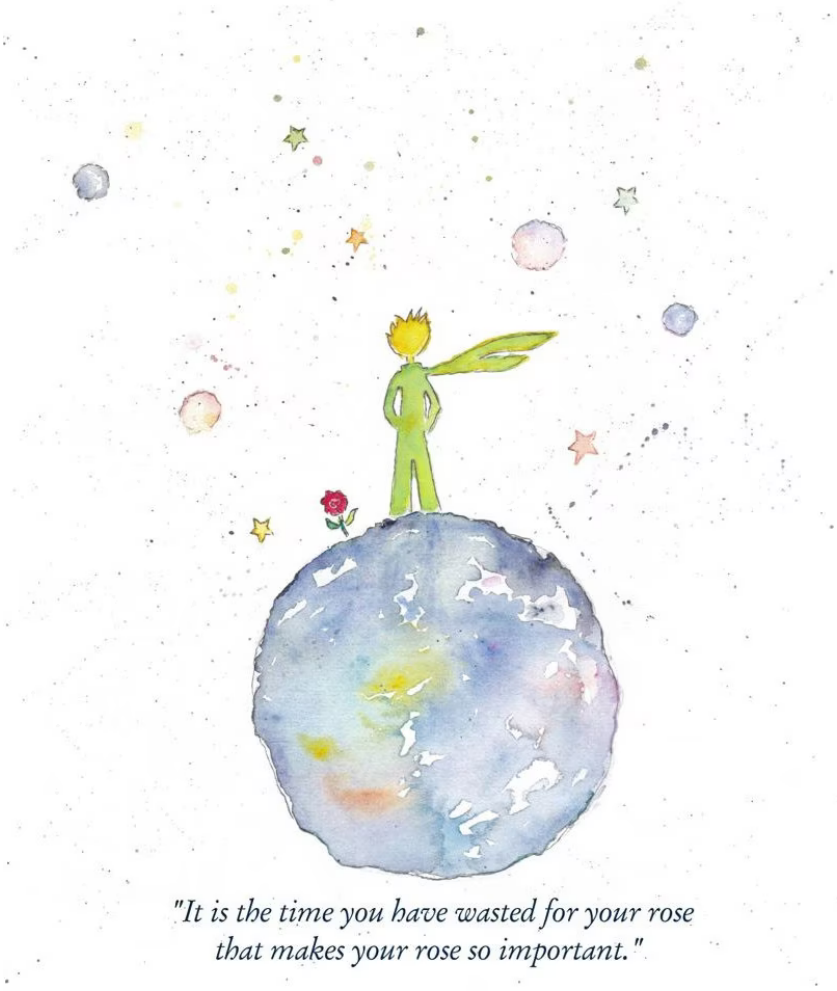
Integrating Bayesian Deep Learning Uncertainties in Medical Image Analysis

Raghav Mehta
Department of Electrical and Computer Engineering
McGill University, Montreal
April 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

To,
poeple who stuck with me in all ups and downs



— *Antoine de Saint-Exupéry,*
The Little Prince

Abstract

Although Deep Learning (DL) models have been shown to perform very well on various medical imaging tasks, inference in the presence of pathology presents several challenges to common models. These challenges impede the integration of DL models into real clinical workflows. Deployment of these models into real clinical contexts requires: (1) that the confidence in DL model predictions be accurately expressed in the form of uncertainties and (2) that they exhibit robustness and fairness across different sub-populations. Quantifying the reliability of DL model predictions in the form of uncertainties could enable clinical review of the most uncertain regions, thereby building trust and paving the way toward clinical translation. Similarly, by embedding uncertainty estimates across cascaded inference tasks, prevalent in medical image analysis, performance on the downstream inference tasks should also be improved. In this thesis, we develop an uncertainty quantification score for the task of Brain Tumour Segmentation. We evaluate the score's usefulness during the two consecutive Brain Tumour Segmentation (BraTS) challenges, BraTS 2019 and BraTS 2020. Overall, our findings confirm the importance and complementary value that uncertainty estimates provide to segmentation algorithms, highlighting the need for uncertainty quantification in medical image analyses. We further show the importance of uncertainty estimates in medical image analysis by propagating uncertainty generated by upstream tasks into the downstream task of interest. Our results on three different clinically relevant tasks indicate that uncertainty propagation helps improve the performance of the downstream task of interest. Additionally, we combine the aspect of uncertainty estimates with fairness across demographic subgroups into the picture. By performing extensive experiments on multiple tasks, we show that popular ML methods for achieving fairness across different subgroups, such as data-balancing and distributionally robust optimization, succeed in terms of the model performances for some of the tasks. However, this can come at the cost of poor uncertainty estimates associated with the model predictions. This tradeoff must be mitigated if fairness models are to be adopted in medical image analysis. In the last part of the thesis, we look at

