tailed paired t-test to compare results between methods. We test the null hypothesis that the results coming from our PCA model have a mean equal to the mean of the other methods, against the alternative that the mean of the PCA model is better[9] than the mean of the others. Table 1 (left) shows the corresponding results. We use the Benjamini-Hochberg procedure [32] for all the tests in this paper, in order to reduce the false discovery rate for multiple comparisons. We select an overall false discovery rate of 0.05 which results in an effective significance level of $\alpha \approx 0.0396$.

We outperform ROBEX, BET and CNN on Dice overlap scores and all distance measures with statistical significance. The comparison results with BSE are not significant due to the outliers of BSE. Our approach performs better than CNN on sensitivity and better than ROBEX and BSE on specificity.

**LPBA40 results:** Figure 5 shows the box-plots summarizing the validation results for the LPBA40 dataset. All five methods perform well. ROBEX, BET and BSE all have a median Dice score between 0.96 and 0.97. Our PCA model obtains the highest median Dice score (above 0.97). All methods except for the CNN approach have a median average surface distance smaller than 1 mm. Table 3 (second top) shows the means and standard deviations for all validation measures for this dataset. Again, all methods have satisfactory mean Dice scores and surface distances with low variances. Compared with other methods, the PCA model achieves the best results.

[9] Better means higher for the Dice overlap scores, smaller for the surface distances.

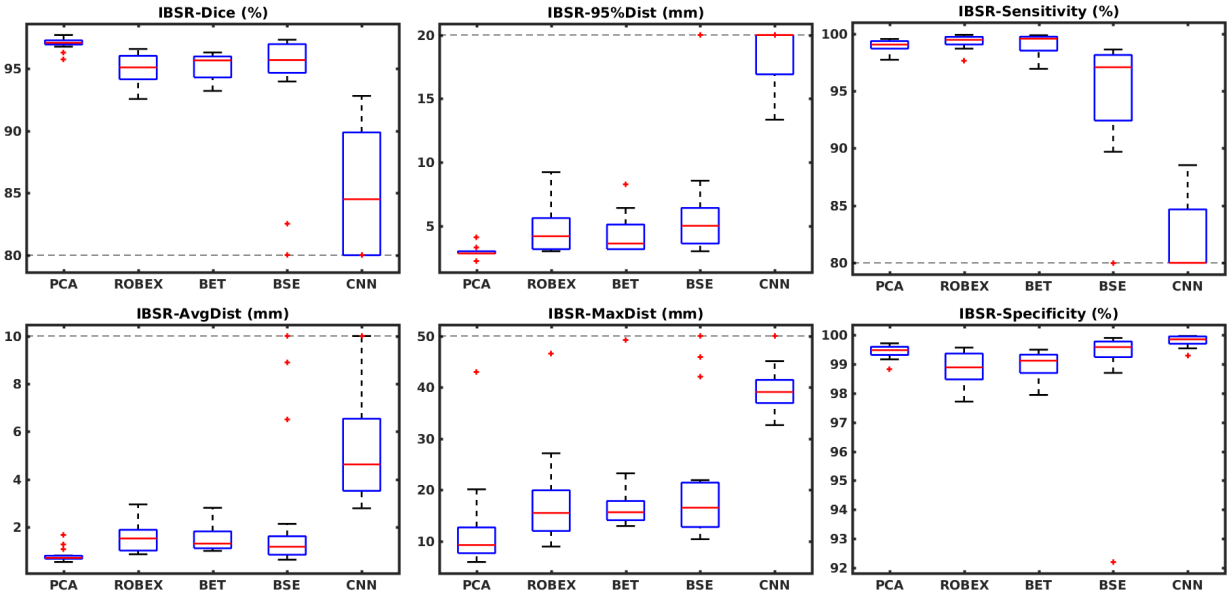



**Figure 4:** Box plot results for the IBSR normal dataset. Due to the poor results of CNN and the outliers of BSE on this dataset, we limit the range of the plots for better visibility. The dashed line indicates this cut-off. ROBEX, BET, and BSE show similar performance, but BSE exhibits two outliers. CNN performs poorly on this dataset. Our PCA model performs best on the Dice scores and surface distances. Although ROBEX and BET show slightly better sensitivity, our method shows better specificity.

