Automatic U-Net based Segmentation Pipeline for Neonatal and Child Brain MRI

Pouria Mashouri[1] POURIA.MASHOURI@UHN.CA
[1] *University Health Network*
Fateme Sadat Haghpanah[1,2,3] FATEME.HAGHPANAH@MAIL.UTORONTO.CA
[2] *The Hospital for Sick Children*
[3] *University of Toronto*
Sam Osia[1,2,3] SAM.OSIA@SICKKIDS.CA
Min Sheng[2] MIN.SHENG@SICKKIDS.CA
Mireille Guillot[2] MIREILLE.GUILLOT.1@ULAVAL.CA
Jarred Garfinkle[2] JARRED.GARFINKLE@MCGILL.CA
Dianna McAllister[1,3] DIANNA.MCALLISTER@MAIL.UTORONTO.CA
Michael Brudno[1,3] BRUDNO@CS.TORONTO.EDU
Ting Guo[2] JESSIE.GUO@SICKKIDS.CA

Editors: Under Review for MIDL 2022

Abstract

Background: Robust segmentation on magnetic resonance images (MRI) is key to assessing brain growth, which is critical to the health and development of children. Machine learning has enabled breakthroughs in automatic brain MRI segmentation, albeit mainly for adults. Given the significant growth and dramatic anatomical variability of the brain during the first years of life, accurate delineation of the developing brain is more challenging. **Objective:** We present a U-Net based automatic segmentation pipeline to robustly segment brain structures on both neonatal and child MR images of children born preterm with enhanced robustness of data input. **Methods:** A total of 300 T1-weighted images were included: 100 early-in-life, 100 term equivalent age (TEA), and 100 8-year (median age at MRI: 32 weeks, 40.3 weeks, and 8.2 years respectively). Each image was accompanied with manually segmented annotations that cover cerebrum, cerebellum, and brainstem. We broke up each 3D image volume along the sagittal axis to 256 2D slices of 256x256 matrices. The U-Net model took as input a single slice, and returned a multi-class prediction denoting the mask of each label. The resultant 256 slices were then re-constructed back to a 3D scan. Our model was trained on a 60/20/20 training/validation/testing split across the age range for 200 epochs. **Results:** When compared to the manually segmented labels, our pipeline achieved an average DICE coefficient of 97.78% (95.40%-98.47%), 94.85% (89.58%-97.91%), and 94.84% (88.39%-97.12%) for cerebrum, cerebellum, and brainstem respectively, and outperformed InfantFS and FreeSurfer, applied to the same data. More importantly, our model had a robust performance across different age groups, demonstrating its applicability to a much wider age range than other existing state-of-the-art methods, as well as robust performance on an independently acquired external cohort. **Conclusions:** Our U-Net based pipeline allows accurate segmentation of brain structures across the first years of life. It is a flexible, robust, and automated tool to assess total and regional brain volumes and growth for studying healthy children and children with medical conditions.

Keywords: Segmentation, Computer Vision, MRI, Radiology, Infant Brain

© 2022 P. Mashouri *et al.*

1. Introduction

Brain growth in the early years of life is critical to the health and development of both healthy children and children with medical conditions (Cayam-Rand et al., 2021). Robust segmentation techniques allow the measurement of brain structural volumes on magnetic resonance images (MRI) and the assessment of brain growth (Thompson et al., 2020); (Cayam-Rand et al., 2021); (Fetit et al., 2020b). However, though generally considered as the gold standard, manual segmentation is labour-intensive and time-consuming, which significantly limits its applicability, especially in larger studies. Automatic segmentation of the developing brain in the first several years of life is challenging, due to (1) dramatic changes of brain tissue contrast on MRI from neonatal period to childhood, (2) significant anatomical variability contributed by increases in gyrification and volumes, and (3) morphological abnormalities preceded by brain injuries, such as severe ventriculomegaly.

A number of automatic segmentation approaches have been developed that allow brain tissue classification (Prastawa et al., 2005); (Xue et al., 2007); (Weisenfeld and Warfield, 2009); (Shi et al., 2011); (Wang et al., 2014); (Wang et al., 2015); (Zhou et al., 2019); (Fetit et al., 2020a) or specific structural delineation (Nishida et al., 2006); (Gousias et al., 2013); (Makropoulos et al., 2014); (Zöllei et al., 2020) in infants. Through machine learning initiatives, breakthroughs in automatic segmentation techniques have been made, however they were mainly designed for adult brain MRI (Fischl, 2012). The application of deep learning has been increasingly recognized in segmenting the developing brain (Mostapha and Styner, 2019); (Zöllei et al., 2020); (Liu et al., 2021). Although some of these methods provide accurate segmentations to infant brain images acquired within a certain age range (Fetit et al., 2020b); (Zöllei et al., 2020); (Liu et al., 2021), they are not readily applicable to segment brain images of school-aged children. Moeskops and colleagues presented a segmentation method using multi-scale convolutional neural network (CNN) and demonstrated excellent accuracy in multiple datasets (Moeskops et al., 2016). However, this method was not tested on images with severe brain anatomical abnormality. In addition, for models trained on a small dataset, overfitting can be a problem, preventing the method from being able to provide accurate brain segmentations for children outside of the training dataset.

