DEVELOPING DEEP LEARNING AND BAYESIAN DEEP LEARNING BASED MODELS FOR MR NEUROIMAGING

By

Gengyan Zhao

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Medical Physics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2019

Date of final oral examination: 05/30/2019

The dissertation is approved by the following members of the Final Oral Committee:

Mary E. Meyerand, Professor, Medical Physics

Rasmus M. Birn, Associate Professor, Psychiatry

Vivek Prabhakaran, Professor, Radiology

Andrew L. Alexander, Professor, Medical Physics

Diego Hernando, Assistant Professor, Medical Physics

Abstract

Magnetic resonance (MR) neuroimaging is an active field in investigating brain structures and functions. After decades of development, the whole pipeline of MR neuroimaging tends to become mature, but many essential steps still faces challenges and difficulties, especially in the accuracy of the image segmentation, image generation and data prediction. Recently, the revival of deep neural networks made immense progress in the field of machine learning. The proposal of Bayesian deep learning further enabled the ability of uncertainty generation in deep learning prediction. In this work, we proposed and developed different kinds of Bayesian neural networks to improve the accuracy of brain segmentation, brain image synthesis, and brain function related behavior prediction. To overcome the challenges in brain segmentation, we proposed a fully-automated brain extraction pipeline combining deep Bayesian convolutional neural network (CNN) and fully connected three-dimensional (3D) conditional random field (CRF). To increase the image synthesis accuracy and improve the calibration of the model uncertainty, we proposed a Bayesian conditional generative adversarial network (GAN). To improve the brain function related behavior prediction, we proposed a Bayesian deep neural network (DNN), and a feature extraction and ranking method for it. Experiments were done on real data to validate the proposed methods. The comparison between our methods and the state-of-the-arts showed that our methods can significantly improve the testing accuracy and the behavior of the model uncertainty generated by the Bayesian neural networks matches our expectation.

Acknowledgements

This work would not be possible without the tremendous support and helpful guidance from my advisors Dr. Beth Meyerand and Dr. Rasmus Birn. I would like to first thank Beth for giving me the opportunity to join her lab and her constant support on my research. She is an excellent leader in the field of biomedical engineering and an amiable advisor in the Applied Neuro MRI Lab. I also want to thank Rasmus for his inspiring coaching. He is an excellent expert in both the methodology and the applications of MRI and functional MRI with undeniable passion for his research. He is also a gentle advisor, who has genuine interest in education. I learned both science and life from him. It's an enjoyable and memorable experience for me to have the opportunity working with them.

I would like to express my gratitude to my thesis committee, Dr. Vivek Prabhakaran, Dr. Andrew L. Alexander, and Dr. Diego Hernando for their comments and advices, which greatly improved this work.

I would like to express my gratitude to Dr. Fang Liu. He is a knowledgeable scientist in the field of medical imaging. A lot of my research was in the collaboration with him. I benefited greatly from his work style and rigorous attitude. To discuss with him is always a fun time for me.

I would like to thank Dr. Veena Nair for her important advices and guidance on many aspects of my work, including the basic MRI/fMRI knowledge, data processing methods, experimental methods and result analysis approaches. She helped me quickly pick up the knowledge and become productive after I joined the lab.

I would like to thank Dr. Bruce Hermann, and Dr. Elizabeth Felton at UW Hospital and Clinics, and Dr. Jeffrey Binder, Dr. Edgar DeYoe, Dr. Andrew Nencka, and Mr. Jedidiah Mathis at Medical College of Wisconsin for their support and collaboration on the Epilepsy Connectome Project.

I would like to thank Dr. Ned Kalin and Dr. Jonathan Oler for their generous sharing of the data and collaboration on the brain segmentation project.

I would like to express my gratitude to Dr. Alan McMillan, Dr. Tyler Bradshaw, Dr. Hyungseok Jang, Dr. Poonam Yadav and Ms. Yilin Liu for all the collaboration with them on multiple deep learning projects.

I would like to express my gratitude to Mr. Gyujoon Hwang and Mr. Cole Cook for their inspiring discussions and collaborations on machine learning experiments and statistics methods. I would like to express my gratitude to Ms. Maribel Torres Velázquez and Ms. Charlene N. Rivera-Bonet for all the support from them as lab members.

I would like to express my gratitude to Dr. Guang-Hong Chen for giving me the opportunity to join his lab. This is how I started my journey in the field of medical imaging. His work style has a huge impact on me, and I learned a lot from his scientific value. I would also like to express my gratitude to the collaborators and friends in Dr. Chen's lab, including Dr. Ke Li, Dr. Yinsheng Li, Dr. Yongshuai Ge, Dr. Ran Zhang, Dr. Kai Niu, Dr. Yinghua Tao, Dr. John Garrett, Dr. Daniel Gomez-Cardona, Dr. Juan Pablo Cruz-Bastida, Mr. John Hayes, and Mr. Xu Ji. I also want to thank Dr. Yijing Wu for the help she gave me in my life.

I also want to thank my friends Mr. Carson Hoffman, Mr. Xiaoke Wang, Dr. Zhaoye Zhou, Dr. Wei Zha, and Ms. Shu Sun for all the fun time we spent together.

Finally, I would like to give special thanks to my family - my parents Dali and Siqing, and my cousin Qiao and her husband Scott for their love and support throughout my study at UW-Madison and my entire life.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2D	two-dimensional
3D	three-dimensional
Adam	adaptive moment estimation
AFNI	Analysis of Functional NeuroImages software suite
AI	artificial intelligence
ASSD	average symmetric surface distance
BET	Brain Extraction Tool
BOLD	blood-oxygenation-level-dependent
BRATS	international multimodal BRAin Tumor Segmentation challenge
BSE	Brain Surface Extractor
BSegNetCRF	Bayesian SegNet and fully connected 3D CRF
CDF	cumulative distribution function
CNN	convolutional neural network
CRF	conditional random field
CT	computed tomography
DC	dice coefficient
DNN	deep neural network
DOF	degrees of freedom
DTI	diffusion tensor imaging

e.g.	exempli gratia (for the sake of an example)
ELBO	evidence lower bound
etc.	et cetera (and the rest of the things)
FA	fractional anisotropy
FC	functional connectivity
FCN	fully convolutional network
FLAIR	Fluid Attenuated Inversion Recovery
fMRI	functional magnetic resonance imaging
GAN	generative adversarial network
GBM	Glioblastoma Multiforme
GPU	graphical processing unit
GRE	gradient echo
HCP	human connectome project
HD	Hausdorff distance
HWA	Hybrid Watershed Algorithm
ICA	independent component analysis
KL	Kullback-Leibler
MAP	maximum a posteriori
MC	Monte Carlo
MPRAGE	Magnetization Prepared Gradient Echo
MR	magnetic resonance
MRI	magnetic resonance imaging

MR-Linac	magnetic resonance - linear accelerator
MSM-ALL	multi-modal surface matching algorithm
NMT	National Institute of Mental Health Macaque Template
nRMSE	normalized root mean square error
nSTD	normalized standard deviation
PCA	principal component analysis
PDF	probability density function
PET	positron emission tomography
RF	radio frequency
RL	reinforcement learning
RMS	root mean square
RNN	recurrent neural networks
ROBEX	Robust Learning-Based Brain Extraction
rs-fMRI	resting state functional magnetic resonance imaging
SE	spin echo
SGD	stochastic gradient decent
s.t.	such that
std	standard deviation
SVM	support vector machine
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Archive
TE	echo time

TR	repetition time

Chapter 1

Introduction

Magnetic resonance (MR) neuroimaging has been widely used in brain structural and functional studies due to its capability of observing brain anatomic structures and tracking brain functional activities. The whole pipeline of MR neuroimaging has been developed for decades and tend to become increasingly mature, which includes the pulse sequence design in data acquisition, the sparse and rapid reconstruction in image generation, the segmentation, registration and noise removal in the image preprocessing, and various statistical models for the final data analysis and prediction. However, the current pipeline for MR neuroimaging is far from perfect and still faces many challenges and difficulties, especially in the accuracy of the image segmentation, image generation and data prediction.

Brain extraction and brain segmentation are the essential steps in magnetic resonance imaging (MRI) and functional magnetic resonance imaging (fMRI) processing. Current methods can segment most healthy human adult's brain correctly, but for patients with brain tumors, trauma or stroke, or nonhuman primates which are also important parts of the neuroscience research,

manually labeling or manually modifying the segmentation results by template-based methods is still necessary.

For image synthesis and image reconstruction in medical imaging, the state-of-the-art methods usually formulate the problem as an optimization problem, whose accuracy highly depends on the accuracy of the forward model. Currently in the field of medical imaging, the forward models are usually based on the physical mechanism of the imaging modality, and often simplified forward models are unavoidable. Even though, these forward models have strict requirements towards the data in the acquisition domain, which makes problems like reconstruction with extremely under-sampled k-space data or synthesizing T2-weighted (T2w) MR image with only T1-weighted (T1w) MR image very challenging.

For the fMRI analysis investigating individual differences in brain activation or functional connectivity, generalized linear models are widely used, but for the fitting and prediction of lots of sociodemographic information, neuropsychological test results, and clinical characteristics, there is still a large room for the prediction models' accuracy to improve. In neuroscience, extracting, ranking and analyzing the important features for the prediction is also very important, and comparing those features between different groups can help us answer many difficult and meaningful neuropsychiatric questions. Thus, improved methods for robust feature extraction and ranking methods from high-accuracy prediction models are still needed.

In the recent past, tremendous progress has been made in artificial intelligence (AI) because of the revival of artificial neural networks and the rapidly advancing parallel computing technique of graphical processing unit (GPU). Among all the machine learning models, deep learning is one of those most successful architectures, which has drawn increasingly attention. This technique has proven its capability in many computer vision applications, such as image classification,

segmentation, and regression, where the performance of traditional methods can hardly compete. In the areas of medical imaging, deep learning has broadly impacted neuroimaging, cancer imaging, and cardiac imaging on the processing performance and the accuracy improvement.

This thesis will describe the advances in three different deep leaning models for different tasks in the field of MR neuroimaging. Among deep learning models, convolutional neural network (CNN) has been proven to be useful in a broad range of image segmentation applications, outperforming traditional state-of-the-art methods. Generative adversarial network (GAN) is a powerful tool in various image synthesis tasks, including image to image translation and image reconstruction. Deep neural network (DNN) is particularly good at dealing with non-image data with high-dimensional feature space, and therefore, is suited for feature-related classification tasks.

The current workflow of deep learning based research usually consists of two steps: a neural network is training with a training dataset, and then the trained neural network is used to make predictions on the data in the testing dataset. By comparing the predicted results and the ground truth of the testing dataset, we can report the accuracy of the results and evaluate the performance of the model. However, if we really want AI to work by itself in the daily clinical routine, there won't be ground truth of the testing data any more. At the same time, the inconsistency within the training dataset and the inconsistency between the training dataset and testing dataset will cause errors in the predictions. In the conventional deep learning, without ground truth, there in no way for us to no whether the predicted results can be trusted or not.

The emergence of Bayesian deep learning provides us the possibility to solve this problem. From the statistical point of view, in the framework of Bayesian deep learning, all the weights and predicted results are treated as random variables following certain statistical distributions. The purpose of the training stage is to get the distributions of the network weights that best explain the

observed data and match our prior knowledge, while the testing stage is equivalent to sampling the posterior distributions of the predicted values or label possibilities. Therefore, from the predicted posterior distributions, we can estimate the model uncertainties for each prediction, which can give us a clue about how confident the model is about each prediction.

1.1 Specific Aims

The objective of this research is to advance the deep learning and Bayesian deep learning based techniques for a better MRI and fMRI neuroimaging processing pipeline, which is more accurate, robust and automatic. To achieve this objective, different kinds of Bayesian deep learning models were developed and applied to the challenging applications of image segmentation, image synthesis, and data classification in MR neuroimaging, to improve the models' prediction accuracy as well as give the predictive models the ability to generate model uncertainty. The specific aims focused on for the completion of the thesis are as follows:

1. *Developing a fully-automated brain extraction pipeline combining deep Bayesian CNN and fully connected three-dimensional (3D) conditional random field (CRF) for the brain extraction in nonhuman primates.*

2. *Developing a Bayesian conditional GAN with concrete dropout and model recalibration for inter-contrast image synthesis, specifically for the image transformation from T1w MR image to T2w MR image.*

3. *Developing a Bayesian Deep Neural Network and corresponding feature extraction and ranking method for brain functional connectivity gender prediction and gender related functional connectivity pattern recognition.*

1.2 Thesis Outline

According to the aims of this work, the remainder of the thesis will be composed in the following structure:

- **Chapter 2** provides a comprehensive review of the related work. This chapter first focuses on the progress of the deep learning techniques including deep neural network, convolutional neural network, generative neural network, and the important regularization technique used by deep learning models, dropout. Next, a review of the framework of Bayesian deep learning are outlined. Finally, a discussion of prior works specifically in brain segmentation, image synthesis in medical imaging, and gender predication and gender difference in neuroimaging are provided.
- **Chapter 3** discusses the framework of the developed motion tracking algorithm. The first section introduces the challenging image segmentation problem of nonhuman primate brain extraction. The following sections discuss the proposed fully-automated brain extraction pipeline combining deep Bayesian CNN and fully connected 3D CRF. The performance of the proposed method is validated with T1w MR brain volumes of 100 nonhuman primates, and is compared with the state-of-the-arts methods. The behavior of the uncertainty generated by Bayesian Neural Network is also shown.
- **Chapter 4** addresses the challenge of inter-modality MR image synthesis with the proposed Bayesian conditional GAN. The first section introduces the challenges in inter-

modality MR image synthesis in the field of medical imaging. Then, the proposed Bayesian conditional GAN with concrete dropout and model recalibration is studied. The method is validated with the T1w to T2w MR image translation with a brain tumor dataset of 102 subjects. Finally, results of the proposed method are compared with the conventional Bayesian neural network with Monte Carlo dropout. The improvement of the calibration of the uncertainty by the uncertainty recalibration method is also illustrated.

- **Chapter 5** discusses the proposed Bayesian deep neural network and the feature extraction and ranking method for brain functional connectivity gender classification. First, the importance of the brain function related behavior classification as well as the importance of feature extraction methods in brain functional related research is discussed. Then, a Bayesian deep neural network and a feature extraction and ranking method for it are proposed to solve the problem. These methods are tested with the resting state functional MRI (rs-fMRI) data of 1003 healthy subjects in the human connectome project (HCP). Finally, the behavior of the uncertainty generated and the robustness of the features extracted are also investigated.
- **Chapter 6** draws a final conclusion of the projects in the thesis and gives a discussion of the potential future work for each aim.

Chapter 2

Learning Models and Applications for MR Neuroimaging

2.1 Machine Learning

Many machine learning models have been invented in the last a few decades. Among them we have the linear regression model for regression problems, and logistic regression, perceptron and support vector machine (SVM) for classification problems. There are also k-means clustering and independent component analysis (ICA) for unsupervised learning, and principal component analysis (PCA) for dimension reduction. In this section, SVM and ICA will be reviewed, since these methods were used in the work of this thesis.

2.1.1 Support Vector Machine

As a linear classifier for a binary classification problem, SVM tries to learn a hyperplane from the training data to separate the two classes with a large margin. For each observation, x is the input

feature vector and $y \in \{-1,1\}$ is the corresponding label. Then the SVM classifier can be formulated as:

$$h_{w,b}(x) = g(w^T x + b) \tag{2.1}$$

where $g(z) = 1$ if $z \geq 0$, and $g(z) = -1$ if $z < 0$. w and b are the weights to learn for the hyperplane $w^T x + b = 0$. Given the separating hyperplane as $w^T x + b = 0$, then $w / \|w\|_2$ is the unit vector perpendicular the hyperplane. Therefore, the geometric margin for the ith observation can be written as:

$$\gamma^{(i)} = y^{(i)} \left[\left(\frac{w}{\|w\|_2} \right)^T x^{(i)} + \frac{b}{\|w\|_2} \right] \tag{2.2}$$

For a confident classifier, we want this margin as large as possible. Thus, we define the geometric margin for the training dataset as the smallest one among the all the $\gamma^{(i)}$, and we can formulate the training of the SVM as an optimization problem:

$$\begin{aligned} &\max_{\gamma,w,b} \ \gamma \\ &s.t. \qquad y^{(i)}(w^T x^{(i)} + b) \geq \gamma, i = 1,...m \\ &\qquad\quad \|w\|_2 = 1 \end{aligned} \tag{2.3}$$

The constraint, $\|w\|_2 = 1$, is added, since the rescaling of w and b won't change the geometric margin. However, this constraint makes the optimization problem non-convex, so the optimization problem is reformulated as the following with the constraint removed:

$$\begin{aligned} &\max_{\gamma,w,b} \ \frac{\hat{\gamma}}{\|w\|_2} \\ &s.t. \qquad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, i = 1,...m \end{aligned} \tag{2.4}$$

Although the non-convex constraint, $\|w\|_2 = 1$, is removed, now the objective, $\frac{\hat{\gamma}}{\|w\|_2}$, is non-convex.

Since rescaling *w* and *b* by a constant is equivalent to rescaling $\hat{\gamma}$ by the same constant, and won't change the geometric margin or the objective. Therefore, by introducing the scaling constraint, $\hat{\gamma} = 1$, the optimization problem can be further reformulated as:

$$\begin{aligned} &\min_{\gamma,w,b} \frac{1}{2}\|w\|_2^2 \\ &s.t. \qquad y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1,...m \end{aligned} \tag{2.5}$$

Now the optimization problem is one with a convex quadratic objective and linear constraints, and it can be solved with quadratic programming. Its solution is the optimal margin classifier.

Sometimes the observed data are not linear separable, and a linear separating hyperplane could also be susceptible to outliers. To make the model more robust to these cases, the optimization problem can be reformulated as:

$$\begin{aligned} &\min_{\gamma,w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m}\xi_i \\ &s.t. \qquad y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, i = 1,...m \\ &\qquad\quad \xi_i \geq 0, i = 1,...m \end{aligned} \tag{2.6}$$

Thus, in the new formulation a margin less than 1 is permitted with a penalty of $C\xi_i$ in the objective function. *C* is the weight to trade off the balance between the two terms in the objective function.

2.1.2 Independent Component Analysis

Independent component analysis (ICA) is a model for recovering the *n* independent sources $s \in \mathbb{R}^n$, given the observations x as mixtures of the sources:

$$x = As \tag{2.7}$$

where A is the mixing matrix. Repeated observations can generate a dataset $\{x^{(i)}; x^{(i)} = As^{(i)}, i = 1,...,m\}$. $s^{(i)}$ is the ith data generated from the n independent sources before mixture, and $x^{(i)}$ is the ith observation after mixture. Each $s^{(i)}$ is an n-dimensional vector, and the goal of the problem is to recover all the $s^{(i)}$, given $x^{(i)}$. Let $W = A^{-1}$ be the unmixing matrix, then if W can be found, the sources can be recovered by $s^{(i)} = Wx^{(i)}$.

Given that each source s_i's probability density function is p_s, and that the sources are independent, the problem can be solved with the maximum likelihood estimation (Bell and Sejnowski, 1995):

$$p(s) = \prod_{i=1}^{n} p_s(s_i) \tag{2.8}$$

With $x = As = W^{-1}s$, the probability density function of x can be solved as:

$$p(x) = \prod_{i=1}^{n} p_s(w_i^T x)|W| \tag{2.9}$$

where w_i^T is the ith row of W, and $|W|$ is the determinant of W. Without any prior knowledge of the sources' probability density functions, the derivative of the sigmoid function, $g(s) = 1/(1+e^{-s})$, can be assigned as the probability density function for each source, which works well for most situations. Given a training set $\{x^{(i)}; i = 1,...,m\}$, the log likelihood can be given as:

$$l(W) = \sum_{i=1}^{m} (\sum_{j=1}^{n} \log g'(w_i^T x) + \log|W|) \tag{2.10}$$

With stochastic gradient ascent the log likelihood can be maximized with respect to W. For each training data point $x^{(i)}$, the update rule for each iteration is:

$$W := W + \alpha \left(\begin{bmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2g(w_n^T x^{(i)}) \end{bmatrix} x^{(i)^T} + (W^T)^{-1} \right) \tag{2.11}$$

There are also some limitations of the ICA model: it can solve neither the permutation of the original sources, nor the correct scaling of the weights in the matrix, W. However, these usually do not matter for applications in reality, like the ICA for signal decomposing in rs-fMRI. The ICA model also cannot solve the problem with sources following Gaussian distributions, and this is usually not the case for the rs-fMRI data.

2.2 Deep Learning

In the recent past, tremendous progress has been made in deep learning as a result of the revival of deep neural networks (Krizhevsky et al., 2012a; LeCun et al., 2015) as well as the rapid advance of parallel computing techniques (Coates et al., 2013; Schmidhuber, 2015). Among many deep learning models, convolutional neural networks (CNN) are good at prediction tasks with images as input; recurrent neural networks (RNN) are frequently used to process tasks with sequential data; deep reinforcement learning (RL) models have been proven effective on sequential decision making and control problems. Multiple deep learning platforms, including TensorFlow (https://www.tensorflow.org/), Caffe (https://caffe.berkeleyvision.org/), PyTorch (https://pytorch.org/) and Theano (http://deeplearning.net/software/theano/) etc., were also released publicly to make the development, training and model sharing of deep learning easier. In this section deep learning models dealing with images and the features extracted from images will be reviewed.

2.2.1 Convolutional Neural Network

Convolutional neural network (CNN) is often used for image processing tasks (LeCun et al., 1999). Convolutional encoder networks can be used for image classification tasks, while convolutional encoder-decoder networks are often used for image segmentation tasks. The encoder network is used to extract features from the input image, and the decoder network is used to generate the target image or mask with the learned features as recovering the original resolution of the input image.

A CNN usually contains several kinds of layers, including convolutional layers, batch normalization layers, activation layers, pooling layers, and fully connected layers. The convolutional layer is the core building block of a CNN. It contains convolutional kernels to extract the features from the input image. All the kernel weights are learned from the training of the CNN. Batch normalization layers (Ioffe and Szegedy, 2015) are usually inserted immediately after convolutional layers and before the activation layers. They are used to force the neuron activations to follow normal distribution throughout the neural network. Therefore, the neuron values locate mostly around the high gradient region of the activation function, and this results in fast convergence of the network parameters during training. Activation layers are used after each convolutional layer or fully connected layer. They are used to add nonlinearity to the model, and this can give the model the ability to imitate the behavior of highly nonlinear functions. Commonly used activation functions include sigmoid, Tanh and ReLU etc. Pooling layers are usually periodically inserted into a CNN between the successive convolutional layers. It is used to reduce the spatial size of the layers with feature-representing neurons. Thus, it has the effect of reducing the total amount of parameters and computation in the neural network, and therefore can prevent overfitting. Fully connected layers use full connections to calculate each neuron in the current

layer with all the neurons in the previous layer. A fully connected layer can be formulated as a matrix multiplication followed by a bias, and is often used to change the spatial size of a layer. For fast converging during training, before being fed into the CNN, input images are usually normalized to zero mean and unit variance.

For both binary classification and multiclass classification tasks, cross-entropy loss is often used as the objective function. For image segmentation tasks, in addition to the cross-entropy loss, a variety of loss functions have been proposed to enhance the segmentation performance, including Dice loss (Milletari et al., 2016), generalized Dice loss (Sudre et al., 2017), sensitivity-specificity loss (Brosch et al., 2015), and generalized Wasserstein Dice loss (Fidon et al., 2018) etc. The stochastic gradient decent (SGD) algorithm is commonly used for the training of CNN and other deep neural networks. Other gradient decent algorithms, like Momentum (Qian, 1999), Adaptive Moment Estimation (Adam) (Kingma and Ba, 2014) and Adadelta (Zeiler, 2012) etc., were also proposed for the fast training of CNN and other deep learning models.

2.2.2 Generative adversarial Network

Generative adversarial network (GAN) usually consists of a generator network and a discriminator network, and is widely used for image synthesis tasks (Goodfellow et al., 2014). For image synthesis, the generator network is usually a convolutional encoder-decoder network used to synthesize images in the output domain from the images in the input domain, while the discriminator network could be a convolutional encoder network used to discriminate the difference between the generated images and the target images. The objective function for GAN usually looks like this:

$$\min_{\theta_g} \max_{\theta_d} \mathbb{E}_{X,Y}[\log D_{\theta_d}(y)] + \mathbb{E}_X[\log(1 - D_{\theta_d}(G_{\theta_g}(x)))] \tag{2.12}$$

where θ_g is the parameter set of the generator network; θ_d is the parameter set of the discriminator network; $x \in X$ is the input image, and $y \in Y$ is the target image for the synthesized image. *G* and *D* are the forward models of the generator network and the discriminator network respectively. The training of GAN optimizes the generator and the discriminator alternatively in each iteration, and the goal is to train a discriminator that can discriminate the difference between the synthesized image and the target image, and at the same time to train a generator that can fool the discriminator. Based on the GAN architecture, conditional GAN (Isola et al., 2016) was proposed and became a more accurate and consistent way to synthesize paired images. It concatenates the input image as the condition for the synthesized image and the ground truth image for the discriminator. This gives the discriminator a clue about the input domain image that the output domain image paired with, and can further improve the image synthesis accuracy for paired data. The objective function used by conditional GAN is:

$$\min_{\theta_g} \max_{\theta_d} \mathbb{E}_{X,Y}[\log D_{\theta_d}(y \mid x)] + \mathbb{E}_X[\log(1 - D_{\theta_d}(G_{\theta_g}(x) \mid x))] \tag{2.13}$$

The l_1 norm between the synthesized image and the ground truth image can also be added to this loss function to further emphasize the similarity between them.

Compared with synthesizing images using only the CNN generator with the l_1 norm or the l^2 norm between the synthesized image and the ground truth image as the objective function, the conditional GAN works much better with the discriminator as part of the objective function. The images synthesized by conditional GAN are sharper with more details captured by the model.

2.2.3 Deep Neural Network

Deep neural network (DNN) consists of many fully connected layers, and usually uses the sigmoid function as the activation function. It is often used for classification tasks with feature vectors as input. For rs-fMRI data, the connectivity matrices can be extracted from the original 4D rs-fMRI data, and then be fed into DNN for classification. Studies have shown high accuracy of DNN for the classification between schizophrenia patients and healthy controls with the rs-fMRI data (Kim et al., 2016).

2.2.4 Dropout

Dropout (Srivastava et al., 2014) is an effective regularization technique for preventing overfitting. In each forward pass during training, the dropout layer randomly sets the activations of the neurons in the previous layer as zero with a probability p, the dropout rate or dropout probability. In this way, all the features in the neural network get the chance to be trained, and the neural network can prevent from only depending on the several important features, which usually causes overfitting. During testing time, no dropout is performed, and the dropout layer rescales the neurons' activations in the previous layer to 1-p times their original values during the forward propagation, This has an effect of keeping the expectation of the activation magnitude in the testing stage at the same level as that in the training stage.

2.3 Bayesian Deep Learning

Being different from conventional deep learning models, the framework of Bayesian deep learning views all the weights to train in the neural network and the values to predict during the testing stage as random variables following certain probability distributions.

2.3.1 Training of Bayesian Deep Learning

The training goal of Bayesian deep learning is to get the posterior $p(\omega \mid X, Y)$. However, usually the true posterior $p(\omega \mid X, Y)$ is intractable analytically, so a variational distribution $q_\theta(\omega)$ parametrized by θ is used to approximate the true posterior. Thus, the training goal of the Bayesian deep learning is transferred to minimize the Kullback-Leibler (KL) divergence with respect to θ (Gal and Ghahramani, 2015):

$$KL(q_\theta(\omega) \,||\, p(\omega \mid X, Y)) = \int q_\theta(\omega) \frac{q_\theta(\omega)}{p(\omega \mid X, Y)} d\omega \tag{2.14}$$

With the techniques in variational inference, it can be proved that minimizing the KL divergence is equivalent to maximizing the evidence lower bound (ELBO). Thus, the optimization problem can be reformulated as maximizing the following objective (Gal and Ghahramani, 2015):

$$\mathcal{L}_{VI}(\theta) := \int q_\theta(\omega) \log p(Y \mid X, \omega) d\omega - KL(q_\theta(\omega) \,||\, p(\omega)) \leq \log p(Y \mid X) \tag{2.15}$$

The first term in this equation drives the approximating distribution $q_\theta(\omega)$ to explain the observed data well, and at the same time the second term encourages $q_\theta(\omega)$ to be like the prior distribution $p(\omega)$. By reparametrizing $q_\theta(\omega)$ with $\omega = g(\theta, \epsilon)$, where θ is a set of parameters and ϵ is a set of random variables having the probability density function of $p(\epsilon)$, the objective can be further reformulated as minimizing the following objective (Gal, 2016):

$$\widehat{\mathcal{L}}_{VI}(\theta) = -\int p(\epsilon) \log p(Y \mid f^{g(\theta,\epsilon)}(X)) d\epsilon + KL(q_\theta(\omega) \,||\, p(\omega)) \tag{2.16}$$

where f is the forward model. Given a training dataset with N observations and using mini-batch optimization, the objective can be rewritten as (Gal, 2016):

$$\widehat{\mathcal{L}}_{VI}(\theta) = -\frac{N}{M} \sum_{i \in S} \int p(\epsilon) \log p(y_i \mid f^{g(\theta,\epsilon)}(x_i)) d\epsilon + KL(q_\theta(\omega) \,||\, p(\omega)) \tag{2.17}$$

where *S* is a random mini-batch with the size of *M*. With the technique of Monte Carlo integral, the expected log likelihood in the objective can be replaced with its stochastic estimator (Gal, 2016):

$$\begin{aligned} &\widehat{\mathcal{L}}_{MC}(\theta) = -\frac{N}{M}\sum_{i\in S}\log p(y_i \mid f^{g(\theta,\epsilon)}(x_i)) + KL(q_\theta(\omega)\,||\,p(\omega)) \\ &s.t. \qquad \mathbb{E}_{S,\epsilon}(\widehat{\mathcal{L}}_{MC}(\theta)) = \widehat{\mathcal{L}}_{VI}(\theta) \end{aligned} \tag{2.18}$$

Given that the prior of the network weights $p(\omega)$ following Gaussian distributions, it can be proved that the training of a Bayesian neural network is equivalent to the training of the conventional neural network with the regularization of dropout and the square of the l^2 norm of the weight matrices (Gal, 2016).

2.3.2 Testing of Bayesian Deep Learning

Instead of a single number, Bayesian deep learning treats each prediction as a posterior distribution. Thus, at testing time, the goal of Bayesian deep learning it to estimate the posterior distribution (Gal and Ghahramani, 2016):

$$p(y^* \mid x^*, X, Y) = \int p(y^* \mid x^*, \omega)\, p(\omega \mid X, Y)\, d\omega \tag{2.19}$$

With the approximating distribution of the weights got during the training stage, this posterior distribution can be approximated as:

$$p(y^* \mid x^*, X, Y) \approx \int p(y^* \mid f^{\omega}(x^*))\, q_\theta^*(\omega)\, d\omega =: q_\theta^*(y^* \mid x^*) \tag{2.20}$$

Given the assumption that $p(y^* \mid f^{\omega}(x^*)) = \mathcal{N}(y^*; f^{\omega}(x^*), \tau^{-1}I)$ with $\tau > 0$, it can be proved that $\mathbb{E}_{q_\theta^*(y^*|x^*)}[y^*]$ can be estimated with the unbiased estimator (Gal and Ghahramani, 2016):

$$\widetilde{\mathbb{E}}[y^*] \coloneqq \frac{1}{T}\sum_{t=1}^{T} f^{\widehat{\omega}_t}(x^*) \underset{T\to\infty}{\to} \mathbb{E}_{q_\theta^*(y^*|x^*)}[y^*] \tag{2.21}$$

With the same assumption, $Var_{q_\theta^*(y^*|x^*)}[y^*]$ can be estimated with the unbiased estimator (Gal and Ghahramani, 2016):

$$\widetilde{Var}[y^*] \coloneqq \tau^{-1} I + \frac{1}{T}\sum_{t=1}^{T} f^{\widehat{\omega}_t}(x^*)^T f^{\widehat{\omega}_t}(x^*) - \widetilde{\mathbb{E}}[y^*]^T \widetilde{\mathbb{E}}[y^*] \underset{T\to\infty}{\to} Var_{q_\theta^*(y^*|x^*)}[y^*] \tag{2.22}$$

This means that the mean and variance of the T stochastic forward passes through the Bayesian neural network can be used to estimate the mean and variance of the true posterior $p(y^* \mid x^*, X, Y)$.

2.4 Physics of MRI

2.4.1 Physics Overview

MRI is based on the quantum mechanics of atomic nuclei. Protons and Neutrons are the nucleons used to make the nucleus of an atom. Both of them have the quantum mechanical property of spin. Spin is measured in the discrete half-integer unit. Only nuclei with non-zero spins have magnetic moments, and therefore can absorb and emit electromagnetic radiation and can have resonance in an external magnetic field. The MR-active nuclei are those that have odd-numbered spins, so the spins of the protons and neutrons don't cancel each other, and result in a net spin. In the clinical MRI, the hydrogen nucleus (proton), ^{1}H, is the most common source of signal, due to its abundance in the human body. ^{1}H is a spin-1/2 nucleus, and there are two spin states, up and down.

Each spin rotates around its own axis, which induces a magnetic field. When the spins are exposed to a strong external magnetic field (B_0), the magnetic field interaction causes the spins to precess. The frequency of the precession is defined by the Larmor equation:

$$\omega_0 = \gamma B_0 \tag{2.23}$$

where ω_0 is the frequency of the precession; γ is the gyromagnetic ratio, which is a constant for every atom; B_0 is the external magnetic field strength. For ^{1}H, $\gamma = 42.57$ MHz/T. Before the external magnetic field is applied, the rotation axes of the spins are randomly aligned. When being exposed to the B_0, the spins starte to precess around the axes along the magnetic axis of B_0: some axes are parallel with B_0, while the others are anti-parallel with B_0. For nuclei having odd-numbered spins, more spins precess around the axes parallel to B_0, since this is a lower energy state. The cumulative effect of this results in a net magnetization vector parallel to B_0.

During an MRI session, the radio frequency (RF) pulses are turned on and off. When the RF pulse has the same frequency as the precessional frequency, the phenomenon of resonance emerges. The RF pulse can transfer energy to the spins, which has two main effects on the spins. First, some spins acquire energy from the RF pulse and move to the higher energy state with being antiparallel to B_0. In consequence, the parallel and antiparallel spins cancel each other and result in reduced longitudinal magnetization or even a growth of longitudinal magnetization in the antiparallel direction. Second, the transference of energy from the RF pulse to the spins causes the spins to precess in phase, and results in a transverse magnetization. The precessing transverse magnetization at the Larmor frequency can be captured by the receiver coil.

As soon as the RF pulse is turned off, the spins start to return to the lower energy state as well as fall out of phase. These are the T1 and T2 relaxations, or spin-lattice and spin-spin relaxations. The recovery of the longitudinal magnetization occurs exponentially with the time constant T1, so it is called T1 relaxation. It is also called spin-lattice relaxation, since the spins return to the low

energy state by emitting energy to their surroundings. The loss of phase coherence of the spins causes the exponential decay of the transverse magnetization with the time constant T2 and thus is called T2 relaxation. The transverse magnetization is dephased due to the interaction between the spins and their magnetic fields, so it is also called spin-spin relaxation. In practice, because of the inhomogeneities of B_0 and the susceptibility boundaries in the sample, there are small differences in the static magnetic field at different locations. This causes the spins to be dephased faster, and is often referred to as T2' relaxation. The combination of T2 and T2' relaxations is referred to as the T2* relaxation, and these time constants have the following relation:

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2'} \tag{2.24}$$

The time constant T1 is much longer than T2, and T2* is always shorter than T2 (Bitar et al., 2006; Currie et al., 2013).

2.4.2 Tissue Contrast

Different tissues have different T1, T2 and proton density properties, and the tissue contrasts in MR images are basically generated by these different tissue properties. In MR pulse sequences, the repetition time (TR) and the echo time (TE) are the two key parameters for the creation of different tissue contrasts. TR is the time cycle between two RF excitation pulses, while TE is the time from the application of the RF excitation pulse to the amplitude peak of the received signal (echo). All types of MR images are affected by the tissue property parameters of T1, T2 and proton density, but the adjustment of TR and TE can emphasize a specific type of contrast mechanism, since TR and TE are sensitive to different spin relaxation processes. The combination of short TR and short TE emphasizes the T1 differences between different tissue types, and can be used for the

T1w MR images. Meanwhile, long TR and long TE lets the longitudinal net magnetization from different tissues fully recover, and can show the differences of the T2 relaxation of different tissues, so this combination is usually used for T2w MR images. When TR is long and TE is short, the amplitude of the signal is mainly determined by the proton density of the tissue, so this combination is often used for the proton-density-weighted (PDw) MRI (Bitar et al., 2006; Currie et al., 2013).

2.4.3 Pulse Sequence

Pulse sequences are the wave forms and their corresponding timing of the gradients and the RF pulses used for MR image acquisition. The spin echo (SE) pulse sequence and the gradient echo (GRE) pulse sequence are the two fundamental pulse sequences of MRI, and other MR pulse sequences can be viewed as the variations of them. In addition, pulse sequences can be two-dimensional (2D) or 3D. 2D pulse sequences use one of the gradients to perform slice selection, while for 3D sequences there is no slice selection, but the phase encoding is performed in two separate directions.

In SE sequence, first a 90-degree RF pulse tips the net magnetization vector into the transverse plane. As the spins go through the T2 and T2* relaxations, the spins are dephased and the net magnetization in the transverse plane decreases. At the time point of half TE a 180-degree RF pulse is applied to flip the spins 180 degrees. Since the spins are still in the same location with the same local magnetic field inhomogeneity, the spins will start to be rephased. At the time point of TE, the phase differences of the spins caused by the T2’ relaxation are eliminated, and the echo is produced and read out, so SE is able to acquire images with T2w contrasts. With the adjustment of the TE and TR, SE can also acquire images with other contrasts, e.g. T2w and PDw etc.