Active Learning (AL) for reduced manual labeling of a dataset. Specifically, we present an information-theoretic active learning framework that guides the optimal selection of images for labeling. Results indicate that the proposed framework outperforms several existing AL methods, and by careful design choices, it can be integrated into existing deep learning classifiers with minimal computational overhead.

Abrégé

Bien qu'il ait été démontré que les modèles d'apprentissage en profondeur (DL) fonctionnent très bien sur diverses tâches d'imagerie médicale, l'inférence en présence de pathologie présente plusieurs défis pour les modèles courants. Ces défis entravent l'intégration des modèles DL dans les flux de travail cliniques réels. Le déploiement de ces modèles dans des contextes cliniques réels nécessite : (1) que la confiance dans les prédictions du modèle DL soit exprimée avec précision sous la forme d'incertitudes, et (2) qu'ils présentent une robustesse et une équité dans différentes sous-populations. La quantification de la fiabilité des prédictions du modèle DL sous la forme d'incertitudes pourrait permettre un examen clinique des régions les plus incertaines, renforçant ainsi la confiance et ouvrant la voie à la traduction clinique. De même, en intégrant les estimations d'incertitude dans les tâches d'inférence en cascade, courantes dans l'analyse d'images médicales, les performances des tâches d'inférence en aval devraient également être améliorées. Dans cette thèse, nous développons un score de quantification de l'incertitude pour la tâche de segmentation des tumeurs cérébrales. Nous évaluons l'utilité du score lors des deux défis consécutifs de segmentation des tumeurs cérébrales (BraTS), BraTS 2019 et BraTS 2020. Dans l'ensemble, nos résultats confirment l'importance et la valeur complémentaire que les estimations d'incertitude apportent aux algorithmes de segmentation, soulignant la nécessité d'une quantification de l'incertitude dans l'imagerie médicale. analyses. Nous montrons en outre l'importance des estimations d'incertitude dans l'analyse d'images médicales en propageant l'incertitude générée par les tâches en amont dans la tâche d'intérêt en aval. Nos résultats sur trois tâches différentes cliniquement pertinentes indiquent que la propagation de l'incertitude contribue à améliorer les performances de la tâche d'intérêt en aval. De plus, nous combinons l'aspect des estimations d'incertitude avec l'équité entre les sous-groupes démographiques dans l'image. En effectuant des expériences approfondies sur plusieurs tâches, nous montrons que les méthodes ML populaires pour atteindre l'équité entre différents sous-groupes, telles que

l'équilibrage des données et l'optimisation robuste de la distribution, réussissent en termes de performances du modèle pour certaines des tâches. Cependant, cela peut se faire au prix de mauvaises estimations de l'incertitude associées aux prévisions du modèle. Ce compromis doit être atténué si des modèles d'équité doivent être adoptés dans l'analyse d'images médicales. Dans la dernière partie de la thèse, nous nous intéressons à l'Active Learning (AL) pour un étiquetage manuel réduit d'un jeu de données. Plus précisément, nous présentons un cadre d'apprentissage actif théorique de l'information qui guide la sélection optimale des images pour l'étiquetage. Les résultats indiquent que le cadre proposé surpasse plusieurs méthodes AL existantes et, grâce à des choix de conception minutieux, il peut être intégré dans les classificateurs d'apprentissage en profondeur existants avec une surcharge de calcul minimale.

Contributions of Author

Related Publications

It should be noted that this is not a manuscript based thesis. However, considerable material from the following papers has been utilised in the thesis.

Peer-Reviewed Journal Articles

- **R. Mehta**, A. Filos, U. Baid, . . . , S. Bakas, Y. Gal, T. Arbel, (95 authors), "QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation - Analysis of Ranking Scores and Benchmarking Results", *The Journal of Machine Learning for Biomedical Imaging (MELBA), August 2022* [161].