In this study, we present a U-Net based automatic segmentation pipeline to segment brain structures on both neonatal and child brain MRIs from as early as 27 weeks of gestation to school age in children born very preterm, while also adding in key steps to enhance robustness of data input. We compare our results to FreeSurfer and InfantFS with suggested parameters, on the same datasets. Finally, we apply this approach to segment the brain of another independent cohort where severe anatomical abnormalities were present, and manually segmented labels were available to test its robustness and generalizability.

2. Methods

2.1. Participants, MRI, and Gold-Standard Brain Segmentation Labels

234 (122 males) very preterm neonates (birth gestation: 24 to 32 weeks) admitted to the NICU at British Columbia's Women's Hospital, Canada (2006–2012) were enrolled in a prospective longitudinal cohort study. 208 neonates acquired early MRI at median postmenstrual age (PMA) of 32 weeks [IQR=30.4–33.6]; 188 at term-equivalent age (TEA)

(median PMA 40.3 weeks, IQR=38.9-42.0); and the first 133 children of this cohort completed the school-age MRI (median age at MRI 8.2 years, IQR=8.1–8.4). Neonatal and 8-year images were acquired with previously reported parameters (Chau et al., 2009) and (Cayam-Rand et al., 2021).

Segmentations of the brain were performed on early-life, TEA and 8-year T1-weighted images using methods described previously (Guo et al., 2015). Thorough manual revisions were done on the cerebrum (Cayam-Rand et al., 2021), cerebellum (Garfinkle et al., 2020), and brainstem (Guillot et al., 2020) labels to ensure accurate delineation of the structures.

The Clinical Research Ethics Board at University of British Columbia and BC Children's and Women's Hospitals approved the study. Written informed consent was obtained from parent/legal guardian, and informed assent from each participating child at 8 years.

2.2. Pre-processing Steps

We first restructured the dataset to be of uniform shape and orientation. All image slices were aligned to the sagittal plane. Although not all scans were captured in the same orientation, given the isotropic voxel size of the raw scans, no data were lost when changing the axes, and retraining the model against other axes is permitted. Next, to achieve a uniform dimension across all axes among all scans, we symmetrically zero-padded each raw scan to achieve the same shape of (256,256,256). Lastly, we scaled the intensity values of each scan to the range of 0-1. We then applied z-score normalization to the entire dataset to minimize the impact of outliers. The mean and standard deviation used for the z-score normalization were computed from training dataset only, but applied to the entire dataset.

2.3. Model Architecture

Our model is a pure U-NET (Ronneberger et al., 2015) based CNN model for brain segmentation. It consists of a U-shaped architecture that uses convolutional layers to build down and condense the input, followed by multiple upsampling layers, which make the final output have the same shape as the input. A key consideration in this project was to reduce the number of inputs. Thus, our model only requires a 3D scan as input, and outputs a 3D mask depicting multiple brain regions, removing any direct dependency on the subject age.

Additionally, the authors acknowledged that any model built on the 3D scan alone will have an underlying reliance/bias related to scan parameters, e.g. slice thickness in the training set. To overcome this without introducing any potential data-loss, each scan was broken down into a set of 2D slices (i.e., a 3D scan of shape 256x256x256 was split into 256 2D slices of shape 256x256), and the model was trained on each 2D slice separately. The predicted masks for all 256 2D slices were then re-constructed back into a 3D volume which holds the final multi-label mask of the subject scan. This structure enhances the robustness of the model by eliminating biases related to scan parameters, such as slice thickness and scan volume/size. Training on 2D scans also allows the model to better learn patterns from similar slices of different subjects (including across age groups), which can be harder to achieve by training on entire 3D scans. Lastly, training a model in 2D space requires significantly less computation and data, permitting better convergence and stabilization.

Model Parameters: The model was implemented with Python v2.7.12, using Tensorflow v1.15.0, and trained on a CentOS v7.6 VM with 170GB RAM, 16 threads, and a

Tesla V100 32vGB GPU. A 5-layer U-Net model was used to prevent overfitting. Batch Normalization was added as a regularizer to help the model train & converge better. Adam was used as the optimizer, along with a batch size of 32 and a learning rate of 1e-5. The SparseCategoricalCrossEntropy loss function was selected to allow for pixel-level learning. The model was set to train for 200 epochs, with EarlyStopping set based on the validation loss. Illustration of the model architecture can be found in Figure 1.