In a GRE sequence, the RF pulse is used to partially tip the net magnetization down to the transverse plane with variable flip angles, and the application of gradient causes the spins to be dephased and rephased in the transverse plane. The gradient reversal only refocuses the spins that were dephased by the gradient instead of the magnetic field inhomogeneity, so the image contrast in GRE sequence is affected by the proton density, T1 and T2* relaxations, but not the T2 relaxation. In a GRE sequence, TE is the time taken from the decay of the signal to the time point when the signal reaches its maximum. Without the 180-degree RF, GRE sequences can take shorter time than the SE sequences, and the combination of short TE and short TR allows rapid signal acquisition. In GRE MRI, T1w, T2*-weighted (T2*w) and PDw images can be obtained through the manipulation of the flip angle, TE and TR (Bitar et al., 2006; Currie et al., 2013).

2.4.4 Signal Localization and Imaging

Gradients are used to change the magnetic field strength linearly along the selected directions. According to the change of the magnetic field strength, the precessional frequency of the spins also changes. In total, there are three kinds of gradients, X, Y and Z, along the orthogonal axes in the 3D space. For 2D MRI, one gradient is used as the slice-selection gradient. It is first applied to select the slice to be imaged. Then, another gradient is applied as the phase-encoding gradient. It causes the spin phase shift proportional to the location in the phase-encoding direction. Finally, the last gradient is used as the frequency encoding gradient. It causes the spins to precess at different frequencies along the frequency encoding direction. In this way, the position information of all the spins are encoded in the signal. The signal can be collected along certain trajectories in the K-space, the source of the signal can be recovered by the 2D Fourier transformation. For 3D MRI, the slice selection gradient is replaced by another phase-encoding gradient, and the signal can be recovered by the 3D Fourier transformation (Bitar et al., 2006; Currie et al., 2013).

2.5 Applications of Deep Learning in MR Neuroimaging

In this section, three key applications in the processing pipeline of MR neuroimaging are discussed, which include image segmentation, image synthesis, and feature classification and extraction. Conventional methods and learning based models used to solve the tasks in these applications are reviewed and discussed.

2.5.1 Brain Extraction in Humans and Nonhuman Primates

A large number of brain extraction methods have been proposed in recent decades, which again emphasize its importance. However, the need for an accurate, robust and sufficiently fast method has not yet been fulfilled. Mainly, these methods can be divided into two categories, edge based methods and template based methods (Roy et al., 2017). Although most of these methods work well for human brains, they encounter challenges when dealing with nonhuman primate brains due to their complex anatomical structure (Wang et al., 2014). A comprehensive review can be found at (Roy et al., 2017; Wang et al., 2014). Due to the sub-optimal performance of existing automated brain extraction routines in rhesus monkeys, prior work from our laboratory has used brain images that were manually extracted by well-trained experts (Oler et al., 2010; Birn et al., 2014; Fox et al., 2015a). This procedure, however, is extremely time consuming and labor intensive.

The Brain Extraction Tool (BET) (Smith, 2002) is based on a deformable tool, which initializes a spherical mesh at the center of gravity of the brain, and then expands it towards the edge of the brain. The whole process is guided by a set of locally adaptive forces determined by surface smoothness and contrast changes in the vicinity of the surface. This toolbox has been reported to be useful for nonhuman primate brain extraction (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET/FAQ).

The Brain Surface Extractor (BSE) was proposed by (Shattuck et al., 2001). It involves anisotropic filtering, Marr-Hildreth edge detection and morphological operations. Serving as an edge-preserving filter, the anisotropic diffusion filtering step is intended to smooth gentle gradients, while preserving sharp gradients, which helps the edge detection. Morphological operations, including erosion and dilation, can further enhance the results from edge detection.

Another popular brain extraction tool is 3dSkullStrip in the Analysis of Functional NeuroImages (AFNI) software suite (https://afni.nimh.nih.gov/afni/). It consists of 3 steps: removing gross non-uniformity artifacts, iteratively expanding a spherical surface to the edge of the brain and creating masks and surfaces of the brain. The second step is a modified version of BET. The improvement includes excluding eyes and ventricles, driving the expansion with data both inside and outside of the surface, and involving 3D edge detection. It has a -monkey option helping with initialization of a surface based on nonhuman primate brains.

The Hybrid Watershed Algorithm (HWA) (Ségonne et al., 2004) combines watershed algorithm with a deformable surface model. Based on the 3D white matter connectivity, the watershed algorithm outputs the initial estimate of the brain volume, and then the deformable model generates a force field to drive a spherical surface to the boundary of the brain. The option -atlas can incorporate a statistical brain atlas generated from accurately segmented human brains to correct the segmentation. The HWA in the software FreeSurfer is not originally designed or optimized for nonhuman primates.

Robust Learning-Based Brain Extraction (ROBEX) (Iglesias et al., 2011) is a more recently published algorithm. In ROBEX a discriminative method is combined with a generative model. A random forest classifier is trained to detect the contour of the brain, after the subject is registered to the template with an affine transformation, and then a point distribution model is fitted to find

the most likely boundary of the brain. Finally, a small deformation optimized by graph cuts serves as the refining step.

Recently the National Institute of Mental Health Macaque Template (NMT) was published by (Seidlitz et al., 2017). It is a high-resolution template for the macaque brain derived from MRI images averaged from 31 subjects. Rhesus macaque brain MRI images can be registered into this template space to extract the brains. After brain extraction in the template space, the brain images can be transformed back to the original image space with the inverse transformation. To increase the accuracy of the registration, both affine and deformable transformations should be involved.

Among many deep learning methods, deep CNN has proven to be very useful in a broad range of computer vision applications, outperforming traditional state-of-the-art methods (Shelhamer et al., 2016; Simonyan and Zisserman, 2014). In the area of image segmentation, many CNN based architectures have been proposed. These methods can broadly be divided into 2D methods and 3D methods. A c (FCN) was proposed by (Long et al., 2015) as a 2D network for the general task of semantic segmentation. SegNet (Badrinarayanan et al., 2015b) was first proposed for road and indoor scene segmentation, and later was combined with a 3D deformable model to solve tissue segmentation in MRI (Liu et al., 2017). UNet (Ronneberger et al., 2015) is a kind of 2D encoder-decoder network proposed for microscopic images, and later expanded to 3D for volumetric data (Çiçek et al., 2016). VNet (Milletari et al., 2016) was proposed as a 3D FCN with dice loss to perform 3D segmentation on MR images. (Wachinger et al., 2017) proposed a 3D patch-based method and arranged 2 networks hierarchically to separate the foreground and then identify 25 brain structures. (Chen et al., 2017) proposed a 3D residual network with multi-modality and multi-level information to identify 3 key structures of the brain. Recently, Several patch-based 3D FCN were also proposed, like LiviaNet to segment the subcortical region of the brain (Dolz et al., 2017)

and DeepMedic to segment brain lesions (Kamnitsas et al., 2017). In the specific field of brain extraction, (Kleesiek et al., 2016) proposed a 3D patch based CNN network for human brain extraction on T1w human brain datasets and a multi-modality human brain dataset with tumors. In a further study, (Salehi et al., 2017) proposed an auto-context CNN where the probability maps output by the CNN are iteratively used as input to the CNN along with the original 2D image patches to refine the results. A more comprehensive review can be found at (Bernal et al., 2017; Craddock et al., 2017).

2.5.2 Image Synthesis in Medical Imaging

Image synthesis is a technique used to translate images in one domain to their corresponding images in another domain, or images in one contrast to their corresponding images in another contrast. It is especially useful in the field of medical imaging. For example, if the MR images with some specific contrast were not collected during data acquisition and were found useful during diagnosis, image synthesis can generate the images with this contrast from other existing images of other contrasts. Images with high resolution can be synthesized from images with low resolution, and artifacts can be removed via image synthesis. CT images can be synthesized through PET or MR images, which can potentially reduce the radiation dose to patients for diagnosis, and increase the accuracy of the processes where CT images are unavailable, such as PET attenuation correction (Liu et al., 2018) and MR only treatment planning of the magnetic resonance - linear accelerator (MR-Linac) (Guerreiro et al., 2017).

Conventional image synthesis approaches can be classified into two categories: registration based image synthesis and intensity transformation based image synthesis. Usually, the problem of image synthesis can be formulated as: given an image a_1 in contrast A, synthesizing its corresponding

image b_1 in contrast B, with an image set, *a*, in contrast A and its corresponding image set, *b*, in contrast B. Usually, the image set *a* and its paired image set *b* are co-registered.

For the registration based image synthesis, the image *a* is registered to a_1 with deformable registration algorithms, and then the obtained deformable transformation is applied to *b* to generate a_1's paired image b_1 in contrast B. This method was first used to synthesize positron emission tomography (PET) images from MR images as a single atlas registration and transformation approach with a single aligned image pair in set *a* and set *b* (Miller et al., 1993). Then, this method was extended to use multiple aligned image pairs in set *a* and set *b*. All the images in a are registered to a_1 with deformable registration. Then the same deformable transformations are applied to the corresponding paired images in *b*, and an intensity fusion is performed to synthesize the intensity of each voxel of b_1. This extended method was first used to synthesize the computed tomography (CT) images from MR images for the attenuation correction in PET reconstruction (Burgos et al., 2014, 2013). Obviously, the performance of registration based image synthesis highly depends on the accuracy of the deformable registration, which is challenging for brain images with many fine structures. Moreover, this method cannot be applied to the situations in which the a_1 image has abnormal anatomy, for example, brain images with stroke, brain tumors or multiple sclerosis. Since the anatomic structures in *a* and *b* do match with those in a_1, the method is not able to generate a_1's corresponding structures in b_1 with contrast B (Burgos et al., 2014; Cardoso et al., 2015; Miller et al., 1993).

Intensity transformation based image synthesis can be viewed as a supervised prediction approach. First, for every voxel location in an image in set *a*, a feature vector is extracted. The feature vector, for example, could be a 3D patch centered at that voxel. Each feature vector has a target intensity value, which is the voxel intensity at the same location in the corresponding image in set *b*. The

training dataset is created by extracting the feature vectors and generating the feature-vector-target-intensity pairs across all in the images in the image sets *a* and *b*. Then, a regression algorithm can be learned to map the feature vectors to their target intensity values. During the prediction stage, feature vectors can be extracted from the image a_1 to generate the corresponding voxel intensities for the image b_1 with the learned regression algorithm. Image analogy was one of the earliest intensity transformation models proposed for image synthesis (Hertzmann et al., 2001). The patch extracted from a_1 is matched to the its *k* nearest-neighbor patches in the patches extracted from *a*. Then the *k* corresponding patches extracted from *b* were combined to generate the patch for b_1. In MR image synthesis, this method was used for the purpose of image registration (Iglesias et al., 2013). MIMECS is another image synthesis method proposed for intensity transformation, which solves the problem with dictionary learning (Roy et al., 2013, 2011). Patches extracted from *a* are treated as a set of bases of a dictionary, and are used to sparsely represent the patches extracted from a_1 with the linear combinations of these bases. Each basis patch extracted from *a* has its corresponding patch extracted from *b*, and the linear combinations of these corresponding patches from *b* with the same weights are used to generate the patches for b_1. Intensity transformation based methods are usually computationally intensive, and cannot synthesize all the contrasts flexibly (Roy et al., 2013). For a more comprehensive review, please refer to (Jog, 2016). As learning based models develop rapidly in the recent past, machine learning and deep learning based techniques were also applied on the task of image synthesis. REPLICA encodes both global and local information in the feature vectors and uses random forests to learn the nonlinear regression for image synthesis (Jog et al., 2017). CNN was used to synthesize the CT images from the corresponding PET images for the attenuation correction in PET reconstruction (Liu et al.,

2018). CT images were also synthesized from MR images by CNNs (Xiang et al., 2018) and GANs (Nie et al., 2018) for patient dose reduction and MR-only treatment planning in radiation therapy.

2.5.3 Functional Connectivity Gender Difference

Morphologically, men in general have a slightly larger brain as well as gray and white matter tissue volumes than women (Ruigrok et al., 2014). In addition, the hippocampus, amygdala, neocortex, insula and many other brain regions in charge of different kinds of cognitive processing are sexually dimorphic (Cahill, 2006). (Sowell et al., 2007) showed a difference in cortical thickness between genders across a large age range, and (Lv et al., 2010) reported significant cortical thickening in the frontal, parietal and occipital lobes in women. In diffusion tensor imaging (DTI) studies, females were found having significantly lower fractional anisotropy (FA) in the right deep temporal region and microstructural organization in multiple white matter regions suggested a sexual dimorphism (Hsu et al., 2008), while (Ingalhalikar et al., 2014) illustrated more inter-hemispheric connectivity in females and more intra-hemispheric connectivity in males. (Feis et al., 2013) achieved very high gender prediction accuracy (96%) when using multimodal anatomical and diffusion MR images.

Structural brain differences can help us understand the gender related psychological and behavioral differences to some extent, and functional brain differences can let us move one step further (Gong et al., 2011). Many studies have shown gender differences in FC derived from fMRI in the last decade. Mainly, these studies can be divided into task based fMRI studies and resting state fMRI (rs-fMRI) studies. (Schmithorst and Holland, 2006) observed a gender-intelligence-age interaction in the FC during the silent verb generation semantic task within a large pediatric group. (Butler et al., 2007) showed the FC difference between men and women in the ventral anterior cingulate

cortex and the dorsal anterior cingulate cortex connection during a visuospatial task. In the last decade, rs-fMRI has drawn enormous attention due to its ability to investigate the spontaneous and intrinsic brain activities. (Bluhm et al., 2008) reported the gender difference in the default mode network. (Biswal et al., 2010) examined the gender difference of resting FC on a large dataset of 1414 subjects collected at 35 international centers, in which men and women showed different connectivity strength in multiple brain connections. Similarly, (Filippi et al., 2013) showed the gender difference between different resting state networks with statistical parametric mapping and ICA. Recently, (Zhang et al., 2016) used regression and graph theory analyses to show the gender differences in resting FC. For a more comprehensive review, please refer to (Gong et al., 2011; Zhang et al., 2018).

Compared with analyzing the FC group mean differences between genders, extracting important features from an accurate prediction model can disclose the direct relationship between FC and genders. (Casanova et al., 2012) used lasso regression and random forest based methods, and reached 62.3% and 65.4% accuracy respectively on a 148-subject dataset. (S. M. Smith et al., 2013) applied leave-one-out training and testing with multivariate linear discriminant analysis to 104 subjects from Human Connectome Project (HCP) and achieved 87% gender prediction accuracy. Recently, (Zhang et al., 2018) achieved 87% accuracy with partial least squares regression on 820 HCP subjects with 10-fold cross validation. All these studies also tried to extract the important features from the prediction models to study the FC characteristics of different genders.

In the recent past, tremendous progress has been made in the field of artificial intelligence because of the resurgence of the deep learning based methods (Krizhevsky et al., 2012a; LeCun et al., 2015) and the rapid advance of parallel computing (Coates et al., 2013; Schmidhuber, 2015). Due to its ability of accurate prediction, deep learning has been quickly applied to image processing in neuro-

imaging and classification in neuroscience. CNN based architectures are effective at brain segmentation or skull stripping, while DNN is broadly used for increasing the prediction accuracy in various neuropsychiatric disorders. (Kim et al., 2016) used DNN to classify schizophrenia patients against healthy controls with an accuracy of 85.8% and investigated multiple DNN configurations' effects on the predicting accuracy. Several groups have applied DNN based methods to the diagnosis of Alzheimer's disease (Liu et al., 2014; Suk et al., 2015; Hu et al., 2016; Bhatkoti and Paul, 2016) and showed improvement against traditional methods. (Hazlett et al., 2017) used a combination of DNN and support vector machine to study the brain development of infants at high risk for autism spectrum disorder. A more comprehensive review can be found in (Vieira et al., 2017).

Chapter 3

Bayesian Convolutional Neural Network for MRI Brain Extraction

3.1 Introduction

Brain extraction, also known as skull stripping, is an essential process in MRI and fMRI. It often serves as the first step in the preprocessing pipeline, since processing software often requires the extracted brains as sources and targets in the registration. By removing the non-brain parts, such as the skull, eyes, muscle, adipose tissue and layers of meninges etc., brain registration achieves improved performance (Wang et al., 2012). Meanwhile, the accuracy of brain extraction is important and can dramatically affect the accuracy of the following processes. Mistakenly removing brain tissues and/or retaining non-brain areas can lead to biased results of further analyses, such as the estimation of cortical thickness, parcel-wise averaged fMRI signal and voxel-based brain morphometry (Fennema-Notestine et al., 2006; Shattuck et al., 2009; van der Kouwe

et al., 2008). Accurate brain extraction is extremely challenging as a result of complex brain anatomical structure, and therefore the improvement of brain extraction still remains an intensively investigated research topic (Roy et al., 2017).

Nonhuman primates have been widely used in neuroimaging as experimental subjects due to their similarity to human beings, especially in intervention studies and studies involving radiation, contrast agent and drugs (Baldwin et al., 1993; Kalin et al., 2007; Fox and Kalin, 2014). The particularity of the nonhuman primate's brain makes the challenge of brain extraction even more difficult. Nonhuman primates' brains are smaller in size than human brains, and have complex tissue structures. The eyes of nonhuman primates are relatively larger than human beings' and surrounded by much more adipose tissue. The adipose tissue behind their eyes are close to the brain, which makes it difficult to be separated. Their frontal lobes are quite narrow and protrude sharply (Rohlfing et al., 2012b), which causes this region to be excluded by many brain extraction packages. As a result, manually examining, refining or even extracting the whole brain is often unavoidable. Therefore, an accurate and robust automatic brain extraction approach for nonhuman primates is highly demanded to mitigate the time-consuming human intervention.

The purpose of this work is to implement and validate deep learning based methods on nonhuman primate brain extraction, and to build a framework for this fully automatic approach. In this study, we propose to improve the brain extraction accuracy using a Bayesian CNN with refinement through fully connected 3D CRF. In comparison to previous brain extraction studies, our study has several novel aspects. Firstly, we evaluated brain extraction using Bayesian SegNet, a Bayesian convolutional encoder-decoder network that involves Monte Carlo dropout layers to provide additional information for model uncertainty evaluation. In our previous study, the basic version of this network, SegNet, was proven to be highly efficient in MRI tissue segmentation (Liu et al.,

2017). Secondly, we incorporated the fully connected 3D CRF as a post-processing step to regularize the result according to the fully 3D anatomical context. Fully connected 3D CRF, as a probabilistic graphic model, is helpful to improve 2D CNN segmentation results by considering the distance and contrast relationships among all the voxel pairs in the whole 3D space. Finally, on a large-scale nonhuman primate dataset, we made a full comparison of the proposed method with current state-of-the-art software packages and well-established deep learning based methods. The accuracy and robustness of these algorithms on challenging nonhuman primate brain extraction were investigated. We hypothesize that a Bayesian deep learning based image segmentation framework with fully connected 3D CRF refinement is suitable for nonhuman primate's brain extraction with improved accuracy and efficiency, and the uncertainty it generated can also reflect the confidence of the model on each prediction.

3.2 Material and Methods

3.2.1 Image Datasets

MRI data of 100 periadolescent rhesus macaques (Macaca mulatta; mean (standard deviation) age = 1.95 (.38) years; 43% female) were collected in a 3T MRI scanner (MR750, GE Healthcare, Waukesha, WI, USA) with a 16-cm quadrature birdcage extremity coil (GE Healthcare, Waukesha, WI, USA) and a stereotactic head-frame integrated with the coil to prevent motion. Immediately prior to the scan, subjects received medetomidine (30 μg/kg i.m.) and a small dose of ketamine (<15 mg/kg) for anesthesia purpose. During the scan, anatomical structures were acquired using a 3D T1-weighted inversion-recovery fast gradient echo sequence with the following imaging parameters: TE = 5.41ms, TR = 11.39, TI = 600ms, Flip Angle = 10°, NEX = 2, FOV = 140 mm, Bandwidth = 61.1 kHz. The whole brain was reconstructed into a 3D volume of 256×224 in-plane

matrix size and 0.27×0.27 mm^2 in-plane pixel size with 248 slices over 124 mm. All the brains were then manually extracted by well-trained image scientists using these T1w images with the software SPAMALIZE (http://psyphz.psych.wisc.edu/~oakes/spam/spam_frames.htm). The data used in this study are a subset of those used in our prior studies (Fox et al., 2015b; Shackman et al., 2017).

3.2.2 Full Brain Extraction Method

The proposed brain extraction pipeline is a combination of Bayesian SegNet (Kendall et al., 2015a) and fully connected 3D CRF (Krähenbühl and Koltun, 2012). As shown in Fig. 3.1 It has a training phase and a testing phase. The 3D brain image volumes and corresponding manual label volumes are treated as a stack of 2D images input to the Bayesian neural network. In the training phase, the process is formulated as an optimization problem to optimize the network parameters by minimizing the difference between the network's output and the manual labels using multinomial logistic loss (Krizhevsky et al., 2012b). In the testing stage, the network with well-trained parameters are used as a pixel-wise segmentation classifier to predict the label probability and generate model uncertainty on each pixel of new brain volumes. Finally, the predicted probabilities and the 3D brain volumes are passed to fully connected 3D CRF for refinement in the whole 3D context.

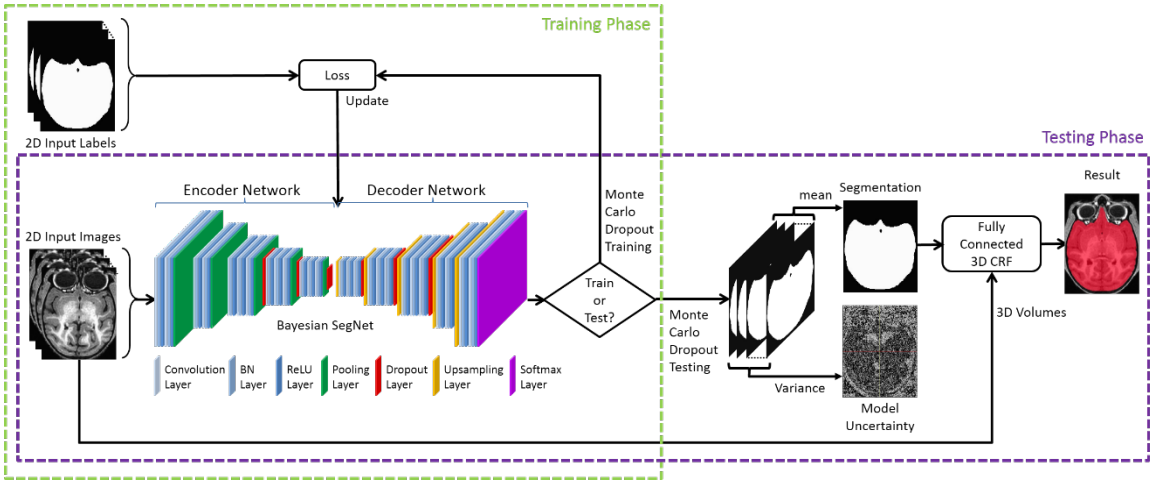

Fig. 3.1. Work flow of the proposed brain extraction method, a combination of Bayesian SegNet and fully connected 3D CRF.

3.2.3 Bayesian Convolutional Neural Network

A convolutional encoder-decoder network, Bayesian SegNet, is used as the core segmentation engine in the brain extraction workflow (Fig. 3.1). This network was first introduced by (Kendall et al., 2015a) and benchmarked on the multiple scene recognition datasets (Everingham et al., 2015) with excellent performance. This network consists of a VGG16 (Simonyan and Zisserman, 2014) encoder network and a reversed decoder network. The encoder network performs the function of feature extraction and data compression, while the decoder network assembles the compressed features to the original image size using extracted features via multi-scale sparse upsampling (Badrinarayanan et al., 2015b). Networks of encoders and decoders were constructed using a series of convolutional layers, batch normalization (Ioffe and Szegedy, 2015), ReLU non-linearity (Nair and Hinton, 2010), and maximum pooling layers or upsampling layers. Compared with other segmentation CNNs, Bayesian SegNet features both dropout training and dropout testing. Dropout training offers the network robustness against overfitting especially on small datasets. Dropout testing predicts both pixel-wise probability maps for all the labels as well as additional

measurement of model uncertainty which is particularly useful for accuracy evaluation. These 2 features are achieved by implementing Bayesian SegNet with Monte Carlo dropout layers as shown in Fig. 3.1. The dropout rate is set beforehand and a certain percentage of neurons in the preceding layer are randomly ignored in every iteration during training or every forward pass during testing (Srivastava et al., 2014).

Given the dataset X and its corresponding label set Y, (Gal and Ghahramani, 2015) showed that Monte Carlo dropout training can be used to evaluate the posterior distribution over the network weights W:

$$p(W|X,Y) \tag{3.1}$$

Since this posterior is not traceable directly from Bayesian SegNet, an approximation can be made by using variational inference (Gal and Ghahramani, 2015; Kendall et al., 2015a), which allows defining an approximating distribution $q(W)$ and inferring it by minimizing the KL divergence (Gal and Ghahramani, 2015):

$$KL(q(W) \| p(W|X,Y)) \tag{3.2}$$

(Gal and Ghahramani, 2015) illustrated that the integral in the KL divergence can be approximated with Monte Carlo integration over the network weights, and the process of minimizing the KL divergence is equivalent to performing Monte Carlo dropout training.

(Gal and Ghahramani, 2015) also showed that after getting the optimal weights, Monte Carlo dropout sampling can also be used in testing. To predict the label y^* for the data x^*, the posterior distribution can be determined through T times Monte Carlo dropout testing. During each testing the network weight subset $\hat{W}_t$ is occupied.

$$p(y^* \mid x^*, X, Y) \approx \int p(y^* \mid x^*, W) q(W) dW \approx \frac{1}{T} \sum_{t=1}^{T} p(y^* \mid x^*, \hat{W}_t)$$
$$\hat{W}_t \sim q(W) \tag{3.3}$$

The integral in the equation is approximated with Monte Carlo integration, which is identical to Monte Carlo dropout sampling of Bayesian SegNet during testing. This can be considered as sampling the posterior distribution over the weights to get the posterior distribution of the predicted label probabilities. The mean of sampled probabilities $p(y^* \mid x^*, \hat{W}_t)$ will be used as the prediction of the probability map for each label, and the variance of them will be used as the model uncertainty on each prediction.

3.2.4 Fully Connected Three-Dimensional Conditional Random Field

The final prediction outputs from the Bayesian SegNet are 2D probability maps for each label. To take into account the 3D contextual relationships among voxels, we propose to incorporate fully connected 3D CRF (Krähenbühl and Koltun, 2012) to refine the results from the Bayesian SegNet. Based on the probability maps from Bayesian SegNet, this approach can maximize the label agreement between voxels having similar contrasts or close to each other in the whole 3D volume by a maximum a posteriori (MAP) inference (He et al., 2004) made in the CRF defined over the full brain volume. Considering x as the label assignment for each voxel, and i, j as the voxel index ranging from 1 to the total number of voxels, to get the MAP inference optimization is carried out to minimize the Gibbs energy in the 3D space:

$$E(x) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \tag{3.4}$$

The probability result on each voxel from the Bayesian SegNet is used to build the unary potential $\psi_u(x_i)$, while the pairwise potential $\psi_p(x_i, x_j)$ depends on each voxel pair's location p_i, p_j and intensity I_i, I_j:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j)\left[\omega_1 \exp\left(-\frac{\|p_i - p_j\|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) + \omega_2 \exp\left(-\frac{\|p_i - p_j\|^2}{2\theta_\gamma^2}\right)\right] \tag{3.5}$$

In the pairwise potential, the appearance kernel and the smoothness kernel are involved (Krähenbühl and Koltun, 2012). The appearance kernel, the first exponential term in Eq. 3.5, assumes voxels close to each other or having similar contrasts tend to share the same label. The extent of each effect is controlled by θ_α or θ_β. The smoothness kernel, the second exponential term, removes isolated small regions (Shotton et al., 2009), and is controlled by θ_γ. ω_1 and ω_2 are the weights for the two kernels. The compatibility function, $\mu(x_i, x_j)$, is set as the Potts model:

$$\mu(x_i, x_j) = 1_{[x_i \neq x_j]} \tag{3.6}$$

To make the complex inference practical given a tremendous number of pairwise potentials in fully connected CRF, we use the highly efficient algorithm proposed by (Krähenbühl and Koltun, 2012), where the pairwise edge potentials are defined as a linear combination of Gaussian Kernels in the feature space. A mean approximation to the CRF distribution is made in the algorithm, and it is optimized through an iterative message passing process. (Krähenbühl and Koltun, 2012) showed that the message passing process can be performed using Gaussian filtering in the feature space. In this way, using highly efficient approximations of high-dimensional filtering, the computational complexity of message passing can be reduced from being quadratic to being linear,

with respect to the number of variables. As a result, the approximate inference algorithm for fully connected 3D CRF is linear with respect to the number of variables and sublinear with respect to the number of edges in the model.

3.2.5 Parameter Selection for Competing Methods

The proposed method was compared to six popular publicly available brain extraction software packages and three state-of-the-art deep learning based methods, including 3dSkullStrip in AFNI (17.0.09; https://afni.nimh.nih.gov/), BET (Smith, 2002) in FSL (5.0.10; https://fsl.fmrib.ox.ac.uk/fsl/fslwiki), BSE (Shattuck et al., 2001) in BrainSuite (v.17a; http://brainsuite.org/), HWA (Ségonne et al., 2004) in FreeSurfer (Stable v6.0; https://surfer.nmr.mgh.harvard.edu/), ROBEX (1.2; https://www.nitrc.org/projects/robex) (Iglesias et al., 2011), NMT (v1.2; https://github.com/jms290/NMT) (Seidlitz et al., 2017), SegNet (https://github.com/alexgkendall/caffe-segnet) (Badrinarayanan et al., 2015c), LiviaNet (https://github.com/josedolz/LiviaNET) (Dolz et al., 2017) and VNet (https://github.com/faustomilletari/VNet) (Milletari et al., 2016). For a direct comparison, SegNet used the same parameters as Bayesian SegNet. LiviaNet used all the default parameters (30 epochs; 20 subepochs per epoch; 1000 samples in each subepoch). VNet used the default parameters, 5000 iterations (500 epochs) and batch size 1 due to the limitation of GPU memory, and the learning rate was changed to 0.00015 accordingly. To determine the parameters of the other software packages, a two-step evaluation strategy was used for each software package, and the parameter selection was done under the assumption that all the subjects are similar enough to one another that they can be properly processed with one set of parameters. We first randomly chose one representative subject, varied each parameter by small increments in either direction from the default values to achieve the best accuracy by careful visual inspection by a well-trained researcher.

The selected parameters were then tested on a second randomly selected subject to verify its validity before application to the rest of the dataset. In our experiments, no obvious difference was observed on the performance of the selected set of parameters between the 2 selected subjects. With the exception of ROBEX and NMT, which do not have any parameters to tune, all the other software packages' parameters studied and their associated values used are shown in Table 3.1. For 3dSkullStrip in AFNI, besides the original method labeled as AFNI in this paper, to achieve a better performance in detecting the protruding frontal lobe, we also performed 3dSkullStrip with a coronal slice thickness reduced to one half of the original value in the headers of the data. This method is labeled as AFNI+ later in this paper. This was done because a common challenge in skull stripping rhesus macaque brains is that the ventral portion of the frontal lobe is quite narrow in the transverse plane, and the resultant high curvature of this region causes many skull stripping algorithms to exclude the anterior portions of the frontal lobe. By reducing the coronal slice thickness by a factor of 2, the curvature is reduced, and the frontal lobe is more easily retained. The slice thickness is then set back to the original value after brain extraction. Note that this procedure does not involve any resampling; only the value of the slice thickness in mm is changed. For the NMT, the AFNI functions, align_epi_anat.py and auto_warp.py (https://afni.nimh.nih.gov/) are used to carry out the 12 degrees of freedom (DOF) affine and deformable registration between the original image and the template for brain extraction in the NMT template space.

Table 3.1. Parameters studied and values used for the competing methods.

Method	Parameter	Description	Range	Optimal value
3dSkullStrip	-push_to_edge	Push to edge aggressively	w/ or w/o	w/
	-monkey	Brain of a monkey	w/ or w/o	w/
	-shrink_fac	Brain VS non-brain intensity threshold	0~1	0.5 for AFNI 0.4 for AFNI+
HWA	-less	Shrink the surface	w/ or w/o	w/o

	-more	Expand the surface	w/ or w/o	w/o
	-atlas	Use the atlas information	w/ or w/o	w/
BET	-f	Fractional intensity threshold	0.1~0.9	0.3
	-g	Vertical gradient	-1~1	-0.5
	-r	Head radius	30~50	35
BSE	-d	Diffusion constant	5~35	25
	-s	Edge detection constant	0.10~0.80	0.69
	-p	Dilate final mask	w/ or w/o	w/o

3.2.6 Metrics for Comparison

Several quantitative metrics commonly used in image segmentation were used to evaluate the performance of all the compared brain extraction methods (Kleesiek et al., 2016; Taha and Hanbury, 2015; Wang et al., 2014). Let M and R represent the brain mask extracted by a specific method and the manually extracted reference serving as the ground truth respectively, then the following metrics can be defined: True Positive: $TP = M \bigcap R$; True Negative: $TN = \overline{M} \bigcap \overline{R}$; False Positive: $FP = M \bigcap \overline{R}$; False Negative: $FN = \overline{M} \bigcap R$; Sensitivity: $Sens = \frac{TP}{TP + FN}$; Specificity: $Spec = \frac{TN}{TN + FP}$; Absolute Error: $E_{abs} = FP \bigcup FN$. We also involved the most commonly used metrics in image segmentation, dice coefficient (DC) (Dice, 1945), maximum symmetric surface distance (or Hausdorff distance, HD) (Huttenlocher et al., 1993) and average symmetric surface distance (ASSD) (Geremia et al., 2011):

$$DC = \frac{2|M \cap R|}{|M| + |R|} = \frac{2TP}{2TP + FP + FN} \tag{3.7}$$

$$HD = \max\left(\max_{m \in \partial(M)} \min_{r \in \partial(R)} \|m - r\|, \max_{r \in \partial(R)} \min_{m \in \partial(M)} \|r - m\| \right) \tag{3.8}$$

$$ASSD = \frac{\sum_{m\in\partial(M)} \min_{r\in\partial(R)} \|m-r\| + \sum_{r\in\partial(R)} \min_{m\in\partial(M)} \|r-m\|}{|\partial(M)| + |\partial(R)|} \tag{3.9}$$

Where $|\bullet|$ means the total voxels in the set, and $\partial(\bullet)$ means the boundary of the set. The Dice coefficient is probably the most widely used metric for image segmentation. It takes the real value within [0,1], where 1 means a perfect segmentation, and 0 means there is no overlap at all. For the segmentation of a region as large as the brain, the Dice coefficient is less sensitive due to the small edge to volume ratio. Hence, we also introduced the surface distance based metrics. HD is defined as the maximum shortest Euclidean distance between two surface sets, while ASSD is defined as the average of these shortest Euclidean distances. HD or ASSD is 0 for a perfect segmentation. Both of these are used as segmentation metrics historically, but since HD is sensitive to outliers, ASSD is usually preferred (Gerig et al., 2001; Zhang and Lu, 2004).

False positive, false negative, and absolute error maps are all spatial error maps. To visualize the systematic spatial error distribution of each method, the averaged error maps are calculated. First, the 12-DOF affine registration and deformable registration are done for each subject's full brain image from the original space to the NMT (Seidlitz et al., 2017) space using AFNI's align_epi_anat.py and auto_warp.py (https://afni.nimh.nih.gov/). Then, each kind of error map for each method are transformed to the NMT space with the transformation matrices calculated in the first step. Next, each specific kind of error map is averaged across all the subjects in the NMT space. Finally, for display purposes, the natural logarithm of the averaged error maps collapsed (averaged) along each axis was plotted (Kleesiek et al., 2016; Wang et al., 2014).

3.2.7 Experiments

3.2.7.1 Brain Extraction for Nonhuman Primates

Before being sent to Bayesian SegNet as inputs, each subject's original 3D image volume was normalized to [0,1] and dissembled along the longitudinal (superior-inferior) axis into a stack of 2D images. Since manual skull stripping was done on the 3D brain volumes that were manually cropped to exclude the body of the monkey and regions far outside the brain, all the 2D images and corresponding manual labels were upsampled to the same size (352×256) with bilinear interpolation and nearest neighbor interpolation respectively. The Bayesian SegNet was trained by SGD algorithm with multinomial logistic loss in 60000 iterations (18 epochs). The learning rate during training was fixed as 0.01 with a momentum of 0.9. In testing, the number of samples were set as 6 based on (Kendall et al., 2015a) and the limitation of GPU memory. The dropout rate for all the dropout layers were set as 0.5 for both MC dropout training and testing (Kendall et al., 2015a). Fully connected 3D CRF was performed for each subject following Bayesian SegNet. The parameters for fully connected 3D CRF were empirically selected in the same manner as was described in Section 3.2.5: $\omega_1 = 3$, $\omega_2 = 1$, $\theta_\alpha = \theta_\gamma = 4$ and $\theta_\beta = 1$ (Eq. 3.5), and a total of 5 iterations were carried out to refine each subject's result. The whole processing pipeline is implemented on the platform of Caffe (Jia et al., 2014) based on the original work of (Kendall et al., 2015; Krähenbühl and Koltun, 2012). The 100-subject dataset was divided into 2 sets by random permutation, resulting in 50 subjects in each half. A two-fold cross-validation was performed between these 2 sets to test the proposed method on all the subjects. In this way, the training and testing phases used independent sets of data. Due to the robustness of deep learning based methods, no registration is used during the entire process. All the training and testing of the proposed method and the evaluation of other compared methods was performed on a workstation hosting 2 Intel Xeon(R) E5-2620 v4 CPUs (8 cores, 16 threads @2.10GHz) with 64 GB DDR4

RAM and an Nvidia GTX980Ti GPU with 6 GB GPU memory. The workstation runs a 64-bit Linux operation system.