The majority of the content of Chapter 3 was published in the Journal of Machine Learning for Biomedical Imaging (MELBA) 2022 (as the above citation). In the published work, I am the lead author. I was responsible for developing a metric to evaluate uncertainties for the brain tumour segmentation task, organizing a sub-challenge on Quantification of Uncertainty in Brain Tumour Segmentation (QU-BraTS) as a part of Brain Tumour Segmentation (BraTS) challenge for two consecutive years (2019 and 2020), writing code to evaluate output provided by challenge participants, performing all the analysis related to the challenge, and writing the original draft of the manuscript. Angelos Filos (Ph.D.

candidate at University of Oxford) helped during the development of the metric and related evaluation code. Dr. Ujjwal Baid (Postdoctoral fellow at University of Pennsylvania) helped during the analysis related to the challenge. Prof. Spyridon Bakas (Assistant Professor at University of Pennsylvania) provided access to the BraTS challenge server and reviewed and edited the manuscript. Prof. Yarin Gal (Associate Professor at University of Oxford) helped during the brainstorming session for developed metrics. Prof. Tal Arbel provided guidance and feedback throughout the duration of the project and reviewed and edited the manuscript. A total of approximately 25 teams participated in the challenge during both years and provided outputs of their algorithm for the analysis conducted in this manuscript. They are included in the manuscript as middle authors.

- **R. Mehta**, T. Christinck, T. Nair, A. Bussy, S. Premasiri, M. Costantino, M. Chakravarty, D. L. Arnold, Y. Gal, T. Arbel, "Propagating Uncertainty Across Cascaded Medical Imaging Tasks for Improved Deep Learning Inference", *IEEE Transactions on Medical Imaging (TMI), Volume: 41, Issue: 2, February 2022* [159].

The majority of the content of Chapter 4 was published in the IEEE Transactions on Medical Imaging (TMI) journal 2022 (as the above citation). In the published work, I am the lead author. I was responsible for all practical considerations of the proposed method during development and experimentation. I took a major role in this publication by leading all experimental design, implementation, and analysis. Thomas Christinck (undergraduate student at Probabilistic Vision Group, McGill University) worked on the part of the project for his undergraduate thesis. Tanya Nair (graduate student at Probabilistic Vision Group, McGill University) helped Thomas Christinck during the project and reviewed and edited the manuscript. Clinical collaborators Aurélie Bussy, Swapna Premasiri, Manuela Costantino, Prof. Mallar Chakravarty (Computational Neuroscientist at Douglas Mental Health University Institute), and Dr. Douglas Arnold (MD, Neurologist at McGill University) provided access to the data used throughout the project and reviewed the manuscript. Prof. Yarin Gal (Associate Professor at University of Oxford) provided inputs during the proposed method's development and reviewed the manuscript.

Prof. Tal Arbel provided guidance and feedback throughout the duration of the project and reviewed and edited the manuscript.

Peer-Reviewed Conference and Workshop Articles

- **R. Mehta**, C. Shui, T. Arbel, "Evaluating the Fairness of Deep Learning Uncertainty Estimates in Medical Image Analysis", *Medical Imaging with Deep Learning (MIDL) 2023* [164].

The majority of the content of Chapter 5 was published in the Medical Imaging with Deep Learning (MIDL) conference 2023 (as the above citation). In the published work, I am the lead author. I was responsible for all practical considerations of the proposed method during development and experimentation. I took a major role in this publication by leading all experimental design, implementation, and analysis. Furthermore, I will present this in person at MIDL 2023 in Nashville, USA. Dr. Changjian Shui (postdoctoral fellow at Probabilistic Vision Group, McGill University) provided consultation and helped with writing. Prof. Tal Arbel provided guidance and feedback throughout the duration of the project and reviewed and edited the manuscript.

- **R. Mehta**, C. Shui, B. Nichyporuk, T. Arbel, "Information Gain Sampling for Active Learning in Medical Image Classification", *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE) Workshop held in conjunction with 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2022* [165].

