ANALYSING THE IMPACT OF GENDER ON TRAINING BRAIN SEGMENTATION NEURAL NETWORKS

David Cooksley, Evan Cooksley, Brendan Mahler, Khoi Nguyen

University of Calgary
Department of Electrical and Computer Engineering
2500 University Drive NW Calgary Alberta

ABSTRACT

A common application of neural networks in the medical field is image segmentation. When interpreting MRI brain images, using neural networks to segment the image can greatly improve the efficiency of the process, but training these models can be time consuming and expensive. In the interest of improving the viability of this process, we investigate the impacts of using a gender-imbalanced data set on the performance of U-net neural image segmentation models. We trained multiple models on all female, all male, and a mixed-gender set of images, and investigated each model's performance on both male and female test images. Our experiment found no significant variation in the Dice coefficient scores of brain segmentation performed by models trained using single or mixed-gender data sets.

1. INTRODUCTION

Magnetic Resonance Imaging (MRI) is a technology that is used to generate detailed pictures of different parts of the human body. One application is using MRI images to analyse brain health, which can help doctors to identify conditions or injuries within the brain. A major issue with MRI scans is the lengthy data acquisition which leads to long wait times and patient discomfort. The estimated average wait time for an MRI in Canada in 2022 is 133 days [1]. Applying deep learning methods to MRI scans can be an effective way to shorten the data acquisition time. MRI's have advantages over other scanning methods, such as CT scans, including no radiation exposure and clearer images. Recent studies indicate that gender may have a substantial influence on human cognitive functions, including emotion, memory, perception, etc. [2]. Understanding these differences can be important in ruling out certain conditions based on gender. Accurate detection of brain tissues is important and MRI images can be used to detect these conditions at an early stage.

Our project looks at understanding the effects of using an unbalanced data set, containing only males or females, to train a neural-network image segmentation model used to identify different sections of the brain. We designed our experiment to train several models on varying input data, and investigated the differences in performance. This experiment could help to understand the best way to gather a data set to train models which could be used to shorten the lengthy wait times for MRI's. Additionally, training a neural network for MRI scans is computationally expensive, taking days of time and/or multiple high-end processors. Understanding the effects of training on male and female brains could help improve this as a more efficient data set could be selected for training. Another advantage is that fewer experts on brain image interpretation would be required, as analysis of the MRI images could be sped up using deep learning [3].

2. RELATED WORK

Brain segmentation is not the only medical field where machine learning can have an impact. Machine learning is becoming much more common in medicine, seeing use in disease diagnosis [4, 5], specifically for diseases such as breast cancer [6] or diabetic retinopathy [7], as well as various forms of medical imaging [8, 9, 10, 11]. It can be applied to electronic health records to create actionable insights, improving patient risk score systems and streamlining hospital operations [12]. Machine learning is not, however, without its problems; there are many ethical concerns with the use of machine learning for medicine, including concerns about the privacy of data sourcing, potential for bias, and the lack of transparency, as discussed by Vayena et al. [13]. Additionally, Cabitza et al. raise concerns about some of the potential unintended consequences of integrating machine learning into medicine, such as reducing the skill of physicians, focusing on data without context, and the inability to account for the intrinsic uncertainty in medicine [14].

3. MATERIALS AND METHODS

3.1. Dataset

The data set used for this study was the Calgary-Campinas-359 (CC-359) data set of multi-vendor, multi-field strength magnetic resonance T1-weighted volumetric brain images

[15]. This data set was composed of healthy brain images from adults aged 29-80, along with corresponding brain masks generated using the STAPLE algorithm [16], as well as white-matter, grey-matter, and cerebrospinal fluid masks (WM-GM-CSF). Images 340-359 were excluded from this study as our available WM-GM-CSF masks did not include these subjects, resulting in a final data set of 340 subjects. The images were acquired on scanners from three vendors (General Electric, Philips, and Siemens) at field strengths of 1.5T and 3T. The data set was mostly balanced between each gender, vendor, and field strength, typically with 30 subjects of each. There were two exceptions: the images from Philips at 1.5T had 26 male and 33 female subjects, and the images from General Electric at 3T had 23 male and 18 female subjects.

3.2. Experimental Setup

This project investigated the influence of gender in developing deep learning models for brain images. Segmentation models were trained using male and female subjects, only male subjects, and only female subjects. These models were then used to make predictions on subjects of both genders to compare the performance of models trained on a single gender to models trained on both genders, and to determine if models trained on solely male subjects would perform adequately for predictions on female subjects and vice-versa.