3.2.7.2 Uncertainty of the Bayesian SegNet

One important aspect of Bayesian SegNet is the output of model uncertainty on predictions. We studied the influence of training set size, label inconsistency and training-testing inconsistency on the model uncertainty. Besides the training set of 50 subjects, we trained the Bayesian SegNet with 25 and 5 subjects to show how the size of the training set can affect the uncertainty. We also trained Bayesian SegNet with 50 subjects, of which either 25 or 5 subjects had sub-optimal labels from AFNI+, while the rest of the labels were the manually-segmented ground truth. This was done to investigate how label consistency affects the uncertainty. Since AFNI+ usually includes some non-brain tissues around the frontal lobe and includes the adipose tissue behind the eyes in nonhuman primates, these labels can be used to simulate the same kind of errors possibly made in manual labels by carelessness or fatigue. To achieve similar level of convergence, the training procedures in this section were also performed with 60000 iterations. Finally, to study the influence of inconsistency between the training set and the testing set, 4 new testing sets were designed to be slightly inconsistent with the training set. The 50 subjects in fold 2 (set #2) were rotated around the longitudinal axis by 10, 20 and 30 degrees, to create 3 new testing sets. Another 50 subjects (mean (standard deviation) age = 3.20 (.86) years; 66% female) were scanned at another site, with an older scanner model (GE Signa 3T, Waukesha, WI, USA), but the same coil model and experimental setup. The brain masks for these 4 new testing sets were all generated by the Bayesian SegNet trained with the original 50 subjects in fold 1 (set #1) (mean (standard deviation) age = 1.98 (.37) years; 38% female) and processed by 3D CRF with the same parameters used in Section 3.2.7.1. These results were then compared to the results of the original 50 subjects in fold 2 (mean

(standard deviation) age = 1.92 (.39) years; 48% female) tested in Section 3.2.7.1. Model uncertainties on these new testing sets were also generated and compared with those on the original fold 2 in Section 3.2.7.1.

3.3 Results

3.3.1 Convergence of Bayesian SegNet during Training

Fig. 3.2 shows the convergence of the Bayesian SegNet on a 50-subject nonhuman primate training set and the convergence speed in one training set. There is no obvious improvement in the loss and accuracy after 18 epochs. Without loss of generality, we used the network weights at the 18th epoch, which is equivalent to 60000 iterations to predict brain masks, and one 18-epoch training over 50 subjects took about 12.3 hours on our workstation.

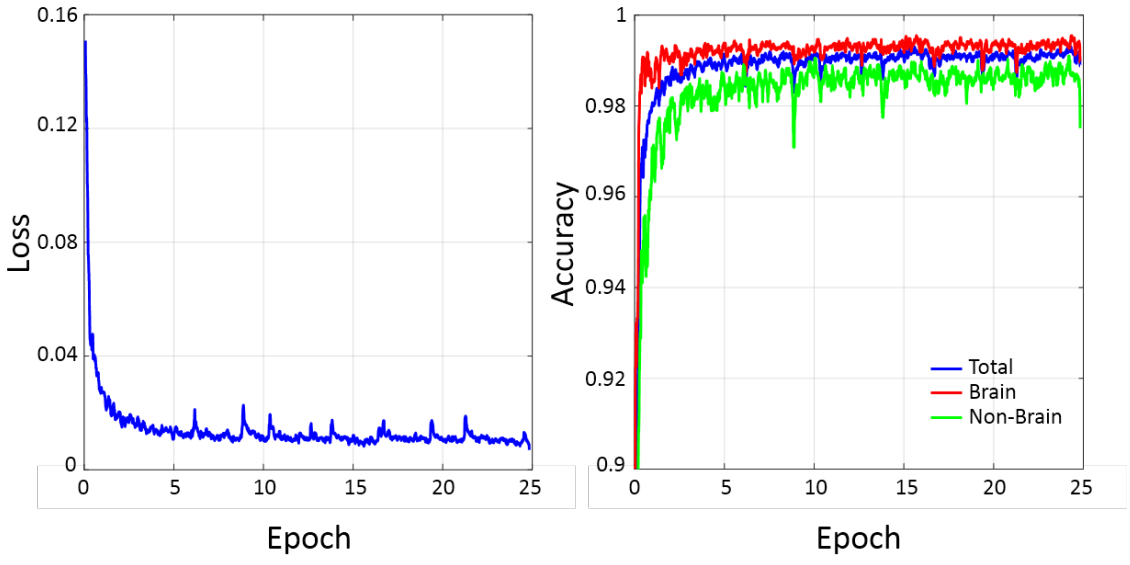

Fig. 3.2. Loss and accuracy for Bayesian SegNet during training against epochs. Loss is the multinomial logistic loss between the output labels and the ground truth labels. Accuracy is defined as the ratio of correctly labeled pixels over total number of pixels for each category.

3.3.2 Brain Extraction for Nonhuman Primates

The performance of the proposed method and nine other state-of-the-art-methods were evaluated on the T1w volumes from 100 subjects. The Dice coefficient and average symmetric surface distance of each individual for each method are plotted in Fig. 3.3. The boxplots are shown in Fig. 3.4, and the corresponding mean values and standard deviations are shown in Table 3.2. Fig. 3.3 shows that the performance of the proposed method, a combination of Bayesian SegNet and fully connected 3D CRF (BSegNetCRF), is the best on both metrics among all the compared methods for each individual's brain extraction. The boxplots in Fig 3.4 shows the median and quartiles of the Dice coefficient and the average symmetric surface distance for each method. Table 3.2 illustrates that BSegNetCRF has not only achieved the best mean values, but also the smallest standard deviation on both metrics. Multiple pairwise Wilcoxon signed rank tests (two-sided) were done to compare the performance of these methods. The performance of BSegNetCRF is better than all other methods, as evaluated on both metrics ($p < 10^{-4}$, Bonferroni corrected). BSegNet is significantly better than SegNet on both metrics ($p < 10^{-4}$, Bonferroni corrected). In the comparison between BSegNet and VNet, VNet's average symmetric surface distance is significantly better than BSegNet's ($p < 10^{-4}$, Bonferroni corrected) but the p-value on Dice Coefficient is 0.0425 before Bonferroni correction, which is insignificant after Bonferroni correction at the 0.05 significance level. Both BSegNet and VNet are better than LiviaNet on both metrics ($p < 10^{-4}$, Bonferroni corrected). The comparisons of different methods on Hausdorff distance, sensitivity and specificity are also shown in Appendix A Fig. A1-A4.

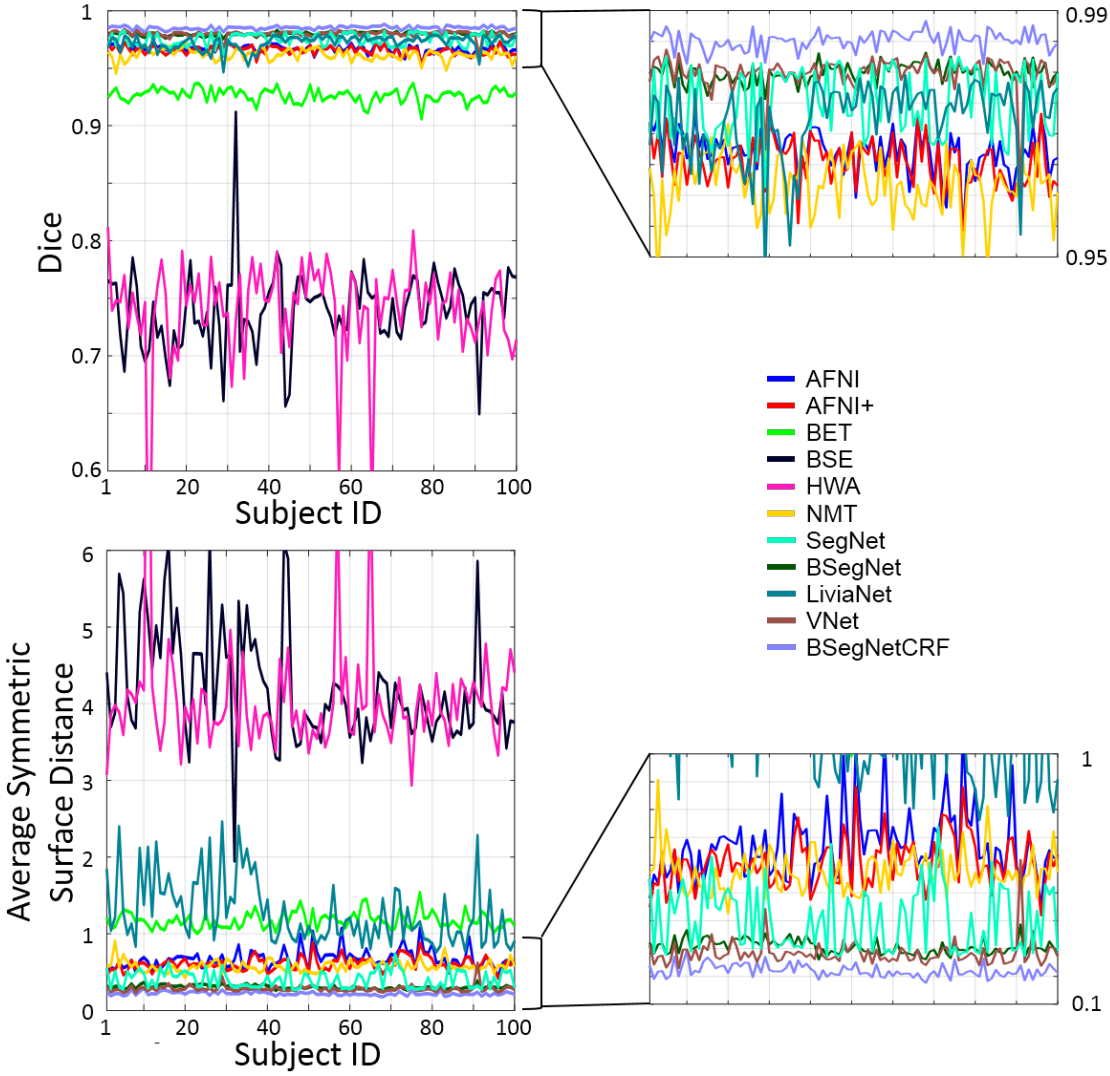

Fig. 3.3. Evaluation scores on each subject from different brain extraction methods. Enlarged figures are on the right. Higher Dice coefficients and lower average symmetric surface distance indicate better agreement between the automatically-defined and manually-labeled (ground truth) brain masks. For all subjects, BSegNetCRF resulted in better brain extraction than all other methods tested.

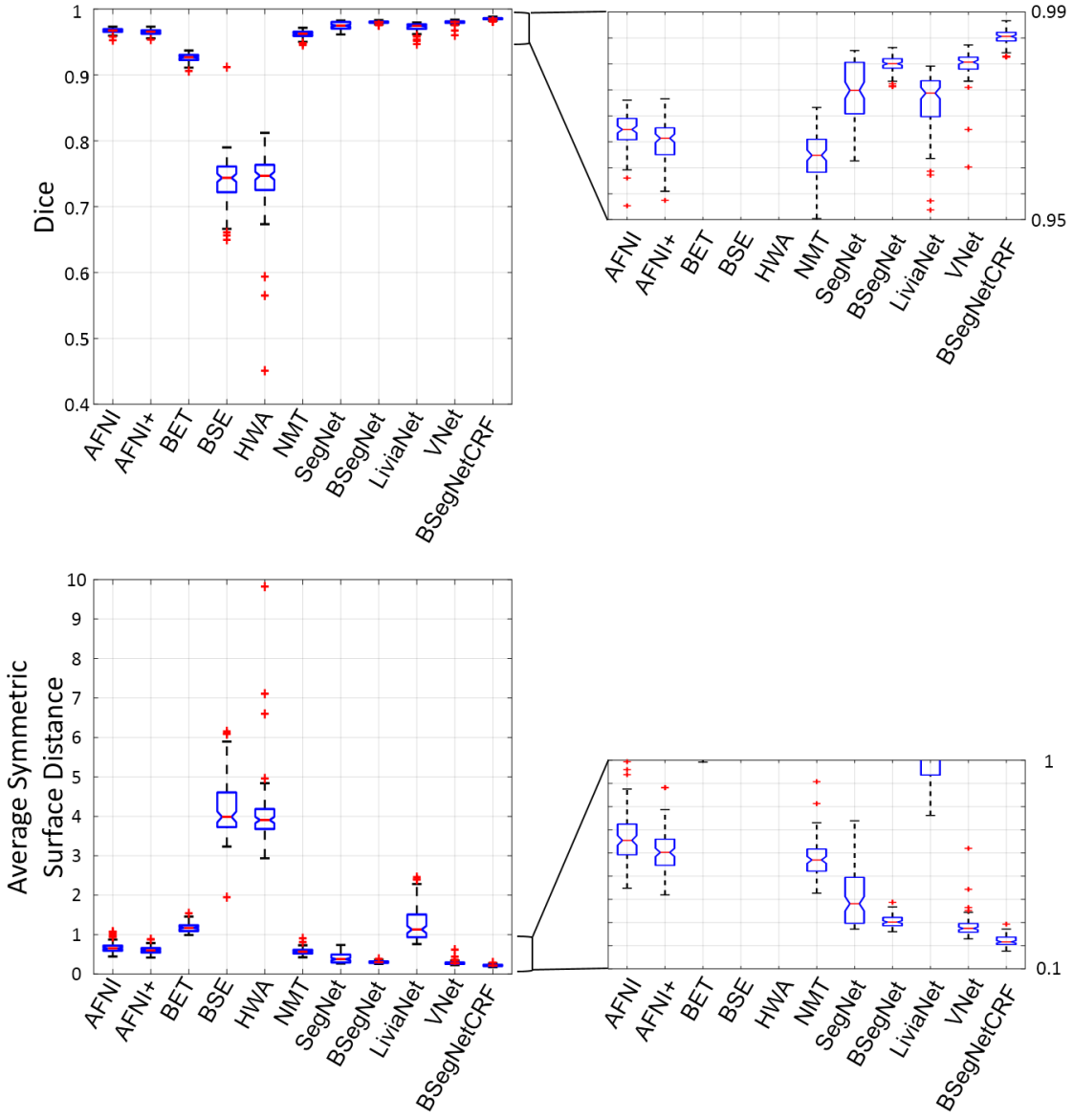

Fig. 3.4. Evaluation scores in boxplots from different brain extraction methods. Enlarged figures are on the right. In the figure points are drawn as outliers with red '+' symbols, if they are greater than $q_3+1.5(q_3-q_1)$ or less than $q_1-1.5(q_3-q_1)$, where q_1 and q_3 are the first and third quartiles respectively.

Table 3.2. Mean and standard deviation of Dice coefficient and ASSD for all 100 subjects. The best result is in bold font.

Method	Dice	ASSD/mm
AFNI	0.967 (±0.003)	0.670 (±0.123)
AFNI+	0.965 (±0.004)	0.609 (±0.088)
BET	0.926 (±0.006)	1.175 (±0.105)
BSE	0.740 (±0.034)	4.200 (±0.738)

HWA	0.739 (±0.046)	4.046 (±0.812)
NMT	0.962 (±0.005)	0.578 (±0.075)
SegNet	0.975 (±0.006)	0.404 (±0.116)
BSegNet	0.980 (±0.002)	0.306 (±0.026)
LiviaNet	0.972 (±0.006)	1.271 (±0.431)
VNet	0.980 (±0.003)	0.283 (±0.046)
BSegNetCRF	**0.985 (±0.002)**	**0.220 (±0.023)**

Fig. 3.5 shows the extracted brain masks from all the methods for a representative subject. AFNI typically cannot catch the complete frontal lobe due to its challenging sharp curvature in nonhuman primates. AFNI+ was used to fix this by reducing the coronal slice thickness to one half. Although AFNI+ can capture the frontal lobe more completely, but it also captures tissues outside the brain. In addition, both AFNI and AFNI+ mistakenly include the adipose tissue behind the eyes. BET misses the frontal and occipital lobes of the brain, and the mask often extends past the upper boundary of the brain. BSE and HWA are not designed for nonhuman primates, and their resultant brain masks include a lot of non-brain tissue. ROBEX failed on all the nonhuman primate data, so it is not shown in the figure. NMT mistakenly includes the adipose tissue behind the eyes as part of the brain mask, and misses some boundaries and overshoots some others. SegNet tends to include the nonbrain tissue around the frontal lobe, eyes and brain stem as part of the brain mask. LiviaNet includes multiple nonbrain regions and misses some small regions within the brain. Results from BSegNet, VNet and BSegNetCRF are very close to the manually labeled ground truth. VNet performs well at the frontal lobe and eyes, but in general it includes slightly more nonbrain voxels close to the boundaries than BSegNetCRF, especially at the area close to the bottom of the brain and the brain stem. BSegNetCRF is also better than BSegNet, especially at excluding the brain stem. Fig. 3.5 shows a comparison of the error maps of these methods on the representative subject. A more comprehensive systematic comparison is shown in Fig. 3.6.

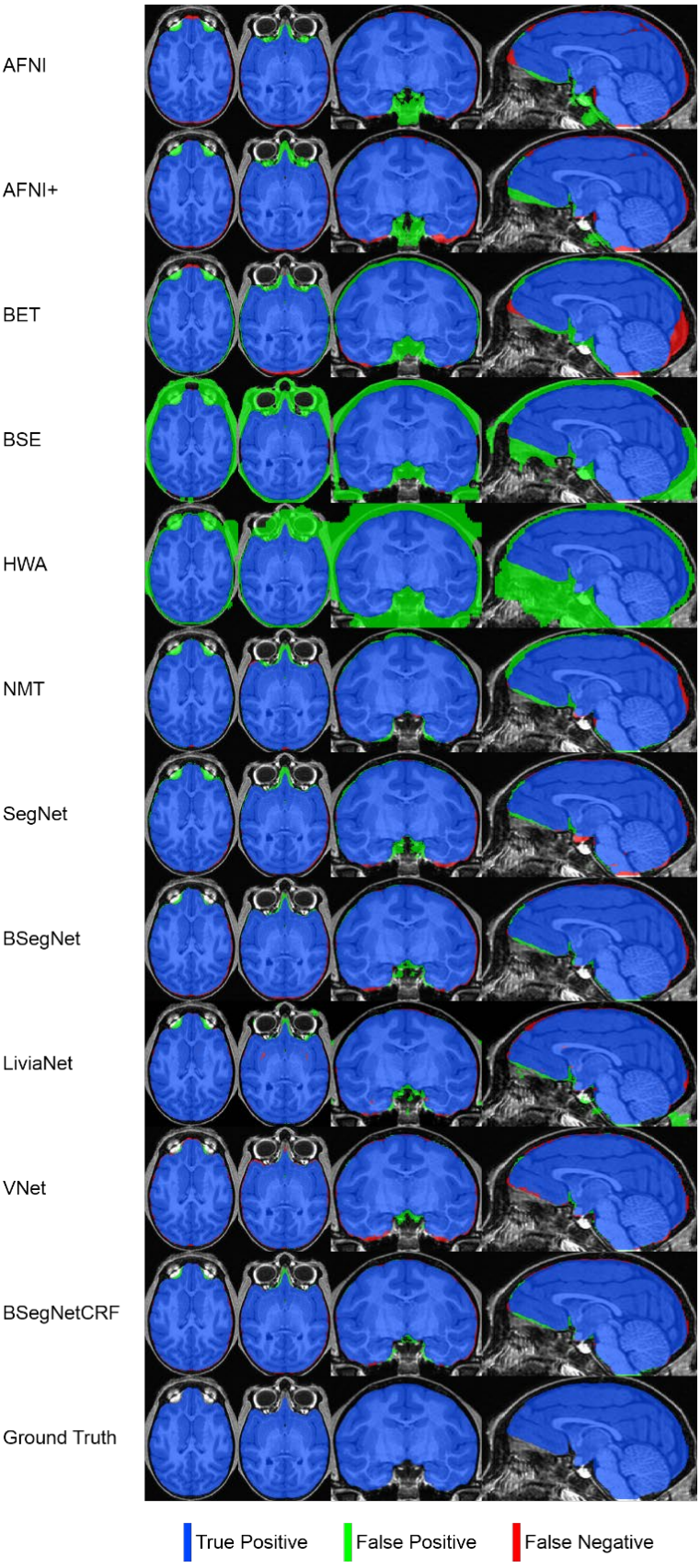
AFNI
AFNI+
BET
BSE
HWA
NMT
SegNet
BSegNet
LiviaNet
VNet
BSegNetCRF
Ground Truth
True Positive
False Positive
False Negative

Fig. 3.5. Comparison of the brain masks extracted by different methods on a typical subject: subject 007.

Fig. 3.6 is the averaged absolute error map of each method in the NMT template space. As mentioned in Section 3.2.6, the natural logarithm of the averaged error maps collapsed (averaged) along each axis is shown for display purposes. In Fig. 3.6's comparison, BSegNetCRF has the best systematic performance with a much smaller error distribution than other methods considering the results in all the voxels for every subject, and the systematic performance improvement by fully connected 3D CRF can also be viewed between the absolute error maps of BSegNet and BSegNetCRF. VNet also has very good performance, but BSegNetCRF is still better than VNet around the bottom area of the brain. The averaged false positive and false negative maps can also be found in the Appendix A Fig. A5 and A6.

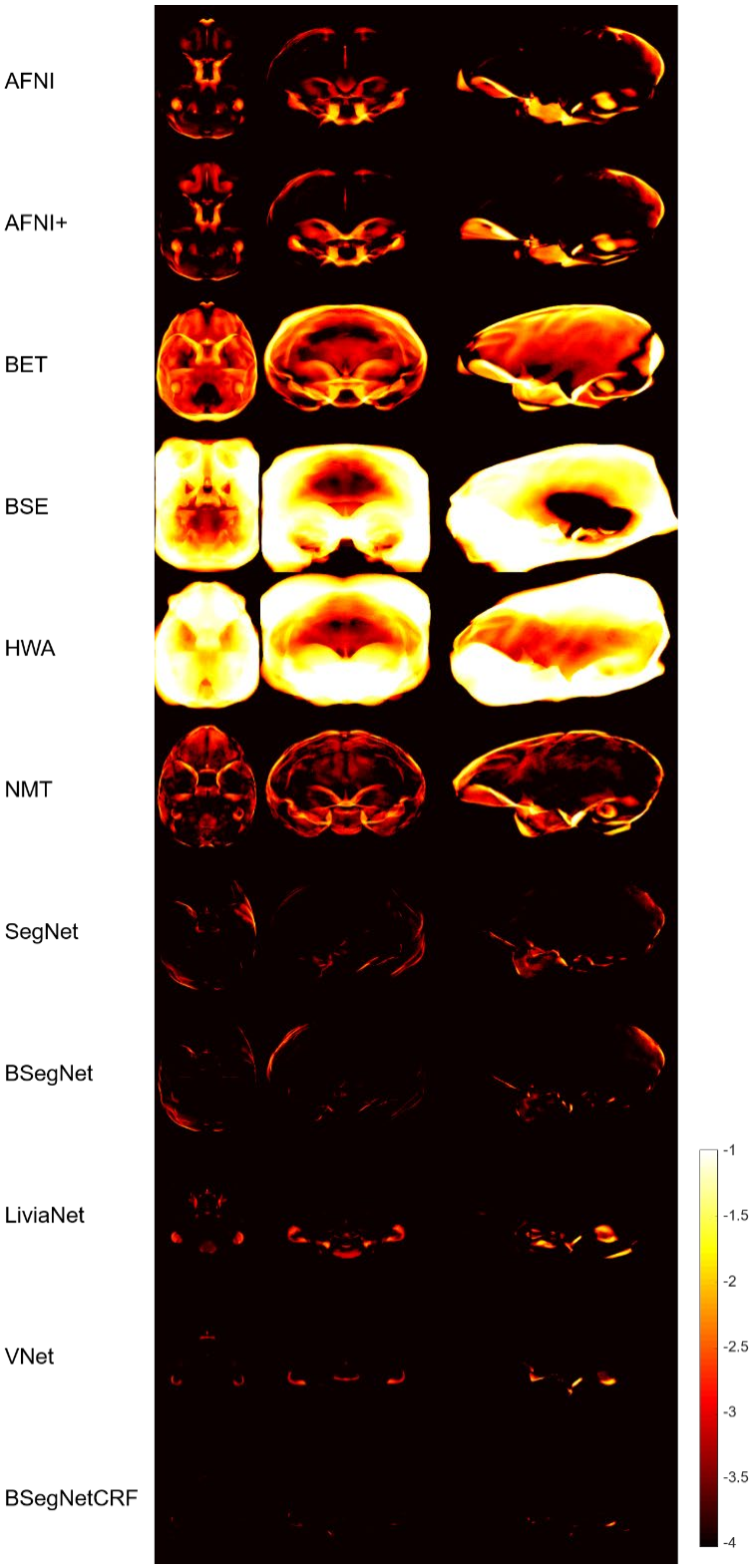

Fig. 3.6. Averaged absolute error maps for compared methods. For display purposes, the natural logarithm of the averaged map collapsed (averaged) along each axis is shown.

As a probabilistic network, Bayesian SegNet is able to output the model uncertainty on the prediction of every voxel's label via Monte Carlo dropout testing. The maximum voxel-labeling uncertainty of Bayesian SegNet within each subject's 3D volume was calculated, and it has a mean value of 0.116 and a standard deviation of 0.023 across all 100 subjects. Fig. 3.7 shows the voxel-labeling uncertainty on the same representative subject. In general, the uncertainty of the brain extraction is very low, and the relatively higher uncertainty regions concentrate around the edges of the brain. The uncertainty map of each subject was also transformed, averaged, collapsed and displayed in the same manner as the averaged absolute error map to calculate and show the averaged uncertainty map in the NMT space (Fig. 3.8). Fig. 3.8 illustrates the systematic uncertainty distribution in the 3D volume over all the subjects. Overall, the uncertainty is very low, and the relative high uncertainty area is at boundary of the brain close to the brain stem. The fully connected 3D CRF successfully helped correct the results from Bayesian SegNet around this area as shown is Fig. 3.5 and 3.6.

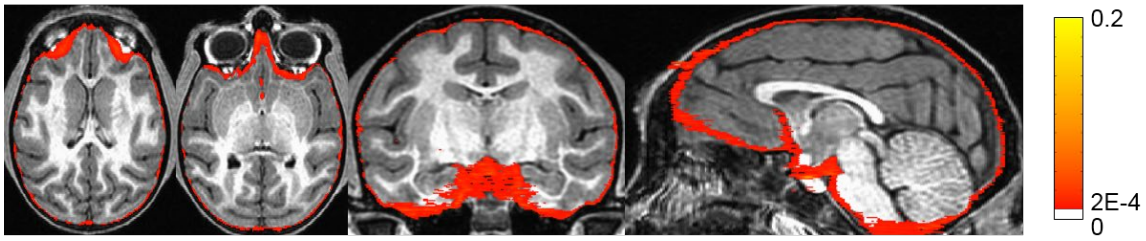

Fig. 3.7. The uncertainty map given by Bayesian SegNet for subject 007.

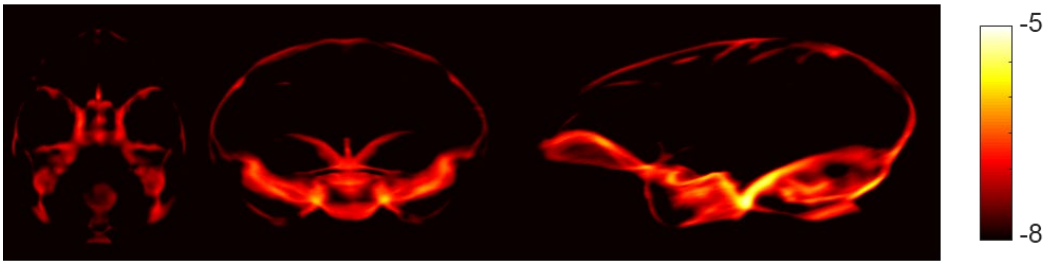

Fig. 3.8. Averaged uncertainty maps from Bayesian SegNet. For display purposes, the natural logarithm of the averaged map collapsed (averaged) along each axis is shown.

In terms of processing time, the prediction for a single subject in the test stage by Bayesian SegNet with Nvidia GTX980Ti GPU is around 40 seconds, and after this the fully connected 3D CRF with an Intel Xeon(R) E5-2620 v4 CPU (8 cores, 16 threads @2.10GHz) is approximately 80 seconds for 5 iterations. Thus, the total time for one prediction is approximately 2 minutes, which is comparable to other edge detecting based methods. However, the template based method, NMT, is very time consuming. It costs around 5 hours even with an OpenMP version AFNI and 2 Intel Xeon(R) E5-2620 v4 CPUs to do the registration for one subject.

3.3.3 Uncertainty of the Bayesian SegNet

Uncertainty maps were also generated by Bayesian SegNet trained with different numbers of subjects to study the effect of training set size on the uncertainty. The uncertainty maps of the representative subject are shown in Fig 3.9, from which it can be seen that as the training set size decreases, the uncertainty increases, especially at the boundaries of the frontal lobe and behind the eyes. The total uncertainty defined as $\sigma_{tot} = \sqrt{\sum_i \sigma_i^2}$ (i is the voxel index) was also calculated for each subject, and the total uncertainties for the 50 subjects in the testing set generated by different training set sizes are shown in the boxplot in Fig. 3.10. In Fig. 3.10, an increase in the total uncertainty can be seen as the training set decreases, and the total uncertainty of every subject tends to deviate more from one to another as the training set size decreases.

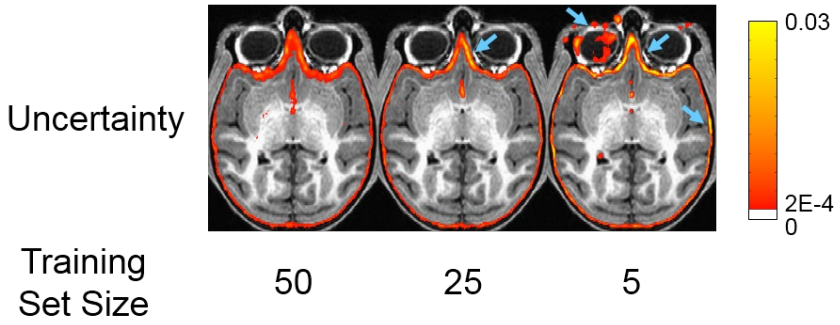

Fig. 3.9. Uncertainty maps on subject 007 generated by Bayesian SegNet trained with different training set sizes. The blue arrows in the figure point out the regions with obvious uncertainty increase.

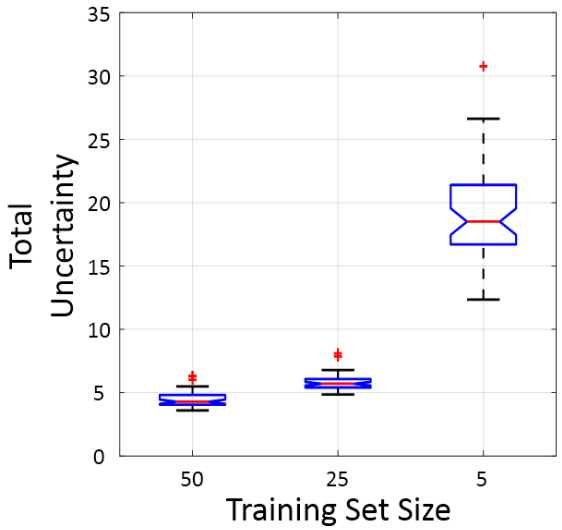

Fig. 3.10. Total uncertainty in boxplots generated by Bayesian SegNet trained with different training set sizes. In the figure points are drawn as outliers with red '+' symbols, if they are greater than q3+1.5(q3-q1) or less than q1-1.5(q3-q1), where q1 and q3 are the first and third quartiles respectively.

The relationship between training label consistency and prediction uncertainty was also studied. Fig. 3.11 shows the uncertainty maps of a representative subject 007 generated by Bayesian SegNet trained with manual labels and labels generated by AFNI+. As the number of AFNI+ labels increases in the 50-subject training set, the uncertainty generated by Bayesian SegNet also increases, especially in the frontal lobe and regions behind the eyes where the AFNI+ labels mismatch the corresponding manual labels. Fig. 3.12 shows a boxplot of the total uncertainty in the ROI behind the eyes against different AFNI+ label numbers in the training set. As the number

of AFNI+ labels in the training set increases, each tested subject's total uncertainty in the inconsistently labeled area also increases and tends to deviate more from one to another.

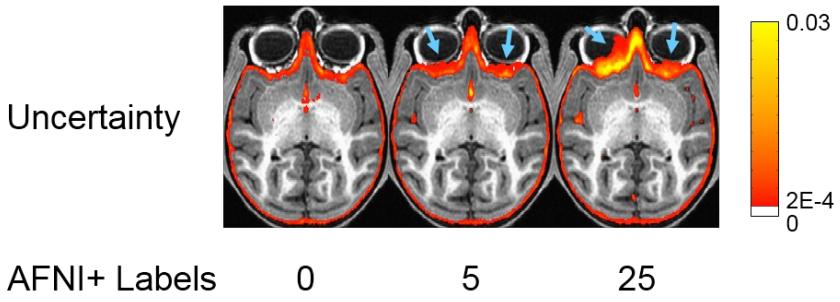

Fig. 3.11. Uncertainty maps on subject 007 generated by Bayesian SegNet trained with 50 subjects in which different numbers of manual labels were replaced by labels generated by AFNI+ for the corresponding subjects. The blue arrows in the figure point out the regions with obvious uncertainty increase.

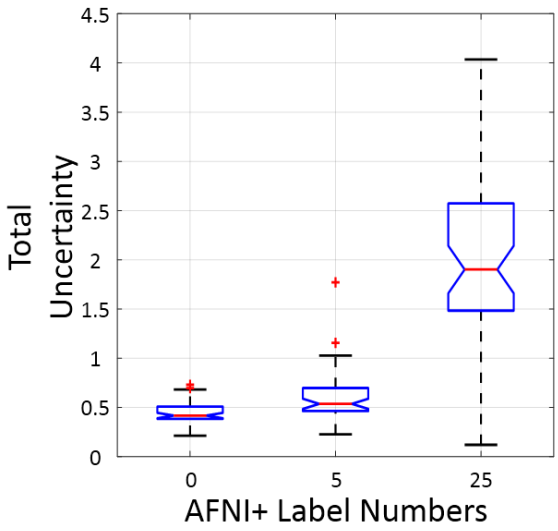

Fig. 3.12. Total uncertainty of the ROI behind eyes in boxplots generated by Bayesian SegNet trained with 50 subjects in which different numbers of manual labels were replaced by labels generated by AFNI+ for the corresponding subjects. In the figure points are drawn as outliers with red '+' symbols, if they are greater than q3+1.5(q3-q1) or less than q1-1.5(q3-q1), where q1 and q3 are the first and third quartiles respectively.

The inconsistency between the training set and the testing set can also cause erroneous results. The uncertainty generated can give a warning about this kind of inconsistency. Fig. 3.13 shows the brain extraction performance of the proposed method with the same parameters (trained with the

50 subjects in fold 1 and using the same 3D CRF parameters) on 5 different testing sets. The original data in fold 2 are most consistent with the training data, and this set had the best performance. When the fold 2 data were rotated by increasingly larger amounts, the inconsistency of them against the training set was larger, and the brain extraction results were worse. Since the data collected on the older scanner were also slightly inconsistent with the training data due to slight contrast differences and age and gender differences between different subject groups, the results were also slightly worse. Fig. 3.14 shows the corresponding uncertainty behavior. When the testing set is inconsistent with the training set, the total uncertainty is higher. The more inconsistency there is, the more the total uncertainty increases.

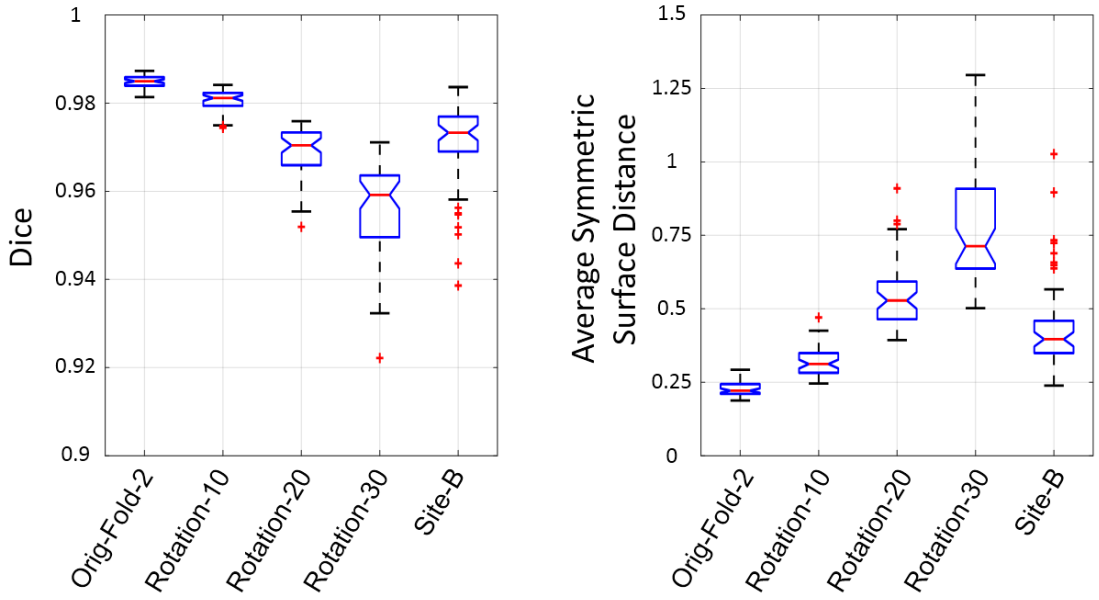

Fig. 3.13. Evaluation scores in boxplots for the original fold 2 data, rotated fold 2 data and data from another site. In the figure points are drawn as outliers with red '+' symbols, if they are greater than q3+1.5(q3-q1) or less than q1-1.5(q3-q1), where q1 and q3 are the first and third quartiles respectively.

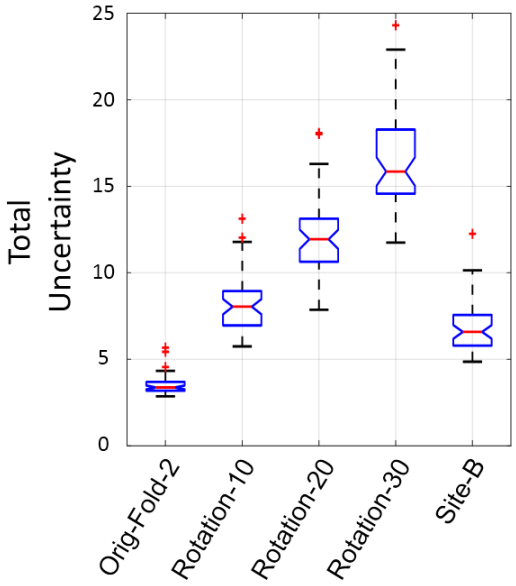

Fig. 3.14. Total uncertainty generated by Bayesian SegNet for the original fold 2 data, rotated fold 2 data and data from another site in boxplots. In the figure points are drawn as outliers with red '+' symbols, if they are greater than q3+1.5(q3-q1) or less than q1-1.5(q3-q1), where q1 and q3 are the first and third quartiles respectively.