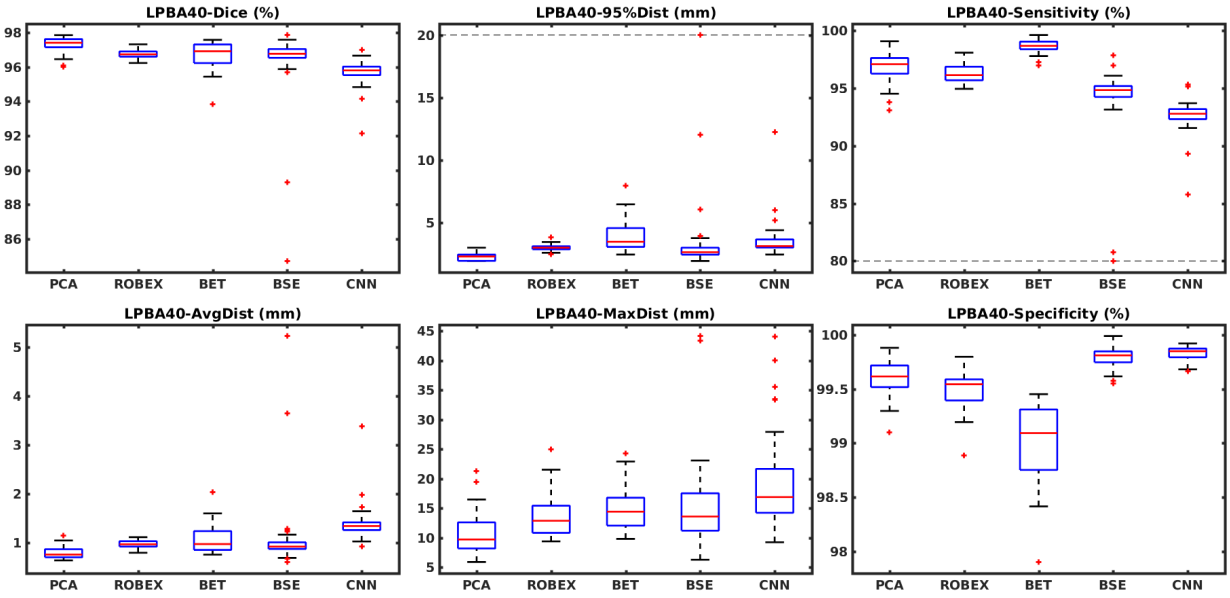



**Figure 5:** Box plot results for the LPBA40 normal dataset. All five methods work well on this dataset. Our PCA model has the best Dice and surface distances. ROBEX, BET and BSE show similar performance but BET exhibits larger variance and BSE exhibits two failed outliers. The CNN model shows overall slightly worse performance than the other methods.

Table 1 (right) shows the one-sided paired t-test results. Again we use the Benjamini-Hochberg procedure, resulting in a significance level $\alpha \approx 0.0395$. All methods perform well on this dataset, but our PCA approach still shows statistically significant improvement. We outperform other methods on

| Dataset: IBSR | | | | |
|---|---|---|---|---|
| | ROBEX | BET | BSE | CNN |
| Dice | 2.47e-8 | 7.19e-10 | 7.13e-2 | 6.05e-6 |
| Avg Dist | 7.53e-8 | 2.46e-9 | 8.46e-2 | 8.83e-7 |
| 95% Dist | 1.15e-5 | 3.41e-5 | 2.88e-2 | 1.19e-8 |
| Max Dist | 2.89-8 | 3.66e-10 | 1.05e-2 | 2.98e-9 |
| Sensitivity | 0.979 | 0.655 | 2.89e-2 | 1.09e-7 |
| Specificity | 1.10e-6 | 5.85e-7 | 0.219 | 1.000 |

| Dataset: LPBA40 | | | | |
|---|---|---|---|---|
| | ROBEX | BET | BSE | CNN |
| Dice | 4.27-12 | 6.22e-5 | 4.11e-3 | 3.79e-15 |
| Avg Dist | 5.67e-12 | 1.09e-6 | 8.70e-3 | 3.52e-12 |
| 95% Dist | 2.18e-15 | 5.67e-10 | 1.64e-2 | 3.49e-6 |
| Max Dist | 3.57e-16 | 2.73-10 | 6.37e-4 | 6.04e-8 |
| Sensitivity | 6.01e-3 | 1.00 | 1.38e-4 | 3.18e-17 |
| Specificity | 7.75e-7 | 9.38e-14 | 1.00 | 1.00 |

**Table 1:** $p$-values for IBSR and LPBA40 datasets. We perform a one-tailed paired t-test, where the null-hypothesis ($\mathcal{H}_0$) is that the results coming from our PCA model have a mean equal to the mean of the compared method, against the alternative ($\mathcal{H}_1$) that the mean of the PCA model is better than the mean of the compared method. Here, *better* means a higher Dice score or lower surface distances. In addition, we perform the Benjamini-Hochberg procedure to reduce the false discovery rate (FDR). We highlight in green the results where our PCA model performs statistically significantly better. The results show that our PCA model outperforms ROBEX, BET, and CNN, but does not show a statistical differences over BSE on IBSR. This may be due to the outliers of BSE.

Dice and surface distances with statistical significance. We perform better than all other methods except BET on sensitivity and except ROBEX and BET on specificity.