The majority of the content of Chapter 6 was published in the UNSURE workshop held in conjunction with MICCAI 2022 conference (as per the above citation). In the published work, I am the lead author. I was responsible for all practical considerations of the proposed method during development and experimentation. I took a major role in this publication by leading all experimental design, implementation, and analysis. Furthermore, I

presented this in person at MICCAI 2022 in Singapore. Dr. Changjian Shui (Postdoctoral fellow at Probabilistic Vision Group, McGill University) and Brennan Nichyporuk (Research Assistant at Probabilistic Vision Group, McGill University) provided consultation and helped with writing. Prof. Tal Arbel provided guidance and feedback throughout the duration of the project and reviewed and edited the manuscript.

- **R. Mehta***, T. Christinck*, T. Nair, P. Lemaitre, D. L. Arnold, T. Arbel, "Propagating Uncertainty Across Cascaded Medical Imaging Tasks for Improved Deep Learning Inference", *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE) Workshop held in conjunction with 22nd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2019, Lecture Notes in Computer Science, Springer, LNCS 11840, pp. 23-32, 2019* [160]. **[Best Paper Award]**

This workshop paper (as the above citation) was published in the UNSURE workshop held in conjunction with MICCAI 2019 conference. This work was the starting point for the related journal article [159], whose content is part of Chapter 4. In the published work, I am the lead author. I was responsible for all practical considerations of the proposed method during development and experimentation. I took a major role in this publication by leading all experimental design, implementation, and analysis. Furthermore, I presented this work in person at MICCAI 2019 in Shenzhen, China. This paper won the **Best Paper Award** during the workshop. Thomas Christinck (undergraduate student at Probabilistic Vision Group, McGill University) worked on the part of the project for his undergraduate thesis. Tanya Nair (graduate student at Probabilistic Vision Group, McGill University) and Dr. Paul Lemaitre (Research Assistant at Probabilistic Vision Group, McGill University) helped Thomas Christinck during the project. Both reviewed and edited the manuscript. Dr. Douglas Arnold (MD, Neurologist at McGill University) provided access to the data used throughout the project and reviewed the manuscript. Prof. Tal Arbel provided guidance and feedback throughout the duration of the project and reviewed and edited the manuscript.

- **R. Mehta**, T. Arbel, "RS-Net: Regression-Segmentation 3D CNN for Synthesis of Full Resolution Missing Brain MRI in the Presence of Tumours", *Simulation and Synthesis in Medical Imaging (SASHIMI) workshop held in conjunction with 21st International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2018, Lecture Notes in Computer Science, Springer, Vol. 11037, pp. 119-129* [158].

This workshop paper (as the above citation) was published in the SASHIMI workshop held in conjunction with MICCAI 2018 conference. In the published work, I am the lead author. I was responsible for all practical considerations of the proposed method during development and experimentation. I took a major role in this publication by leading all experimental design, implementation, and analysis. Furthermore, I presented this work in person at MICCAI 2018 in Granada, Spain. Prof. Tal Arbel provided guidance and feedback throughout the duration of the project and reviewed and edited the manuscript.

Peer Reviewed Short Papers

- **R. Mehta**, A. Filos, Y. Gal, T. Arbel, "Uncertainty Evaluation Metrics for Brain Tumour Segmentation", *Medical Imaging with Deep Learning (MIDL) 2020* [162].

This short paper was published in the Medical Imaging with Deep Learning (MIDL) 2020 conference (as the above citation). This work served as the starting point for the related journal article [161], whose content is part of Chapter 3. In the published work, I am the lead author. I was responsible for developing a metric to evaluate uncertainties for the brain tumour segmentation task and writing the original draft of the manuscript. Furthermore, I presented this work virtually at MIDL 2020 in Montreal, Canada. Angelos Filos (Ph.D. candidate at University of Oxford) helped during the development of the metric and related evaluation code. Prof. Yarin Gal (Associate Professor at University of Oxford) helped during the brainstorming session for developed metrics. Prof. Tal Arbel provided guidance and feedback throughout the duration of the project and reviewed and edited the manuscript.

Other Publications

Apart from the above papers, I have contributed significantly, in the course of my doctoral research, to the papers listed below.