3.4 Discussion

A new fully-automated brain extraction method is proposed as a combination of deep probabilistic neural network and fully connected 3D conditional random field, for the challenging task of brain extraction in nonhuman primates. The brain extraction results of the 100-subject dataset suggest that the proposed method can achieve higher accuracy and superior performance compared to state-of-the-art methods, as is measured by many different metrics. In addition, the proposed method is also highly time-efficient for a single prediction of a couple of minutes, with the facilitation of parallel computation.

The difficulties of nonhuman primate brain extraction are mainly due to the unique anatomical structures, especially the adipose tissue behind the eyes, the sharp curvature of the frontal lobe and

generally more muscular and bony structures (Rohlfing et al., 2012a). For the competing nondeep-learning based methods, the brain extraction results are in good agreement with previous published studies (Wang et al., 2014). Edge detection methods and gradient based methods, like BSE and AFNI, can easily fail on the adipose tissue behind the eyes because of their proximity to the brain and high contrast to surrounding tissues (Iglesias et al., 2011; Wang et al., 2014). Algorithms involving surface expansion or deformable techniques, like BET and HWA, usually reach a result that either misses the frontal lobe, or includes areas out of the brain, because of the sharp curvature of the frontal lobe (Fennema-Notestine et al., 2006; Shattuck et al., 2001). Template registration based methods, like ROBEX and NMT, highly depend on the accuracy of registration. Since every subject must have its own anatomical specificity, which can be highly unique, given significant differences in age, gender and health conditions, the registration often has flaws. (Roy et al., 2017).

In the deep learning based methods, BSegNet is better than SegNet due to the involvement of Monte Carlo dropout training and testing (Kendall et al., 2015b), and BSegNetCRF is better than BSegNet since fully connected 3D CRF makes it possible to refine the probability results in a fully 3D context. BSegNetCRF, LiviaNet and VNet are all 3D methods using different strategies to take the 3D context into consideration. BSegNetCRF refines the results from a 2D neural network with 3D CRF; LiviaNet unstacks the original 3D volumes into small 3D patches for the 3D network; VNet downsamples the original 3D volumes and processes the whole downsampled 3D volume with the 3D network. The results show that BSegNetCRF outperforms LiviaNet and VNet in this application, and there could be multiple possible reasons for this. Directly processing large 3D images on current GPUs is very challenging due to the limit of GPU memory, so the design of a 3D network has to be relatively light and shallow to reduce the memory request from the network parameters. Even so, to process the 3D input, either a patch-based strategy or downsampling still

need to be used. A 3D Patch-based strategy reduces the network's receptive field, so it is usually used for segmenting small structures. For regions as large as nonhuman primate brains, small regional errors can be caused. Meanwhile, downsampling can affect the results directly. LiviaNet and VNet also use different network designs, such as number of layers, stride and loss function, and these differences could also be the potential reasons for the performance differences in this application.

Overall, compared with nondeep-learning based methods, deep learning uses the training process to capture the features of the dataset with the help of a subset of manually labeled brains as the prior knowledge. This gives deep learning the ability to segment very complex structures. For brain extraction, it can exclude the complex ventricle structures within the brain in a manner as the manual labels are defined in the training stage (Kleesiek et al., 2016). Meanwhile, deep learning also offers more flexibility in brain extraction since one can define the brain region as preferred by the training labels, for example, including or excluding certain parts of the brain, like the brain stem or cerebellum.

As a convolutional encoder-decoder network, Bayesian SegNet reaches a balance of being both deep and light, which makes it a powerful tool in brain extraction. For being deep, it has 13 convolutional layers, 13 deconvolutional layers and 26 corresponding ReLU layers, which makes it deep enough to extract high level features with a considerable receptive field, while possessing sufficient nonlinearity to build the transformation from the original images to the brain extraction labels (Badrinarayanan et al., 2015a). However, usually a deep neural network suffers from an enormous number of parameters to train, which has a high cost in terms of GPU memory, and time in training and predicting, and can also result in overfitting. In terms of also being light, Bayesian SegNet elegantly uses the pooling indices in the maximum pooling layers to perform the nonlinear

upsampling in the corresponding upsampling layers. This eliminates the need for training in all the upsampling steps and makes the upsampled maps sparse. Moreover, it also uses small convolutional filters and no fully connected layers. All these features make it small in terms of the number of trainable parameters and efficient in terms of both the memory cost and computational time (Badrinarayanan et al., 2015b; Liu et al., 2017).

Being different from other deep learning based neural networks applied to brain extraction, Bayesian SegNet is a probabilistic neural network, so it has the ability to provide the uncertainty of the network on each prediction, as well as predict accurate labels for all pixels (Kendall et al., 2015a). It is important for a predictive system to generate model uncertainty as a part of the output, since meaningful uncertainty measurement is important for decision-making, especially biomedical applications where accuracy is extremely important. To replace any manual procedures with deep learning based methods means the conventional ground truth is not available any more in a real prediction, so an output including the confidence of the model on each specific case becomes very important. The uncertainty offered by Bayesian SegNet meets this very need. In a routine procedure implemented by Bayesian SegNet, every output uncertainty will be checked automatically against an empirical threshold to determine whether the result can be trusted or human intervention should be started. In our exploration of the uncertainty generated by Bayesian SegNet, we demonstrated that the behavior of the uncertainty generated by Monte Carlo dropout sampling matches our expectation very well. The uncertainty tends to increase and deviate more from subject to subject, as the size of the training set decreases, the inconsistency of training labels increases, or the inconsistency between the training set and testing set increases (Fig. 3.9-3.12). Considering each training process takes about 12 hours, we only studied a few training set sizes and mismatching label numbers. In addition, although the training process always drives the loss

to converge, the training procedure itself is a stochastic process. A more thorough study of the behavior of the uncertainty, and to what extent the training process affects the uncertainty, are future research topics. In terms of the inconsistency between training and testing sets, there are several possible solutions to improve the robustness of the method. One is to make the training set more like the testing test. For example, to predict randomly rotated data, the training set can also be randomly rotated before training, or all the training and testing data can be aligned to a template before training and testing to make the model more robust. The other solution is to use transfer learning (Pan and Yang, 2010). To predict the data from another site or using another pulse sequence, the trained network can be further finetuned with a small subset of this kind of inconsistent data before the real prediction.

The combination of fully connected 3D CRF takes the probability maps from Bayesian SegNet's 2D predictions and moves forward to predictions in a fully 3D context. Because of the limitation of current GPU memory, and the huge data size of brain images, it is currently challenging to make fully 3D predictions through deep learning on a single GPU (Wachinger et al., 2017). Thus, we chose the 2D neural network to meet the GPU memory limitation without compromising the network performance, while making the whole 2D slice available for training and predicting. Then, we involve fully connected 3D CRF to implement a complete 3D prediction taking into account all the information from the entire original brain volume. Results shown from Fig. 3.3, 3.4 and 3.6 demonstrate the improvement made by the combination of this fully 3D process to the deep learning alone method. In addition, there can be errors in the direct results from deep learning based methods due to the imperfection and inconsistency of manual labeling. Fully-connected 3D CRF can fix these errors to some extent by taking into account the contrast and distance of all the connections in the original 3D image (shown in Fig. 3.5 and 3.6). The parameters of fully

connected 3D CRF were empirically selected. To achieve the optimal performance, the weights for the result of deep learning, image intensity and voxel distance need to reach a balance. If the result of deep learning is over weighed, the effect of CRF's refinement won't be realized, and the result will be similar to that from deep learning. If the image intensity or voxel distance is over weighed, then the deep learning portion will be underemphasized, and the result could be worse than that from deep learning. As each of the parameters changes, the Dice coefficient will change gradually.

There are also some limitations in our study. To simulate the typical parameter-selecting procedure, we didn't carry out subject specific parameter selection, nor did we perform a grid search in the parameter space. It is possible that the performance of these methods can be improved with these strategies, however, they are very time consuming, and thus, impractical (Kleesiek et al., 2016). Another limitation is that we only studied the periadolescent rhesus monkeys. The structure of a nonhuman primate's skull and the amount of muscle tissue change dramatically across development, and infant monkeys are being used more and more in neuroscience studies (Kourtzi et al., 2006; Livingstone et al., 2017). Currently we are collecting brain MR images of rhesus monkeys across the whole age spectrum to test our method and investigate how transfer learning can be applied across different age groups. Future study also includes the possibility of combining the third-dimensional information (Xu et al., 2017) and image noise information (Kendall and Gal, 2017) of MRI brain volumes into the network. These approaches may further improve the performance of brain extraction.

3.5 Conclusion

In conclusion, we proposed and evaluated a new fully-automated brain extraction method integrating Bayesian SegNet and fully connected 3D CRF for nonhuman primate MRI brain images. Being different from previous designs, our approach is not only able to generate accurate and rapid brain extraction in a fully 3D context, but it also involves a probabilistic convolutional neural network that can output the uncertainty of the network on each prediction. This can greatly facilitate current large-scale MRI based neuroscience, neuroimaging and psychiatry studies on nonhuman primates.

Chapter 4

Bayesian Conditional GAN for MRI Brain Image Synthesis

4.1 Introduction

Image synthesis is an important topic in the field of medical imaging, and many techniques in all levels of medical image processing can be categorized into the field of image synthesis. With image synthesis techniques, MR images can be reconstructed from the data collected in the k-space (Zhu et al., 2018), image denoising can be achieved by generating images with low noise from the images with high noise (Jiang et al., 2018), and the resolution of an image can be improved from a low resolution image, which is also called super-resolution (Sanchez and Vilaplana, 2018). Sparse reconstruction can significantly shorten the scanning time, and freeze the motion for dynamic imaging, while denoising and super-resolution can immensely improve the image quality for diagnosis. With image synthesis techniques, we can even further generate the image of one

modality from the image in another modality, or synthesis images across different contrast mechanisms, for example: generating CT images from MR images (Guerreiro et al., 2017; Roy et al., 2014), or synthesizing T2w MR images from T1w MR images (Jog et al., 2017). The former example can reduce the patient's radiation dose and generating CT images when CT scanner is not available, for example in a PET/MR scanner (Liu et al., 2018). The later example can reduce the total scan time of the clinical protocol.

In medical neuroimaging, compared with other modalities, MRI can deliver high resolution, 3D images with a variety of contrast mechanisms in a radiation free manner. Many pulse sequences have been designed to capture different tissue contrast mechanisms for different diagnosis purposes. For example, the pulse sequence of Magnetization Prepared Gradient Echo (MPRAGE) (Deichmann et al., 2000) provides a heavily T1-weighted contrast, which is useful to visualize the cortex and subcortical structures, while the Fluid Attenuated Inversion Recovery (FLAIR) (Hajnal et al., 1992) is a kind of T2w sequence, which is good at catching the white matter lesions in normal white matter and is widely used for imaging multi-sclerosis patients (Simon et al., 2006). Depending on certain diseases and the diagnosis requirements, multiple MR pulse sequences can be acquired during a single MRI session to get a comprehensive information about the brain anatomy and function. In this situation, MR inter-contrast image thesis has the potential to reduce the total scan time and cost. If an unacquired MR contrast is found to be useful, inter-contrast image synthesis is also able to generate it retrospectively.

The purpose of this work is to build a deep learning based model to increase the accuracy in image synthesis as well as generating model uncertainty for each synthesized image in MR contrast transformation. In this study, we propose Bayesian conditional GAN with concrete dropout and a

model recalibration method to increase the accuracy in image synthesis and improve the calibration of the uncertainty generated.

In comparison to previous image synthesis studies, our study has several novel aspects. First, we propose Bayesian conditional GAN as the main image synthesis engine for this task. As a Bayesian neural network it can not only synthesize the image in the target contrast but also generate the uncertainty map for the synthesized image as well. The uncertainty map can be the source that our judgement on whether the synthesized image can be trusted or not is based on. Second, we use concrete dropout in the Bayesian neural network instead of the conventional Monte Carlo dropout (Gal and Ghahramani, 2016). As a gradient-tuned dropout, the dropout rate of concrete dropout can converge to its optimal value during the training stage. This eliminates the complex grid search procedure to find the best dropout rate for each Monte Carlo dropout layer, and also ends up with a better calibrated uncertainty. Finally, we incorporate a model recalibration method as a post-processing approach in the model to further improve the calibration of the posterior distribution of the predicted voxel values and the corresponding model uncertainties.

The accuracy and robustness of the model were evaluated on the challenging application of T1w to T2w brain tumor image synthesis. We hypothesize that the Bayesian conditional GAN with concrete dropout and model recalibration is suitable for the inter-contrast MR brain image synthesis with accurate predictions and well-calibrated uncertainties, which can reflect the confidence levels of the model on the predictions.

4.2 Material and Methods

4.2.1 Dataset and Preprocessing

This work used the T1w and T2w MR brain volumes of 102 pre-operative subjects of The Cancer Genome Atlas (TCGA, cancergenome.nih.gov) Glioblastoma Multiforme (GBM) collection. The data were released by the international multimodal BRAin Tumor Segmentation challenge (BRATS 2018, https://www.med.upenn.edu/sbia/brats2018/) (Menze et al., 2015) through The Cancer Imaging Archive (TCIA, www.cancerimagingarchive.net). The brain images were collected from 8 institutions with 3T scanners of different vendors and having different MR imaging sequence implementations. The data were distributed after preprocessing. All the brain volumes were co-registered to the same anatomical template with affine registration, and then were resampled into the same 1 mm^3 resolution. Finally, all the brain volumes were skull-stripped. Detailed patient information, scanner information and imaging information for each image can be found in (Bakas et al., 2017).

4.2.2 Bayesian Conditional GAN

Conditional GAN (Isola et al., 2016) is an accurate and consistent approach to synthesize images. To make it also have the ability of generating model uncertainty for each prediction, we propose to convert it into a Bayesian neural network. In the framework of Bayesian deep learning, all the variables in the neural network and the predicted results are treated as random variables following certain distributions. The training purpose of Bayesian deep learning is to estimate the posterior distribution $p(\omega \mid X, Y)$ of the weights in the neural network, which is usually intractable. Thus, a weight distribution $q_\theta(\omega)$ with the a parameter set θ is used to approximate the intractable

posterior, and the training procedure is equivalent to minimizing the KL divergence between them (Gal and Ghahramani, 2015).

$$\mathcal{L}(\theta) = KL(q_\theta(\omega) \,||\, p(\omega \mid X, Y)) \tag{4.1}$$

With the techniques in variational inference and Monte Carlo integration, the KL divergence in Eq. 4.1 can be simplified as the following loss function (Gal and Ghahramani, 2015):

$$\widehat{\mathcal{L}}(\theta) = -\frac{N}{M}\sum_{i \in S} \log p(y_i \mid f^{g(\theta,\epsilon)}(x_i)) + KL(q_\theta(\omega) \,||\, p(\omega)) \tag{4.2}$$

where N is the total number of observations and M is the number of observations used in the current training step. $g(\theta, \epsilon)$ is the collection of the weight random variables in the Bayesian neural network with the collection of the weight matrices, θ, and the collection of the random variables, ϵ, which follow Bernoulli distributions. x_i is the ith input data and y_i is the ground truth of the ith input data. $f^{g(\theta,\epsilon)}(x_i)$ is the predicted result from the forward pass of the Bayesian neural network. It can be proved that the training of a Bayesian neural network is equivalent to the training of a conventional neural network with the dropout regularization and the square of the l^2 norm regularization of the weight matrices in the conventional neural network (Gal and Ghahramani, 2015). Therefore, by plugging in dropout layers, the conventional conditional GAN can be changed into a Bayesian conditional GAN.

During the testing stage, unlike the dropout layers in a conventional neural network, the dropout layers in a Bayesian neural network will still function in the forward passes. This procedure is called dropout testing, and it is equivalent to sampling the posterior distribution of the predicted random variables (Gal and Ghahramani, 2016).

$$p(y^* \mid x^*, X, Y) \approx \int p(y^* \mid x^*, \omega) q_\theta^*(\omega) d\omega =: q_\theta^*(y^* \mid x^*) \tag{4.3}$$

Given the assumption that these predicted random variables follow normal distributions, we can use the mean and the variance of the predicted values from multiple forward passes as the unbiased estimators of the mean and variance of the distributions of the predicted random variables. The mean can be used as the final prediction result of the input data, and variance can be used as the model uncertainty for the prediction (Gal and Ghahramani, 2016).

4.2.3 Concrete dropout

One drawback of the conventional Monte Carlo dropout is that the dropout probability is a hyperparameter, and need to be tuned manually. Grid-searching over the entire space of dropout probabilities for all the dropout layer is exhausting work and will cost an immense amount of computation power. Moreover, the predicted posterior distribution $p(y^* \mid x^*, X, Y)$ in dropout testing can be affected by the dropout probability. Thus, the accuracy of the final predicted result and the calibration of the model uncertainty can also be greatly influenced by the hand-tuned dropout probability.

Given the assumption that the prior of the weights, $p(\omega)$, also follows normal distribution, the KL divergence in the simplified loss function, Eq. 4.2, can be proved to be proportional to the following equation (Gal et al., 2017):

$$KL(q_\theta(\omega) \,||\, p(\omega)) \propto \sum_{DropoutLayers} \frac{1}{2} l^2 (1-p) \left\| M \right\|_2^2 - K\mathcal{H}(p) \tag{4.4}$$

where the sum is calculated over all the dropout layers; l is a hyperparameter; p is the dropout probability; M is the weight matrix in the layer before the dropout layer; K is the number of input

channels for the dropout layer; $\mathcal{H}(p)$ is the entropy of a Bernoulli random variable with the probability p:

$$\mathcal{H}(p) := -p \log p - (1-p)\log(1-p) \tag{4.5}$$

The entropy term in the regularization only depends on the dropout probability, which means it will be completely ignored when the weights of the network are the only variables to be optimized. However, it enables a gradient-tuned dropout probability when the dropout probability is part of the optimization variables. This means the dropout probability will converge to its optimal value as the training proceeds. This spares the effort of a grid-search for the optimal dropout probability for each dropout layer, and will also result in a better calibration of the predicted posterior and the model uncertainty. The whole structure of a Bayesian conditional GAN with concrete dropout is illustrated in Fig. 4.1.

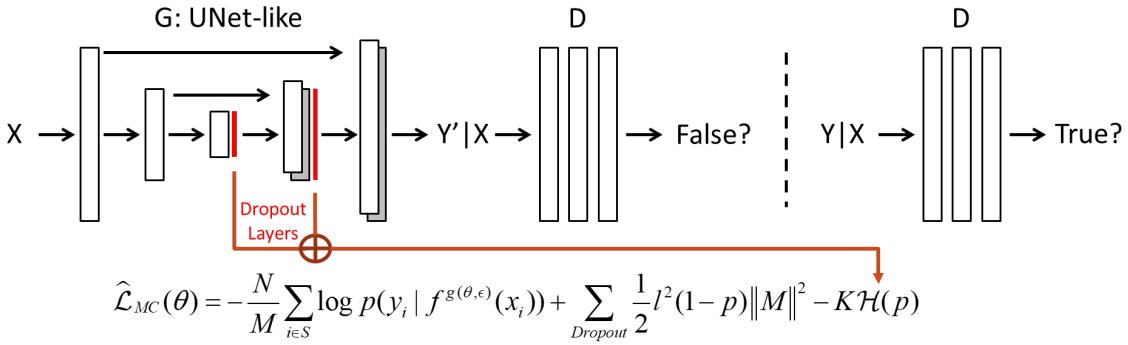

Fig. 4.1. Illustration of the structure of a Bayesian conditional GAN.

Since the derivative of the KL divergence term, $KL(q_\theta(\omega) \,\|\, p(\omega))$, need to be calculated with respect to p during the back propagation, we use the concrete distribution relaxation of dropout's discrete Bernoulli distribution, which reparametrizes the distribution as the following (Gal et al., 2017):

$$\tilde{z} = sigmoid(\frac{1}{t}(\log p - \log(1-p) + \log(u) - \log(1-u))) \tag{4.6}$$

where t is the hyperparameter, temperature, and $u \sim \mathcal{U}(0,1)$. This concrete relaxation of the dropout operation is referred to as concrete dropout (Gal et al., 2017). With it we can optimize the dropout probability in the training stage of a Bayesian neural network.

4.2.4 Model Recalibration

The image synthesis in this work is achieved by the proposed Bayesian conditional GAN, and the predicted posterior and the corresponding model uncertainty are generated by the concrete dropout. However, as a variation inference technique, Bayesian deep learning cannot guarantee the absolute accuracy of the predicted posterior and the model uncertainty. Thus, to further improve the accuracy of the predicted posterior and the model uncertainty, we propose to incorporate a model recalibration procedure for Bayesian deep learning.

Based on its probabilistic definition in Bayesian statistics, a 95% credible interval should be able to catch the value of interest (e.g. ground truth) with a 95% probability. With probabilistic calibration and a calibration dataset, the predicted posterior can be mapped to the true distribution, which can accurately reflect the probabilistic definition of credible interval. In the Bayesian deep learning model for image synthesis, for each voxel t the model generates a posterior distribution, a probability density function (PDF), targeting the ground truth value y_t during the testing stage. The PDF can be converted to a cumulative distribution function (CDF), F_t, which can be used to perform the model's probabilistic recalibration (Kuleshov et al., 2018):

$$p \to f = \frac{\sum_{t=1}^{T} \mathbb{I}\{y_t \le F_t^{-1}(p)\}}{T} \text{ for all } p \in [0,1] \tag{4.7}$$

where $F_t^{-1}(p) := \inf\{y : p \le F_t(y)\}$ is the quantile function. From Eq. 4.7 it can be seen that the model recalibration procedure maps the probability of the CDF of the predicted posterior to the real probability that the ground truth values of the voxels in the calibration dataset fall in the corresponding credible interval. This exactly follows the probabilistic definition of credible interval. The model recalibration procedure should calibrate all the *p* values for the Bayesian deep learning model according to the calibration dataset after the training of the Bayesian deep learning model. Then, during the testing time the predicted posteriors from the Bayesian deep learning model can be mapped to the calibrated posteriors.

4.2.5 Experiments

Before doing image synthesis, each subject's original 3D brain volume was dissembled along the longitudinal (superior-inferior) axis into a stack of 2D images, and then each 2D image was normalized to [0,1]. Before sent to Bayesian conditional GAN, all the T1w images and T2w images were resampled to the size of 286×286 with bilinear interpolation. Since all the subjects have brain tumors at different locations with different sizes, a 256×256 window was randomly shifted within the 286×286 brain image to cut a 256×256 image to send to the Bayesian conditional GAN at each iteration as a data augmentation approach. The Bayesian conditional GAN used a UNet-like (Ronneberger et al., 2015) convolutional encoder-decoder network as the generator and a CNN with 5 convolutional layers as the discriminator. Concrete dropout layers were plugged into the network structure after the 2nd, 3rd and 4th transposed convolutional layers. Batch normalization was used in the network, and a batch size of 16 was used during the training stage. A combination

of the conditional GAN loss, the l_1 norm between the synthesized image and the ground truth, and the regularization of the KL divergence term, $KL(q_\theta(\omega) \| p(\omega))$, Eq. 4.4, was used as the loss function. A weight of 100 was used as the weights for both the l_1 norm and the KL divergence term in the loss function. Within the KL divergence term a weight of 1e-6 was used for the network weight regularization term and a weight of 1e-5 was used for the concrete dropout regularization term. The hyperparameter, temperature, *t*, in the concrete distribution relaxation was set as 0.1. The Bayesian conditional GAN was trained by the Adaptive Moment Estimation (ADAM) (Kingma and Ba, 2014) algorithm with a fixed learning rate of 0.0002 and the momentum: $\beta_1 = 0.5$ and $\beta_2 = 0.999$. After 40 epochs of training, there was no obvious improvement of the loss, and the network weights from the 40th epoch was used for the following tests. After prediction, all the synthesized images and the ground truth images were normalized to the range of [0,255] for later visualization and result analysis.

The whole processing pipeline was implemented on the platform of PyTorch (Paszke et al., 2017) based on the original work of conditional GAN (Isola et al., 2016). All the training and testing of the proposed method and the evaluation of other compared methods was performed on a workstation hosting 2 Intel Xeon(R) E5-2620 v4 CPUs (8 cores, 16 threads @2.10GHz) with 64 GB DDR4 RAM and an Nvidia TITAN Xp GPU with 12 GB GPU memory. The workstation runs a 64-bit Linux operation system.

In the 102 subjects, 82 subjects were used for training and 20 subjects were used for testing. The image synthesis performances and the generated model uncertainties of the Bayesian conditional GANs with the concrete dropout and the conventional Monte Carlo dropout were compared. For Monte Carlo dropout a dropout rate of 0.5 was used. In the neural network structure, the positions

of the Monte Carlo dropout layers were the same as those of the concrete dropout layers, but no KL divergence term was used in the loss function for the Monte Carlo dropout. In the model recalibration procedure, the training dataset was used for the model recalibration. No obvious overfitting was observed.

4.3 Results

4.3.1 Image Synthesis: Prediction Accuracy and Model Uncertainty

The image synthesis accuracy of the proposed Bayesian conditional GAN with concrete dropout was evaluated and compared with that of the Bayesian conditional GAN with Monte Carlo dropout. In Fig. 4.2 the boxplots of the root mean square (RMS) errors of each subject's synthesized brain volumes were shown. Overall, the brain volumes synthesized with concrete dropout is more accurate than those with Monte Carlo dropout. A two-sided paired t-test was performed to compare the performance of the two methods, and the p-value of 0.0186 shows that the synthesized brain images with concrete dropout are significantly more accurate than the ones synthesized with Monte Carlo dropout at the 0.05 significance level.

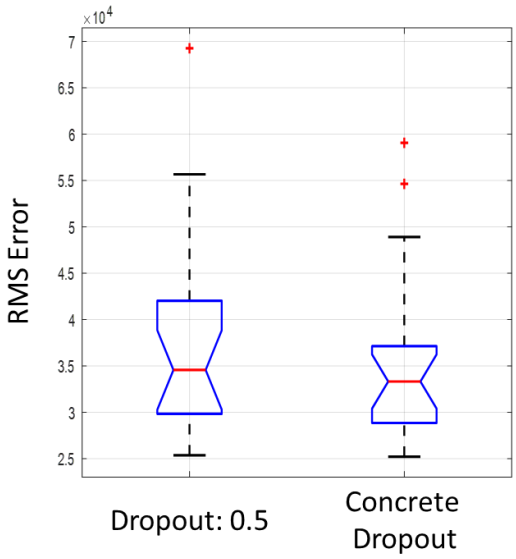

Fig. 4.2. Accuracy of the synthesized images. The boxplots of the RMS errors of the synthesized brain volumes for the subjects in the testing dataset are shown. The accuracy of the synthesized images by Bayesian conditional GAN with Monte Carlo dropout and concrete dropout were compared. The dropout rate of the Monte Carlo dropout was set as 0.5. In this figure, the points with red '+' symbols are drawn as outliers, if they are greater than q3+1.5(q3-q1) or less than q1-1.5(q3-q1), where q1 and q3 are the first and third quartiles respectively.

Fig. 4.3 shows the image synthesis results of a representative subject at 3 different slices. The original T1w image, the ground truth T2w image, the synthesized T2w image, the absolute error map between the prediction and the ground truth, and the uncertainty map for both methods are shown. In general, the synthesized T2w images can accurately reflect the brain anatomic structures in the real T2w images. However, there are also regions with relatively large image synthesis errors. For example, in slice 1 the upper left edge of the brain and the lower right spot in the brain marked by the green arrows in the absolute error maps show relatively large synthesis errors, and the uncertainty maps generated by both methods catch these regions with large distribution standard deviations or uncertainties, which are also marked by the green arrows. In slice 2, it can be observed that over all the uncertainty maps generated by both methods can catch the regions with large errors in the absolute error maps. However, the uncertainty map generated by Monte Carlo dropout has a hot spot around the upper left edge of the brain, but the error in corresponding region

in the error map is not very large. In addition, if we look into the details of the absolute error maps and the uncertainty maps of both methods, the model uncertainty of each voxel may not always be directly proportional to the corresponding synthesis error in the error map. This is mainly because by definition higher model uncertainty only means that the model is not very confident about the prediction and there is a higher chance for the model to make a prediction with a large error, but this does not mean that the model will definitely make a prediction with a large error at this voxel. In slice 3, we can see that both methods' uncertainty maps catch the high error region in the center of the tumor, but miss the high error region around the edge of the tumor. The possible reason for this is that the brain tumor of every subject in the training dataset has a different shape, location and contrast, and it's challenging for the model to catch the consistency among them.

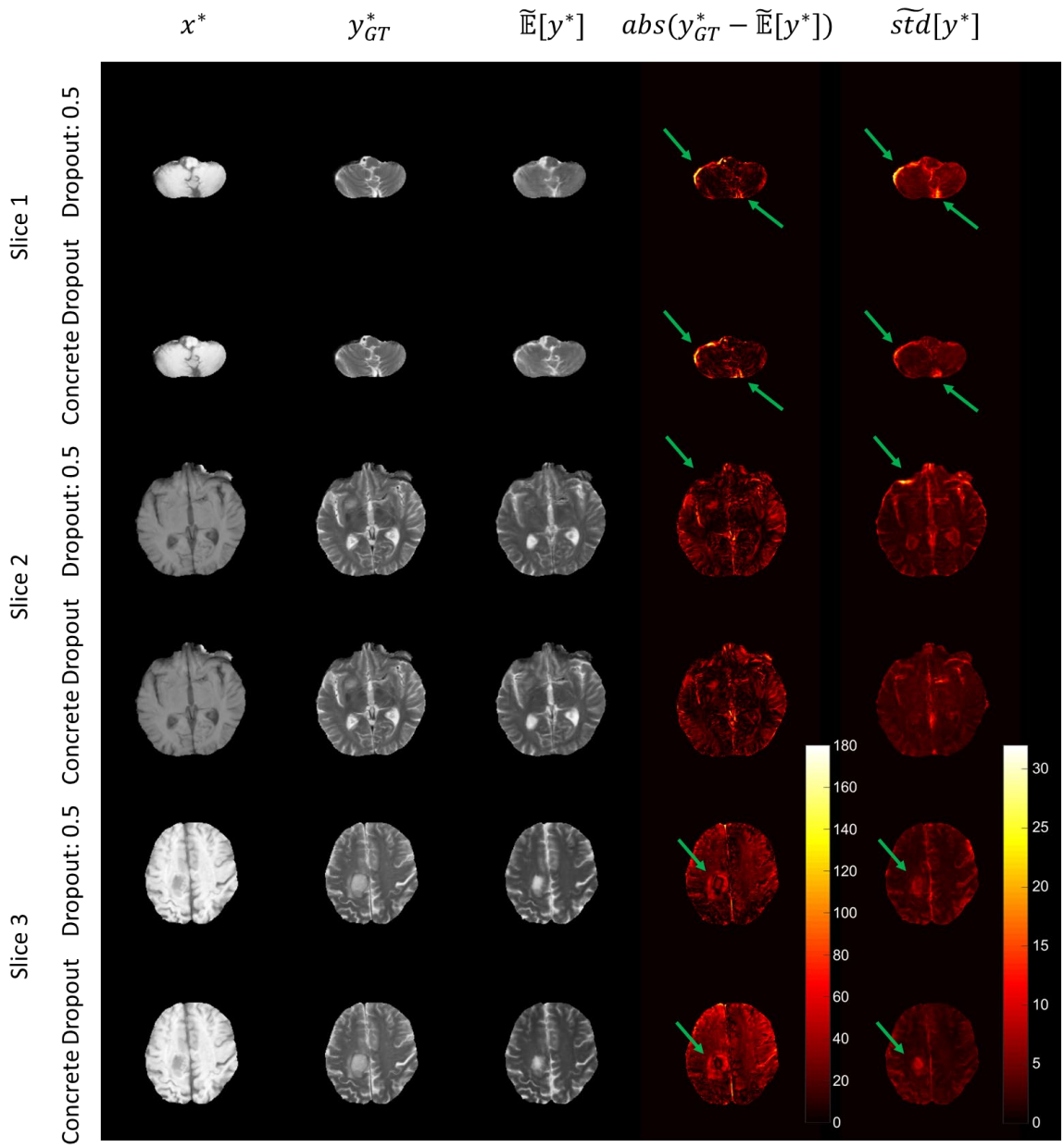

Fig. 4.3. Image synthesis results of a representative subject at 3 different slices. The original T1w image, the ground truth T2w image, the synthesized T2w image, the absolute error map between the prediction and the ground truth and the uncertainty map for both methods are shown.

4.3.2 Relationship between Prediction Accuracy and Model Uncertainty

By definition a voxel with high model uncertainty only means the model is not confident about the prediction on it, and the voxel has a higher chance to get a large prediction error. With this

definition, we cannot relate model uncertainty to prediction accuracy directly, but their relation should become stronger with large number of predictions and dimension reduction calculations, e.g. sum or average. Therefore, we plot out the accuracy versus uncertainty relations at 3 different levels in Fig. 4.4. For the voxel level, we plot the absolute error between the ground truth voxel value and the predicted value against the standard deviation (std) of the predicted posterior distribution. Since for each 3D brain volume the number of voxels in the brain region is huge, we only plot out the data in the brain region for a representative subject. For the slice level and the volume level, we plot the normalized root mean square error (nRMSE) of the predicted slice or 3D volume respectively against the normalized std (nSTD) of that slice or 3D volume. The definition of nRMSE and nSTD are:

$$nRMSE = \frac{\left\| y - \tilde{y} \right\|_{2,\phi}}{\left\| y \right\|_{2,\phi}} \tag{4.8}$$

$$nSTD = \frac{\left\| std \right\|_{2,\phi}}{\sqrt{N_\phi}} \tag{4.9}$$

where y and $\tilde{y}$ are the ground truth and the predicted value for a single voxel respectively; $\left\| \bullet \right\|_{2,\phi}$ denotes the l^2 norm over the brain region ϕ in that slice or 3D volume; *std* is the standard deviation of the predicted posterior distribution of a single voxel, and it is used as the model uncertainty for that voxel; N_ϕ is the total number of voxels in the brain region ϕ in that slice or 3D volume. From Fig. 4.4 we can find that, as the analysis goes to a more summary level, the relation between the prediction accuracy and the generated uncertainty becomes stronger. This holds true for both methods. This means that as a clue for the prediction error, the uncertainty of a voxel may not

work very well, but the uncertainty of a subject's 3D image volume has much stronger proportional relation with the prediction error.

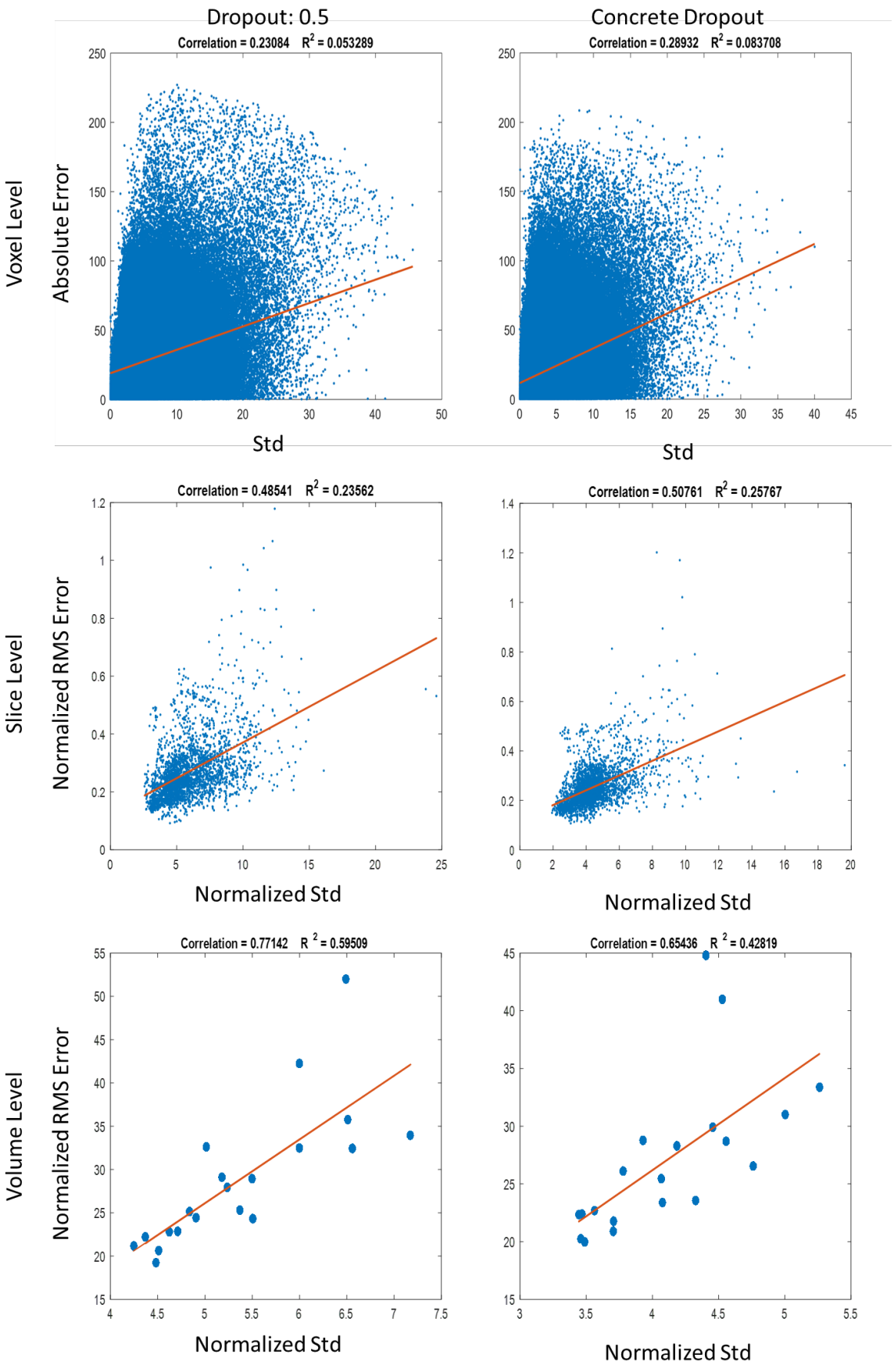
Dropout: 0.5
Concrete Dropout
Voxel Level
Slice Level
Volume Level
Absolute Error
Normalized RMS Error
Normalized RMS Error
Correlation = 0.23084 R2 = 0.053289
Correlation = 0.28932 R2 = 0.083708
Correlation = 0.48541 R2 = 0.23562
Correlation = 0.50761 R2 = 0.25767
Correlation = 0.77142 R 2 = 0.59509
Correlation = 0.65436 R 2 = 0.42819
Std
Std
Normalized Std
Normalized Std
Normalized Std
Normalized Std

Fig. 4.4. Relationship between the prediction accuracy and the model uncertainty at different levels. For both methods the error VS uncertainty observations were plotted at the voxel level, the slice level and the volume level. For the voxel level, the absolute error and the sdt of the predicted posterior are used. For the slice level and the volume level, the nRMSE and the normalized std are used. The voxel level plots only include all the voxels from a representative subject. The slice level plots and the volume level plots include all the slices and volumes in the tesing dataset respectively.