Figure 9 (top) visualizes the average brain mask errors for ISBR and LPBA40. All images are first affinely registered to the atlas. Then we transform the gold-standard expert segmentations as well as the automatically obtained brain masks of the different methods to atlas space. We compare the segmentations by counting the average over- and under-segmentation errors over all cases at each voxel. This results in a visualization for areas of likely mis-segmentation. Our PCA model, ROBEX and BET perform well on these two datasets. ROBEX and BET consistently show localized errors, e.g., at the boundary of the parietal lobe, the occipital lobe and the cerebellum. While BSE and CNN perform well on the LPBA40 dataset, they perform poorly on the IBSR dataset. This is in particular the case for the CNN approach.

### *3.4. Datasets with strong pathologies: BRATS/TBI*

**BRATS results:** Figure 6 shows the box-plots for the validation measures for the BRATS dataset. BSE and CNN, using their default settings, do not work well on the BRATS dataset. This may be because of the data quality of the BRATS data. Many of the BRATS images have relatively low out-of-plane resolutions. BSE results may be improved by a better parameter setting. However, as our goal is to evaluate all methods with the same parameter setting across all datasets, we do not explore dataset specific parameter tuning. BET shows good performance, but suffers from a few outliers. ROBEX works generally well, with a median Dice score around 0.95 and an average distance error of 1.5 mm. However, as for IBSR and LPBA40, our PCA model performs generally the best with a median Dice score 0.96 and a 1 mm average distance error. The PCA model results also show lower variance, as shown in table 3 (second bottom), underlining the very consistent behavior of our approach. Table 2 (left) shows (via a one-sided paired t-test with a correction for multiple comparisons using a false discovery rate of 0.05) that our model has statistically significantly better performance than ROBEX on all measures. The improvements over BET are not statistically significant. However, our approach is statistically significantly better on all measures (except specificity) than BSE and CNN.

**TBI results:** Figure 7 shows the box-plots for the results on our TBI dataset. Our PCA model still outperforms all other methods. Our method achieves the largest Dice scores and the lowest surface errors among all methods (best mean and lowest variance in table 3 (bottom)). Table 2 (right) shows the one-sided paired t-test results with multiple comparisons correction (with a false discovery rate of 0.05). Our model performs significantly better than the other methods on almost all measures.

Finally, Figure 9 (bottom) shows the average segmentation errors on the BRATS and TBI datasets: our PCA method shows less errors than other methods in these two abnormal datasets. ROBEX and BET exhibit large errors at the boundary of the brain. CNN and BSE particularly show large errors for the BRATS dataset presumably again due to the coarse resolution of the BRATS data.

In addition to extracting the brain from pathological datasets, our method also allows for the estimation of a corresponding quasi-normal image in atlas space, although this is not the main goal of

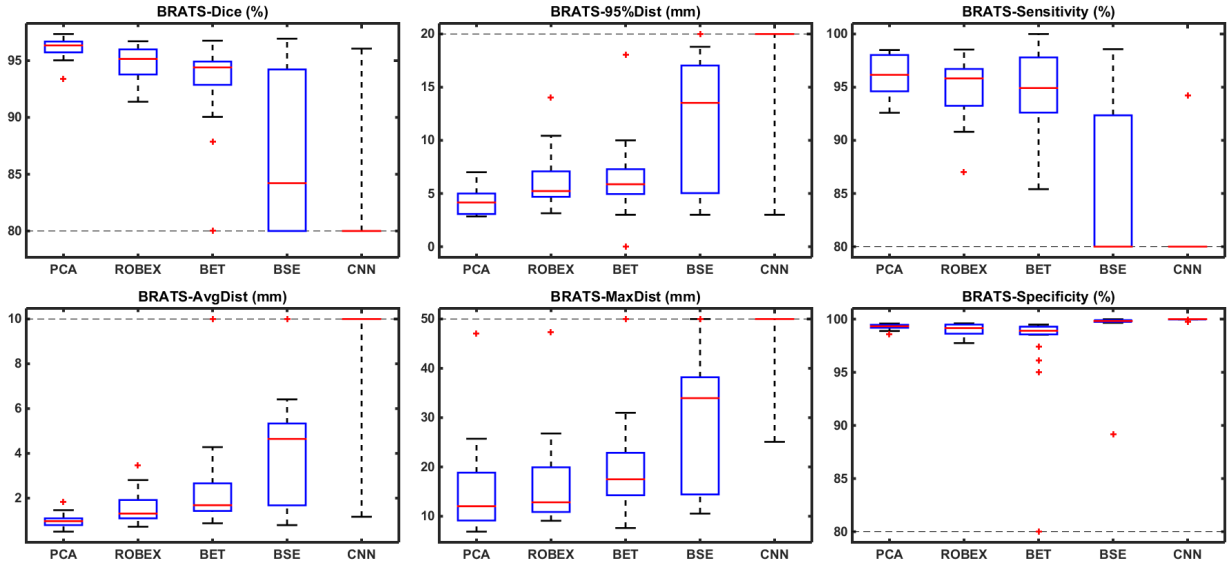



**Figure 6:** Box plot results for the BRATS tumor dataset. BSE and CNN fail on this dataset. BET shows better performance, but also exhibits outliers. ROBEX and our PCA model work well on this dataset. Overall our model works best.