- C. Shui*, J. Szeto*, **R. Mehta**, D. L. Arnold, T. Arbel., "Mitigating Calibration Bias Without Fixed Attribute Grouping for Improved Fairness in Medical Imaging Analysis", *Medical Image Computing and Computer Assisted Intervention (MICCAI) conference 2023.*, [230].
- J. Durso-Finley, J. P. Falet, **R. Mehta**, D. L. Arnold, N. Pawlowski, T. Arbel., "Improving Image-Based Precision Medicine with Uncertainty-Aware Causal Models", *Medical Image Computing and Computer Assisted Intervention (MICCAI) conference 2023.* [62].
- **R. Mehta**, V. Albiero, L. Chen, I. Evtimov, T. Glaser, Z. Li, T. Hassner, "You Only Need a Good Embeddings Extractor to Fix Spurious Correlations", *Workshop on Responsible Computer Vision (RCV) – European Conference on Computer Vision (ECCV) 2022* [156].
- B. Nichyporuk*, J. Cardinell*, J. Szeto, **R. Mehta**, J.P. Falet, D. L. Arnold, S. Tsaftaris, T. Arbel, "Rethinking Generalization: The Impact of Annotation Style on Medical Image Segmentation", *The Journal of Machine Learning for Biomedical Imaging (MELBA), December 2022* [184].
- B. Nichyporuk, J. Cardinell, J. Szeto, **R. Mehta**, D. L. Arnold, S. Tsaftaris, T. Arbel, "Cohort Bias Adaptation in Federated Datasets for Lesion Segmentation", *Domain Adaptation and Representation Transfer (DART) 2021 workshop held in conjunction with 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2021, Lecture Notes in Computer Science, Springer, LNCS 12968, pp. 101-111, 2021* [185]. **[Best Paper Award]**
- S. Vadacchino, **R. Mehta**, N. M. Sepahvand, B. Nichyporuk, J. J. Clark, T. Arbel,

"HAD-Net: A Hierarchical Adversarial Knowledge Distillation Network for Improved Enhanced Tumour Segmentation Without Post-Contrast Images", *Medical Imaging with Deep Learning (MIDL) 2021* [257].

- B. Kaur, P. Lemaitre, **R. Mehta**, N.M. Sepahvand, D. Precup, D. L. Arnold, T. Arbel, "Improving Pathological Structure Segmentation Via Transfer Learning Across Diseases", *Domain Adaptation and Representation Transfer (DART): Learning Transferable, Interpretable, and Robust Representation Workshop held in conjunction with 22nd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2019, Lecture Notes in Computer Science, Springer, LNCS 11795, pp. 90-98, 2019* [119].

- **R. Mehta**, T. Arbel, "RS-Net: Regression-Segmentation 3D CNN for Synthesis of Full Resolution Missing Brain MRI in the Presence of Tumours", *Medical Imaging meets NeurIPS (Med-NeurIPS) 2018 workshop held in conjunction with 32nd Conference on Neural Information Processing Systems (NeurIPS) 2018*.

- **R. Mehta**, T. Arbel, "3D U-net for Brain Tumour Segmentation", *Multimodal Brain Tumour Segmentation (BraTS) challenge 2018 held in conjunction with 21st International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2018, Lecture Notes in Computer Science, Springer, LNCS 11384, pp. 254-266, 2018* [157].

Acknowledgements

> Family [and friends are] ~~is~~ like the forest: if you are outside it is dense; if you are inside you see that each tree has its own position.
>
> — *Akan proverb*

First and foremost, I would like to thank my supervisor Prof. Tal Arbel for her guidance throughout the last six years. She has always supported me during my whole Ph.D. journey. She provided opportunities for me to develop as an independent researcher and teacher. She invested ample time in polishing my writing during countless submissions and never got annoyed by my stupid mistakes in writing, especially my amazing ability to always miss articles in sentences. I will fondly remember our paired writing sessions, which started with writing a paper from scratch eighteen hours before the submission deadline. Her remarkable ability to always find light in any situation, allowing me to grow at my own pace, and also giving me ample opportunities to find my own place, has made my Ph.D. journey worthwhile. I will remember this time for the rest of my life. Thank you for everything.

I also wish to express my gratitude to Prof. Doina Precup and Prof. Christine Tardif for serving on my supervisory committee. I thank the above, the thesis examiners, and the

anonymous reviewers for their time and effort in reviewing my work and providing valuable feedback.

Next, I thank Prof. Jayanthi Sivaswamy. Under her watchful eyes at IIIT-Hyderabad, I saw myself transforming from a shy and bumbling student to a methodological and confident researcher. Everything I learned from her has carried me for the last six years and hopefully will help me further in my career. Similarly, I would like to thank Prof. Mehfuza Holia and Prof. Tanmay Pawar from BVM. Their teaching during my bachelor's sparked my enthusiasm for coding and image processing, which led to me doing a Ph.D. in a related field. Thank you for being amazing teachers. I am also grateful to Bharat Mehta Sir, who taught me mathematics in my 10th standard and showed me how good I could be with proper guidance. Thank you for your teaching.