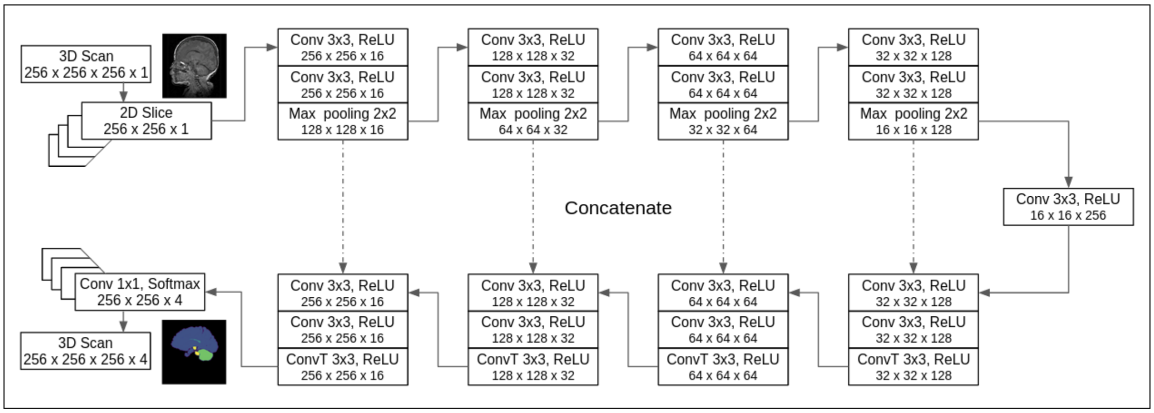


Figure 1: Machine Learning Model Architecture. Each 2D slice is passed through several convolutional layers followed by the same number of upsampling layers, and will end with an outputted 2D slice of the same shape as the input. The final output image has 4 dimensions depicting the prediction of each label: background, Cerebrum, Cerebellum, and Brainstem.