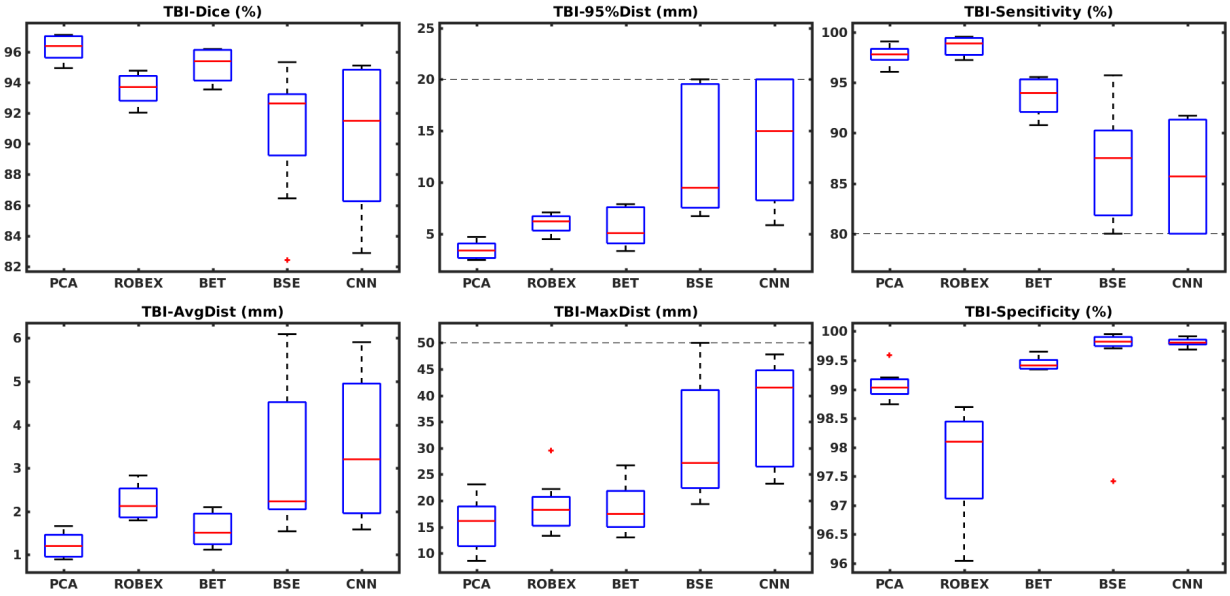
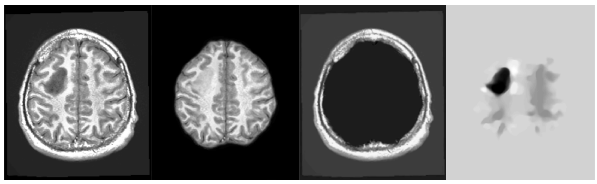



**Figure 7:** Box plot results for the TBI dataset. Our PCA model shows the best evaluation results. BET and ROBEX also perform reasonably well. BSE and CNN do not perform well on this dataset.

this paper. Figure 8 shows an example of the reconstructed quasi-normal image ($L$) for an image of the BRATS dataset, as well as an estimation of the pathology (pathology image $T$ and non-brain image $S$). Compared to the original image, the pathology shown in the quasi-normal image has been greatly reduced. Hence this image can be used for the registration with a normal image or a normal atlas. This has been shown to improve registration accuracy for the registration of pathological images [12]. Furthermore, an estimate of the pathology (here a tumor) is also obtained which may be useful for further analysis. Note that in this example image the total variation term captures more than just the



(a) (b) (c) (d)

**Figure 8:** Example BRATS image with its decomposition result in atlas space. (a) Input image after pre-processing; (b) quasi-normal image $L + M$; (c) non-brain image $S$; (d) pathology image $T$.

tumor. This may be due to inconsistencies in the image appearance between the normal images (obtained from OASIS data) and the test dataset. As our goal is atlas alignment rather than quasi-normal image reconstruction or pathology segmentation, such a decomposition is acceptable, although we could improve this by tuning the parameters or applying regularization steps as in [12].