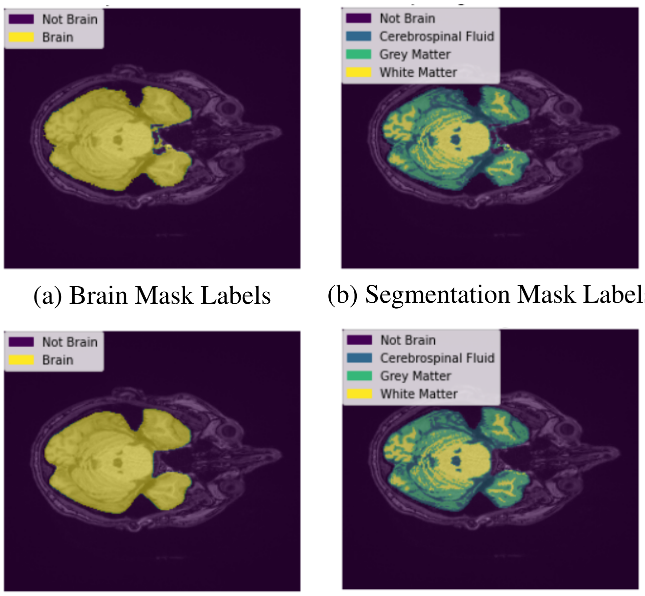
The model used for this project was a 2D multitask U-Net model that was first used to create a segmentation mask for the brain, then using the brain mask created additional segmentation masks for white matter, grey matter, and cerebrospinal fluid. Model and visualization code were adapted from work by R.M. Souza [17]. A visualization of the input labeled data and the output labeled masks can be seen in Figure 1.

The loss function used was the negative dice coefficient, which compares the set of predictions (X) with the set of actual labels (Y), producing a value between 0 and 1. For the multi-label task involving WM-GM-CSF, the dice coefficient was evaluated individually for each label.

$$DiceCoefficient = \frac{2|X \cap Y|}{|X| + |Y|}$$

Due to lack of computational resources, a study involving a 3-dimensional convolutional neural network was not feasible for this project. Instead, the problem was treated as a 2-dimensional problem by splitting each 3D image into 2D slices in the axial plane, resulting in over 100 2D images produced for each volume. Slices consisting of blank images were removed. The images were then scaled using min-max normalization.

For the model trained under the full data set, the training, validation, and test sets were split using a ratio of approximately 70:15:15. This resulted in 244 training subjects, 48 validation subjects, and 48 test subjects. Due to the gender split, the training sets for the single gender models had half as many images compared to the combined set. As such, another model was trained using both genders with half of the training set size in order to have a more fair comparison with the models trained on a single gender. Each set was selected randomly, making sure to keep an equal balance between vendors and field strength, to control for any variation due to those factors.



(a) Brain Mask Labels (b) Segmentation Mask Labels

(c) Brain Mask Predictions (d) Segmentation Mask Predictions

Fig. 1. Example Brain Labels and Predictions

4. RESULTS AND DISCUSSION

4.1. Experiment Results

A summary of the Dice coefficient scores obtained by the female-trained model can be found in Figure 2, and those obtained by the male-trained model can be found in Figure 3. Both models performed nearly identically on all four target masks for both the female and the male test sets. While the female-trained model performed better on the female test set by a slight margin, and the male-trained model performed slightly better on the male test set, it is not statistically significant for any of the scores ($p \geq 0.12$).

For the models trained on balanced data sets, the scores obtained by the model trained on the full data set can be found in Figure 4, and the scores obtained by the model trained on the half-sized data set can be found in Figure 5. The model trained on the full set of data performed slightly better than the other three models on the grey matter and cerebrospinal fluid masks, but beyond that there is no significant difference between the models' performances. The only model with any improvement in performance is the one with extra training