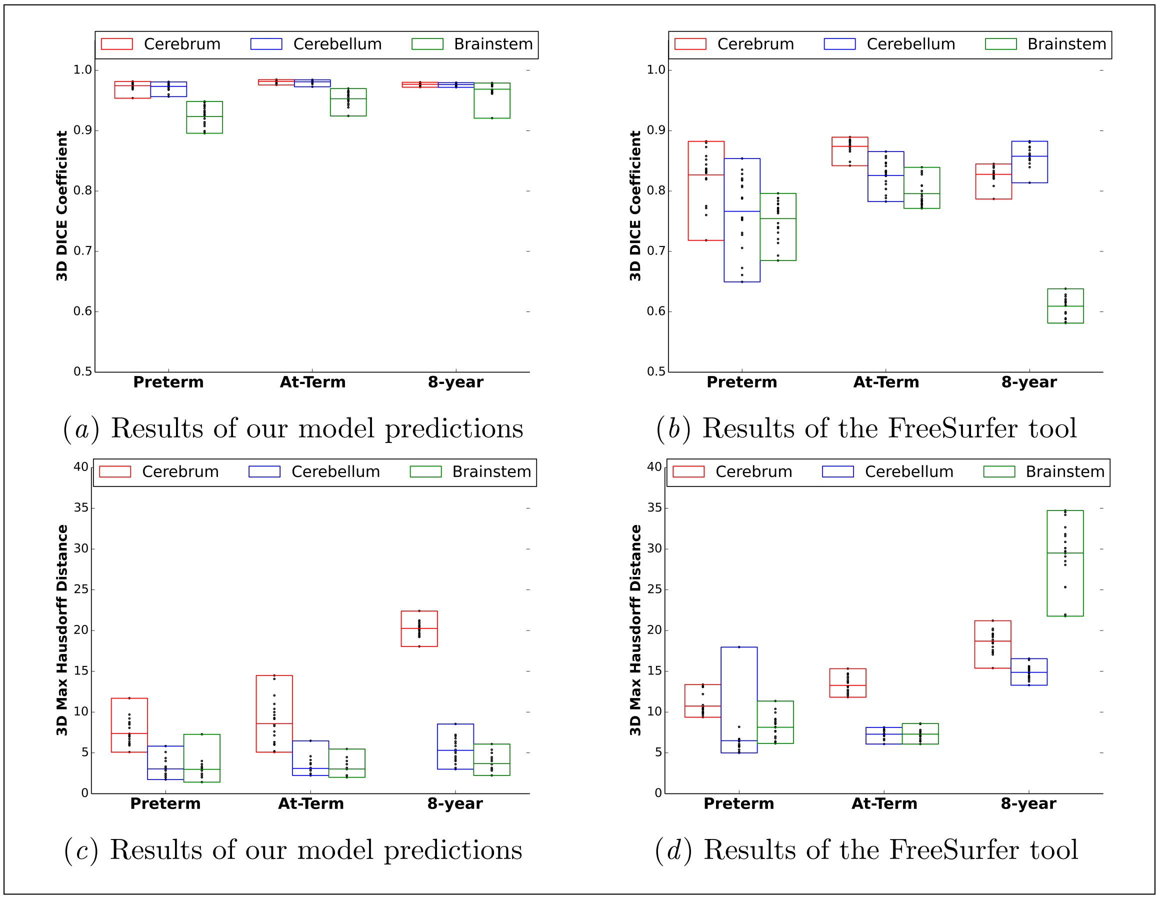
2.4. Post-Processing Steps

The model occasionally made very questionable predictions, e.g. predicting a region far away from the patient's skull. We added a step to circumvent such issues. The largest predicted region was identified as a 3D blob, and labels of any blob not connected to the largest blob and whose center lied further than 40 pixels from the closest edge of the largest blob were removed. More details can be found under Appendix A.

2.5. Evaluation Metrics

The 3D Dice Similarity Coefficient (DSC) (Zou et al., 2004) and the Hausdorff distance (HD) (Taha and Hanbury, 2015) were used to assess the quality of the segmentation predictions. The 3D DSC is a widely-used metric for evaluating spatial overlap between objects, and reflects the "closeness" between two objects. The HD however, measures the maximal contour distance between two objects, and is more sensitive to outliers. They provide complementary information that help paint a more complete picture of the model performance.

The segmentation accuracy of each brain region label was evaluated independently to better determine how well the model performed for each region. Details regarding DSC and HD can be found under Appendix B and C.



(a) Results of our model predictions

(b) Results of the FreeSurfer tool

(c) Results of our model predictions

(d) Results of the FreeSurfer tool

Figure 2: Visualization of results across age groups. Each black dot represents an individual scan. The boxes represent the bounds of each label, with the bisecting line at the mean.

2.6. Comparison with the FreeSurfer Platform

FreeSurfer is a well established platform for brain segmentation (Fischl, 2012). We segmented our test dataset using FreeSurfer and compared with our method. As our dataset spans from 27 weeks gestation to 8 years of age, InfantFS (Zöllei et al., 2020) was used on images under the age of 2 while FreeSurfer was used for 8-year images.

2.7. Cross Validation on another Dataset of an Independent Cohort

The generalizability of our method was evaluated on an independent Preterm Care (PC) cohort where the cerebrum labels were available. This cohort includes 179 neonates born at 24-32 weeks' gestation with serial brain MRI acquired at early life (median PMA at MRI: 32.9 weeks) and at TEA (median PMA: 41.3 weeks) on a Siemens 3T Tim Trio scanner at SickKids Hospital, Canada. 3D T1-weighted images were acquired using a FLASH sequence (TR: 36ms, TE: 9.2ms, 1mm isotropic). Ventriculomegaly was identified in 30 neonates of whom six had manually segmented labels, and were used to further evaluate our model.

Table 1: U-Net based Model Accurately Predicts Brain Labels across Age Groups

A: 3D DICE Coefficients

		Cerebrum	**Cerebellum**	**Brainstem**
preterm	Our model	97.46% (95.40-98.16)	92.36% (89.58-94.85)	92.79% (88.39-95.11)
	InfantFS	82.69% (71.84-88.24)	76.65% (64.96-85.40)	75.44% (68.49-79.62)
at-term	Our model	98.19% (97.59-98.47)	95.31% (92.45-96.99)	94.42% (91.86-95.33)
	InfantFS	87.42% (84.21-88.95)	82.59% (78.27-86.55)	79.58% (77.14-83.94)
8-year	Our model	97.70% (97.22-98.03)	96.88% (92.08-97.91)	96.24% (95.17-97.12)
	FreeSurfer	82.78% (78.69-84.51)	85.78% (81.38-88.26)	60.94% (58.13-63.82)
PC preterm	Our model	91.24% (88.23-95.38)	-	-
PC at-term	Our model	92.24% (84.38-96.70)	-	-

B: 3D Hausdorff Distance

		Cerebrum	**Cerebellum**	**Brainstem**
preterm	Our model	07.38 (05.10-11.70)	03.04 (01.73-05.83)	02.99 (01.41-07.28)
	InfantFS	10.75 (09.38-13.38)	06.50 (05.00-17.97)	08.14 (06.16-11.36)
at-term	Our model	08.59 (05.10-14.49)	03.10 (02.24-06.48)	03.04 (02.00-05.48)
	InfantFS	13.28 (11.83-15.33)	07.30 (06.08-08.12)	07.30 (06.08-08.60)
8-year	Our model	20.27 (18.06-22.41)	05.32 (03.00-08.54)	03.68 (02.24-06.08)
	FreeSurfer	18.71 (15.39-21.21)	14.88 (13.30-16.55)	29.52 (21.77-34.73)
PC preterm	Our model	12.39 (07.87-18.11)	-	-
PC at-term	Our model	16.15 (13.08-19.52)	-	-

3. Experiments & Results

3.1. Training Parameters

Three hundred scans and labels, split evenly for early life, term and 8 years of age were randomly selected from our database, with 100 from each group. An even distribution of data, a 60/20/20 split was applied across each group to create 180 training, 60 validation, and 60 test points. The validation set was used purely as a parameter to detect early-stopping, while the test dataset was held-out for final evaluation. The model was set to train for up to 200 epochs, and early-stopping took effect at 83 epochs.