## 4. Discussion

We presented a PCA-based model specifically designed for brain extraction from pathological images. The model decomposes an image into three parts. Non-brain tissue outside of the brain is captured by a sparse term, normal brain tissue is reconstructed as a quasi-normal image close to a normal PCA space, and brain pathologies are captured by a total-variation term. The quasi-normal image allows for registration to an atlas space, which in turn allows registering the original image to atlas space and hence to perform brain extraction. Although our approach is designed for reliable brain extraction from strongly pathological images, it also performs well for brain extraction from normal images, or from images with subtle pathologies. In fact, we validated our brain extraction method using four different datasets (two of them with strong pathologies: brain tumors and traumatic brain injuries). On all four datasets our approach either performs best or is among the best methods using a fixed set of parameters. Hence, our approach can achieve good brain extraction results on a variety of different datasets and, unlike some of the competing methods, can tolerate pathologies as well as differing image appearances much better. Future work could focus on the reconstruction of the quasi-normal image using regularization steps or using the quasi-normal image to compare images longitudinally, for example the chronic and the acute phases of TBI. Our software is freely available as open source code at `https://github.com/uncbiag/pstrip`.


## Acknowledgements

Research reported in this publication was supported by the National Institutes of Health (NIH) and the National Science Foundation (NSF) under award numbers NIH R41 NS091792, NSF ECCS-1148870, and EECS-1711776. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the NSF.


## References


[1] S. M. Smith, Fast robust automated brain extraction, Human Brain Mapping 17 (3) (2002) 143–155.

[2] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, R. M. Leahy, Magnetic resonance image tissue classification using a partial volume model, NeuroImage 13 (5) (2001) 856–876.

[3] J. E. Iglesias, C.-Y. Liu, P. M. Thompson, Z. Tu, Robust brain extraction across datasets and comparison with publicly available methods, IEEE Transactions on Medical Imaging 30 (9) (2011) 1617–1634.

[4] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, A. Biller, Deep MRI brain extraction: a 3D convolutional neural network for skull stripping, NeuroImage 129 (2016) 460–469.

[5] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, A. W. Toga, Construction of a 3D probabilistic atlas of human cortical structures, Neuroimage 39 (3) (2008) 1064–1080.

[6] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, R. L. Buckner, Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults, Journal of Cognitive Neuroscience 19 (9) (2007) 1498–1507.

[7] F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, B. Fischl, A hybrid approach to the skull stripping problem in MRI, NeuroImage 22 (3) (2004) 1060–1075.

[8] S. F. Eskildsen, P. Coupé, V. Fonov, J. V. Manjón, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Østergaard, D. L. Collins, A. D. N. Initiative, et al., Beast: brain extraction based on nonlocal segmentation technique, NeuroImage 59 (3) (2012) 2362–2373.

[9] Analysis of Functional Neuro Images (AFNI), `https://afni.nimh.nih.gov`.

[10] J. Doshi, G. Erus, Y. Ou, B. Gaonkar, C. Davatzikos, Multi-atlas skull-stripping, Academic Radiology 20 (12) (2013) 1566–1576.

[11] X. Liu, M. Niethammer, R. Kwitt, M. McCormick, S. Aylward, Low-rank to the rescue–atlas-based analyses in the presence of pathologies, in: MICCAI, 2014, pp. 97–104.

| Dataset: BRATS | | | | |
|---|---|---|---|---|
| | ROBEX | BET | BSE | CNN |
| Dice | 1.25e-5 | 4.72e-2 | 6.82e-6 | 3.01e-10 |
| Avg Dist | 8.44e-5 | 0.129 | 2.33e-4 | 9.11e-7 |
| 95% Dist | 2.14e-4 | 9.63e-3 | 1.92e-3 | 6.22e-8 |
| Max Dist | 1.35e-2 | 1.80e-2 | 1.89e-3 | 1.40e-8 |
| Sensitivity | 2.59e-2 | 7.78e-2 | 2.08e-6 | 2.80e-12 |
| Specificity | 2.49e-2 | 0.135 | 0.505 | 1.000 |

| Dataset: TBI | | | | |
|---|---|---|---|---|
| | ROBEX | BET | BSE | CNN |
| Dice | 5.91e-4 | 2.23e-3 | 2.37e-3 | 4.63e-3 |
| Avg Dist | 7.49e-4 | 5.13e-3 | 4.06e-3 | 3.05e-3 |
| 95% Dist | 2.49e-4 | 2.81e-3 | 2.37e-3 | 1.90e-3 |
| Max Dist | 1.31e-2 | 3.95e-2 | 9.11e-3 | 1.44e-3 |
| Sensitivity | 0.987 | 9.06e-5 | 1.46e-3 | 7.87e-4 |
| Specificity | 2.63e-3 | 0.999 | 0.903 | 1.000 |

**Table 2:** $p$-values for BRATS and TBI datasets. See caption of table 1 for a more detailed description. Our method shows significant improvement over ROBEX, BSE and CNN for both datasets. The improvement over BET is not significant on BRATS, but it is significant for the TBI dataset.


[12] X. Han, X. Yang, S. Aylward, R. Kwitt, M. Niethammer, Efficient registration of pathological images: A joint pca/image-reconstruction approach, in: ISBI, 2017, pp. 10–14.

[13] V. S. Fonov, A. C. Evans, R. C. McKinstry, C. Almli, D. Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, NeuroImage 47 (2009) 39–41.