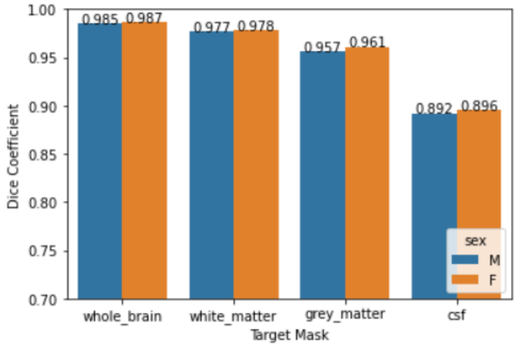


Fig. 2. Female-trained Model Test Scores by Gender

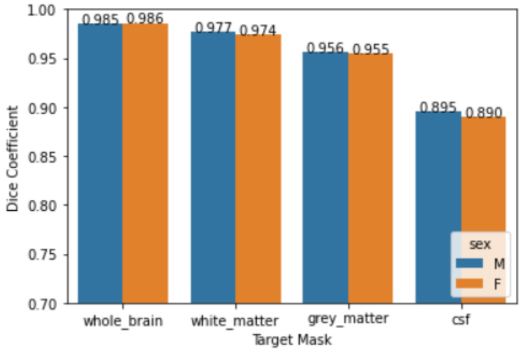


Fig. 3. Male-trained Model Test Scores by Gender

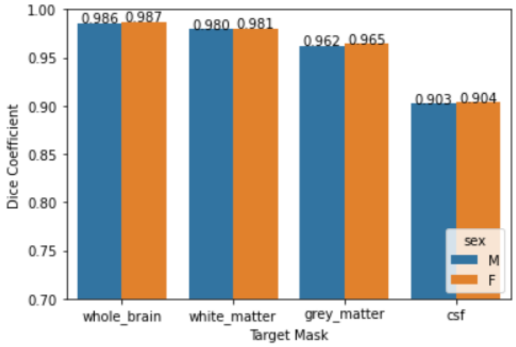


Fig. 4. Full-set-trained Model Test Scores by Gender

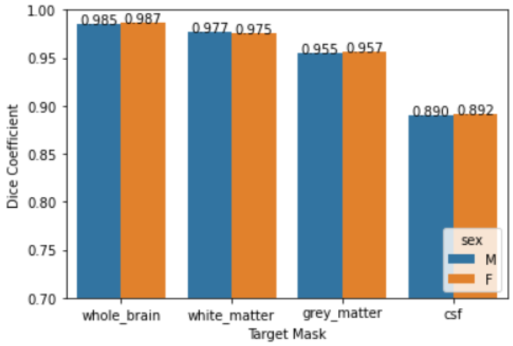


Fig. 5. Half-set-trained Model Test Scores by Gender

data. It is clear from these results that there is no significant difference in the performance of models trained on all female, all male, or a gender-balanced data set.

4.2. Discussion

As is further visualized in Figure 6, which shows the mean Dice coefficient of each model across the entire test set for each target mask, there is no apparent benefit or detriment to training this U-net model on only images of brains from subjects of a single gender. This suggests that there is a small enough difference between male and female brains that it doesn't matter for segmentation problems. However, this result is far from conclusive; there is definitely room for future experimentation along this line. In our results for the whole brain and white matter masks, the model trained on twice as much data did not perform notably better than the other models. This suggests that we have more data than is strictly necessary to train the models. It is conceivable that we could see more variance in the performance of the single-gender models if they are trained on a smaller sample set.

Another factor to consider is that this model is trained entirely on images of healthy, older adult brains. It is possible that there could be more substantial differences in how neurodegenerative diseases present in male vs. female brains. Additionally, younger subjects could potentially have more notable differences, as female brains generally develop faster than male ones [18]. Another possible line of experimentation could be in using different styles of neural networks. While the U-net model that we used saw no effect from training on a single gender, it's possible that another model could see more of a difference.

5. CONCLUSION

Our study trained a u-net segmentation model on four different data sets: an only male data set, an only female data set, a similar sized mixed data set, and a double sized mixed data set. We found that there were minimal differences between models trained on a single gender vs. models trained on both genders. While there were small improvements on the scores for models evaluating the gender they were trained on, they were not statistically significant ($p \geq 0.12$). There were minimal improvements in score for the model trained on a larger data set, despite there being twice as much data. Further study is needed to determine if another model, a smaller training set, or different demographics in the input data set would result in more varied results.