3.2. Evaluation of the Dice Similarity Coefficients

On a held-out test set, the trained model achieved an average DSC of 97.78% (95.40%-98.47%) for predicting the cerebrum label, 94.85% (89.58%-97.91%) for the cerebellum label, and 94.84% (88.39%-97.12%) for the brainstem label. Details can be found in Table 1a, Figure 2a. The model, understandably, performed better on the cerebrum as it is much larger, giving the machine learning model more data to train on. However, the cerebellum & brainstem label segmentations still performed objectively well. More importantly, our model performed similarly well across age groups. There is roughly a 1%, 4% and 4% delta between the averages of the cerebrum, cerebellum and brainstem labels respectively.

To better visualize the model predictions, a random slice from the scans that achieved the median DSC in each age group are plotted in Figure 3.

3.3. Evaluation of Hausdorff Distances

The model achieved, on average, a Hausdorff Distance (HD) of 12.08 (5.10-22.41) for predicting the cerebrum label, 3.82 (1.73-8.54) for the cerebellum label, and 3.24 (1.41-7.28) for the brainstem label. More details can be found in Table 1b and Figure 2c.

Unlike the DSC, the HD is not normalized and scales with the overall scan volume. On average, the brain volume of an at-term scan and that of an 8-year scan are 1.91 and 9.55 times to that of a preterm scan. However, the average HD measured on the term and 8-year scans grew by only 1.20 and 2.76 times, indicating that the model performance is in fact improving on larger scans - an effect which is visible when looking at the DSC as well.

3.4. Evaluating against an Independent Cohort

The presented model was then evaluated using data from an independent cohort, Preterm Care (PC) Cohort, which consisted of 6 scans of neonates with ventriculomegaly, where the only label available was the cerebrum. The 6 scans were split between preterm and at-term subjects. The model was able to achieve an average DSC of 91.74% (84.38%-96.70%), and an average HD of 14.27 (7.87-19.52). The breakdown of the model performance on each age group can be found under Table 1.

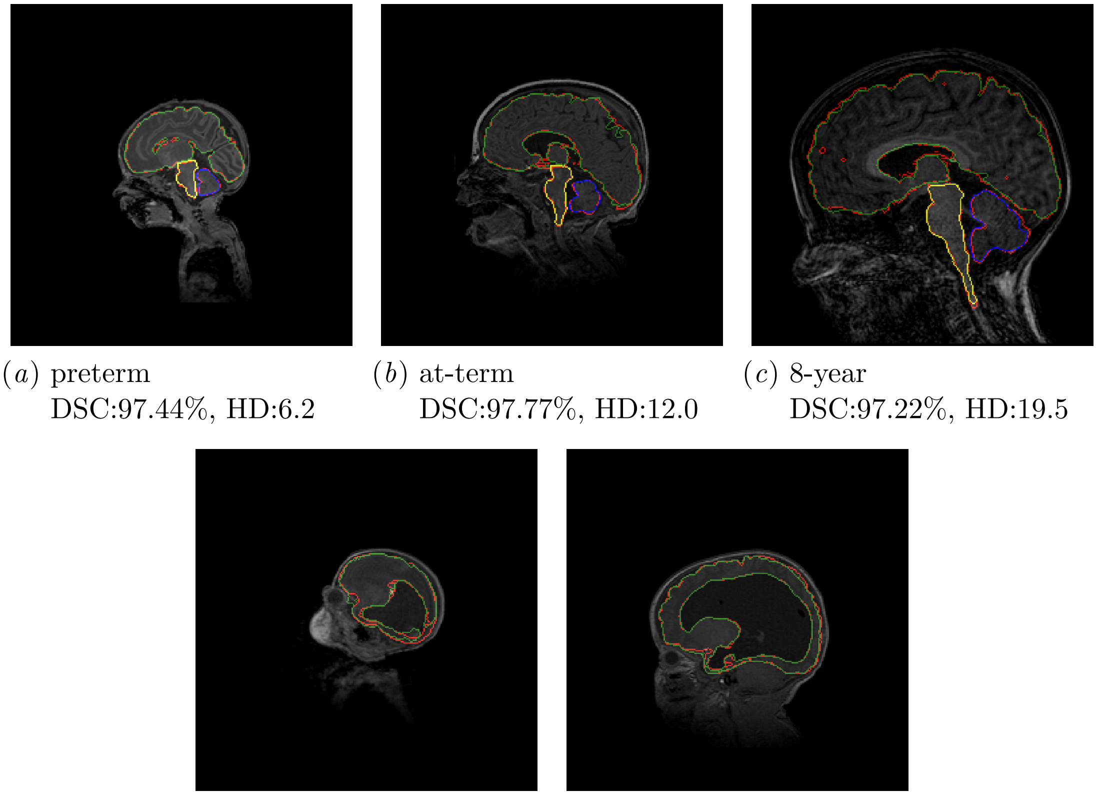
3.5. Comparison to InfantFS and FreeSurfer