[14] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, S. Ourselin, Fast free-form deformation using graphics processing units, Computer Methods and Programs in Biomedicine 98 (3) (2010) 278–284.

[15] T. F. Cootes, G. J. Edwards, C. J. Taylor, Active appearance models, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 681–685.

[16] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Transactions on Medical Imaging 34 (10) (2015) 1993–2024.

[17] M. Brett, A. P. Leff, C. Rorden, J. Ashburner, Spatial normalization of brain images with focal lesions using cost function masking, NeuroImage 14 (2) (2001) 486–500.

[18] M. Niethammer, G. L. Hart, D. F. Pace, P. M. Vespa, A. Irimia, J. D. Van Horn, S. R. Aylward, Geometric metamorphosis, in: MICCAI, 2011, pp. 639–646.

[19] X. Yang, X. Han, E. Park, S. Aylward, R. Kwitt, M. Niethammer, Registration of pathological images, in: SASHIMI, 2016, pp. 97–107.

[20] M. Holmes, A. Gray, C. Isbell, Fast SVD for large-scale matrices, in: Workshop on Efficient Machine Learning at NIPS, 2007, pp. 249–252.

[21] L. I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D: Nonlinear Phenomena 60 (1-4) (1992) 259–268.

[22] M. Modat, D. M. Cash, P. Daga, G. P. Winston, J. S. Duncan, S. Ourselin, Global image registration using a symmetric block-matching approach, Journal of Medical Imaging 1 (2) (2014) 024003–024003.

[23] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, J. C. Gee, N4ITK: improved N3 bias correction, IEEE Transactions on Medical Imaging 29 (6) (2010) 1310–1320.

[24] J. G. Sled, A. P. Zijdenbos, A. C. Evans, A nonparametric method for automatic correction of intensity nonuniformity in MRI data, IEEE Transactions on Medical Imaging 17 (1) (1998) 87–97.

[25] B. C. Lowekamp, D. T. Chen, L. Ibáñez, D. Blezek, The design of simpleitk, Frontiers in Neuroinformatics 7.

[26] X. Liu, M. Niethammer, R. Kwitt, N. Singh, M. McCormick, S. Aylward, Low-rank atlas image analyses in the presence of pathologies, IEEE Transactions on Medical Imaging 34 (12) (2015) 2583–2591.

[27] T. Goldstein, M. Li, X. Yuan, E. Esser, R. Baraniuk, Adaptive primal-dual hybrid gradient methods for saddle-point problems, `arXiv:1305.0546`.

[28] J. Nickolls, I. Buck, M. Garland, K. Skadron, Scalable parallel programming with CUDA, Queue 6 (2) (2008) 40–53.

[29] L. E. Givon, T. Unterthiner, N. B. Erichson, D. W. Chiang, E. Larson, L. Pfister, S. Dieleman, G. R. Lee, S. van der Walt, B. Menn, T. M. Moldovan, F. Bastien, X. Shi, J. Schlüter, B. Thomas, C. Capdevila, A. Rubinsteyn, M. M. Forbes, J. Frelinger, T. Klein, B. Merry, L. Pastewka, S. Taylor, A. Bergeron, N. H. Ukani, F. Wang, Y. Zhou, scikit-cuda 0.5.1: a Python interface to GPU-powered libraries, `http://dx.doi.org/10.5281/zenodo.40565` (12 2015). `doi:10.5281/zenodo.40565`.
URL `http://dx.doi.org/10.5281/zenodo.40565`

[30] V. Yeghiazaryan, I. Voiculescu, An overview of current evaluation methods used in medical image segmentation, Tech. rep., Tech. Rep. CS-RR-15-08, Department of Computer Science, University of Oxford, Oxford, UK (2015).

[31] M. Sonka, J. M. Fitzpatrick, Handbook of medical imaging(Volume 2, Medical image processing and analysis), SPIE Publications, 2000.

[32] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society. Series B (Methodological) (1995) 289–300.

| Dataset: IBSR | | | | | |
|---|---|---|---|---|---|
| | PCA | ROBEX | BET | BSE | CNN |
| Dice(%) | 96.99±0.53 | 94.98±1.17 | 95.16±0.96 | 89.54±21.76 | 82.91±16.63 |
| Avg Dist(mm) | 0.79±0.27 | 1.51±0.56 | 1.49±0.47 | 4.16±10.53 | 5.43±4.73 |
| 95% Dist(mm) | 2.84±0.43 | 4.50±1.58 | 4.22±1.39 | 12.27±20.83 | 22.25±9.41 |
| Max Dist(mm) | 11.97±8.14 | 17.30±8.40 | 17.91±7.85 | 24.93±23.32 | 39.86±10.73 |
| Sensitivity(%) | 98.99±0.46 | 99.33±0.54 | 99.09±0.93 | 88.68±22.86 | 74.76±19.25 |
| Specificity(%) | 99.44±0.21 | ±0.51 | 98.98±0.46 | 99.15±1.67 | 99.78±0.22 |