4.3.3 Model Uncertainty Evaluation

Since the model uncertainty generated is not necessarily directly proportional to the prediction error by its definition. The better metrics to evaluate the uncertainty generated are the precision recall plot and the uncertainty calibration plot. Fig. 4.5 is the precision recall plot for both methods, which shows how the prediction RMS error of the predictive model changes as the voxels with uncertainty larger than various percentile thresholds are removed. For example, on the recall axis, the point 1 means that all the voxels in the testing dataset are taken into account when the RMS error is calculated, and 0.9 means the RMS error is calculated without the top 10% voxels with the largest model uncertainty values in the testing dataset. First, as we can see, for both curves the RMS error decreases as voxels with uncertainties larger than various percentile thresholds are removed from the RMS error calculation. Since both curves monotonically decreasing as voxels with relatively large uncertainties are removed, this means that the RMS error of voxels with smaller uncertainties are also smaller, and that for both methods the RMS error correlates with the model uncertainty generated with a large number of predictions and the dimension reduction calculation, RMS error. In addition, both curves decrease faster, and the absolute values of the gradients become larger, as the same number of voxels with higher model uncertainties are removed. Second, at each recall value – each model uncertainty percentile threshold – the prediction RMS error achieved by the concrete dropout is smaller than that by the Monte Carlo dropout. This means that the model of Bayesian conditional GAN with concrete dropout is more

accurate than the model with Monte Carlo dropout in this application at each model uncertainty percentile threshold.

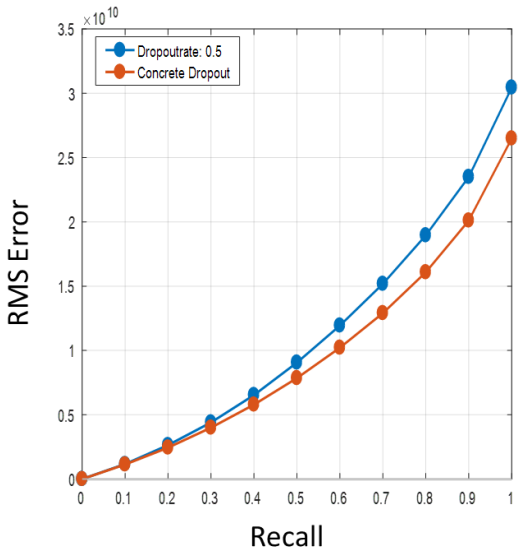

Fig. 4.5. Precision recall plot for Bayesian conditional GANs with Monte Carlo dropout and concrete dropout.

To analyze the quality of the uncertainty generated by the Bayesian conditional GANs, we studied the uncertainty calibration plot of the models on the testing data. To generate the uncertainty calibration plot, we select a number of probabilities with equal intervals in the CDF of the predicted posterior for each voxel. Then the frequency of the ground truth values of all the voxels falling below the corresponding quantiles of each selected probability is treated as the true probability. The definition of the selected probability and the true probability is the same as the p and f in Eq. 4.7. Then, we can plot out the true probabilities against the selected probabilities as the uncertainty calibration plot. The uncertainty generated with better quality should be closer to the diagonal line, $y = x$. Fig. 4.6 shows the uncertainty calibration plots for Bayesian conditional GANs with Monte Carlo dropout and concrete dropout. As we can see the uncertainty calibration plots from both methods have error in their scales, and this can be corrected by the model recalibration procedure

later. However, the uncertainty calibration plot of the Monte Carlo dropout misses the central point (0.5, 0.5), which is marked by a green point in the figure, and the whole curve is not symmetric to the central point. This means that compared to the true posterior distribution, the posterior predicted by the Monte Carlo dropout is shifted and distorted. First, by distortion, the predicted posterior doesn't align with the normal distribution assumption well. Second, the median of the predicted posterior is not the median of the true distribution. If we use the median, which is the same value as the mean in a normal distribution, as the final prediction result, the prediction will end up with a large prediction error. In contrast, the uncertainty calibration plot of the concrete dropout catches the central point and is symmetric to the central point. Although it also has an error in the scale, this can be corrected with the model recalibration procedure.

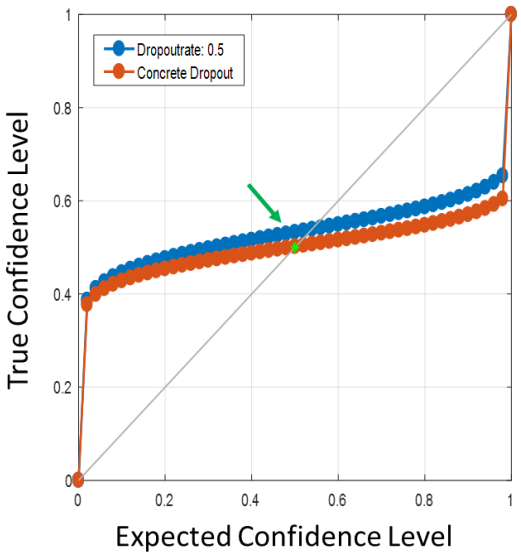

Fig. 4.6. Uncertainty calibration plot for Bayesian conditional GANs with Monte Carlo dropout and concrete dropout. A perfect calibration of the uncertainty corresponds to the diagonal line, $y = x$, shown in gray in the figure. The central point (0.5, 0.5) is marked as a green asterisk. The expected confindence level is the probability in the predicted posterior distribution, and the true confindence level is the oberserved true frequency in the data.

4.3.4 Model Recalibration

To further improve the quality of the uncertainty generated by the models we incorporated the model recalibration procedure after training the Bayesian conditional GANs. Instead of using a separate calibration dataset, we used the training dataset for model recalibration, and no obvious overfitting was observed. After training, predictions were made on the training dataset. With the ground truth of the training dataset the relationship between the selected probability p in the predicted posterior and the true probability, the observed frequency, f, can be calibrated according to Eq. 4.7. Then, in the prediction on the testing dataset, the probabilities in the predicted posterior was mapped to the true probabilities according to the calibrated relationship between them. In Fig. 4.7, the uncertainty calibration plots before and after the model recalibration procedure are illustrated for both methods. As we can see, for both methods the uncertainty calibration plots are closer to the diagonal line after model recalibration. This means that the model recalibration approach can improve the accuracy of the predicted posterior as well as the generated model uncertainty. By calculating the RMS error between the uncertainty calibration plot after model recalibration and the $y = x$ line, we can see that after model recalibration the calibration error of concrete dropout (RMS error = 0.2868) is still smaller than that of Monte Carlo dropout (RMS error = 0.3783).

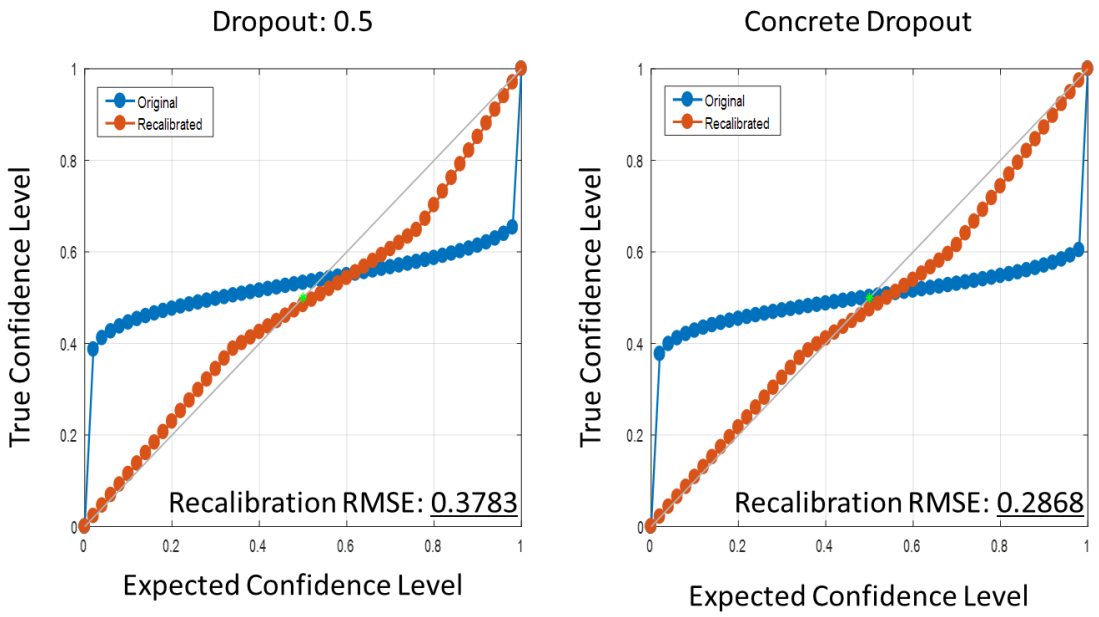

Fig. 4.7. Uncertainty calibration plot for Bayesian conditional GANs with Monte Carlo dropout and concrete dropout, before and after model recalibration. The point (0.5, 0.5) is marked as a green asterisk. The expected confindence level is the probability in the predicted distribution, and the true confindence level is the oberserved true frequency in the data.

4.4 Discussion

A new image synthesis model is proposed using Bayesian conditional GAN as the main image synthesis engine for the challenging application of MR brain tumor image synthesis. With the framework of Bayesian deep learning the proposed method can generate the posterior distribution of each voxel value for the synthesized image, which gives it the ability to make predictions as well as generate model uncertainties for the predictions. The use of concrete dropout enables the gradient-tuned dropout probability and results in more accurate prediction and uncertainty. By involving the model recalibration approach the calibration quality of the predicted posterior and the generated model uncertainty is further improved. The proposed method was applied to the challenging task of MR brain tumor image synthesis. In comparison with the model with the

conventional Monte Carlo dropout, the superior performance of the proposed method was validated.

As we can see in Fig. 4.2 and 4.5, the prediction accuracy of the Bayesian conditional GAN with concrete dropout is significantly better than that with Monte Carlo dropout. The possible reason is that the concrete dropout technique offers each dropout layer a gradient-tuned dropout probability which can converge to its optimal value during the training of the network. This optimal dropout probability can help the Bayesian deep learning model reach a higher accuracy than the dropout rate set by experience in the Monte Carlo dropout. At the same time, the concrete dropout changes the hyperparameter, dropout rate, in the forward model to a variable in the loss function for training and saves the time for grid-searching the optimal value of dropout rate. This also makes the model more robust than the model with empirical dropout rate.

In more details, the optimal dropout probability reached by concrete dropout can make the Bayesian deep learning model generate more accurate posterior distributions during the testing time, which is shown in Fig. 4.6, and that is why it can end up with a higher prediction accuracy. In Fig. 4.6, the error in the scale of calibration plot is not important, since we can still get a correct prediction when using the mean of the posterior as the final prediction. When the std of the posterior is used as the uncertainty, the uncertainty is only affected by a scale value. However, the shift or distortion of the calibration plot will end up with a wrong final prediction if the mean or median of the posterior is used as the final prediction result, since the final prediction result calculated is no more the median in the true distribution. The distortion of the calibration plot will also make the std uncertainty inaccurate and hard to correct. In Fig. 4.7, we can see that the error in the scale of the calibration plot can be easily corrected with the model recalibration procedure.

The shift and distortion in the calibration plot can also be corrected to some extent, but still will result in a slightly larger RMS error in the calibration plot after model recalibration.

The current workflow of artificial intelligence based research in the field of medical imaging looks like this: a deep learning model is trained on a training dataset, and then it will be applied to a testing dataset to make predictions. The prediction results will be compared with the ground truth given by radiologists. Obviously, various kinds of inconsistency will cause errors in the prediction: e.g. the inconsistency within the training dataset, the inconsistency between the training dataset and the testing dataset, etc. Therefore, if a trained AI is distributed to different hospitals and works by itself in the daily clinical routine, we won't have the ground truth any more, and we won't know whether the predictions made by the AI can be trusted or not. Bayesian deep learning based models solve this problem by generating model uncertainty information for each prediction. Whenever the model uncertainty is above a certain threshold, human intervention can be started to double check the case, or more information of the patient can be required for the AI to make a more confident prediction. By its definition, the model uncertainty may not directly reflect whether a prediction has a small error or a large error, but it suggests the possibility of a small or large error in the prediction. This means by having a large number of predictions and with the dimension reduction calculation we can get an averaged uncertainty having stronger correlation with the prediction accuracy. In Fig. 4.3 and 4.4, it is shown that at the voxel level the model uncertainty may not be directly proportional to the prediction absolute error, which will make it harder for us to locate the voxels with possible large prediction errors. However, at the slice level and the volume level the proportional relationship is much stronger, we can at least locate the slice or the brain volume that likely have a large prediction error.

There are also some limitations in this study. To have a more complete understanding of the performance we still need to compare our methods with many conventional image synthesis models. Another limitation is that we only applied our method to the application of MR brain tumor image synthesis from T1w to T2w images. The performance of the method can be different on different applications, so more experiments are still needed to verify the performance and the characteristics of the proposed method. In addition, by including more information in the input data, deep learning based methods have a higher chance to generate more accurate results. For example, instead of using only T1w MR images to synthesize T2w images. T1w and T2-FLAIR images can be used together to synthesize T2w images, or T1w, T2w and PDw images can be used together to synthesize T2-FLAIR images. Since more information is offered, and T1, T2 and proton density are the 3 basic tissue properties for MR tissue contrasts, the image synthesis accuracy achieved by deep neural network can be higher. These studies will be included in our future work.

4.5 Conclusion

This study presents a new Bayesian deep learning based image synthesis model, Bayesian conditional GAN, which can not only accurately synthesize MR neuroimages but also generate uncertainty maps for the synthesized images. The model takes advantage of the concrete relaxation of the Bernoulli distribution and the KL divergence regularization term in the loss function of Bayesian deep learning for gradient-tuned dropout probabilities, and ends up with higher image synthesis accuracy and more accurate model uncertainty. Moreover, the incorporation of the model recalibration method further improves the model uncertainty calibration. The successful

application of the proposed method to the MR brain tumor image synthesis suggests that the method can be further applied to other fields of medical image synthesis.

Chapter 5

Bayesian Deep Neural Network for Brain Functional Connectivity Gender Prediction

5.1 Introduction

The difference in brain structure and function between genders has been a durative topic in the field of neuroscience, and has an important role in determining differences between genders in various psychological and behavioral processes (Gong et al., 2011). For example, it has been known for a long time that males and females are different in memory, language, emotion, perception, navigation and other cognitive categories (Cahill, 2006). At the same time, both structural and functional brain differences between genders have been found in various modalities (Cosgrove et al., 2007; Gong et al., 2011), and brain gender differences widely exist not only in healthy subjects but also in subjects with different kinds of brain disorders, including autism (Alaerts et al., 2016), depression (Orgo et al., 2016) and Alzheimer's disease (Malpetti et al., 2017)

etc. Thus, analyzing the brain differences between genders can help improve the understanding of the neurobiological mechanisms behind different gender-related psychological and behavioral processes. Furthermore, for brain disorders that manifest differences between genders, knowledge of brain differences may help to provide a deeper understanding of the psychopathology and aid in the development of related treatments.

In the recent past, tremendous progress has been made in the field of artificial intelligence because of the surge of the deep learning methods (Krizhevsky et al., 2012a; LeCun et al., 2015) and the rapid advance of parallel computing (Coates et al., 2013; Schmidhuber, 2015). Due to its ability of accurate prediction, deep learning has been quickly applied to the field of MR neuroimaging. CNN has been successfully applied to segmentation and classification tasks based on structural MR images. Several patch-based 3D CNNs were proposed to segment the subcortical regions of the brain (Dolz et al., 2017) and brain lesions (Kamnitsas et al., 2017). Bayesian CNN was combined with fully connected conditional random field to perform brain extraction and generate corresponding uncertainty maps for non-human primates (Zhao et al., 2018). CNN was also combined with support vector machine to classify the overall survival time of brain tumor patients (Nie et al., 2016).

At the same time DNN is more suitable for the prediction with brain functional connectivity (FC) as input. (Kim et al., 2016) used DNN to classify schizophrenia patients against healthy controls with an accuracy of 85.8% and investigated multiple DNN configurations' effects on the predicting accuracy. Several groups have applied DNN based methods to the diagnosis of Alzheimer's disease (Liu et al., 2014; Suk et al., 2015; Hu et al., 2016; Bhatkoti and Paul, 2016) and showed improvement over traditional methods. (Hazlett et al., 2017) used a combination of DNN and

support vector machine to study the brain development of infants at high risk for autism spectrum disorder. A more comprehensive review can be found in (Vieira et al., 2017).

In neuroscience it is not only the prediction accuracy, but also the underlying features that drive the accurate predictions, that are of great interest. However, DNN is often treated as a black box, and it is challenging to understand its classification mechanism and extract the important features. In this study, we propose a novel feature extraction and ranking method for DNN and apply it to the study of brain FC gender difference to verify the method's effectiveness and study its characteristics.

There are two purposes of this work. One is to use deep learning and Bayesian deep learning based methods to perform gender prediction, related feature extraction and model uncertainty generation from the resting state brain FC at multiple scales. The other is to build a framework for testing the reliability, repeatability and robustness of the prediction accuracy, the extracted FC features and the model uncertainty. Gender prediction can also serve as a basic testbed to verify and compare these methods' performance to conventional methods and study the characteristics of these methods for further neuroscience applications.

First, we propose to utilize the highly nonlinear model of DNN to predict gender from brain FC. The trend of the prediction accuracy of DNN at different scales of brain FC was investigated in comparison with that of the linear SVM. Second, instead of treating each connection in the connectivity matrix as a single feature, with DNN the features are extracted as connectivity patterns in the whole connectivity matrix. We propose a method to rank the extracted high-level male and female features based on their contributions to the prediction and studied how much prediction accuracy the several most important features can preserve. Third, Bayesian deep learning was further applied to the FC based gender prediction. The prediction accuracies at

multiple dropout rates were compared with the conventional weight averaging technique. The behavior of the model uncertainty generated by Bayesian deep learning for each prediction was also studied. Finally, the repeatability and robustness of all the results were tested through 50 randomly permuted cross validations in all the DNN structures and scales of brain FC studied. All the tests were done on the high-quality large-scale dataset of the HCP (Van Essen et al., 2013) S1200 release (https://www.humanconnectome.org) at multiple brain connectivity scales derived from different numbers of ICA components. A full comparison was also made between different DNN structures and commonly used machine learning methods.

5.2 Material and Methods

5.2.1 Dataset and Preprocessing

The rs-fMRI data of 1003 healthy adults (age: 22~37; 53.2% female) having 4 complete runs in the HCP S1200 release were used for this study (S. Smith et al., 2013). According to the HCP data dictionary, the term "gender" is used instead of "sex" (https://wiki.humanconnectome.org). While "gender" and "sex" are different, participants in the HCP were asked a number of demographic questions, including "gender" (HCP S1200 release reference manual, https://www.humanconnectome.org). We therefore use the term "gender" in this study to conform with the HCP nomenclature. All the data were collected on the customized Siemens Skyra 3T MRI scanner for the HCP with a standard 32-channel Siemens receive head coil. A multi-band gradient-echo EPI sequence was used for the rs-fMRI with the imaging parameters: TE = 33.1ms, TR = 720ms, Flip Angle = 52°, Multi-band factor = 8. Each subject underwent 4 rs-fMRI runs in 2 sessions. The duration of each run was 14 minutes 33 seconds, and each run was reconstructed into 1200 3D volumes of 104×90 in-plane matrix size and 2×2 mm^2 in-plane pixel size with 72

slices and 2 mm slice thickness. Table 5.1 is a summary of the demographics, brain volumes and motion measurements of the male and female groups of subjects involved in this study.

Table 5.1. Summary of the demographics and measurements of the subjects used.

	Male (Mean ± std)	**Female (Mean ± std)**	**p-value**
Demographics			
Number (%)	469 (46.76%)	534 (53.24%)	
Ethnicity (Caucasian/%)	365 (77.83%)	393 (73.60%)	
Age (year)	27.87 ± 3.65	29.45 ± 3.61	<0.001
Education (year)*	14.87 ± 1.74	15.05 ± 1.79	0.092
Handedness	61.04 ± 43.48	70.97 ± 43.23	<0.001
Weight (pound)*	189.64 ± 34.77	155.95 ± 35.50	<0.001
Blood pressure: systolic (mmHg)*	128.00 ± 13.25	119.40 ± 13.00	<0.001
Blood pressure: diastolic (mmHg)*	78.20 ± 10.38	74.74 ± 10.46	<0.001
Neuropsychological measurement			
Fluid intelligence: PMAT24_A_CR*	17.72 ± 4.59	16.41 ± 4.73	<0.001
Fluid intelligence: PMAT24_A_SI*	2.42 ± 3.63	3.40 ± 3.90	<0.001
Fluid intelligence: PMAT24_A_RTCR*	17235 ± 9421	14754 ± 8825	<0.001
Brain volume (cm^3) (Gray matter + White Matter + CSF)	1215 ± 96	1063 ± 85	<0.001
Motion (mm) (Movement_RelativeRMS_mean)	0.0862 ± 0.0346	0.0876 ± 0.0346	0.218

* Missing values from some subjects were removed in the calculations.

Handedness: [-100,100], positive numbers indicate more right-handedness; Fluid intelligence: measured with Penn Progressive Matrices (Bilker et al., 2012); PMAT24_A_CR: number of correct responses; PMAT24_A_SI: total skipped items; PMAT24_A_RTCR: median reaction time for correct responses; Brain volume: results from FreeSurfer (http://freesurfer.net/); Motion: temporal mean of the root mean squire of the relative motion, results from HCP minimal preprocessing pipeline (Glasser et al., 2013); for more detailed definitions, please refer to the HCP Data Dictionary for the 1200 Subjects Release (https://wiki.humanconnectome.org).

All the data were preprocessed by the HCP minimal preprocessing pipeline (Glasser et al., 2013) including distortion correction, field map correction, motion correction and spatial normalization. After the gradient distortion correction, head motion of the rs-fMRI data was corrected by registration to the single-band reference image collected at the beginning of each run. Then the

spatial distortion caused by B0 was corrected with field maps, and the single-band reference image was further registered to the T1w structural image. All the preceding transforms were concatenated and applied to the original rs-fMRI images and the images were resampled into 2mm MNI space. Finally, global intensity normalization was done, and non-brain areas were masked out (S. Smith et al., 2013). The HCP multi-modal surface matching algorithm (MSM-ALL) (Robinson et al., 2014) was used during the preprocessing to improve the inter-subject registration of cerebral cortex with the areal features derived from myelin maps, resting state network and rs-fMRI visuotopic maps. The artifacts within the rs-fMRI data were further removed with ICA-FIX (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014) and regression of 24 confound time series derived from motion estimation (including 6 rigid-body parameters, their derivatives and the squares of all the 12).

Spatial group-ICA (Hyvarinen, 1999; Beckmann and Smith, 2004) was applied to the data in the grayordinate space following group-PCA (Smith et al., 2014) to get ICA based parcellations. Since larger number of components in ICA leads to more and smaller parcels and vice versa, group-ICA was used to decompose the data into several different levels: 25, 50, 100, 200 and 300 components, to analyze brain connectivity at different scales. The associated average blood-oxygenation-level-dependent (BOLD) time series for each subject's ICA components can then be obtained through multiple-spatial-regression of the 4D rs-fMRI data against the group-ICA spatial maps. Then the brain connectivity matrix for each subject was calculated using the Pearson's correlation coefficient from the concatenated 4 runs' ICA component time series.

The preprocessing steps were performed by the HCP, and the connectivity matrices released in the HCP S1200 Extensively Processed fMRI Data (https://www.humanconnectome.org/study/hcp-young-adult/document/extensively-processed-fmri-data-documentation) were used in this study.

Each connectivity matrix was Fisher's r-to-z transformed and normalized to a matrix with zero mean and unit variance to be used as the input of the deep learning classifier. For comparison, the connectivity matrices from the template based parcellation were also calculated. After the ICA-FIX step, the average BOLD time series were extracted from the 330-parcel HCP multi-modal cortical parcellation (Glasser et al., 2016) and the 49-parcel FreeSurfer (http://freesurfer.net/) subcortical parcellation in the grayordinate space (https://balsa.wustl.edu/WK71) (Van Essen et al., 2017). Then, the matrices were Fisher's r-to-z transformed and normalized as above.

5.2.2 Deep Neural Network

Deep learning is a kind of mathematical model originally inspired by the biological neural system and then widely used for artificial intelligence tasks due to its extraordinary performance (LeCun et al., 2015). As a kind of artificial neural network, DNN usually contains multiple hidden layers that are fully connected layers. The structure of a typical DNN with three hidden layers is shown in Fig. 5.1. To simulate the behavior of biological neurons (London and Häusser, 2005), the value of the kth neuron in layer l+1 of DNN, A_k^{l+1}, equals the nonlinearly transformed linear combination of neurons A^l in layer l with a bias b_k^l. The sigmoid function $sgm(x)$ is a commonly used nonlinear transformation in DNN, since it maps a real-valued input to (0,1), which can represent the firing rate of a neuron.

$$\begin{aligned} A_k^{l+1} &= sgm(W_k^{l+1,l} A^l + b_k^{l+1,l}) \\ A_k^1 &= sgm(W_k^{1,0} X + b_k^{1,0}) \\ sgm(x) &= 1/(1+e^{-x}) \end{aligned} \tag{5.1}$$

$W_k^{l+1,l}$ and $b_k^{l+1,l}$ are the weight matrix and bias, respectively, for layer l pointing to the kth neuron in layer $l+1$.

In the last layer L, for classification the softmax function is used as the nonlinear transformation to map the real-valued score for each class into normalized estimated class probabilities, and a categorical cross entropy loss function with elastic net regularization of the weights is used to train the DNN.

$$\begin{aligned} \hat{W}, \hat{b} &= \arg\min_{W,b} -\sum_i \sum_k p_i(k) \log q_i(k) + \sum_l \beta^{l+1,l} \left\| W^{l+1,l} \right\|_1 + \sum_l \frac{\gamma^{l+1,l}}{2} \left\| W^{l+1,l} \right\|_2^2 \\ q(k) &= e^{S(k)} / \sum_k e^{S(k)} \end{aligned} \tag{5.2}$$

$\sum_k p_i(k) \log q_i(k)$ is the cross entropy between the estimated class probability $q(k)$ and the true probability distribution $p(k)$ for the ith input in the batch. β and γ are the weights for the l_1 norm and l^2 norm regularization of the weights in the DNN respectively. $S(k) = W_k^{L,L-1} A^{L-1} + b_k^{L,L-1}$ is the real-valued score for the kth class in the last layer. In addition to preventing overfitting, the l_1 norm and l^2 norm regularization is also used based on the assumptions of the human neural system. l^2 norm regularization is used to penalize peaky weight vectors and favor diffuse weight vectors, since we are looking for connectivity patterns for males and females and want the whole connectivity matrix taken into consideration. Meanwhile, l_1 norm is used to drive the weight vector sparse, since only a subset of the connections is assumed to be finally relatively useful for the gender prediction. In addition, a dropout layer is also used following each hidden layer to prevent overfitting.

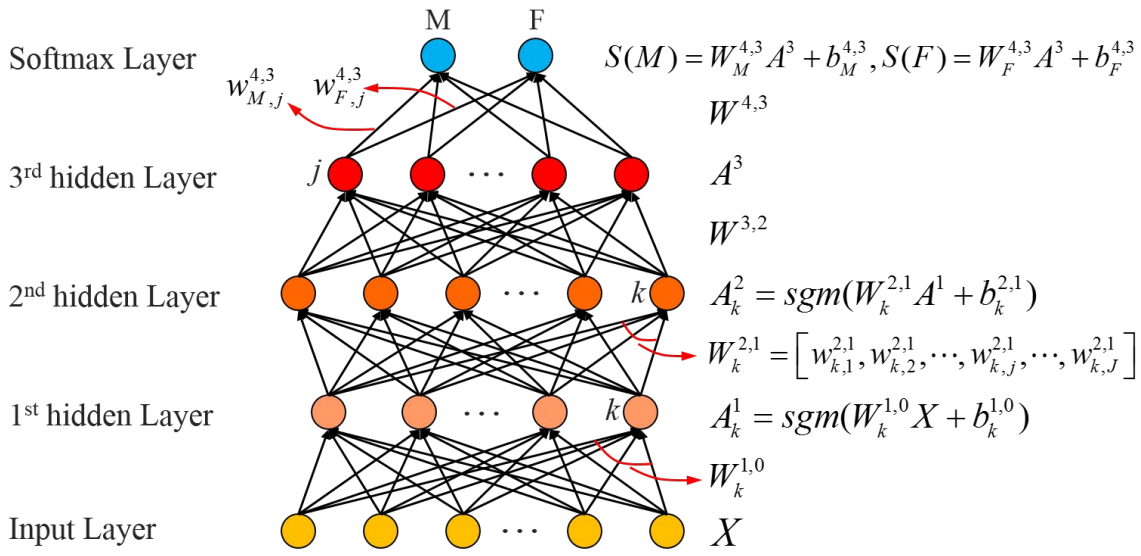

Fig. 5.1. Illustration of the structure of a typical DNN with 3 hidden layers for classifing 2 classes.

5.2.3 Deep Neural Network Feature Extraction and Ranking

In the field of neuroscience, knowing which features matter most in a prediction is as important as reaching a high accuracy in the prediction, and at the same time reaching a high accuracy is a necessary condition to make the features extracted meaningful. In this study instead of extracting the features based on comparing the group differences of all the input features, we use the framework of extracting features from a high-accuracy prediction model to extract the most important FC patterns from the DNN used for gender classification. Based on the DNN high-level feature extraction approach in the previous study (Kim et al., 2016), we further developed the feature extraction mechanism in DNN, and proposed a method to extract and rank the features based on their contributions to the final classification.

Unlike the feature extraction approaches in other statistical or machine learning models (e.g. logistic regression, SVM etc.), which usually use the weight for each individual predictor variable to reflect the importance of the predictor variable in the prediction, our method looks for the most

important predictor variable patterns in the input for each class as features. Specifically, in the experiments in this study, our method extracts and ranks the male and female FC patterns of the whole connectivity matrix instead of several individual functional connections.

The weight vectors between the input layer and the 1st hidden layer can be considered as convolutional kernels. There are the same number of convolutional kernels as the neurons in the 1st hidden layer, and these kernels are trained to extract the FC patterns which are most helpful on classifying different classes from the input connectivity matrices. $W_k^{1,0}$ is the convolutional kernel pointing to the *k*th neuron in the 1st hidden layer, and only the connectivity matrices X having a similar pattern can make the product $W_k^{1,0}X$ large. $b_k^{1,0}$ is the bias added to the product towards the *k*th 1st-hidden-layer neuron. The biases are also trained to optimize the classification task, and each of them serves as the baseline of the corresponding product between the input and the convolutional kernel. Then, the sigmoid function is used to map the real-value convolution result to the activation A_k^1, which is in the range (0,1). The more an activation is close to 1, the more the FC pattern in the input connectivity matrix is similar to the trained convolutional kernel's pattern.

The whole set of activations of the 1st hidden layer can be understood as the evaluation on how much each kind of pattern $W_k^{1,0}$ a specific input X has, while between the 1st and 2nd hidden layers the convolutional kernels $W_k^{2,1}$ are used to look for a certain combination of these patterns for the activation A_k^2. $w_{k,j}^{2,1}$ is a weight in the convolution kernel $W_k^{2,1}$ pointing to the *k*th neuron in the 2nd hidden layer from the *j*th neuron in the 1st hidden layer. The absolute value of the weight $w_{k,j}^{2,1}$ determines the importance of the corresponding pattern $W_j^{1,0}$, and the sign of $w_{k,j}^{2,1}$ means whether the pattern $W_j^{1,0}$ is preferred to appear in the combination or not. If the absolute value of $w_{k,j}^{2,1}$ is

very small, then regardless of whether the corresponding pattern $W_j^{1,0}$ appears or not in the input, the product $w_{k,j}^{2,1}A_j^1$ will not make much difference on the activation A_k^2 in the higher layer. Since all the activations are positive values in the range (0,1), a large positive weight $w_{k,j}^{2,1}$ will generate a relatively large positive product $w_{k,j}^{2,1}A_j^1$ and lead the activation A_k^2 in the higher layer closer to 1, while a large negative weight value will lead the activation more towards 0. When it comes to the higher hidden layers, the higher level of combinations of the patterns are evaluated for the classification task.

The key part of the DNN structure for extracting and ranking the features for each class in the proposed method is between the last hidden layer and the softmax layer. In the application of gender classification, there are only two classes, male and female. Each activation, A_j^L, in the last hidden layer L will be multiplied by the weights $w_{M,j}^{L+1,L}$ and $w_{F,j}^{L+1,L}$ respectively towards the male and female neurons, and the baselines $b_{M,j}^{L+1,L}$ and $b_{F,j}^{L+1,L}$ will also be added to them to get the scores $S(M) = W_M^{L+1,L}A^L + b_M^{L+1,L}$ and $S(F) = W_F^{L+1,L}A^L + b_F^{L+1,L}$. The two scores compete with each other and the input will be classified into the class with the larger score, which means the difference between each $w_{M,j}^{L+1,L}$ and $w_{F,j}^{L+1,L}$ pair determines whether the high-level FC pattern represented by A_j^L is used as a male feature or a female feature, and how important it is for the whole DNN classification model. If the difference $w_{M,j}^{L+1,L} - w_{F,j}^{L+1,L}$ is positive, then the corresponding high-level FC pattern is the feature looked for in a subject's connectivity matrix to identify one as a male and vice versa. Meanwhile, the larger the absolute value of the difference is, the more important feature the corresponding FC pattern is in the DNN model for identifying genders, since compared with other FC patterns, the appearance of this FC pattern in a subject's connectivity matrix contributes

more to the total difference between the final scores. Therefore, the absolute value of the difference $\left|w_{M,j}^{L+1,L} - w_{F,j}^{L+1,L}\right|$ can be sorted to rank the importance of the high-level features for both male and female classes. Since a high-level feature F_k^{l+1} can be treated as the combination of lower-level features:

$$\begin{aligned} F_k^{l+1} &= \sum_{j \in J} w_{k,j}^{l+1,l} F_j^l \\ F_k^1 &= W_k^{1,0} \end{aligned} \tag{5.3}$$

then any high-level feature can be finally represented by the combination of the FC patterns at the original input level (Kim et al., 2016). F_k^1 is the *k*th feature at the original input level, which equals to the *k*th convolutional kernel between the input layer and the 1st hidden layer. *J* is a set of features and their corresponding weights. It can be the universal set including all the features and weights, or a subset including only the weights having relatively large absolute values by a threshold and the corresponding features.

5.2.4 Bayesian Deep Learning and Bayesian Deep Neural Network

Bayesian DNN is implemented by adding dropout layers in the network structure and using dropout in both training and testing (Srivastava et al., 2014; Gal and Ghahramani, 2016). A dropout rate in the range (0,1) is set beforehand for each dropout layer, and the neurons in the preceding layer are randomly set to zero at the dropout rate in every iteration during training or in every forward pass during testing. Dropout training offers Bayesian DNNs extra regularization against overfitting (Srivastava et al., 2014), while dropout testing can be viewed as sampling the posterior distribution of the predicted label probabilities over the Bayesian DNN weights (Gal and Ghahramani, 2016).

In dropout training, given the training dataset X and the label set Y, the posterior distribution of the network weights W, $p(W|X,Y)$, can be obtained through an approximating distribution, $q(W)$, made with variational inference, and $q(W)$ can be inferred by minimizing the KL divergence (Gal and Ghahramani, 2015):

$$KL(q(W) \| p(W|X,Y)) \tag{5.4}$$

In dropout testing, to predict the label y^* for the data x^*, the posterior distribution can be inferred by Monte Carlo (MC) dropout testing. In practice, T times of stochastic forward passes are performed and each time a network weight subset $\hat{W}_t$ is sampled (Gal and Ghahramani, 2015).

$$\begin{aligned} &p(y^* \mid x^*, X, Y) \approx \int p(y^* \mid x^*, W) q(W) dW \approx \frac{1}{T} \sum_{t=1}^{T} p(y^* \mid x^*, \hat{W}_t) \\ &\hat{W}_t \sim q(W) \end{aligned} \tag{5.5}$$

The mean of the sampled probabilities is used as each label's predicted probability, and the variance of them is used as the model uncertainty on each prediction. The uncertainty generated by MC dropout testing can reflect the confidence of the model on each prediction and is particularly useful for the applicability evaluation of the trained model on the testing dataset. In comparison, the conventional weight averaging testing uses the weights multiplied by the corresponding retaining probabilities (1 - dropout rate) during testing time and can only make predictions (Srivastava et al., 2014; Gal and Ghahramani, 2016).

5.2.5 Experiments

5.2.5.1 Gender Prediction from Multi-Scale Functional Connectivity

To verify and study the performance and characteristics of DNN models and the proposed feature extraction and ranking method, we used the application of FC gender prediction and gender related FC feature extraction as a testbed for experiments. First, to understand the prediction performance of DNN models at different scales of brain connectivity, we used the connectivity matrices derived from different numbers of ICA components: 25, 50, 100, 200, 300 components, as the input for the DNN classifier. In case the structure of the DNN also has a significant influence on the prediction accuracy, the performances of DNNs with different numbers of hidden layers and different numbers of neurons in each layer were also studied. DNNs with 1, 2, 3 hidden layers and 20, 50, 100, 200 neurons in the first hidden layer, and half number of neurons in the following hidden layers were used for training and testing. For example, the DNN with 3 hidden layers and 20 neurons in the 1^{st} hidden layer has 10 neurons in the 2^{nd} and 3^{rd} hidden layers, while the DNN with 1 hidden layer and 20 neurons only has a 1^{st} hidden layer with 20 neurons. Since the number of neurons in the following hidden layers controls the number of combined features used for classification, using a smaller number of neurons in the higher hidden layers is based on the assumption that a smaller number of 1^{st} hidden layer features' combinations are enough as higher-level features for the classification, and this can dramatically reduce the parameter space of the DNN model. The training and testing were carried out in a two-fold cross validation manner on the 1003 subjects (502 subjects in one fold and 501 in another), and 50 randomized permutations were performed on how to divide the training and testing datasets. A three-way ANOVA using an overdispersed binominal logistic regression was used to test the factors of number of ICA components, number of DNN hidden layers and number of neurons in the DNN layer and their interactions on prediction accuracy. Post hoc testing was also performed following ANOVA to

determine the prediction accuracies of which parameter combinations were significantly different from one another.