The same test dataset was segmented by InfantFS and FreeSurfer. The preterm and at-term scans were segmented by InfantFS, while the 8-year scans were segmented by the adult version of FreeSurfer. On test data, the InfantFS/FreeSurfer tool achieved an average DSC of 84.30% (71.84%-88.95%) for predicting the cerebrum label, 81.67% (64.96%-88.26%) for the cerebellum label, and 71.99% (58.13%-83.94%) for the brainstem labels, respectively. When compared to the segmentations of our model, the InfantFS/FreeSurfer predictions were, on average, 13-23% lower in DSC. More detailed analysis of the InfantFS/FreeSurfer predictions can be found in Figure 2b.

Similarly, InfantFS/FreeSurfer achieved a lower performance on the HD, with an average of 14.25 (9.38-21.21) on the cerebrum label, 9.56 (5.00-17.97) on the cerebellum label, and 14.99 (6.08-34.73) on the brainstem label. Comparing these values across the age groups, the average HD grew by 1.3 times between the preterm and at-term age groups, and 4.0 times between the at-term and 8-year age groups. This shows that the InfantFS/FreeSurfer tool performs more equally across the age groups, however still under-performs our model (Table 1b and Figure 2d).

4. Discussion

This paper presents a U-Net based pipeline that can robustly and accurately segments multiple structures of the developing brain in very preterm children (born 24-32 weeks gestation) on both their neonatal and school-age images. Our model outperformed the FreeSurfer/InfantFS tools across the age spectrum. It is noteworthy that our model achieved



(*a*) preterm
DSC:97.44%, HD:6.2

(*b*) at-term
DSC:97.77%, HD:12.0

(*c*) 8-year
DSC:97.22%, HD:19.5

(*d*) PC preterm
DSC:90.12%, HD:18.1

(*e*) PC at-term
DSC:95.65%, HD:13.1

Figure 3: Visualization of the scans that achieved the median DICE coefficient in each age group. The labels are as follows: red is the ground-truth for all labels, green is the predicted cerebrum, blue is the predicted cerebellum, and yellow is the predicted brainstem.

very similar segmentation accuracy for images from the three age groups, demonstrating its general applicability to brains of large anatomical variability. Furthermore, our model can robustly segment brain images with severe anatomical abnormality (e.g. severe ventriculomegaly).

Our pipeline employs preprocessing steps that centre the brain and normalize image intensity to minimize bias from outliers on model training. Training on 2D slices eliminates biases introduced by scan parameters, such as slice thickness, permits better pattern learning, and is much less computationally demanding. While our model unifies the size of the data through symmetrically zero-padding the scan without altering the resolution. The model was trained on MRI scans alone, and the authors are confident to expand the model to predict more classes (i.e. more regions & sub-regions), as well as a wider age demographic. Our code and model will be made freely available (open source) on Github, with plans to release a Dockerized version as well.

References

Dalit Cayam-Rand, Ting Guo, Anne Synnes, Vann Chau, Connor Mabbott, Isabel Benavente-Fernández, Ruth E Grunau, and Steven P Miller. Interaction between preterm white matter injury and childhood thalamic growth. *Annals of neurology*, 90(4):584–594, 2021.

Vann Chau, Kenneth J Poskitt, Deborah E McFadden, Tim Bowen-Roberts, Anne Synnes, Rollin Brant, Michael A Sargent, Wendy Soulikias, and Steven P Miller. Effect of chorioamnionitis on brain development and injury in premature newborns. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 66(2):155–164, 2009.

Ahmed E Fetit, Amir Alansary, Lucilio Cordero-Grande, John Cupitt, Alice B Davidson, A David Edwards, Joseph V Hajnal, Emer Hughes, Konstantinos Kamnitsas, Vanessa Kyriakopoulou, et al. A deep learning approach to segmentation of the developing cortex in fetal brain mri with minimal manual labeling. In *Medical Imaging with Deep Learning*, pages 241–261. PMLR, 2020a.