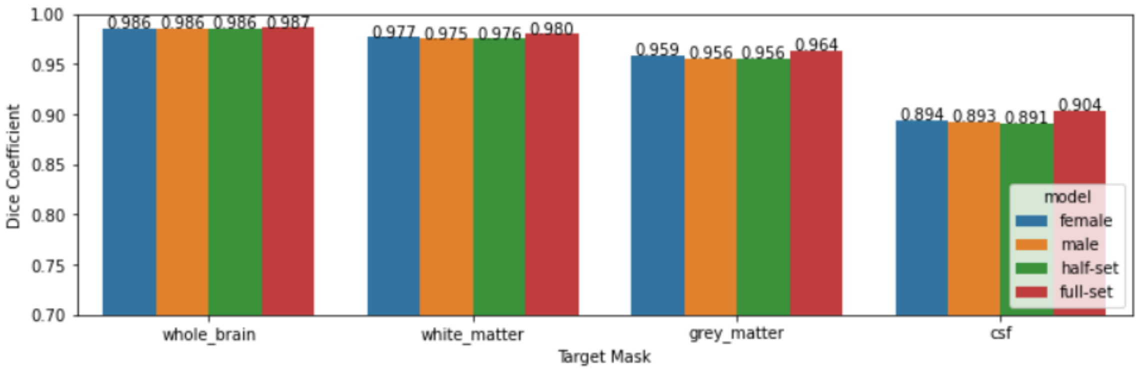


Fig. 6. Summary of Model Scores

6. PROJECT REPOSITORY

Project code can be found at https://github.com/brendan-mahler/enel645-final-project

7. REFERENCES

[1] Greg Sutherland, Nigel Russel, Robyn Gibbard, and Alexandru Dobrescu, "The value of radiology, part ii," 2019.

[2] Larry Cahill, "Why sex matters for neuroscience," *Nature reviews. Neuroscience*, vol. 7, pp. 477–84, 07 2006.

[3] Patricia Johnson, Michael Recht, and Florian Knoll, "Improving the speed of mri with artificial intelligence," *Seminars in Musculoskeletal Radiology*, vol. 24, pp. 012–020, 02 2020.

[4] Kunio Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4, pp. 198–211, 2007, Computer-aided Diagnosis (CAD) and Image-guided Decision Support.

[5] Igor Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[6] Jenni A. M. Sidey-Gibbons and Chris J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *BMC Medical Research Methodology*, vol. 19, no. 1, 2019.

[7] Priya R. and Aruna P., "Diagnosis of diabetic retinopathy using machine learning techniques," *ICTACT Journal on Soft Computing*, vol. 03, no. 04, pp. 563–575, 2013.

[8] Maryellen L. Giger, "Machine learning in medical imaging," *Journal of the American College of Radiology*, vol. 15, no. 3, Part B, pp. 512–520, 2018, Data Science: Big Data Machine Learning and Artificial Intelligence.

[9] Miles Wernick, Yongyi Yang, Jovan Brankov, Grigori Yourganov, and Stephen Strother, "Machine learning in medical imaging," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 25–38, 2010.

[10] June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim, "Deep learning in medical imaging: General overview," *kjr*, vol. 18, no. 4, pp. 570–584, 2017.

[11] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert Müller, "Introduction to machine learning for brain imaging," *NeuroImage*, vol. 56, no. 2, pp. 387–399, 2011, Multivariate Decoding and Brain Reading.

[12] Alison Callahan and Nigam H. Shah, "Chapter 19 - machine learning in healthcare," in *Key Advances in Clinical Informatics*, Aziz Sheikh, Kathrin M. Cresswell, Adam Wright, and David W. Bates, Eds., pp. 279–291. Academic Press, 2017.

[13] Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen, "Machine learning in medicine: Addressing ethical challenges," *PLOS Medicine*, vol. 15, no. 11, pp. e1002689, 2018.

[14] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini, "Unintended Consequences of Machine Learning in Medicine," *JAMA*, vol. 318, no. 6, pp. 517–518, 08 2017.

[15] Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Letícia Rittner, Richard Frayne, and Roberto Lotufo, "An open,

multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement," *NeuroImage*, vol. 170, pp. 482–494, 2018, Segmenting the Brain.

[16] S.K. Warfield, K.H. Zou, and W.M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[17] R.M. Souza, "Enel 645 tutorial 12: Medical image segmentation - unet," https://github.com/rmsouza01/ENEL645/tree/master/SLURM/unet-segmentatio, Feb 2022, accessed: 2022-03-16.

[18] Rhoshel K. Lenroot and Jay N. Giedd, "Sex differences in the adolescent brain," *Brain and Cognition*, vol. 72, no. 1, pp. 46–55, 2010.