Labmates are among the most important people to make any Ph.D. journey enjoyable. First and foremost, I wish to thank Lex, Naz, and Tanya for welcoming me to PVG with open arms. Our usual monthly visits to Thomson House were really enjoyable and eye-opening. Thank you for making me feel like a part of a group despite all the cultural differences. I will always cherish our talks and fun board game nights. Special thanks to Naz and Tanya for our memorable trip to Spain. Thank you, Josh and Barleen, for our exciting trip to Shenzhen. Similarly, thanks, Amar, for our Singapore trip. Next, I would like to thank all the remaining members of PVG from the last six years: Ian, Paul, Qing, Thomas, Brennan, Adrian, Justin, Eric, Savario, Jean-Pierre, Julien, Jillian, Chelsea, Nima, Changjian, and Bonaventure. Everyone has made our weekly lab and journal club meetings fruitful and intellectually stimulating. Special thanks to everyone at Meta Responsible AI team who made my internship experience memorable: Tal, Ivan, Albert, Tamar, Vitor, Li, and Aram. I not only learned a lot about robustness and fairness from them but also learned how to conduct good research and the ability to pivot projects.

During my Ph.D., I was fortunate to work with some truly intelligent and hard-working

colleagues who have taught me different research-related things. I want to thank everyone who co-authored a paper with me during my Ph.D.: Prof. Yarin Gal, Angelos Filos, Thomas Christinck, Tanya Nair, Paul Lemaitre, Swapna Parmeshwari, Aurelie Bussy, Prof. Mallar Chakravarty, Barleen Kaur, Nazanin Sepahvand, Dr. Douglus L. Arnold, Prof. Doina Precup, Saverio Vadacchino, Brennan Nichyporuk, Prof. James J. Clark, Jillian Cardinal, Justin Szeto, Prof. Sotos Tsfatris, Prof. Spyridon Bakas, Ujjwal Baid, Changjian Shui, Jean-Pierre Falet, Nick Pawlowski, and Joshua Durso-Finley. I would also like to thank everyone who participated in the QU-BraTS challenge and helped us in making it a success.

Next, I would like to thank my friends who made my Ph.D. journey enjoyable and kept me sane during difficult times. First and foremost, thank you Karthik, for staying by my side during all the ups and downs of my professional and personal life and never giving up on me. Also, special mention to Vasudha for tolerating our (Karthik's and mine) poor jokes over the last year. Similarly, I would like to thank Nausheen for showing me that you can simultaneously be independent and vulnerable; and that relying on people you trust is okay. Thank you, Sukesh, for making me understand the importance of taking things slowly and continuing with life. Thank you, Chetan, for being a real-life example of how to make lemonade when life unexpectedly gives you lemon. Thank you, Rutuja, for organizing all the trips and celebrations and making Montreal feel like home. Thank you, Sumedh, for always making me laugh with your sense of humour. Special mention to Nihira for making me smile over the last four months. Thank you, Nehal, for sharing the joy of books and music. Thank you, Samrudhdhi, for our shared TAship and Ph.D. journey experiences. Thank you, Utpal and Riya, for showing me that distance (both in space and time) doesn't change a friendship when you have developed a true bond in the past. Also, thanks to many other friends who were part of my life in one way or another over the last six years: Shaheen, Shruti, Snehil, Aditya, Sayantan, Soumyajit, Rahul, Smita, Ibtihel, Michael, Tabish, Balamurali, Saypraseuth, Fabi, Yogesh, Louis, Jigar, Shreya, Priyanka, Bhavin, Kanwal, and anyone else I am missing.