Ahmed E Fetit, John Cupitt, Turkay Kart, and Daniel Rueckert. Training deep segmentation networks on texture-encoded input: application to neuroimaging of the developing neonatal brain. In *Medical Imaging with Deep Learning*, pages 230–240. PMLR, 2020b.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Jarred Garfinkle, Ting Guo, Anne Synnes, Vann Chau, Helen M Branson, Steven Ufkes, Emily WY Tam, Ruth E Grunau, and Steven P Miller. Location and size of preterm cerebellar hemorrhage and childhood development. *Annals of Neurology*, 88(6):1095–1108, 2020.

Ioannis S Gousias, Alexander Hammers, Serena J Counsell, Latha Srinivasan, Mary A Rutherford, Rolf A Heckemann, Jo V Hajnal, Daniel Rueckert, and A David Edwards. Magnetic resonance imaging of the newborn brain: automatic segmentation of brain images into 50 anatomical regions. *PloS one*, 8(4):e59990, 2013.

Mireille Guillot, Ting Guo, Steven Ufkes, Juliane Schneider, Anne Synnes, Vann Chau, Ruth E Grunau, and Steven P Miller. Mechanical ventilation duration, brainstem development, and neurodevelopment in children born preterm: a prospective cohort study. *The Journal of pediatrics*, 226:87–95, 2020.

Ting Guo, Julie L Winterburn, Jon Pipitone, Emma G Duerden, Min Tae M Park, Vann Chau, Kenneth J Poskitt, Ruth E Grunau, Anne Synnes, Steven P Miller, et al. Automatic segmentation of the hippocampus for preterm neonates from early-in-life to term-equivalent age. *NeuroImage: Clinical*, 9:176–193, 2015.

Meichen Liu, Xin Yan, Chenhui Wang, and Kejun Wang. Segmentation mask-guided person image generation. *Applied Intelligence*, 51(2):1161–1176, 2021.

Antonios Makropoulos, Ioannis S Gousias, Christian Ledig, Paul Aljabar, Ahmed Serag, Joseph V Hajnal, A David Edwards, Serena J Counsell, and Daniel Rueckert. Automatic whole brain mri segmentation of the developing neonatal brain. *IEEE transactions on medical imaging*, 33(9):1818–1831, 2014.

Pim Moeskops, Max A Viergever, Adriënne M Mendrik, Linda S De Vries, Manon JNL Benders, and Ivana Išgum. Automatic segmentation of mr brain images with a convolutional neural network. *IEEE transactions on medical imaging*, 35(5):1252–1261, 2016.

Mahmoud Mostapha and Martin Styner. Role of deep learning in infant brain mri analysis. *Magnetic resonance imaging*, 64:171–189, 2019.

Mitsuhiro Nishida, Nikolaos Makris, David N Kennedy, Mark Vangel, Bruce Fischl, Kalpathy S Krishnamoorthy, Verne S Caviness, and P Ellen Grant. Detailed semiautomated mri based morphometry of the neonatal brain: preliminary results. *Neuroimage*, 32(3): 1041–1049, 2006.

Marcel Prastawa, John H Gilmore, Weili Lin, and Guido Gerig. Automatic segmentation of mr images of the developing newborn brain. *Medical image analysis*, 9(5):457–466, 2005.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

Lin Shi, Defeng Wang, Winnie CW Chu, Geoffrey R Burwell, Tien-Tsin Wong, Pheng Ann Heng, and Jack CY Cheng. Automatic mri segmentation and morphoanatomy analysis of the vestibular system in adolescent idiopathic scoliosis. *Neuroimage*, 54:S180–S188, 2011.

Abdel Aziz Taha and Allan Hanbury. An efficient algorithm for calculating the exact hausdorff distance, 2015.

Deanne K Thompson, Lillian G Matthews, Bonnie Alexander, Katherine J Lee, Claire E Kelly, Chris L Adamson, Rod W Hunt, Jeanie LY Cheong, Megan Spencer-Smith, Jeffrey J Neil, et al. Tracking regional brain growth up to age 13 in children born term and very preterm. *Nature communications*, 11(1):1–11, 2020.

Li Wang, Feng Shi, Gang Li, Yaozong Gao, Weili Lin, John H Gilmore, and Dinggang Shen. Segmentation of neonatal brain mr images using patch-driven level sets. *NeuroImage*, 84: 141–158, 2014.

Li Wang, Yaozong Gao, Feng Shi, Gang Li, John H Gilmore, Weili Lin, and Dinggang Shen. Links: Learning-based multi-source integration framework for segmentation of infant brain images. *NeuroImage*, 108:160–172, 2015.

Neil I Weisenfeld and Simon K Warfield. Automatic segmentation of newborn brain mri. *Neuroimage*, 47(2):564–572, 2009.