| Dataset: LPBA40 | | | | | |
|---|---|---|---|---|---|
| | PCA | ROBEX | BET | BSE | CNN |
| Dice(%) | 97.32±0.42 | 96.74±0.24 | 96.70±0.78 | 96.29±2.26 | 95.70±0.74 |
| Avg Dist(mm) | 0.79±0.12 | 0.97±0.07 | 1.06±0.27 | 1.11±0.81 | 1.32±0.33 |
| 95% Dist(mm) | 2.27±0.32 | 2.96±0.26 | 3.92±1.24 | 3.46±3.38 | 3.56±1.56 |
| Max Dist(mm) | 10.83±3.76 | 13.81±3.47 | 15.14±3.75 | 15.54±7.74 | 18.22±6.14 |
| Sensitivity(%) | 96.81±1.23 | 96.33±0.85 | 98.66±0.54 | 94.02±4.10 | 92.62±1.46 |
| Specificity(%) | 99.61±0.16 | 99.49±0.16 | 99.04±0.34 | 99.79±0.09 | 99.83±0.07 |

| Dataset: BRATS | | | | | |
|---|---|---|---|---|---|
| | PCA | ROBEX | BET | BSE | CNN |
| Dice(%) | 96.16±0.92 | 94.83±1.49 | 90.95±13.41 | 84.91±8.89 | 21.89±29.54 |
| Avg Dist(mm) | 1.00±0.31 | 1.54±0.70 | 7.58±25.30 | 4.37±3.61 | 44.87±29.05 |
| 95% Dist(mm) | 4.35±1.27 | 6.03±2.50 | 6.18±3.53 | 13.92±13.00 | 73.85±38.77 |
| Max Dist(mm) | 15.26±9.32 | 16.42±8.80 | 22.78±22.61 | 32.02±22.38 | 86.60±36.92 |
| Sensitivity(%) | 96.17±1.84 | 94.95±2.88 | 94.77±3.82 | 77.80±13.43 | 16.17±24.73 |
| Specificity(%) | 99.29±0.25 | 98.98±0.65 | 93.69±22.08 | 99.29±2.38 | 99.97±0.05 |

| Dataset: TBI | | | | | |
|---|---|---|---|---|---|
| | PCA | ROBEX | BET | BSE | CNN |
| Dice(%) | 96.28±0.85 | 93.60±1.00 | 95.14±1.12 | 91.00±4.31 | 90.40±5.07 |
| Avg Dist(mm) | 1.22±0.30 | 2.20±0.40 | 1.57±0.40 | 3.15±1.66 | 3.46±1.75 |
| 95% Dist(mm) | 3.41±0.85 | 5.99±0.97 | 5.57±1.91 | 13.07±7.11 | 16.04±8.72 |
| Max Dist(mm) | 15.54±5.03 | 18.89±5.12 | 18.54±4.59 | 31.96±12.71 | 37.06±10.09 |
| Sensitivity(%) | 97.76±0.92 | 98.64±0.93 | 93.65±1.87 | 85.65±8.17 | 83.77±8.58 |
| Specificity(%) | 99.07±0.26 | 97.75±0.93 | 99.44±0.11 | 99.54±0.86 | 99.81±0.07 |

**Table 3:** Means and standard deviations for validation measures for all the methods and all the datasets. We highlight the best results in red based on the mean values. Among all datasets, our PCA model has the best mean and lowest variance on Dice overlap scores and surface distances, except for LPBA40 where ROBEX shows lower variance for the average surface distances.