Then, to compare with conventional machine learning models, SVMs with linear and order-2 polynominal kernels were also studied for gender prediction. To compare with the ICA based functional connectivity, connectivity matrices derived from the template based parcellation, the combination of the 330-parcel HCP multi-modal cortical parcellation (Glasser et al., 2016) and the 49-parcel FreeSurfer (http://freesurfer.net/) subcortical parcellation in the grayordinate space (https://balsa.wustl.edu/WK71) (Van Essen et al., 2017), were also used to predict gender. 50 randomized permutations of 2-fold cross validations were also done respectively for these comparative methods, and related statistical tests were also performed.

All the DNNs with different structures were trained with the SGD algorithm. The learning rate varies from 0.05 to 0.8 and number of iterations varies from 40 to 300, depending on the DNN structure and the size of the input vector (the number of input FC connections in different numbers of ICA components) to make all the training procedures converge around the loss of 0.1. β and γ were set as 10^{-6} and 10^{-4} respectively while a dropout rate of 0.2 was set for the last hidden layer in all the DNN models. All DNN models were implemented in Python with the libraries Keras and Tensorflow, and the SVM models were implemented with the package scikit-learn. All the training and testing were performed on a workstation hosting 2 Intel Xeon(R) E5-2620 v4 CPUs (8 cores, 16 threads @2.10GHz) with 64 GB DDR4 RAM and two GPUs: an Nvidia GTX980Ti GPU with 6 GB memory and an Nvidia TITAN Xp GPU with 12 GB memory. A 64-bit Linux operation system ran on the workstation.

5.2.5.2 DNN Feature Extraction and Robustness Evaluation

After the gender prediction and related statistical analysis, we extracted and ranked the FC features for gender classification with the proposed method, studied their robustness and analyzed the feature differences between males and females. First, to verify the proposed feature-ranking method, we used several individual high-level male and female feature pairs ranked at different importance levels to make predictions on the training dataset itself, and compared the cross entropy loss achieved by each of these feature pairs. Second, we studied the relationship between the prediction accuracy and the number of learned important feature pairs involved in the prediction by investigating how much prediction accuracy the DNN model can preserve by using only a few most important (only the several highly important) high-level feature pairs during prediction. Third, the repeatability of the features learned from each DNN structure was also studied. All these experiments were done on all the DNN structures and all the numbers of ICA components in all the 50 randomly permuted cross validations. Finally, based on the results of these experiments, the extracted FC features from the selected DNN models were plotted to visualize the differences between males and females.

5.2.5.3 Bayesian Deep Learning and Monte Carlo Dropout Testing

To study the performance of Bayesian deep learning, MC dropout was performed on the previously trained 3-hidden-layer DNNs with the 3rd hidden layer dropped out in dropout testing. The prediction accuracies of several different dropout rates were studied, and the prediction accuracy of weight averaging testing was also compared. The behavior of the uncertainty generated by Bayesian deep learning on the application of image segmentation was already studied in previous research (Zhao et al., 2018). To verify the validity of the model uncertainty generated by Bayesian DNN in the classification application, in this study we also investigated the model uncertainty's behavior by making tests on the data generated with varying levels of male/female uncertainty. In

the testing dataset of each cross validation fold, 200 female subjects and 200 male subjects were randomly selected and referred to as the female subset and the male subset. Then, 200 0.75-female-0.25-male subjects were made up by linearly combining the connectivity matrices in the female and male subsets and referred to as the 0.75-female-0.25-male subset. Each subject's connectivity matrix in the 0.75-female-0.25-male subset is the sum of 0.75 times a random connectivity matrix from the female subset and 0.25 times a random connectivity matrix from the male subset, and each connectivity matrix in the female and male subsets was used only once. In a similar manner, the 0.5-female-0.5-male subset and the 0.25-female-0.75-male subset were also created. In each cross validation, the prediction accuracy and corresponding model uncertainty on all the 5 subsets were generated by the Bayesian DNNs with the 3rd hidden layer followed by a dropout layer. Based on the previous experiment, to keep both the prediction accuracy and enough variation a dropout rate of 0.5 was selected. Since the Bayesian deep learning models are trained with only the real female and male subjects, and the made-up subjects have more uncertainty in the data themselves, on the made-up subsets the prediction accuracies of the Bayesian DNNs are supposed to be lower and the corresponding model uncertainties are supposed to be higher.

5.3 Results

5.3.1 Gender Prediction from Multi-Scale Functional Connectivity

The performances of DNNs with various model structures on gender prediction at different scales of brain FC were evaluated, and the SVMs with linear kernel and order-2 polynomial kernel were also compared. In Fig. 5.2 the mean and standard deviation of the accuracies across the 50 randomized cross validation permutations are plotted. The accuracy in each permutation is the prediction accuracy on the 1003 HCP subjects evaluated from the 2-fold cross validation. As is

seen in Fig. 5.2, as the number of ICA components increases, the scale of the brain FC under investigation becomes finer, and the gender prediction accuracy increases and gradually plateaus. For the predictions from the same scale of FC but different DNN structures, the prediction accuracies are relatively close. The accuracy of the SVM prediction also goes up and gradually plateaus as the FC scale becomes more and more detailed. In the comparison between DNN and linear SVM, DNNs' predictions have an advantage with small numbers of ICA components. As the number of ICA components increase, the prediction accuracy of the linear SVM catches up, and becomes better than the DNNs with large numbers of ICA components. The order-2 polynomial SVM is slightly better than the linear SVM for 25-component ICA. For larger numbers of ICA components the linear SVM is better than the order-2 polynomial SVM. The prediction accuracies for DNNs and SVMs on the brain FC derived from the template based parcellation were also shown in Fig. 5.2 for reference. Although there are 330 cortical parcels and 49 subcortical parcels in total, the prediction accuracies are much smaller than the predictions for the 100-component ICA. All the exact mean and standard deviation values in Fig. 5.2 are shown in the Appendix B Table B1 and B2.

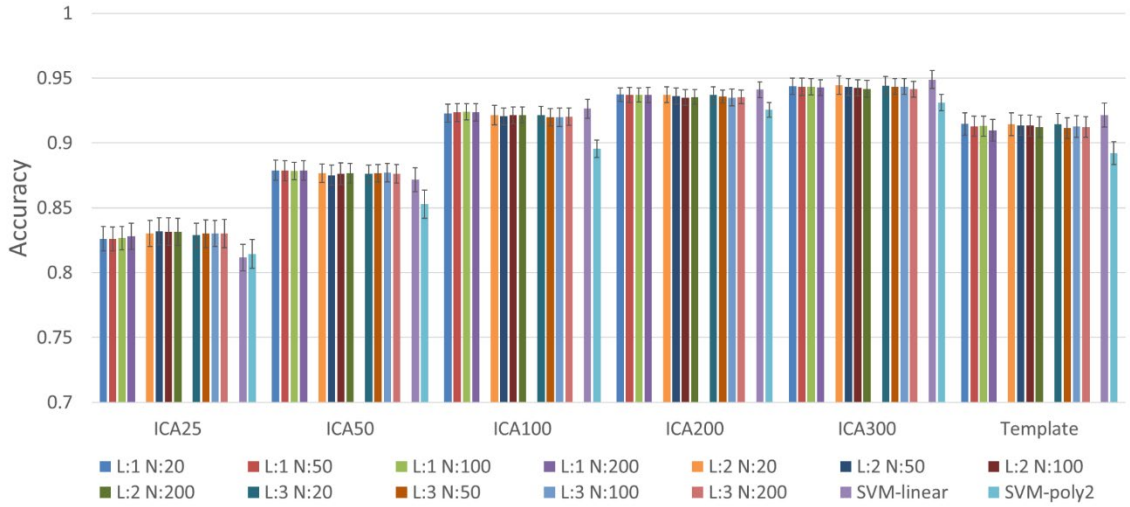

Fig. 5.2. The mean and standard deviation of the prediction accuracies across the 50 cross validation permutations for each kind of predictive model and each kind of input. L – number of hidden layers in the DNN model; N – number of neurons in the first hidden layer of the DNN, the number of neurons in each of the rest hidden layers is half of this number.

The corresponding statistical tests were also carried out to show the significance of the differences between the reported accuracies. First, a three-way ANOVA test using an overdispersed binomial model was performed on the accuracies generated by DNNs on different scales of FC derived from different numbers of ICA components. Due to the inclusion of interactions in the model, type II ANOVA was used on the 3 factors: the ICA component number, the DNN hidden layer number and the DNN neuron number. The ANOVA test was performed in a hierarchical manner with the full model including all the interactions at the beginning. The factor of ICA component number is significant with a p-value $<< 0.001$, and interaction between the ICA component number and layer number is significant with a p-value $<< 0.001$. (All the p-values reported in this study are two-sided.) Other factors and interactions are not significant at the 0.05 significance level. Multiple comparison was done for the significant interaction in the post hoc analysis with the Tukey adjustment. For all the 105 pairwise comparisons, the differences between different ICA component numbers are significant with Tukey corrected p-values < 0.01, regardless of the other factors. The differences within the group having the same ICA component number are mainly insignificant at the 0.05 significance level, except that in the 25-component ICA group the difference between the 2-layer and 1-layer DNNs is significant (Tukey corrected p-value < 0.01), and in the 100-component ICA group the difference between 2-layer and 1-layer DNNs (Tukey corrected p-value $= 0.0366$) and the difference between 3-layer and 1-layer DNNs (Tukey corrected p-value < 0.01) are also significant. Second, another two-way ANOVA test using an overdispersed binomial model was carried out to compare the 1-hidden-layer DNN with 20 neurons against the linear SVM across all numbers of ICA components. Type II ANOVA was used

on the 2 factors of the ICA component number and the prediction method (1-hidden-layer 2-neuron DNN vs. linear SVM). Both factors and the interaction are significant (p-values << 0.001). To investigate the significance of the difference between DNN and linear SVM at each FC scale, uncorrected post hoc testing p-values were reported in Table 5.2. With the conservative Bonferroni correction, the difference for the 100-component ICA is insignificant at the 0.05 significance level, and difference for the 200-component ICA is also insignificant at the 0.01 significance level. The mean difference and the p-value in the table also show that, for small number of ICA components (large scale FC) DNN is more accurate than linear SVM. As the number of ICA components increases, the accuracy of linear SVM catches up and finally surpasses DNN. This means different models can have different advantages at different FC scales.

Table 5.2. Multiple comparisons between DNN and linear SVM at each FC scale.

ICA Components	Prediction Accuracy Mean Difference (DNN-SVM)	P-value (uncorrected)
25	1.456E-02	5.214E-13
50	7.298E-03	2.447E-05
100	-3.490E-03	1.157E-02
200	-3.829E-03	2.215E-03
300	-5.184E-03	1.127E-05

5.3.2 DNN Feature Extraction and Robustness Evaluation

The proposed feature ranking method is verified by the comparison of the cross entropy loss obtained from each feature pair in predicting the training data. The male and female features in the last hidden layer of each DNN were extracted and ranked by their contributions to the difference between the final male and female scores. Then, predictions were made with the 1^{st}, 2^{nd}, 3^{rd}, 4^{th} and 5^{th} most highly ranked male-female feature pair in each DNN respectively on the training data

themselves. The cross entropy loss achieved by each of these high-level feature pairs for each DNN structure and each number of ICA components in all the 50 randomly permuted cross validations the are shown in Fig. 5.3. In all the situations, as the rank of the feature pair goes higher, the lower cross entropy loss it can get when predicting the training data. This means the more highly ranked feature is better at making the predicted label distribution similar to the ground truth distribution and, therefore, the importance and contribution of the features ranked by the proposed method in the DNN model are verified.

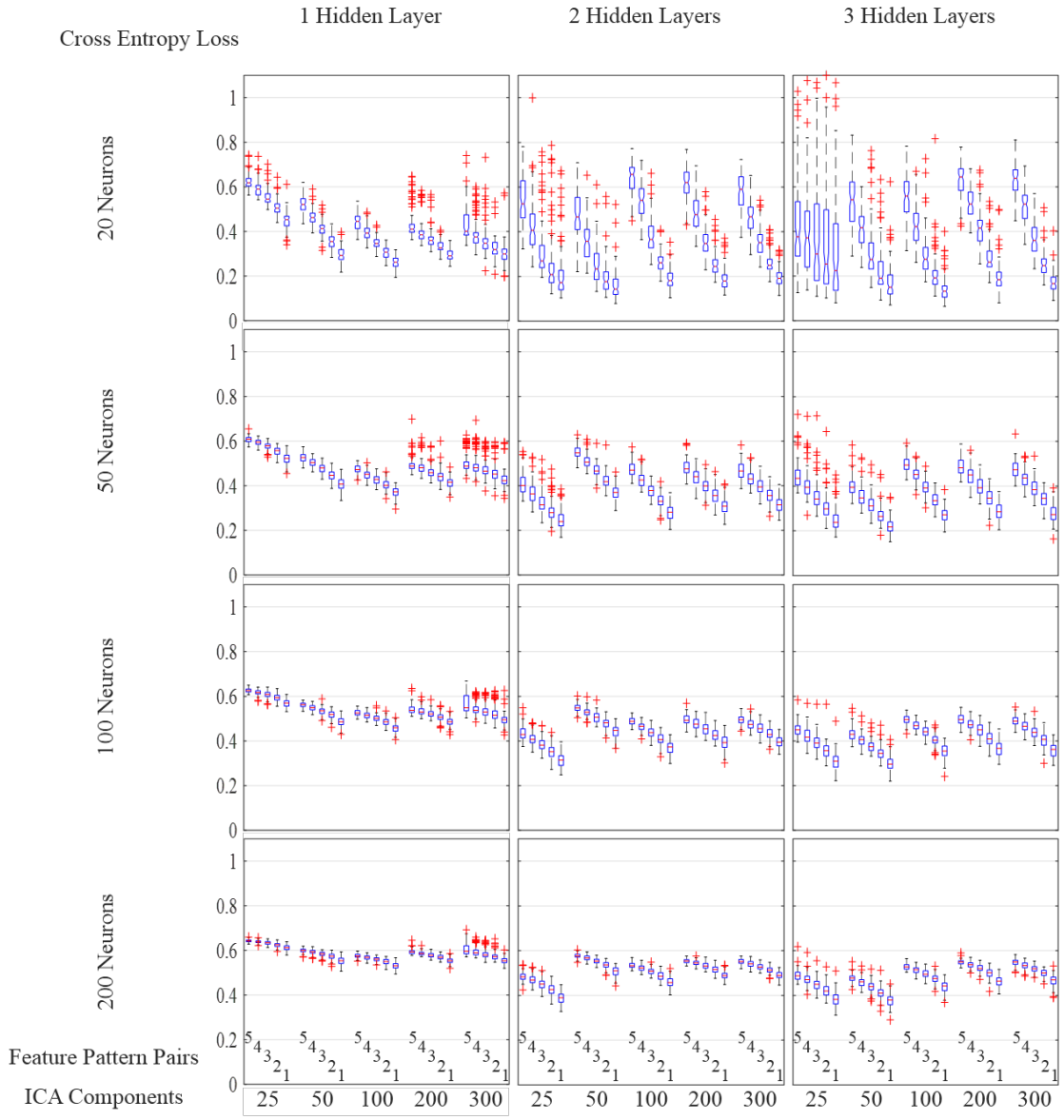

Fig. 5.3. The cross entropy loss achieved by a single high-level male and female feature pair of different importance on the training dataset. '1', '2', '3', '4', '5' are the importance levels of the feature pairs, which mean the most, 2nd, 3rd, 4th and 5th highly ranked feature pair in the last hidden layer. In this figure and all the figures having boxplots, the points with red '+' symbols are drawn as outliers, if they are greater than q3+1.5(q3-q1) or less than q1-1.5(q3-q1), where q1 and q3 are the first and third quartiles respectively.

The relationship between the prediction accuracy and number of learned features used in the prediction was also studied, and the results are shown in Fig. 5.4. For each kind of DNN structure and number of ICA components, only a subset of the most important male and female features in the last hidden layer learned from the training dataset were kept in the DNN model for the prediction on the testing dataset. The most important 1, 2, 5 and 10 male and female feature pairs, which have the largest weight differences towards the final male and female neurons were used for each prediction in all the 50 randomly permuted cross validations, and the resulting accuracies were compared with those from the predictions with all the feature pairs. As is shown in Fig. 5.4, in general for all the DNN structures and numbers of ICA components, as the number of important feature pairs used in the prediction increases accumulatively, the more accuracy the DNN can recover compared to the prediction made by all the features. This is because all the features in a certain DNN are trained to work together to reach the highest prediction accuracy (lowest loss), and the more complete the DNN model is, the more accuracy can be preserved. It is shown that in Fig. 5.4 several predictions with fewer features have slightly higher accuracy than the predictions made by more features. This could be caused by the slight inconsistency of the internal features between the training and testing datasets. In the Appendix B Fig. B1 and B2, predictions were also made with the same subsets of learned features on the training dataset. As the number of features used increases, the prediction accuracy increases monotonically and the prediction loss decreases monotonically. Fig. 5.4 also shows that in the task of predicting gender from brain FC, the most important feature pair can usually recover the majority of the prediction accuracy achieved by all

the features. (For example, for the 1-hidden-layer 20-neuron DNN at the scale of 25 ICA components, a median accuracy of 0.795 is achieved by using only the single most important feature pair vs. a median accuracy of 0.827 when using all the features.)

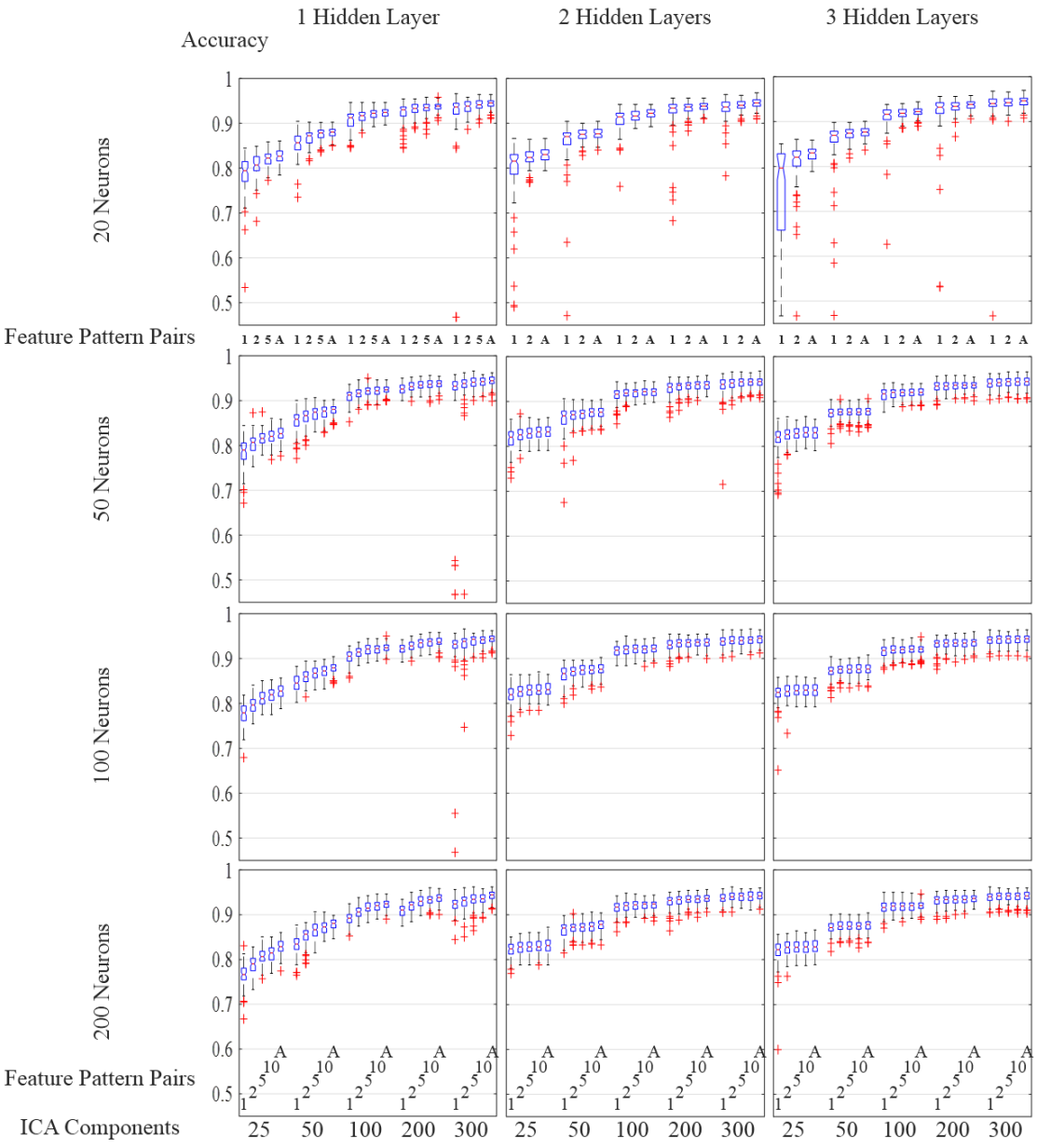

Fig. 5.4. Prediction accuracy recovered by the several most important high-level male and female feature pairs in the predicitons on the testing dataset. '1', '2', '5' and '10' mean that the predictions were made by the most important 1, 2, 5, 10 male and female feature pairs in the last hidden layer respectively. 'A' means the predictions were made by all the features.

To study the robustness and repeatability of the most important high-level male and female features extracted by different neural network structures for each FC scale, the correlations between the features across all the 50 randomly permuted cross validations were calculated and are shown in Fig. 5.5. As the neural network gets deeper, the correlation becomes higher, which means that the repeatability of the most important high-level feature is higher. This is true for all the FC scales and numbers of network neurons, although for the 25-ICA-component FC, the differences in correlations from a 2-hidden-layer networks compared to a 3-hidden layer networks are minimal. In addition, as the number of ICA components increases, the correlation generally decreases.

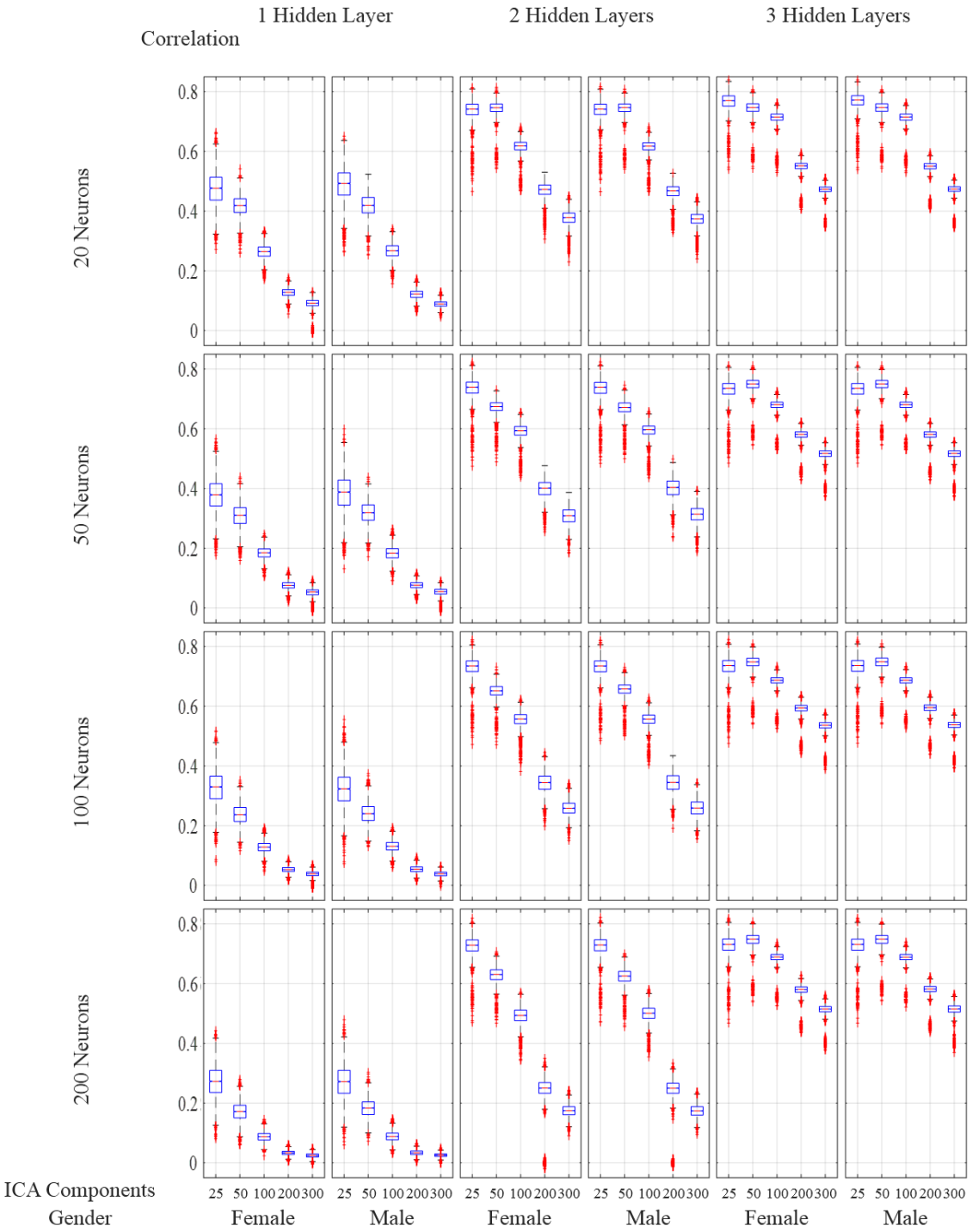

Fig. 5.5. Correlations of the most important high-level features across all the 50 randomly permuted cross validations for each neural network structure, number of ICA component and gender. For each number of neuron, hidden layer and ICA component, male and female features in the highest level hidden layer were extracted and ranked with the proposed method. The correlations of the highliest ranked features were calculated across the 50 randomly permuted 2-fold cross valadtions (each boxplot shows the correlations between 100 extracted features).

Considering both the accuracy that can be preserved by the most important feature pairs and the repeatability of these feature pairs in the randomly permuted cross validations, the most important high-level brain FC feature pairs extracted by the 200-neuron 2-hidden-layer DNN from the 25-component ICA connectivity matrices and those by the 200-neuron 3-hidden-layer DNN from the 50-component ICA connectivity matrices are shown in Fig. 5.6. First, from Fig. 5.6 it can be seen that the group means of the male and female connectivity matrices are similar, and the connectivity patterns of the group means are basically consistent for different number of ICA components. Second, the extracted high-level features by DNN of different genders basically have reversed patterns. This is because DNN is trained to discriminate these 2 genders by looking for the relative pattern difference in their connectivity inputs. This is true regardless of the number of ICA components. Third, the difference of the extracted most important high-level features between genders captures some of the characteristic patterns in input group mean difference, but does not exactly match the input group mean difference. The feature extracted by DNN is trained to maximize the gender prediction accuracy by searching for certain FC patterns in every input, but the group mean difference cannot reflect the variance or subject-level fingerprint patterns across all the subjects.

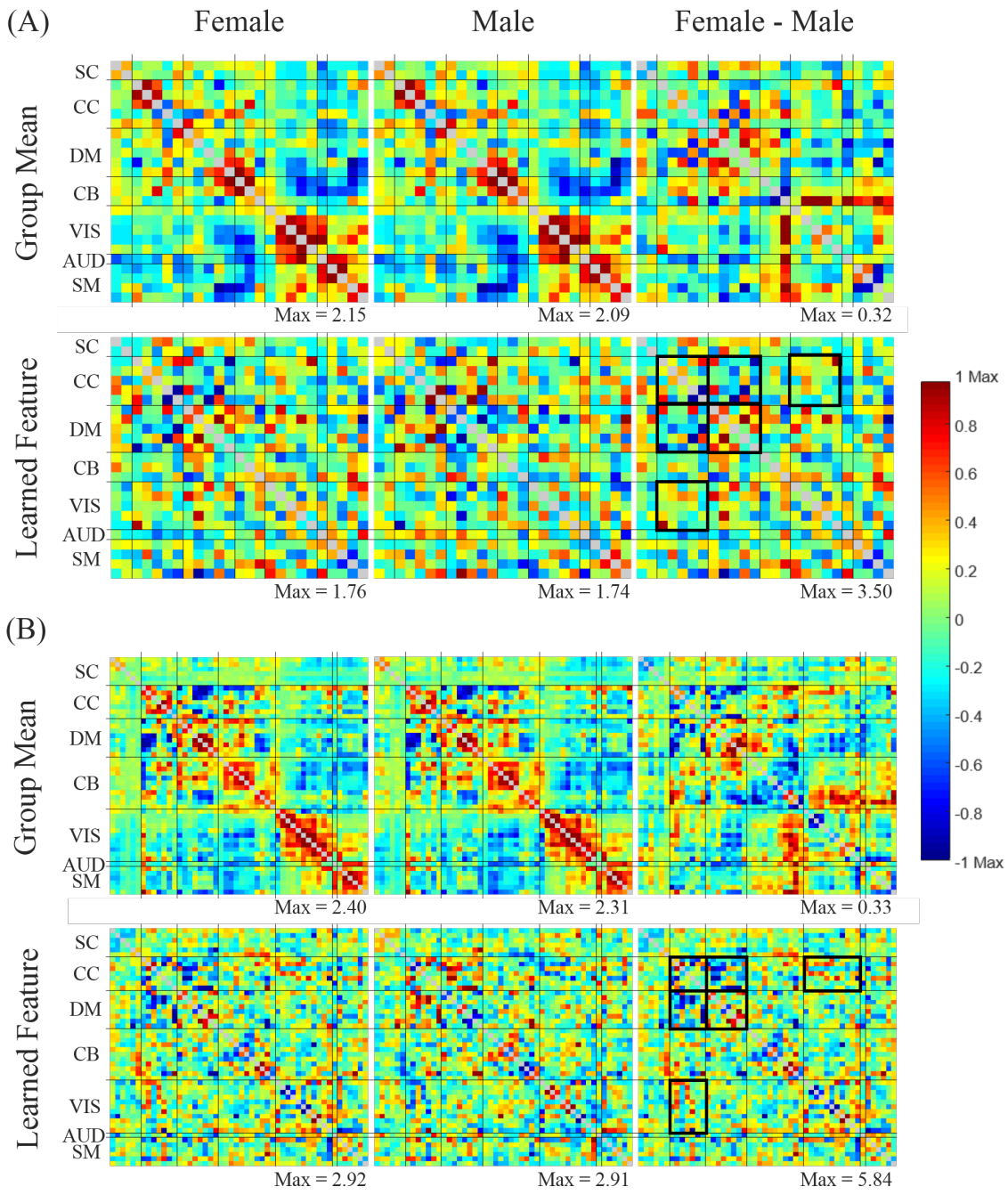

Fig. 5.6. The most important high-level brain FC feature pairs extracted by DNN. (A) DNN: 200-neuron 2-hidden-layer network; input: 25-component ICA connectivity matrices. (B) DNN: 200-neuron 3-hidden-layer network; input: 50-component ICA connectivity matrices. Each group mean of the input is also shown for comparison. SC, subcortical; CC, cognitive control; DM, default-mode; CB, cerebellar; VIS, visual; AUD, auditory; SM, somatomotor.

In Fig. 5.6 the most important male and female features extracted by DNN are the most highly ranked high-level FC patterns DNN looks for to make accuracy predictions. For both 25 and 50

ICA components, females have stronger DM-DM, CC-VIS connections, relatively weaker CC-DM connections, and both stronger and weaker connections in CC-CC (in the black boxes in Fig. 5.6). These findings are consistent with the previous studies (S. M. Smith et al., 2013; Zhang et al., 2018). In Fig. 5.6 the most important female-male feature differences learned by DNN also show some different patterns for these 2 different FC input scales, for example: the 50-ICA-component feature difference shows stronger and weaker female connections in CB-CB, VIS-VIS, CB-CC, VIS-CC and SM-VIS, whereas the 25-ICA-component feature difference shows stronger female connections in VIS-CC, VIS-DM and a weaker female connection in SM-CC.

5.3.3 Monte Carlo Dropout Testing and Bayesian Deep Learning

The prediction accuracies of Bayesian deep learning with MC dropout testing at the dropout rates: 0.2, 0.5, 0.9 and the extreme dropout rate R2, which only retains 2 neurons, were studied and compared with the prediction accuracy of weight averaging testing. The results in Fig. 5.7 show that in general the accuracies achieved by MC dropout testing at low dropout rates (0.2, 0.5) are comparable with the accuracy of weight averaging testing on gender classification. As the dropout rate goes up to extreme large values (0.9, R2), the accuracy of MC dropout testing tends to decrease. The decreasing is especially obvious for the Bayesian DNN with large number of neurons at the dropout rate R2.

Further, the behavior of the uncertainty generated by Bayesian deep learning was also studied by performing MC dropout on the real and made up subsets in dropout testing. Fig. 5.8 shows that for the 3-hidden-layer Bayesian DNNs trained with the real female and male data, as the uncertainty within the testing data increases, the prediction accuracy decreases and the corresponding

uncertainty increases. This is true for all the numbers of neurons and all the numbers of ICA components.

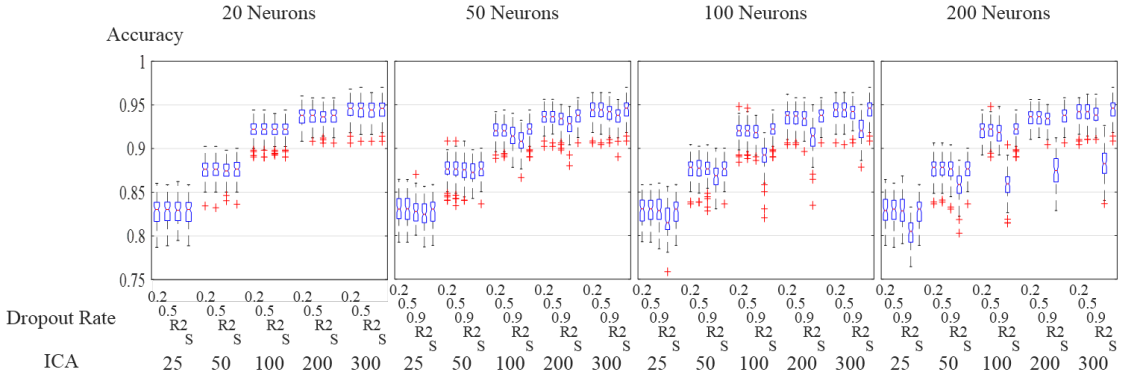

Fig. 5.7. Prediction accuracy VS dropout rate of MC dropout testing in 3-hidden-layer Bayesian DNNs with the 3rd hidden layer dropped out in all the 50 randomly permuted cross validations for each number of neurons and each number of ICA components. R2 means the dropout rate was set to only retaining 2 neurons; S stands for the result by weight averaging technique during testing. The networks used are previously trained 3-hidden-layer DNNs with the dropout rate of 0.2 on the 3rd hidden layers in dropout training.

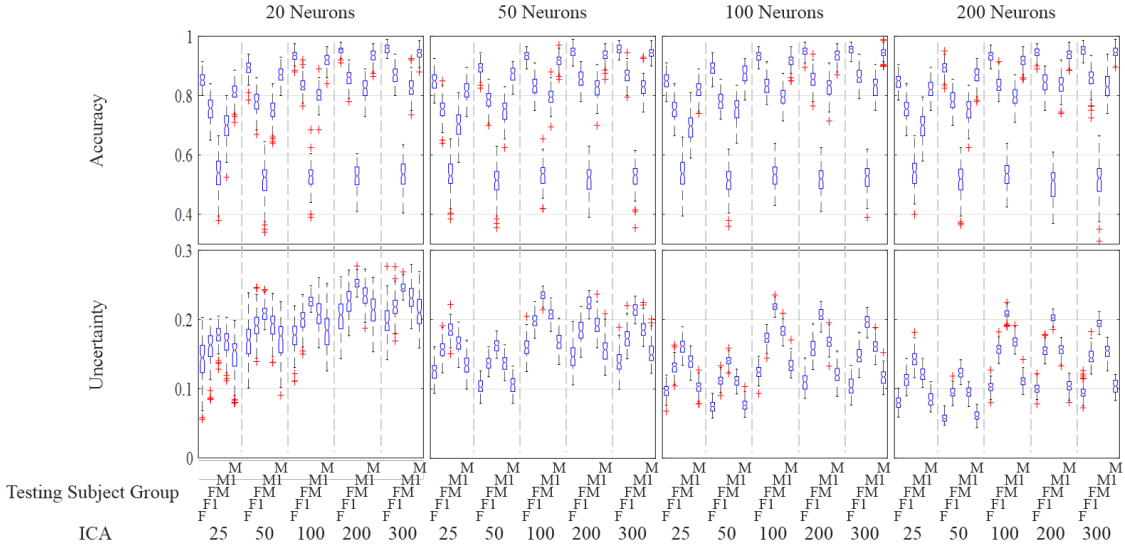

Fig. 5.8. Prediction accuracy and model uncertainty of Bayesian DNN with MC dropout testing on different testing subsets in all the 50 randomly permuted cross validations for each number of neurons and each number of ICA components. F – real female subsets; F1 – 0.75-female-0.25-male subsets; FM – 0.5-female-0.5-male subsets; M1 – 0.25-female-0.75-male subsets; M – real male subsets. For the purpose of calculation, the labels of the FM subsets were set as female. Previously trained 3-hidden-layer DNNs were used for this experiment, and in MC dropout testing 0.5 dropout rate was set on the 3rd hidden layers.