Wen Xue, Lars Zender, Cornelius Miething, Ross A Dickins, Eva Hernando, Valery Krizhanovsky, Carlos Cordon-Cardo, and Scott W Lowe. Senescence and tumour clearance is triggered by p53 restoration in murine liver carcinomas. *Nature*, 445(7128):656–660, 2007.

Sihang Zhou, Dong Nie, Ehsan Adeli, Jianping Yin, Jun Lian, and Dinggang Shen. High-resolution encoder–decoder networks for low-contrast medical image segmentation. *IEEE Transactions on Image Processing*, 29:461–475, 2019.

Lilla Zöllei, Juan Eugenio Iglesias, Yangming Ou, P Ellen Grant, and Bruce Fischl. Infant freesurfer: An automated segmentation and surface extraction pipeline for t1-weighted neuroimaging data of infants 0–2 years. *Neuroimage*, 218:116946, 2020.

Kelly H Zou, Simon K Warfield, Aditya Bharatha, Clare MC Tempany, Michael R Kaus, Steven J Haker, William M Wells III, Ferenc A Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports, 2004.

Appendix A. Outlier Correction Post-processing Step

One of the key post-processing steps performed in this paper was for outlier correction. This occurred when the model predicted certain regions that were quite far from the subject's skull, thus being an invalid prediction. Such cases are very obvious to identify visually, however we wanted to automate the identification & correction of such issues to help make the pipeline smoother.

Such a step would be trivial to perform on an adult dataset, as adult skull sizes are generally the same size (or close enough). However, due to the fact that our age demographic ranges from preterm neonates all the way up to 8-years of age, there is such a significant amount of variability to the skull size. This makes it very hard to estimate where the skull is, and by extension where the boundary of ”valid” predictions would be.

Our solution was a fairly elegant one. We simply viewed the full 3D predicted mask and marked the largest ”blob” as our main object, and marked any other 3D ”blob” that was unconnected to this main object and was over 40 pixels away radially (edge-to-edge) as an invalid prediction. A visual representation of this step can be seen under Figure 4.

This one step helped stabilize our evaluation metrics significantly. Although the DICE coefficients did not change by much (since the ”invalid” predictions were often just a few stray pixels), it did heavily impact our Hausdorff distance, resulting in a 3-4x reduction there.

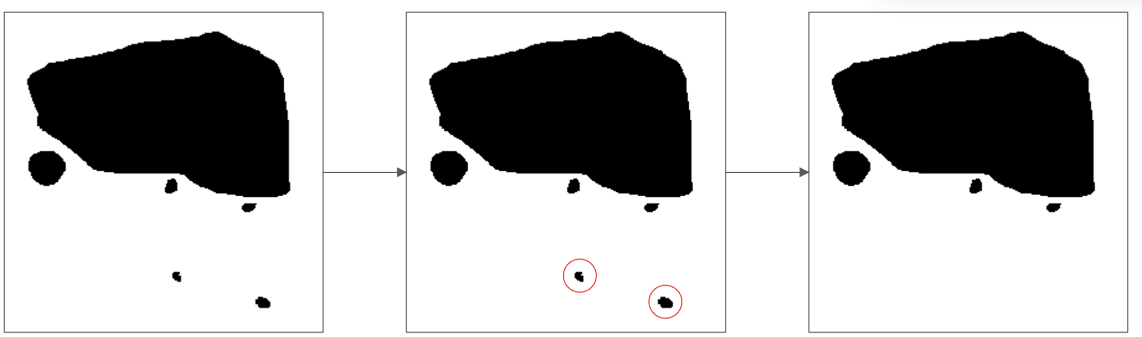


Figure 4: A 2D visual representation of how the outlier correction operates. All 'blobs' that are non-connected to the primary 'blob' and are over 20% further radially get marked as an outlier, and unset

Appendix B. Dice Similarity Coefficient

The Dice Similarity Coefficient (DSC) can be formally defined as:

$$DSC = \frac{2 \mid GT \cap PM \mid}{\mid GT \mid + \mid PM \mid} x100 \tag{1}$$

In the above formula, GT and PM are the ground-truth and prediction model respectively. A DSC of 0 indicates that there is no overlap between GT and PM, while a maximum value of 100 indicates that the different masks perfectly overlap.

Appendix C. Hausdorff Distance

The Hausdorff Distance (HD) is a measure of the maximal contour distance between two objects. More formally:

$$d(x \rightarrow y) = max(d_i^{x \rightarrow y}), i = 1...N_x \tag{2}$$

$$HD(GT, PM) = max[d(GT \rightarrow PM), d(PM \rightarrow GT)] \tag{3}$$

where GT and PM are the ground-truth and prediction model respectively.

HD is the greatest of all distances from a 3D point (voxel) in GT to the closest voxel in PM. A smaller HD indicates that the predicted and ground-truth brain masks are more similar, while a larger Hausdorff distance indicates that the two masks have a greater number of differences between them.