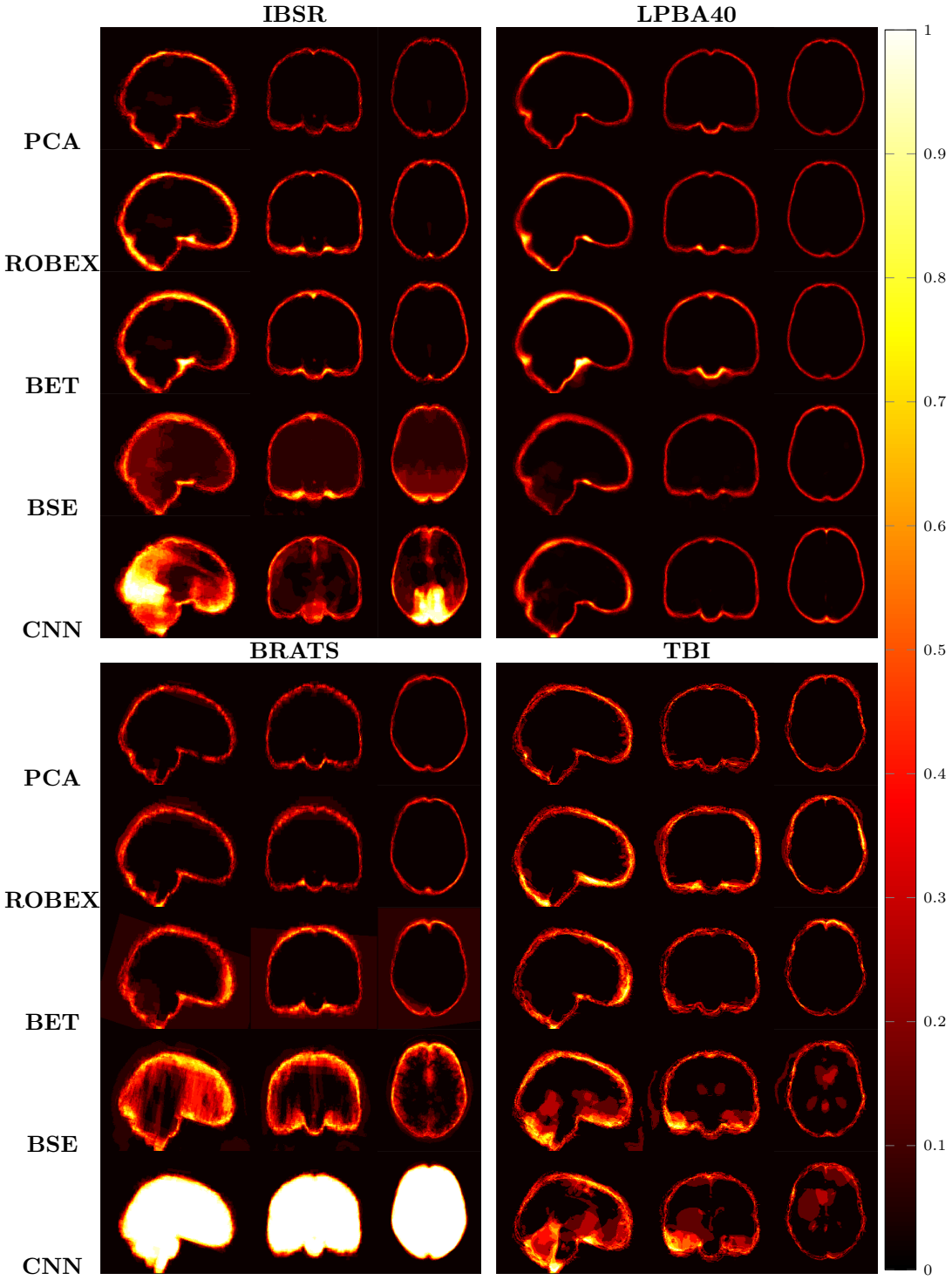



**Figure 9:** Example of 3D volumes of average errors for the normal IBSR and LPBA40 datasets as well as for the pathological BRATS and TBI datasets. Images and their brain masks are first affinely aligned to the atlas. At each location we then calculate the proportion of segmentation errors among all the segmented cases of a dataset (both over- and under-segmentation errors). Lower values are better (a 0 indicates perfect results over all images) and higher values indicate poorer performance (a value of 1 indicates failure on all cases). Clearly, BSE and CNN struggle with the BRATS dataset whereas our PCA method shows good performance across all datasets.

## Appendix A. `NiftyReg` Settings

This section introduces the settings for `NiftyReg` used in this paper. We mainly use the affine registration `reg_aladin` and the B-spline registration `reg_f3d`.

*Affine Registration:.* For affine registration, we use `reg_aladin` in `NiftyReg`. The options for affine registration are `-ref, -flo, -aff, -res`, which stand for reference image, floating image, affine transform output, warped result image, respectively. If the symmetric version is disabled, we add "`-noSym`". If center of gravity is used for the initial transformation, we add "`-cog`".

*B-spline Registration:.* For b-spline registration, we use `reg_f3d` in `NiftyReg`. In addition to the options as shown in affine (except for `reg_f3d` we use `-cpp` for output transform), we also use options `-sx 10, --lncc 40, -pad 0`, which include local normalized cross correlation with standard deviation of the Gaussian kernel of 40, grid spacing of 10 mm along all axes, and padding 0.

## Appendix B. Methods settings

This section introduces the settings that are used for all methods.

*PCA.* We use $\lambda = 0.1$ for the sparse penalty and $\gamma = 0.5$ for the total variation penalty.

*Robex/CNN.* ROBEX and CNN do not require parameter tuning. Therefore we use the default settings, and for ROBEX we add a seed value of `1` for all datasets.

*BET.* We use the suggested parameter settings in the literature [3][4] for the IBSR and LPBA40 datasets. For the BRATS and TBI datasets, we choose the option `"-B"` for BET, which corrects the bias field and "cleans-up" the neck.

*BSE.* We use the suggested parameter settings in the literature [3][4] for the IBSR and LPBA40 datasets. For the BRATS and TBI datasets, we use the default settings.