5.4 Discussion

In this present study, a novel deep learning based feature extraction and ranking method was developed and applied to the task of gender prediction and related feature extraction from resting state brain functional connectivity. First, the classification performance of DNN was evaluated with different network depths and different number of neurons at different scales of brain FC. The prediction accuracy of DNN was also compared with that of SVMs at different brain FC scales. Second, the proposed feature extraction and ranking method gives a better understanding of DNN's prediction mechanism, and is able to extract the highly important FC features for the prediction to build the relationship between brain FC and genders. The robustness and repeatability of the features from different network structures and scales of brain FC were also studied. Third, Bayesian deep learning was applied to the brain FC gender prediction. The prediction accuracy with different dropout rates in MC dropout testing and the behavior of the uncertainty generated by Bayesian DNN were also studied at different brain FC scales. In summary, the experiments on the 1003-subject HCP dataset suggest that the results are highly related to the scale of brain FC under investigation.

Since gender is one of the very basic physiological and psychological properties of human beings (Gong et al., 2011), gender prediction from brain FC on the large-scale high-quality HCP data can serve as a very basic case to verify and study the performance and characteristics of DNN for further neuroscience applications. From the prediction results in Fig. 5.2 and the following statistical analysis, it can be seen that the results from DNNs with different depths and different number of neurons are very similar. In the comparison between DNN and linear SVM, the results show that DNN is much more accurate than linear SVM when the number of ICA components is small. As the number of ICA components increases the accuracy of linear SVM catches up. This

illustrates that the nonlinear DNN and the linear SVM have different advantages at different FC scales. The possible reason for this is the change of the amount of signal and noise in the FC input. When the number of ICA components is small, the signal and noise in the FC is highly reduced, and in this situation the nonlinear DNN model is better at taking the limited signal for accurate prediction. When the number of ICA components is large, the signal is highly redundant but the noise is also greater, and in this case the linear model has more advantages. This trend about the linear and nonlinear methods can also be verified through the comparison between the linear SVM and the nonlinear (order 2 polynomial) SVM. Therefore, this study shows that no model is the best for every situation, and that for predictions using brain FC as input the scale of brain FC should be taken into consideration. In the comparison between the ICA based FC and the template based FC, the ICA based FC can reach a better result with a smaller number of components, which means in terms of increment of the signal noise ratio ICA is more efficient than the template based parcellation.

In the proposed feature ranking method, the importance of the extracted high-level features were ranked by their contributions to the difference between the male score and the female score in the softmax layer. Since the cross entropy loss in a prediction reflects how the predicted probability distribution over the classes is similar to the ground truth probability distribution, experiments were performed to compare the cross entropy loss achieved by each high-level feature pair on the training data. Fig. 5.3 shows that a more highly ranked high-level feature pair can reach a lower cross entropy loss for various DNN structures and inputs, which proves that the importance and contribution of the more highly ranked feature pair is higher. This result verifies the effectiveness and robustness of the proposed DNN feature extraction and ranking method.

Fig. 5.4 shows that as the number of the important high-level features involved in the DNN model increases accumulatively according to their ranks, the prediction accuracy that can be recovered towards the full DNN model also increases. For the application of FC gender prediction, the most important feature pair in various network structures can recover the majority of the accuracy achieved by the full DNN model for all scales of FC input. This is especially true for the networks with more layers and more neurons, which means the redundancy levels of the high-level features in these networks are relatively high. Thus, the ranking method offers us a way to focus on the most important feature pairs in the application of FC gender prediction instead of checking all the hundreds of features learned by DNN. Also, Fig. 5.4 shows with more lower-level features the most important high-level feature pair in a deep network can recover the accuracy of the whole DNN model better. For the 2-hidden-layer and 3-hidden-layer DNNs, there are large improvements of prediction accuracies recovered with the most important feature pair from 20-neuron DNNs to 50-neuron DNNs. However, the improvement is not as clear for further increasing the number of low-level features (from 50 to 100 and 200 neurons). This is likely because for a specific application, there is a certain number of useful lower-level features making the main contribution to the final prediction accuracy, and each high-level feature is a combination of all the low-level features. When the number of lower-level neurons is not enough, increasing it will let a high-level feature combine more useful low-level features to make more accurate predictions. After the certain number is reached, adding more low-level features only increases the redundancy of the network and does not result in much increase in the prediction accuracy recovered by a high-level feature.

The features extracted by DNN during gender classification can only be generalized as the features for brain FC gender classification if they are highly robust and repeatable to different subject

groups, otherwise they are more reflective of the features for a specific sample group rather than the features for the whole population. Thus, we investigated the repeatability of the most important feature pairs in all the randomly permuted cross validations for each kind of DNN structure and input FC scale. Fig. 5.5 shows that as the network gets deeper, the correlation of the most important high-level feature pair gets higher, which means the repeatability of it gets higher. This is because in DNN features of the input are learned hierarchically by hidden layers of different depths, and each high-level feature learned by a deeper hidden layer is constituted of many different low-level features (Lee et al., 2008, 2009; Kim et al., 2016). Since the number of classes to predict is relatively small, in real world applications the number of distinct low-level features needed are much more than that of distinct high level features to make accurate predictions. Thus, the most important high-level features across different training group are more likely to be similar. Also, as the number of ICA components increases, the repeatability of the most important high-level feature generally decreases. This could be because as the dimension of the input increases, more information is contained in the data, and more detailed differences among different training datasets can be described by the high-dimensional input. This result also shows a potential downside to using higher numbers of ICA components in classification. That is, while higher numbers of ICA components result in improved classification accuracy, the repeatability of the features that drive this classification is lower. Conversely, using a lower number of ICA components for classification results in slightly lower classification accuracy, but the patterns of connectivity that drive the classification are more consistent and thus more meaningful for the subsequent interpretation of group differences.

Although for the application of FC gender prediction DNNs of different depths can achieve similar accuracies in Fig. 5.2, Fig. 5.4 and 5.5 show that deeper network structures still have advantages

on the accuracy recovery and repeatability of the most important feature pair. Based on the those results in Fig. 5.4 and 5.5, the most important male and female high-level features from two selected DNNs and brain FC scales are shown in Fig 5.6. First, it can be seen that the characteristic patterns presented by the most important female-male feature differences in both FC scales have similar parts. Both of them captured the relatively high DM-DM and CC-VIS connections, relatively low CC-DM connections and both high and low connections in CC-CC as the female FC features, which means these are very important features for the FC gender classification. These patterns are consistent with the results from previous studies (S. M. Smith et al., 2013; Zhang et al., 2018). Also, the most important female-male feature differences are not all the same across these 2 FC scales, and there are several possible reasons: the input FC group mean patterns are not all the same across these 2 FC scales; different parts of FC can have different importance in the predictions at different FC scales. To achieve the highest prediction accuracy, all the learned features should be combined together to make the decision, since all the features are trained to work together to achieve the best performance. In addition, in Fig. 5.6 the female-male feature differences extracted by DNN capture some of the patterns in the female-male group mean differences, but not all the large group mean differences. This is because the features extracted by the DNN are the highly important ones for the model to make accurate classification, but not all the large group mean differences are important for the classification, since no standard deviation information is reflected in the group mean differences.

Bayesian DNN can not only make predictions but also report the model uncertainty on each prediction. The prediction accuracies of Bayesian DNN at different dropout rates in Fig. 5.7 show that in the application of FC gender prediction Bayesian DNN with MC dropout testing can reach the same level of prediction accuracy as the conventional DNN with weight averaging testing as

long as the dropout rate is not too low. Thus, 0.5 dropout rate was selected based on the tradeoff between accuracy and variation to test the behavior of the uncertainty generated. As is seen in Fig. 5.8, the model uncertainty increases as the uncertainty within the testing data increases, which matches the expectation. Since the uncertainty generated by Bayesian DNN can help with picking up the testing subjects whose predictions the model is not very confident about based on the knowledge learned from the training data. In real neuroimaging and neuroscience predication practices, such as classifying the Alzheimer's disease, locating the source of epilepsy, or predicting the effect of a treatment, we can be aware of the uncertainty level of each prediction, and start other procedures to consolidate the highly uncertain prediction results.

There are also some limitations in our study. Based on the results from the previous studies (S. M. Smith et al., 2013; Zhang et al., 2018), we did not regress out the potential confounds from the HCP brain FC data when classifying gender from brain FC. (Zhang et al., 2018) showed in their study that regressing out the confounds, such as brain volume, in gender prediction with partial least squares regression did not affect the predictive performance much, and the important FC connections involved in the gender prediction and the brain volume prediction were generally different. (S. M. Smith et al., 2013) also discussed in their study that since males and females all have stronger and weaker connections in the important FC features involved in the gender prediction, which is also seen in our study, the gender classification is unlikely to be driven primarily by the gross group differences, such as the difference in brain volume. In our study each input connectivity matrix was also normalized to zero mean and unit variance with no gross asymmetry between the male and female groups.

5.5 Conclusion

The present study successfully verified the feasibility of using DNN to predict gender from the resting state brain FC. The predictions under different scales of brain FC illustrate that the performance of DNN is highly related to the scale of brain FC, and the comparison between the nonlinear DNN model and the linear SVM model suggests that to achieve the best prediction result during model selection, the scale of the input brain FC should be taken into consideration. The proposed DNN feature extraction and ranking method provides a new understanding of the DNN model and manages to build the relationship between the input brain FC and the output gender. Bayesian DNN was also applied for the first time to a neuroscience classification problem, and showed how well the model uncertainty generated reacts to the uncertainty within the data. We believe that based on the success of the DNN and Bayesian DNN models in the FC gender prediction on the large-scale HCP data, these models can be further applied to other fields of neuroscience.

Chapter 6

Conclusion and Future Works

This work developed deep learning and Bayesian deep learning based models for MR neuroimaging. The original purpose of this work is to make the whole MR neuroimaging pipeline more accurate, robust and efficient. With the facilitation of deep learning models, the prediction accuracy of various applications in neuroimaging was improved compared with the conventional methods. Bayesian deep learning models further give deep learning models the ability of generating model uncertainty for the predictions, which makes these procedures more robust to meet the strict requirement of medical imaging applications. Last but not the least, deep learning and Bayesian deep learning based models make the processing pipeline in MR neuroimaging fully automated. Together with the help of the advance technology in parallel computing and GPU, these models can immensely improve the efficiency in medical imaging.

Specifically, the development of the deep learning and Bayesian deep learning models in this work focus on 3 computer vision applications in the field of MR neuroimaging: image segmentation, image synthesis, and feature extraction.

First, this work presented a novel tool combining Bayesian CNN and fully connected 3D CRF to perform brain extraction on nonhuman primates. Bayesian CNN's prediction accuracy is significantly higher than the traditional gradient-based and registration-based brain extraction methods, and the conventional CNN models. The refinement of the results by fully connected 3D CRF further improved its accuracy. Another important additional value offered by Bayesian CNN is its ability to generate model uncertainty for each prediction, which can give us a clue whether the predicted result can be trusted or not, and it is extremely important for the applications in medical imaging. We studied the behavior of the uncertainty generated in various situations, and it is always able to reflect the inconsistency within the training data or between the training data and the testing data. Base on this model's success in nonhuman primate brain extraction, we believe its application can be expanded to other image segmentation applications in medical imaging.

Second, we proposed a novel image synthesis model – Bayesian conditional GAN with concrete dropout and model recalibration, and applied it on the challenging task of brain tumor image synthesis. The verification of the proposed method through synthesizing T2w MR brain tumor images from the corresponding T1w images proves that Bayesian conditional GAN is an accurate and consistent approach for the task of image synthesis. With concrete dropout, the gradient-tuned dropout probability is enabled, and it is much more efficient than the hand-tuned dropout rate in Monte Carlo dropout. Concrete dropout also results in higher predication accuracy and better calibrated uncertainty. The involvement of the model recalibration approach further improves the calibration of the uncertainty generated by Bayesian deep learning. The relationship between the prediction accuracy and the generated model uncertainty was also studied. Although by definition it is not necessary that they are directly proportional to each other, with the increase of the number

of observations and the dimension reduction calculation, a stronger correlation was found between them.

Third, deep neural network models were investigated to study the functional connectivity gender difference. The performance of DNN and SVM were compared on the brain functional connectivity gender prediction with brain connectivity as input at various levels. DNN is more accurate than SVM when the number of ICA components is small, and SVM's accuracy catches up as the number of ICA components increases. In the research field of neuroscience, the model's ability to extract useful features that make accurate predictions is as important as making accurate predictions itself. Thus, we proposed a feature extraction and ranking method for DNN, which can extract and rank the connectivity patterns based on their contributions to the final prediction. The contributions of the features ranked at different levels were validated through the cross entropy loss achieved by them. The robustness of the features extracted by different DNN structures at different connectivity levels was also studied. Dropout testing was also applied on gender prediction to generate model uncertainties for predictions. The behavior of the model uncertainty also matches our expectation.

6.1 Future Work

Although this work includes many necessary parts towards the original goal of making the whole processing pipeline of neuroimaging more accurate, robust and efficient. There are still many more topics that could be investigated to expand this work and strengthen the conclusions in the applications of image segmentation, synthesis and classification in the field of medical imaging. The following sections will discuss the potential deep learning research directions that can further improve MR neuroimaging.

6.1.1 Image Segmentation

Currently, deep learning based image segmentation methods highly depend on the characteristic of the data in the training set. For example, if the age group of the training dataset is slightly different from that of the testing dataset, the prediction error for the testing dataset will increase dramatically. Data augmentation and transfer learning are possible solutions to fix this problem to some extent. To effectively transfer a learned model to slightly different dataset will require investigation on efficient methods in data augmentation and transfer learning.

Different segmentation tasks also have different requirements on the features of deep learning models. Amygdala is a very important anatomic structure in the brain, and the change of its shape and volume is highly related to many brain disorders. The automatic segmentation of amygdala is very challenging, since it is a very small structure. Limited by the GPU memory CNN can hardly have a large receptive field as well as keep a high resolution at the same time. Therefore, a duo-pathway CNN may be the solution for the segmentation of ultra-small structures. One pathway is aimed at keeping the original resolution with smaller receptive field, and another pathway is aimed at enlarging the receptive field to get more contexts for the target with compromised resolution.

There are also many sources of uncertainty in the final prediction result. The Bayesian deep learning model we used only formulates the model uncertainty. Noise in the data may also cause errors and uncertainty in the prediction (Kendall and Gal, 2017). To have a more thorough uncertainty model will also help to make the uncertainty generated more accurate and have stronger correlation with the prediction accuracy.

Currently, due to the limitation on GPU memory, there is no way for deep learning based models to hold the full 3D brain volumes during training. Thus, the images are usually processed in slices or little cubes, and then refined in the full 3D space with algorithms having lighter memory

requirement. Either the slice based or cube based deep learning models will lose information in the full 3D. Therefore, Duo-pathway or multi-pathway deep learning model can be used to take advantage of the information in each dividing way.

6.1.2 Image synthesis

Image synthesis is a very important technique in the field of medical imaging. With it inter-modality image translation, image denoising, artifact removing, super resolution and even sparse reconstruction can all be achieved. We only investigated the image synthesis from T1w to T2w brain MR image. The model proposed in this thesis can be adjusted to all the applications mentioned above. For example, CT images can also be synthesized from MR images. In this way CT images can be generated for MR-Linac, and it can be used for treatment planning in human oncology without the acquisition of real CT images and the registration between CT images and MR images. Using MR images to synthesize CT images can reduce the patient dose at the same time. CT images can also be synthesized from the PET or MR images acquired from a PET/MR scanner, and then the synthesized CT images can be used for the PET attenuation correction. For image reconstruction, with the help of deep learning, the data needed in the acquisition domain can be dramatically reduced, and thus the acquisition can be fastened. In this way, dynamic imaging will become possible.

In addition to the frame work of Bayesian deep learning with concrete dropout other models, including ensemble methods (Lakshminarayanan et al., 2016), can also generate uncertainty for each prediction. A comparison between Bayesian deep learning based methods, other uncertainty generation models and the conventional models will give us a clearer understanding about their performances and the pros and cons for each method. Moreover, in this work, we only validated the proposed method on the image synthesis of brain tumor data, experiments on traumatic brain

data, brain data with brain disorders and healthy brain data will also give us a better overall understanding about the performance of the proposed method.

6.1.3 Classification

In this work, the DNN was compared with SVM on brain connectivity gender classification, and both models have advantages in different situations. Applying DNN based models to classify brain disorders versus healthy controls with the brain connectivity from rs-fMRI may let us find more key applications for this model, as the reported findings in the classification of schizophrenia (Kim et al., 2016). For the proposed feature extraction and ranking method, it can help with extracting the important connectivity patterns the model looks for in each subject's connectivity matrix for making correct predictions. Therefore, this method can also be used to extract important functional connectivity patterns in different age groups and in different species to help us better understand the related brain functional connectivity changes in each stage of the development of the brain and the brain functional connectivity differences across species.

Bibliography

Alaerts, K., Swinnen, S.P., Wenderoth, N., 2016. Sex differences in autism: a resting-state fMRI investigation of functional brain connectivity in males and females. Soc. Cogn. Affect. Neurosci. 11, 1002–1016. https://doi.org/10.1093/scan/nsw027

Badrinarayanan, V., Handa, A., Cipolla, R., 2015a. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling. ArXiv150507293 Cs.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2015b. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. ArXiv151100561 Cs.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2015c. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. ArXiv151100561 Cs.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data 4, 170117. https://doi.org/10.1038/sdata.2017.117

Baldwin, R.M., Zea-Ponce, Y., Zoghbi, sami S., Laurelle, M., Al-Tikriti, M.S., Sybirska, E.H., Malison, R.T., Neumeyer, J.L., Milius, R.A., Wang, S., Stabin, M., Smith, E.O., Charney, D.S., Hoffer, P.B., Innis, R.B., 1993. Evaluation of the monoamine uptake site ligand [131I]methyl 3β-(4-Iodophenyl)-tropane-2β-carboxylate ([123I]β-CIT) in non-human primates: Pharmacokinetics, biodistribution and SPECT brain imaging coregistered with MRI. Nucl. Med. Biol. 20, 597–606. https://doi.org/10.1016/0969-8051(93)90028-S

Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans. Med. Imaging 23, 137–152. https://doi.org/10.1109/TMI.2003.822821

Bell, A.J., Sejnowski, T.J., 1995. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Comput. 7, 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Bernal, J., Kushibar, K., Asfaw, D.S., Valverde, S., Oliver, A., Martí, R., Lladó, X., 2017. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. ArXiv171203747 Cs.

Bhatkoti, P., Paul, M., 2016. Early diagnosis of Alzheimer's disease: A multi-class deep learning framework with modified k-sparse autoencoder classification, in: 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ). Presented at the 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1–5. https://doi.org/10.1109/IVCNZ.2016.7804459

Bilker, W.B., Hansen, J.A., Brensinger, C.M., Richard, J., Gur, R.E., Gur, R.C., 2012. Development of Abbreviated Nine-Item Forms of the Raven's Standard Progressive Matrices Test. Assessment 19, 354–369. https://doi.org/10.1177/1073191112446655

Birn, R.M., Shackman, A.J., Oler, J.A., Williams, L.E., McFarlin, D.R., Rogers, G.M., Shelton, S.E., Alexander, A.L., Pine, D.S., Slattery, M.J., Davidson, R.J., Fox, A.S., Kalin, N.H., 2014. Evolutionarily conserved prefrontal-amygdalar dysfunction in early-life anxiety. Mol. Psychiatry 19, 915–922. https://doi.org/10.1038/mp.2014.46

Biswal, B.B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., Dogonowski, A.-M., Ernst, M., Fair, D., Hampson, M., Hoptman, M.J., Hyde, J.S., Kiviniemi, V.J., Kötter, R., Li, S.-J., Lin, C.-P., Lowe, M.J., Mackay, C., Madden, D.J., Madsen, K.H., Margulies, D.S., Mayberg, H.S., McMahon, K., Monk, C.S., Mostofsky, S.H., Nagel, B.J., Pekar, J.J., Peltier, S.J., Petersen, S.E., Riedl, V., Rombouts, S.A.R.B., Rypma, B., Schlaggar, B.L., Schmidt, S., Seidler, R.D., Siegle, G.J., Sorg, C., Teng, G.-J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.-C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.-F., Zhang, H.-Y., Castellanos, F.X., Milham, M.P., 2010. Toward discovery science of human brain function. Proc. Natl. Acad. Sci. 107, 4734–4739. https://doi.org/10.1073/pnas.0911855107

Bitar, R., Leung, G., Perng, R., Tadros, S., Moody, A.R., Sarrazin, J., McGregor, C., Christakis, M., Symons, S., Nelson, A., Roberts, T.P., 2006. MR Pulse Sequences: What Every Radiologist Wants to Know but Is Afraid to Ask. RadioGraphics 26, 513–537. https://doi.org/10.1148/rg.262055063

Bluhm, R.L., Osuch, E.A., Lanius, R.A., Boksman, K., Neufeld, R.W. j, Théberge, J., Williamson, P., 2008. Default mode network connectivity: effects of age, sex, and analytic approach. Neuroreport 19, 887–891. https://doi.org/10.1097/WNR.0b013e328300ebbf

Brosch, T., Yoo, Y., Tang, L.Y.W., Li, D.K.B., Traboulsee, A., Tam, R., 2015. Deep Convolutional Encoder Networks for Multiple Sclerosis Lesion Segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science. Presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, pp. 3–11. https://doi.org/10.1007/978-3-319-24574-4_1

Burgos, N., Cardoso, M.J., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Duncan, J.S., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S., 2013. Attenuation correction synthesis for hybrid PET-MR scanners. Med. Image Comput. Comput.-Assist. Interv. MICCAI Int. Conf. Med. Image Comput. Comput.-Assist. Interv. 16, 147–154.

Burgos, N., Cardoso, M.J., Thielemans, K., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Ahmed, R., Mahoney, C.J., Schott, J.M., Duncan, J.S., Atkinson, D., Arridge, S.R., Hutton, B.F., Ourselin, S., 2014. Attenuation correction synthesis for hybrid PET-MR scanners: application to brain studies. IEEE Trans. Med. Imaging 33, 2332–2341. https://doi.org/10.1109/TMI.2014.2340135

Butler, T., Imperato-McGinley, J., Pan, H., Voyer, D., Cunningham-Bussel, A.C., Chang, L., Zhu, Y.-S., Cordero, J.J., Stern, E., Silbersweig, D., 2007. Sex specificity of ventral anterior cingulate cortex suppression during a cognitive task. Hum. Brain Mapp. 28, 1206–1212. https://doi.org/10.1002/hbm.20340

Cahill, L., 2006. Why sex matters for neuroscience. Nat. Rev. Neurosci. 7, 477–484. https://doi.org/10.1038/nrn1909

Cardoso, M.J., Sudre, C.H., Modat, M., Ourselin, S., 2015. Template-Based Multimodal Joint Generative Model of Brain Data. Inf. Process. Med. Imaging Proc. Conf. 24, 17–29.

Casanova, R., Whitlow, C.T., Wagner, B., Espeland, M.A., Maldjian, J.A., 2012. Combining Graph and Machine Learning Methods to Analyze Differences in Functional Connectivity Across Sex. Open Neuroimaging J. 6. https://doi.org/10.2174/1874440001206010001

Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.-A., 2017. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage. https://doi.org/10.1016/j.neuroimage.2017.04.041

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. ArXiv160606650 Cs.

Coates, A., Huval, B., Wang, T., Wu, D., Catanzaro, B., Andrew, N., 2013. Deep learning with COTS HPC systems, in: International Conference on Machine Learning. pp. 1337–1345.

Cosgrove, K.P., Mazure, C.M., Staley, J.K., 2007. Evolving Knowledge of Sex Differences in Brain Structure, Function, and Chemistry. Biol. Psychiatry 62, 847–855. https://doi.org/10.1016/j.biopsych.2007.03.001

Craddock, R.C., Bellec, P., Jbabdi, S., 2017. Neuroimage special issue on brain segmentation and parcellation - Editorial. NeuroImage. https://doi.org/10.1016/j.neuroimage.2017.11.063

Currie, S., Hoggard, N., Craven, I.J., Hadjivassiliou, M., Wilkinson, I.D., 2013. Understanding MRI: basic MR physics for physicians. Postgrad. Med. J. 89, 209–223. https://doi.org/10.1136/postgradmedj-2012-131342

Deichmann, R., Good, C.D., Josephs, O., Ashburner, J., Turner, R., 2000. Optimization of 3-D MP-RAGE Sequences for Structural Brain Imaging. NeuroImage 12, 112–127. https://doi.org/10.1006/nimg.2000.0601

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Dolz, J., Desrosiers, C., Ben Ayed, I., 2017. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. NeuroImage. https://doi.org/10.1016/j.neuroimage.2017.04.039

Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A., 2015. The Pascal Visual Object Classes Challenge: A Retrospective. Int. J. Comput. Vis. 111, 98–136. https://doi.org/10.1007/s11263-014-0733-5

Feis, D.-L., Brodersen, K.H., von Cramon, D.Y., Luders, E., Tittgemeyer, M., 2013. Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. NeuroImage 70, 250–257. https://doi.org/10.1016/j.neuroimage.2012.12.068

Fennema-Notestine, C., Ozyurt, I.B., Clark, C.P., Morris, S., Bischoff-Grethe, A., Bondi, M.W., Jernigan, T.L., Fischl, B., Segonne, F., Shattuck, D.W., Leahy, R.M., Rex, D.E., Toga, A.W., Zou, K.H., BIRN, M., Brown, G.G., 2006. Quantitative Evaluation of Automated Skull-Stripping Methods Applied to Contemporary and Legacy Images: Effects of Diagnosis, Bias Correction, and Slice Location. Hum. Brain Mapp. 27, 99–113. https://doi.org/10.1002/hbm.20161

Fidon, L., Li, W., Garcia-Peraza-Herrera, L.C., Ekanayake, J., Kitchen, N., Ourselin, S., Vercauteren, T., 2018. Generalised Wasserstein Dice Score for Imbalanced Multi-class Segmentation using Holistic Convolutional Networks. ArXiv170700478 Cs 10670, 64–76. https://doi.org/10.1007/978-3-319-75238-9_6

Filippi, M., Valsasina, P., Misci, P., Falini, A., Comi, G., Rocca, M.A., 2013. The organization of intrinsic brain activity differs between genders: A resting-state fMRI study in a large cohort of young healthy subjects. Hum. Brain Mapp. 34, 1330–1343. https://doi.org/10.1002/hbm.21514

Fox, A.S., Kalin, N.H., 2014. A translational neuroscience approach to understanding the development of social anxiety disorder and its pathophysiology. Am. J. Psychiatry 171, 1162–1173. https://doi.org/10.1176/appi.ajp.2014.14040449

Fox, A.S., Oler, J.A., Shackman, A.J., Shelton, S.E., Raveendran, M., McKay, D.R., Converse, A.K., Alexander, A., Davidson, R.J., Blangero, J., Rogers, J., Kalin, N.H., 2015a. Intergenerational neural mediators of early-life anxious temperament. Proc. Natl. Acad. Sci. U. S. A. 112, 9118–9122. https://doi.org/10.1073/pnas.1508593112

Fox, A.S., Oler, J.A., Shackman, A.J., Shelton, S.E., Raveendran, M., McKay, D.R., Converse, A.K., Alexander, A., Davidson, R.J., Blangero, J., Rogers, J., Kalin, N.H., 2015b. Intergenerational neural mediators of early-life anxious temperament. Proc. Natl. Acad. Sci. U. S. A. 112, 9118–9122. https://doi.org/10.1073/pnas.1508593112

Gal, Y., 2016. Uncertainty in Deep Learning. University of Cambridge.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in: PMLR. Presented at the International Conference on Machine Learning, pp. 1050–1059.

Gal, Y., Ghahramani, Z., 2015. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. ArXiv150602158 Cs Stat.

Gal, Y., Hron, J., Kendall, A., 2017. Concrete Dropout. ArXiv170507832 Stat.

Geremia, E., Clatz, O., Menze, B.H., Konukoglu, E., Criminisi, A., Ayache, N., 2011. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. NeuroImage 57, 378–390.

Gerig, G., Jomier, M., Chakos, M., 2001. Valmet: A New Validation Tool for Assessing and Improving 3D Object Segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001. Presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Berlin, Heidelberg, pp. 516–523. https://doi.org/10.1007/3-540-45468-3_62

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., Smith, S.M., Essen, D.C.V., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178. https://doi.org/10.1038/nature18933

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Van Essen, D.C., Jenkinson, M., 2013. The Minimal Preprocessing Pipelines for the Human Connectome Project. NeuroImage 80, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Gong, G., He, Y., Evans, A.C., 2011. Brain Connectivity: Gender Makes a Difference. The Neuroscientist 17, 575–591. https://doi.org/10.1177/1073858410386492

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680.

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C.F., Auerbach, E.J., Douaud, G., Sexton, C.E., Zsoldos, E., Ebmeier, K.P., Filippini, N., Mackay, C.E., Moeller, S., Xu, J., Yacoub, E., Baselli, G., Ugurbil, K., Miller, K.L., Smith, S.M., 2014. ICA-based artefact and accelerated fMRI acquisition for improved Resting State Network imaging. NeuroImage 95, 232–247. https://doi.org/10.1016/j.neuroimage.2014.03.034

Guerreiro, F., Burgos, N., Dunlop, A., Wong, K., Petkar, I., Nutting, C., Harrington, K., Bhide, S., Newbold, K., Dearnaley, D., deSouza, N.M., Morgan, V.A., McClelland, J., Nill, S., Cardoso, M.J., Ourselin, S., Oelfke, U., Knopf, A.C., 2017. Evaluation of a multi-atlas CT synthesis approach for MRI-only radiotherapy treatment planning. Phys. Med. 35, 7–17. https://doi.org/10.1016/j.ejmp.2017.02.017

Hajnal, J.V., Bryant, D.J., Kasuboski, L., Pattany, P.M., De Coene, B., Lewis, P.D., Pennock, J.M., Oatridge, A., Young, I.R., Bydder, G.M., 1992. Use of fluid attenuated inversion recovery (FLAIR) pulse sequences in MRI of the brain. J. Comput. Assist. Tomogr. 16, 841–844.

Hazlett, H.C., Gu, H., Munsell, B.C., Kim, S.H., Styner, M., Wolff, J.J., Elison, J.T., Swanson, M.R., Zhu, H., Botteron, K.N., Collins, D.L., Constantino, J.N., Dager, S.R., Estes, A.M., Evans, A.C., Fonov, V.S., Gerig, G., Kostopoulos, P., McKinstry, R.C., Pandey, J., Paterson, S., Pruett, J.R., Schultz, R.T., Shaw, D.W., Zwaigenbaum, L., Piven, J., IBIS Network, Clinical Sites, Data Coordinating Center, Image Processing Core, Statistical Analysis, 2017. Early brain development in infants at high risk for autism spectrum disorder. Nature 542, 348–351. https://doi.org/10.1038/nature21369

He, X., Zemel, R.S., Carreira-Perpinan, M.A., 2004. Multiscale conditional random fields for image labeling, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. Presented at the Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., pp. II-695-II-702 Vol.2. https://doi.org/10.1109/CVPR.2004.1315232

Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H., 2001. Image Analogies, in: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '01. ACM, New York, NY, USA, pp. 327–340. https://doi.org/10.1145/383259.383295

Hsu, J.-L., Leemans, A., Bai, C.-H., Lee, C.-H., Tsai, Y.-F., Chiu, H.-C., Chen, W.-H., 2008. Gender differences and age-related white matter changes of the human brain: A diffusion tensor imaging study. NeuroImage 39, 566–577. https://doi.org/10.1016/j.neuroimage.2007.09.017

Hu, C., Ju, R., Shen, Y., Zhou, P., Li, Q., 2016. Clinical decision support for Alzheimer's disease based on deep learning and brain network, in: 2016 IEEE International Conference on Communications (ICC). Presented at the 2016 IEEE International Conference on Communications (ICC), pp. 1–6. https://doi.org/10.1109/ICC.2016.7510831

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15, 850–863.

Hyvarinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. 10, 626–634. https://doi.org/10.1109/72.761722

Iglesias, J.E., Konukoglu, E., Zikic, D., Glocker, B., Van Leemput, K., Fischl, B., 2013. Is Synthesizing MRI Contrast Useful for Inter-modality Analysis?, in: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 631–638.

Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z., 2011. Robust Brain Extraction Across Datasets and Comparison With Publicly Available Methods. IEEE Trans. Med. Imaging 30, 1617–1634. https://doi.org/10.1109/TMI.2011.2138152

Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Hakonarson, H., Gur, R.E., Gur, R.C., Verma, R., 2014. Sex differences in the structural connectome of the human brain. Proc. Natl. Acad. Sci. 111, 823–828. https://doi.org/10.1073/pnas.1316909110

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ArXiv150203167 Cs.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2016. Image-to-Image Translation with Conditional Adversarial Networks. ArXiv161107004 Cs.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM International Conference on Multimedia. ACM, pp. 675–678.

Jiang, D., Dou, W., Vosters, L., Xu, X., Sun, Y., Tan, T., 2018. Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network. Jpn. J. Radiol. 36, 566–574. https://doi.org/10.1007/s11604-018-0758-8

Jog, A., 2016. Image Synthesis in Magnetic Resonance Neuroimaging. Johns Hopkins University.

Jog, A., Carass, A., Roy, S., Pham, D.L., Prince, J.L., 2017. Random Forest Regression for Magnetic Resonance Image Synthesis. Med. Image Anal. 35, 475–488. https://doi.org/10.1016/j.media.2016.08.009

Kalin, N.H., Shelton, S.E., Davidson, R.J., 2007. Role of the Primate Orbitofrontal Cortex in Mediating Anxious Temperament. Biol. Psychiatry 62, 1134–1139. https://doi.org/10.1016/j.biopsych.2007.04.004

Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78. https://doi.org/10.1016/j.media.2016.10.004

Kendall, A., Badrinarayanan, V., Cipolla, R., 2015a. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. ArXiv151102680 Cs.

Kendall, A., Badrinarayanan, V., Cipolla, R., 2015b. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. ArXiv151102680 Cs.

Kendall, A., Gal, Y., 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? ArXiv170304977 Cs.

Kim, J., Calhoun, V.D., Shim, E., Lee, J.-H., 2016. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. NeuroImage 124, Part A, 127–146. https://doi.org/10.1016/j.neuroimage.2015.05.018

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. ArXiv14126980 Cs.

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., Biller, A., 2016. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. NeuroImage 129, 460–469. https://doi.org/10.1016/j.neuroimage.2016.01.024

Kourtzi, Z., Augath, M., Logothetis, N.K., Movshon, J.A., Kiorpes, L., 2006. Development of visually evoked cortical activity in infant macaque monkeys studied longitudinally with fMRI. Magn. Reson. Imaging 24, 359–366. https://doi.org/10.1016/j.mri.2005.12.025

Krähenbühl, P., Koltun, V., 2012. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. ArXiv12105644 Cs.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012a. ImageNet Classification with Deep Convolutional Neural Networks, in: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012b. Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems. pp. 1097–1105.

Kuleshov, V., Fenner, N., Ermon, S., 2018. Accurate Uncertainties for Deep Learning Using Calibrated Regression. ArXiv180700263 Cs Stat.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2016. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. ArXiv161201474 Cs Stat.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Haffner, P., Bottou, L., Bengio, Y., 1999. Object Recognition with Gradient-Based Learning, in: Forsyth, D.A., Mundy, J.L., di Gesú, V., Cipolla, R. (Eds.), Shape, Contour and Grouping in Computer Vision, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 319–345. https://doi.org/10.1007/3-540-46805-6_19

Lee, H., Ekanadham, C., Ng, A.Y., 2008. Sparse deep belief net model for visual area V2, in: Platt, J.C., Koller, D., Singer, Y., Roweis, S.T. (Eds.), Advances in Neural Information Processing Systems 20. Curran Associates, Inc., pp. 873–880.

Lee, H., Grosse, R., Ranganath, R., Ng, A.Y., 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. ACM Press, pp. 1–8. https://doi.org/10.1145/1553374.1553453

Liu, F., Jang, H., Kijowski, R., Zhao, G., Bradshaw, T., McMillan, A., 2018. A Deep Learning Approach for 18F-FDG PET Attenuation Correction. Eur. J. Nucl. Med. Mol. Imaging.

Liu, F., Zhou, Z., Jang, H., Samsonov, A., Zhao, G., Kijowski, R., 2017. Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. Magn. Reson. Med. https://doi.org/10.1002/mrm.26841

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., Feng, D., 2014. Early diagnosis of Alzheimer's disease with deep learning, in: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI). Presented at the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp. 1015–1018. https://doi.org/10.1109/ISBI.2014.6868045

Livingstone, M.S., Vincent, J.L., Arcaro, M.J., Srihasam, K., Schade, P.F., Savage, T., 2017. Development of the macaque face-patch system. Nat. Commun. 8. https://doi.org/10.1038/ncomms14897

London, M., Häusser, M., 2005. Dendritic Computation. Annu. Rev. Neurosci. 28, 503–532. https://doi.org/10.1146/annurev.neuro.28.061604.135703

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Lv, B., Li, J., He, H., Li, M., Zhao, M., Ai, L., Yan, F., Xian, J., Wang, Z., 2010. Gender consistency and difference in healthy adults revealed by cortical thickness. NeuroImage 53, 373–382. https://doi.org/10.1016/j.neuroimage.2010.05.020

Malpetti, M., Ballarini, T., Presotto, L., Garibotto, V., Tettamanti, M., Perani, D., Null, N., 2017. Gender differences in healthy aging and Alzheimer's Dementia: A 18F-FDG-PET study of brain and cognitive reserve. Hum. Brain Mapp. 38, 4212–4227. https://doi.org/10.1002/hbm.23659

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.-C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K., 2015. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans. Med. Imaging 34, 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

Miller, M.I., Christensen, G.E., Amit, Y., Grenander, U., 1993. Mathematical textbook of deformable neuroanatomies. Proc. Natl. Acad. Sci. U. S. A. 90, 11944–11948.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. ArXiv160604797 Cs.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10). pp. 807–814.

Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., Wang, Q., Shen, D., 2018. Medical Image Synthesis with Deep Convolutional Adversarial Networks. IEEE Trans. Biomed. Eng. 65, 2720–2730. https://doi.org/10.1109/TBME.2018.2814538

Nie, D., Zhang, H., Adeli, E., Liu, L., Shen, D., 2016. 3D Deep Learning for Multi-modal Imaging-Guided Survival Time Prediction of Brain Tumor Patients. Med. Image Comput. Comput.-Assist. Interv. MICCAI Int. Conf. Med. Image Comput. Comput.-Assist. Interv. 9901, 212–220. https://doi.org/10.1007/978-3-319-46723-8_25

Oler, J.A., Fox, A.S., Shelton, S.E., Rogers, J., Dyer, T.D., Davidson, R.J., Shelledy, W., Oakes, T.R., Blangero, J., Kalin, N.H., 2010. Amygdalar and hippocampal substrates of anxious temperament differ in their heritability. Nature 466, 864–868. https://doi.org/10.1038/nature09282

Orgo, L., Bachmann, M., Kalev, K., Hinrikus, H., Järvelaid, M., 2016. Brain functional connectivity in depression: Gender differences in EEG, in: 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES). Presented at the 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), pp. 270–273. https://doi.org/10.1109/IECBES.2016.7843456

Pan, S.J., Yang, Q., 2010. A Survey on Transfer Learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch.

Qian, N., 1999. On the momentum term in gradient descent learning algorithms. Neural Netw. 12, 145–151. https://doi.org/10.1016/S0893-6080(98)00116-6

Robinson, E.C., Jbabdi, S., Glasser, M.F., Andersson, J., Burgess, G.C., Harms, M.P., Smith, Stephen.M., Van Essen, D.C., Jenkinson, M., 2014. MSM: a new flexible framework for Multimodal Surface Matching☆. NeuroImage 100, 414–426. https://doi.org/10.1016/j.neuroimage.2014.05.069

Rohlfing, T., Kroenke, C.D., Sullivan, E.V., Dubach, M.F., Bowden, D.M., Grant, K., Pfefferbaum, A., 2012a. The INIA19 Template and NeuroMaps Atlas for Primate Brain Image Parcellation and Spatial Normalization. Front. Neuroinformatics 6. https://doi.org/10.3389/fninf.2012.00027

Rohlfing, T., Kroenke, C.D., Sullivan, E.V., Dubach, M.F., Bowden, D.M., Grant, K.A., Pfefferbaum, A., 2012b. The INIA19 Template and NeuroMaps Atlas for Primate Brain Image Parcellation and Spatial Normalization. Front. Neuroinformatics 6. https://doi.org/10.3389/fninf.2012.00027

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. ArXiv150504597 Cs.

Roy, S., Butman, J.A., Pham, D.L., 2017. Robust skull stripping using multiple MR image contrasts insensitive to pathology. NeuroImage 146, 132–147. https://doi.org/10.1016/j.neuroimage.2016.11.017

Roy, S., Carass, A., Jog, A., Prince, J.L., Lee, J., 2014. MR to CT Registration of Brains using Image Synthesis. Proc. SPIE 9034. https://doi.org/10.1117/12.2043954

Roy, S., Carass, A., Prince, J., 2011. A compressed sensing approach for MR tissue contrast synthesis. Inf. Process. Med. Imaging Proc. Conf. 22, 371–383.

Roy, S., Carass, A., Prince, J.L., 2013. Magnetic Resonance Image Example-Based Contrast Synthesis. IEEE Trans. Med. Imaging 32, 2348–2363. https://doi.org/10.1109/TMI.2013.2282126

Ruigrok, A.N.V., Salimi-Khorshidi, G., Lai, M.-C., Baron-Cohen, S., Lombardo, M.V., Tait, R.J., Suckling, J., 2014. A meta-analysis of sex differences in human brain structure. Neurosci. Biobehav. Rev. 39, 34–50. https://doi.org/10.1016/j.neubiorev.2013.12.004

Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Auto-context Convolutional Neural Network (Auto-Net) for Brain Extraction in Magnetic Resonance Imaging. IEEE Trans. Med. Imaging PP, 1–1. https://doi.org/10.1109/TMI.2017.2721362

Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., Smith, S.M., 2014. Automatic Denoising of Functional MRI Data: Combining Independent Component Analysis and Hierarchical Fusion of Classifiers. NeuroImage 90, 449–468. https://doi.org/10.1016/j.neuroimage.2013.11.046

Sanchez, I., Vilaplana, V., 2018. Brain MRI super-resolution using 3D generative adversarial networks.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Netw. 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schmithorst, V.J., Holland, S.K., 2006. Functional MRI evidence for disparate developmental processes underlying intelligence in boys and girls. NeuroImage 31, 1366–1379. https://doi.org/10.1016/j.neuroimage.2006.01.010

Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. NeuroImage 22, 1060–1075. https://doi.org/10.1016/j.neuroimage.2004.03.032

Seidlitz, J., Sponheim, C., Glen, D., Ye, F.Q., Saleem, K.S., Leopold, D.A., Ungerleider, L., Messinger, A., 2017. A population MRI brain template and analysis tools for the macaque. NeuroImage. https://doi.org/10.1016/j.neuroimage.2017.04.063

Shackman, A.J., Fox, A.S., Oler, J.A., Shelton, S.E., Oakes, T.R., Davidson, R.J., Kalin, N.H., 2017. Heightened extended amygdala metabolism following threat characterizes the early phenotypic risk to develop anxiety-related psychopathology. Mol. Psychiatry 22, 724–732. https://doi.org/10.1038/mp.2016.132

Shattuck, D.W., Prasad, G., Mirza, M., Narr, K.L., Toga, A.W., 2009. Online resource for validation of brain segmentation methods. NeuroImage 45, 431–439. https://doi.org/10.1016/j.neuroimage.2008.10.066

Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic Resonance Image Tissue Classification Using a Partial Volume Model. NeuroImage 13, 856–876. https://doi.org/10.1006/nimg.2000.0730

Shelhamer, E., Long, J., Darrell, T., 2016. Fully Convolutional Networks for Semantic Segmentation. ArXiv160506211 Cs.

Shotton, J., Winn, J., Rother, C., Criminisi, A., 2009. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. Int. J. Comput. Vis. 81, 2–23.

Simon, J.H., Li, D., Traboulsee, A., Coyle, P.K., Arnold, D.L., Barkhof, F., Frank, J.A., Grossman, R., Paty, D.W., Radue, E.W., Wolinsky, J.S., 2006. Standardized MR imaging protocol for multiple sclerosis: Consortium of MS Centers consensus guidelines. AJNR Am. J. Neuroradiol. 27, 455–461.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv14091556 Cs.

Smith, S., Andersson, J., Auerbach, E.J., Beckmann, C.F., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D.A., Griffanti, L., Harms, M.P., Kelly, M., Laumann, T., Miller, K.L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A.Z., Vu, A., Woolrich, M.W., Xu, J., Yacoub, E., Ugurbil, K., Van Essen, D., Glasser, M.F., 2013. Resting-state fMRI in the Human Connectome Project. NeuroImage 80, 144–168. https://doi.org/10.1016/j.neuroimage.2013.05.039

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155. https://doi.org/10.1002/hbm.10062

Smith, S.M., Hyvärinen, A., Varoquaux, G., Miller, K.L., Beckmann, C.F., 2014. Group-PCA for very large fMRI datasets. Neuroimage 101, 738–749. https://doi.org/10.1016/j.neuroimage.2014.07.051

Smith, S.M., Vidaurre, D., Beckmann, C.F., Glasser, M.F., Jenkinson, M., Miller, K.L., Nichols, T.E., Robinson, E.C., Salimi-Khorshidi, G., Woolrich, M.W., Barch, D.M., Uğurbil, K., Van Essen, D.C., 2013. Functional connectomics from resting-state fMRI. Trends Cogn. Sci., Special Issue: The Connectome 17, 666–682. https://doi.org/10.1016/j.tics.2013.09.016

Sowell, E.R., Peterson, B.S., Kan, E., Woods, R.P., Yoshii, J., Bansal, R., Xu, D., Zhu, H., Thompson, P.M., Toga, A.W., 2007. Sex Differences in Cortical Thickness Mapped in 176 Healthy Individuals between 7 and 87 Years of Age. Cereb. Cortex N. Y. N 1991 17, 1550–1560. https://doi.org/10.1093/cercor/bhl066

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J., 2017. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. ArXiv170703237 Cs 10553, 240–248. https://doi.org/10.1007/978-3-319-67558-9_28

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, T.A.D.N., 2015. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. 220, 841–859. https://doi.org/10.1007/s00429-013-0687-3

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging 15. https://doi.org/10.1186/s12880-015-0068-x

van der Kouwe, A.J.W., Benner, T., Salat, D.H., Fischl, B., 2008. Brain morphometry with multiecho MPRAGE. NeuroImage 40, 559–569. https://doi.org/10.1016/j.neuroimage.2007.12.025

Van Essen, D.C., Smith, J., Glasser, M.F., Elam, J., Donahue, C.J., Dierker, D.L., Reid, E.K., Coalson, T., Harwell, J., 2017. The Brain Analysis Library of Spatial maps and Atlases (BALSA) Database. NeuroImage 144, 270–274. https://doi.org/10.1016/j.neuroimage.2016.04.002

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., 2013. The WU-Minn Human Connectome Project: An Overview. NeuroImage 80, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Vieira, S., Pinaya, W.H.L., Mechelli, A., 2017. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neurosci. Biobehav. Rev. 74, Part A, 58–75. https://doi.org/10.1016/j.neubiorev.2017.01.002

Wachinger, C., Reuter, M., Klein, T., 2017. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. NeuroImage. https://doi.org/10.1016/j.neuroimage.2017.02.035

Wang, Y., Jia, H., Yap, P.-T., Cheng, B., Wee, C.-Y., Guo, L., Shen, D., 2012. Groupwise Segmentation Improves Neuroimaging Classification Accuracy, in: Yap, P.-T., Liu, T., Shen, D., Westin, C.-F., Shen, L. (Eds.), Multimodal Brain Image Analysis: Second International Workshop, MBIA 2012, Held in Conjunction with MICCAI 2012, Nice, France, October 1-5, 2012. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 185–193. https://doi.org/10.1007/978-3-642-33530-3_16

Wang, Y., Nie, J., Yap, P.-T., Li, G., Shi, F., Geng, X., Guo, L., Shen, D., 2014. Knowledge-Guided Robust MRI Brain Extraction for Diverse Large-Scale Neuroimaging Studies on Humans and Non-Human Primates. PLoS ONE 9. https://doi.org/10.1371/journal.pone.0077810

Xiang, L., Wang, Q., Nie, D., Zhang, L., Jin, X., Qiao, Y., Shen, D., 2018. Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image. Med. Image Anal. 47, 31–44. https://doi.org/10.1016/j.media.2018.03.011

Xu, Y., Géraud, T., Bloch, I., 2017. From Neonatal to Adult Brain MR Image Segmentation in a Few Seconds Using 3D-Like Fully Convolutional Network and Transfer Learning, in: Proceedings of the 23rd IEEE International Conference on Image Processing (ICIP). Presented at the ICIP, pp. 4417--4421.

Zeiler, M.D., 2012. ADADELTA: An Adaptive Learning Rate Method. ArXiv12125701 Cs.

Zhang, C., Cahill, N.D., Arbabshirani, M.R., White, T., Baum, S.A., Michael, A.M., 2016. Sex and Age Effects of Functional Connectivity in Early Adulthood. Brain Connect. 6, 700–713. https://doi.org/10.1089/brain.2016.0429

Zhang, C., Dougherty, C.C., Baum, S.A., White, T., Michael, A.M., 2018. Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity. Hum. Brain Mapp. 39, 1765–1776. https://doi.org/10.1002/hbm.23950

Zhang, D., Lu, G., 2004. Review of shape representation and description techniques. Pattern Recognit. 37, 1–19.

Zhao, G., Liu, F., Oler, J.A., Meyerand, M.E., Kalin, N.H., Birn, R.M., 2018. Bayesian convolutional neural network based MRI brain extraction on nonhuman primates. NeuroImage 175, 32–44. https://doi.org/10.1016/j.neuroimage.2018.03.065

Zhu, B., Liu, J.Z., Cauley, S.F., Rosen, B.R., Rosen, M.S., 2018. Image reconstruction by domain-transform manifold learning. Nature 555, 487–492. https://doi.org/10.1038/nature25988

Appendix A: Additional Figures for MRI Brain Extraction

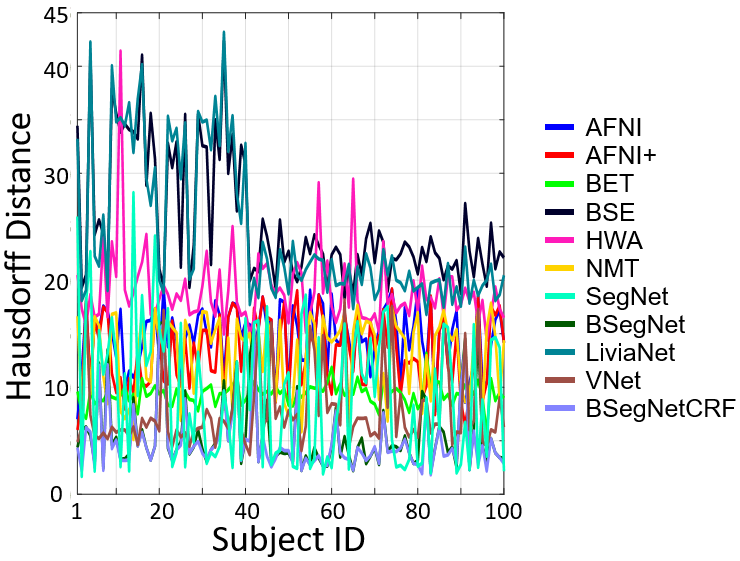

Fig. A1. Hausdorff distance on each subject from different brain extraction methods.

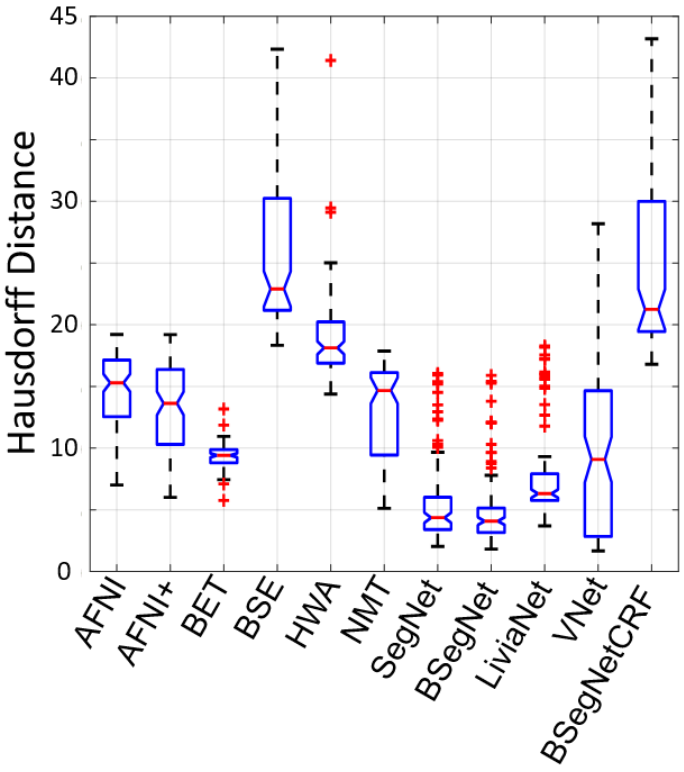

Fig. A2. Hausdorff distance in boxplots from different brain extraction methods. In the figure points are drawn as outliers with red '+' symbols, if they are greater than q3+1.5(q3-q1) or less than q1-1.5(q3-q1), where q1 and q3 are the first and third quartiles respectively.

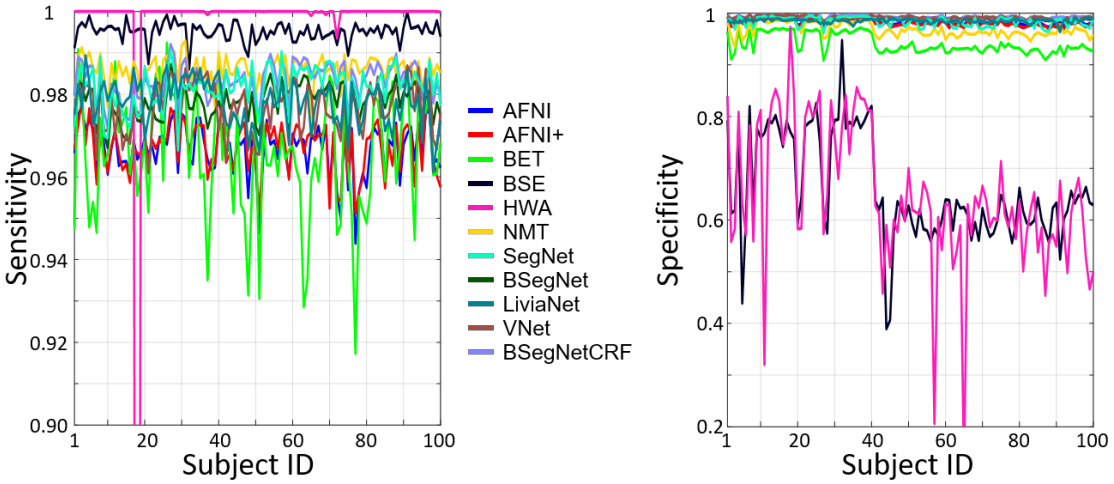

Fig. A3. Sensitivity and specificity on each subject from different brain extraction methods.

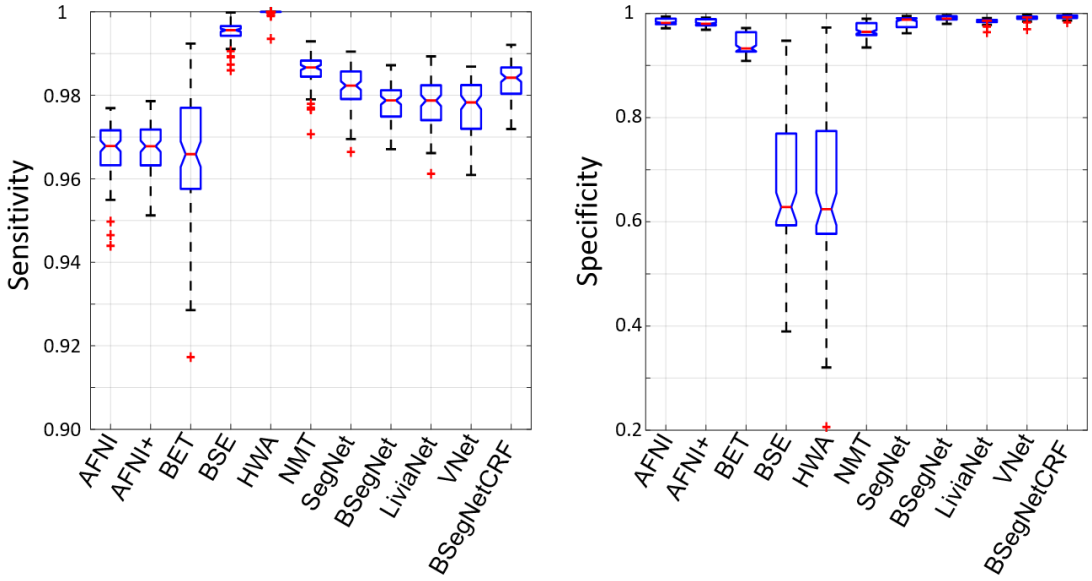

Fig. A4. Sensitivity and specificity in boxplots from different brain extraction methods. In the figure points are drawn as outliers with red '+' symbols, if they are greater than q3+1.5(q3-q1) or less than q1-1.5(q3-q1), where q1 and q3 are the first and third quartiles respectively.

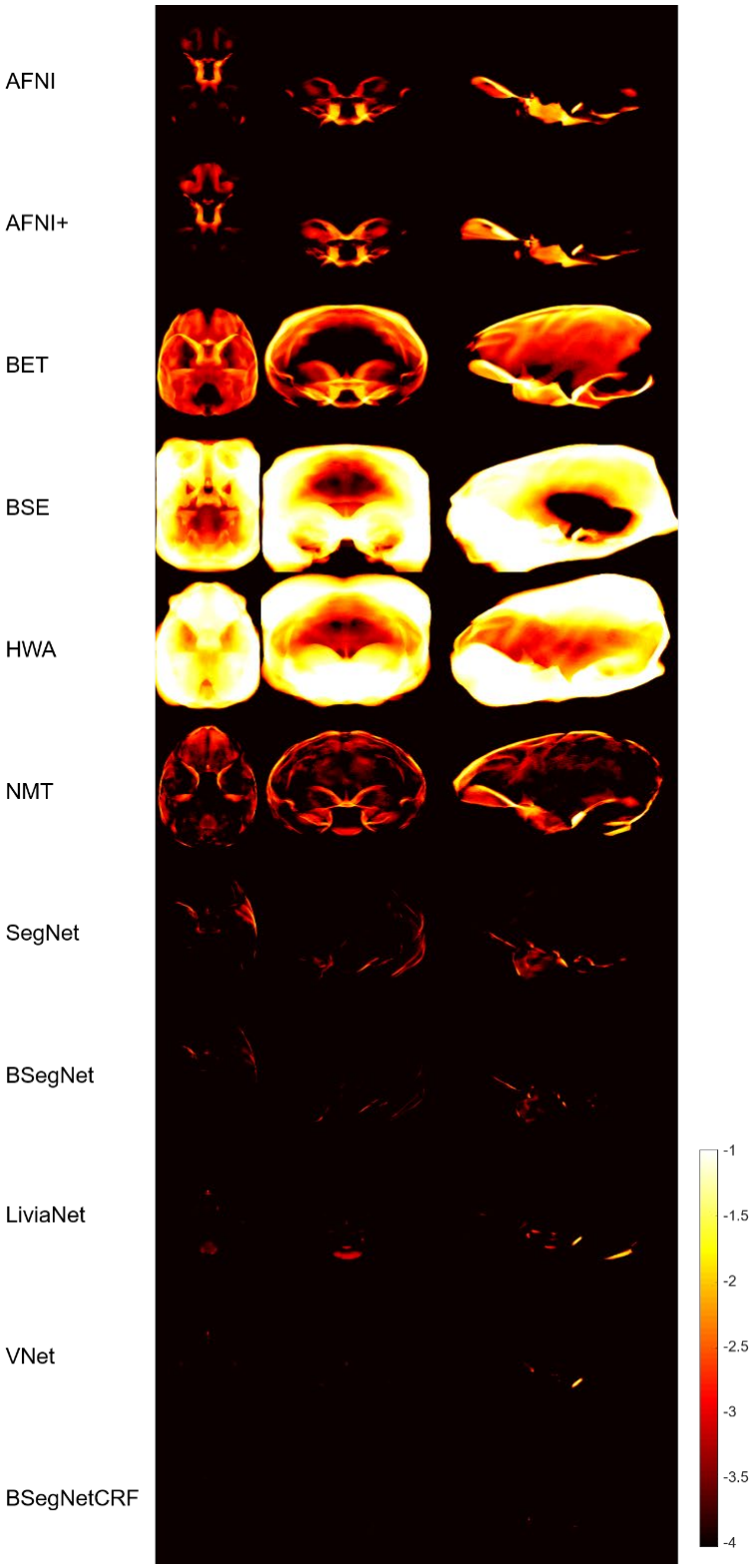

Fig. A5. Averaged false positive map for compared methods. For display purposes, the natural logarithm of the averaged map collapsed (averaged) along each axis is shown.

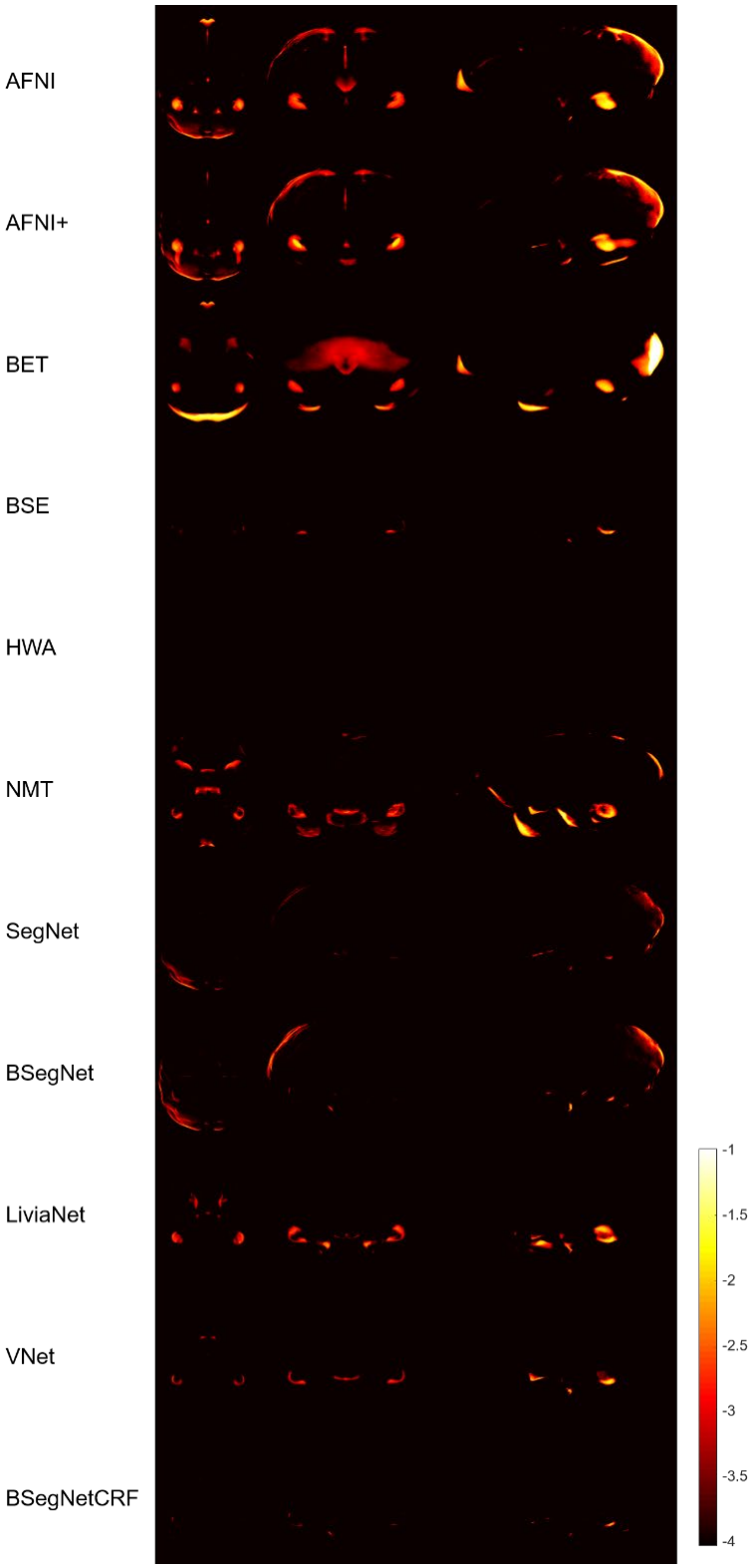

Fig. A6. Averaged false negative map for compared methods. For display purposes, the natural logarithm of the averaged map collapsed (averaged) along each axis is shown.

Appendix B: Additional Tables and Figures for Functional Connectivity Gender Prediction

Table B1. The mean of the prediction accuracies across the 50 randomized cross validation permutations for each kind of predictive model and each kind of input. L – number of hidden layers in the DNN model; N – number of neurons in the first hidden layer of the DNN, the number of neurons in each of the rest hidden layers is half of this number.

	ICA25	**ICA50**	**ICA100**	**ICA200**	**ICA300**	**Template**
L:1 N:20	0.8262	0.8790	0.9229	0.9373	0.9437	0.9146
L:1 N:50	0.8260	0.8786	0.9235	0.9370	0.9434	0.9129
L:1 N:100	0.8268	0.8783	0.9239	0.9369	0.9433	0.9130
L:1 N:200	0.8281	0.8786	0.9237	0.9370	0.9427	0.9097
L:2 N:20	0.8301	0.8767	0.9213	0.9371	0.9446	0.9144
L:2 N:50	0.8318	0.8751	0.9205	0.9361	0.9432	0.9137
L:2 N:100	0.8316	0.8764	0.9213	0.9350	0.9425	0.9134
L:2 N:200	0.8314	0.8766	0.9215	0.9355	0.9418	0.9122
L:3 N:20	0.8288	0.8761	0.9214	0.9368	0.9443	0.9142
L:3 N:50	0.8300	0.8767	0.9198	0.9357	0.9434	0.9114
L:3 N:100	0.8301	0.8770	0.9198	0.9350	0.9433	0.9126
L:3 N:200	0.8300	0.8763	0.9202	0.9354	0.9414	0.9123
SVM-linear	0.8116	0.8717	0.9264	0.9411	0.9489	0.9214
SVM-poly2	0.8144	0.8527	0.8956	0.9256	0.9311	0.8922

Table B2. The standard deviation of the prediction accuracies across the 50 randomized cross validation permutations for each kind of predictive model and each kind of input. L – number of hidden layers in the DNN model; N – number of neurons in the first hidden layer of the DNN, the number of neurons in each of the rest hidden layers is half of this number.

	ICA25	**ICA50**	**ICA100**	**ICA200**	**ICA300**	**Template**
L:1 N:20	0.0094	0.0076	0.0068	0.0053	0.0061	0.0086
L:1 N:50	0.0091	0.0077	0.0068	0.0060	0.0067	0.0079
L:1 N:100	0.0091	0.0067	0.0062	0.0054	0.0061	0.0078
L:1 N:200	0.0100	0.0076	0.0066	0.0060	0.0061	0.0084
L:2 N:20	0.0101	0.0072	0.0075	0.0060	0.0070	0.0089
L:2 N:50	0.0106	0.0078	0.0065	0.0062	0.0064	0.0078
L:2 N:100	0.0106	0.0084	0.0067	0.0061	0.0063	0.0081
L:2 N:200	0.0103	0.0074	0.0061	0.0056	0.0063	0.0080
L:3 N:20	0.0094	0.0068	0.0067	0.0063	0.0069	0.0086
L:3 N:50	0.0106	0.0069	0.0068	0.0050	0.0060	0.0080
L:3 N:100	0.0100	0.0071	0.0071	0.0066	0.0061	0.0084
L:3 N:200	0.0109	0.0070	0.0067	0.0053	0.0061	0.0081
SVM-linear	0.0102	0.0090	0.0074	0.0061	0.0069	0.0093
SVM-poly2	0.0111	0.0109	0.0068	0.0057	0.0064	0.0087

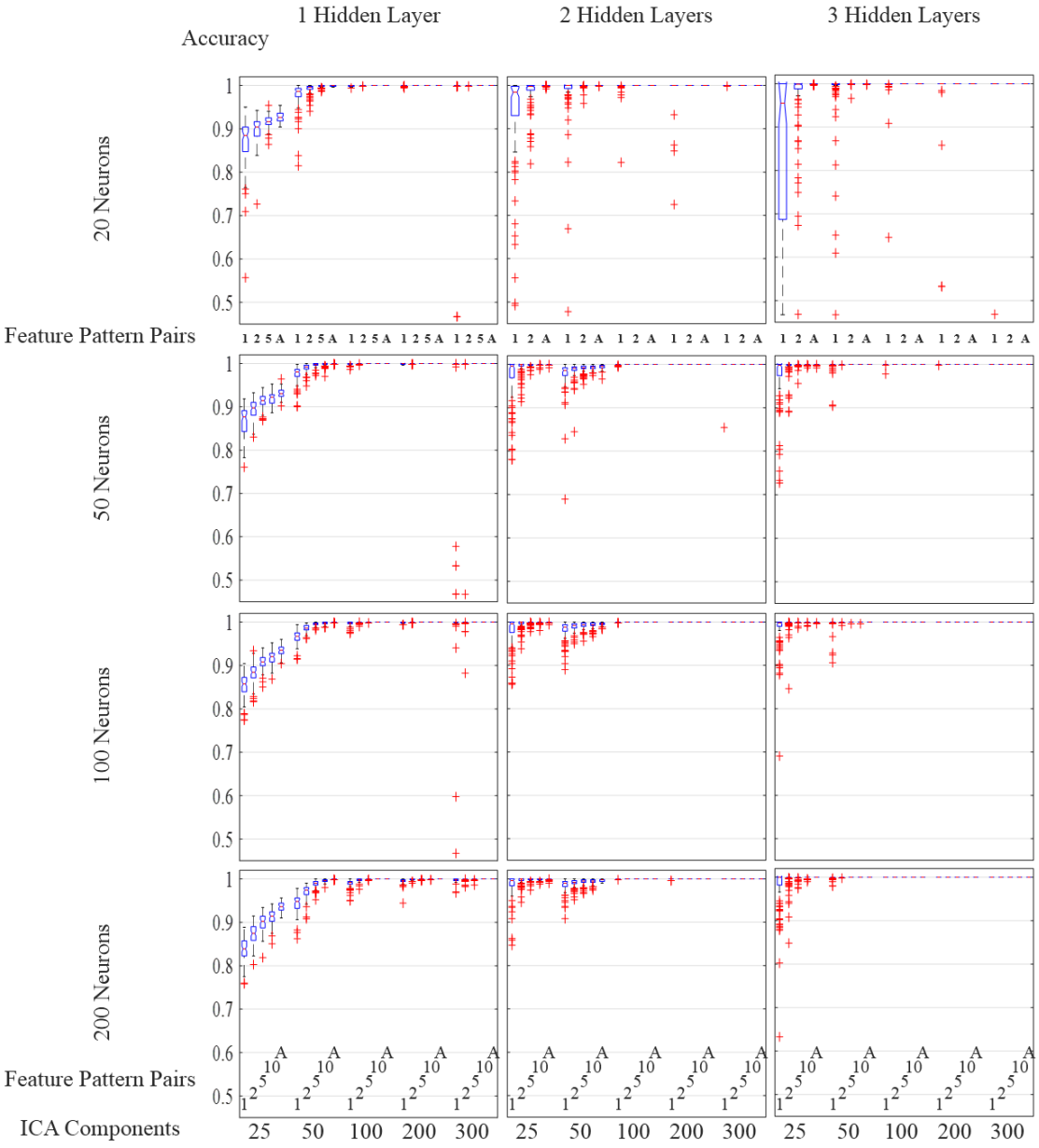

Fig. B1. Prediction accuracy recovered by the several most important high-level male and female feature pairs in the predictions on the training dataset. '1', '2', '5' and '10' mean that the predictions were made by the most important 1, 2, 5, 10 male and female feature pairs in the last hidden layer respectively. 'A' means the predictions were made by all the features..

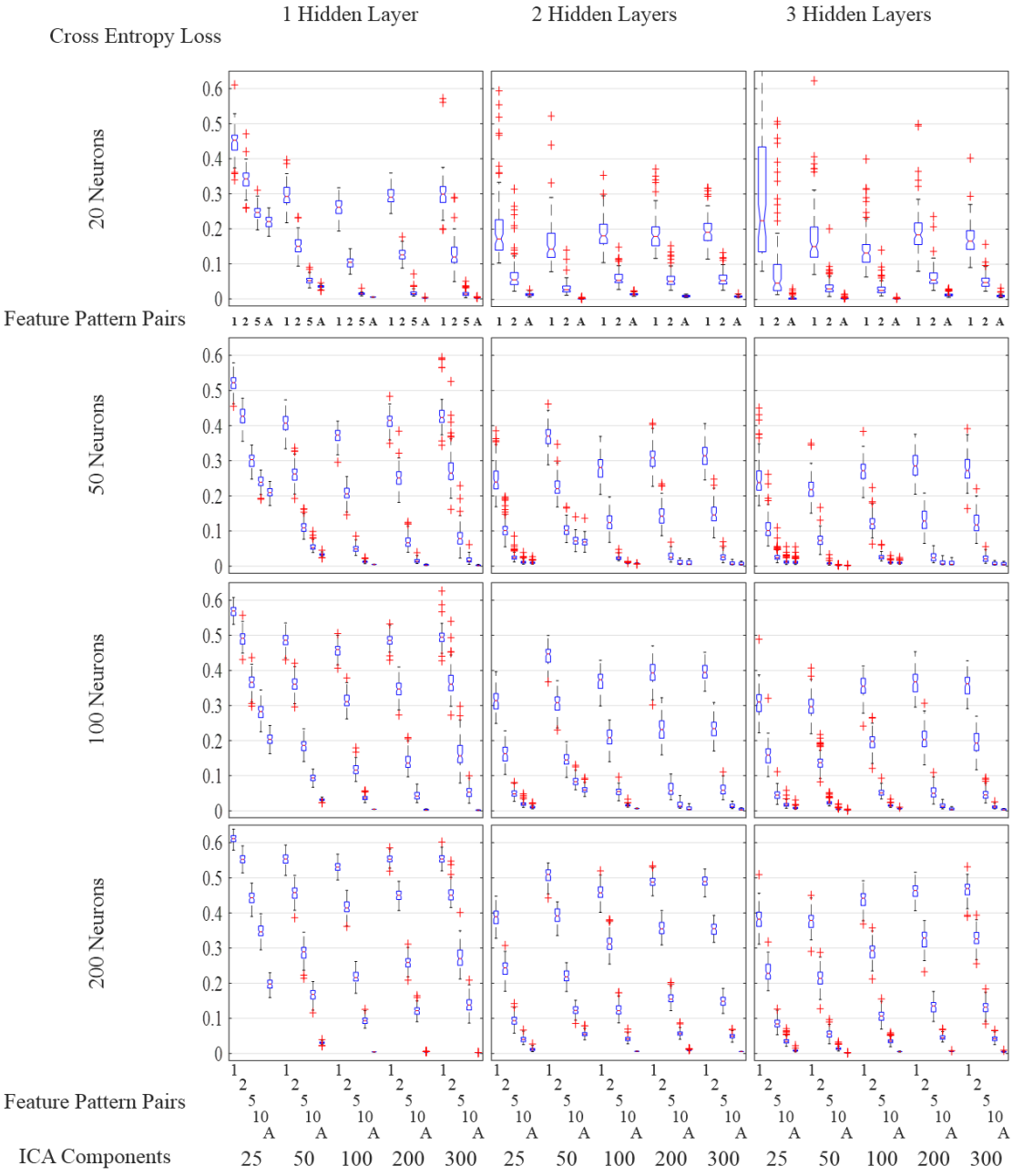

Fig. B2. The cross entropy loss achieved by the several most important high-level male and female feature pairs in the predictions on the training dataset. '1', '2', '5' and '10' mean that the predictions were made by the most important 1, 2, 5, 10 male and female feature pairs in the last hidden layer respectively. 'A' means the predictions were made by all the features.