Family is the most important part of anyone's life. They not only support you when you are successful and happy but also when you are dejected and sad. First and foremost, I want to thank my brother Parth, who always dived deep into the unknown, showed me a way through his experiences, and provided unconditional support. Special thanks to Divya for filling in a missing piece in our family. Next, I would like to thank my twin sister Toral, for being my biggest supporter, emotional mentor, and someone I can call whenever I feel down. Thank you, Neelka and Atul Masa, for welcoming me to Canada with your frequent parcels when I initially arrived and understanding the importance of giving me my space during the later stages. Thank you, Mama and Mami, for our almost weekly calls, always listening to my rants, and still providing some important and sound advice. I will always cherish our trips and laughs during the summer of 2019 and 2022. Next, I thank Aben and Nani for our not-so-frequent but important calls. Last but not least, I cannot thank my parents enough for never questioning my choice of staying away from the family for the last 9 years, never asking me when I would complete my Ph.D., and always being understanding and kind. This thesis is for both of you.

Contents

List of Figures

List of Tables

List of Acronyms

Abbreviation	Meaning
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
ReLU	Rectified Linear Unit
BDL	Bayesian Deep Learning
BraTS	Brain Tumour Segmentation
QU − BraTS	Quantification of Uncertainty in Brain Tumour Segmentation
AD	Alzheimer's Disease
EIG	Expected Information Gain
AEIG	Adapted Expected Information Gain
UIG	Uniform Information Gain
CFIG	Class-Frequency Information Gain
MS	Multiple Sclerosis
MRI	Magnetic Resonance Imaging
RRMS	Relapsing-Remitting Multiple Sclerosis
CN	Cognitive Normal
MCI	Mild Cognitive Impairment

Abbreviation	Meaning
ADAS	Alzheimer's Disease Assessment Scale
MMSE	Mini-Mental State Examination
MC − Dropout	Monte-Carlo Dropout
KL	Kullback-Leibler
MI	Mutual Information
DR	Diabetic Retinopathy
ECE	Expected Calibration Error
MCE	Maximum Calibration Error
U − E	Uncertainty-Error
cVAE	conditional variational auto-encoder
PhiSeg	probabilistic hierarchical segmentation
GroupDRO	group distributionally robust optimization
AL	Active Learning
VR	Variation Ratio
MSD	Mean Standard Deviation
WT	Whole Tumour
TC	Tumour Core
ET	Enhancing Tumour
TP	True Positive
TN	True Negative
FTP	Filtered True Positive
FTN	Filtered True Negative
TPR	True Positive Rate
FPR	False Positive Rate
FLAIR	Fluid Attenuated Inversion Recovery
GT	Ground Truth
DSC	Dice Score

Abbreviation	Meaning
CRS	Cumulative Ranking Score
NRS	Normalized Ranking Score
FRS	Final Ranking Score
FG	Fairness Gap
EM	Evaluation Metric
ISIC	International Skin Imaging Collaboration
HGG	High Grade Glioma
LGG	Low Grade Glioma
RMSE	Root Mean Squared Error
ROC – AUC	Area Under the Receiver Operating Characteristic Curve

1

Introduction

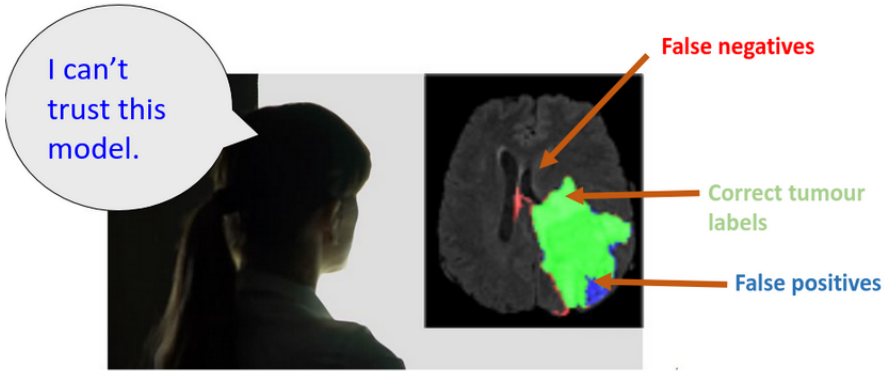
> Life was like this game [pachinko] where players could adjust the dials yet also expect the uncertainty of factors they couldn't control.
>
> — *Min Jin Lee*, Pachinko

Deep Learning has recently become omnipresent in almost all applied machined learning fields. For example, in the field of computer vision, Deep Learning (DL) has outperformed many classical machine learning algorithms in a varied range of problems like classification [132], segmentation [144], object detection [201], depth regression [64], etc. Though these results are promising, most methods only provide a single prediction for

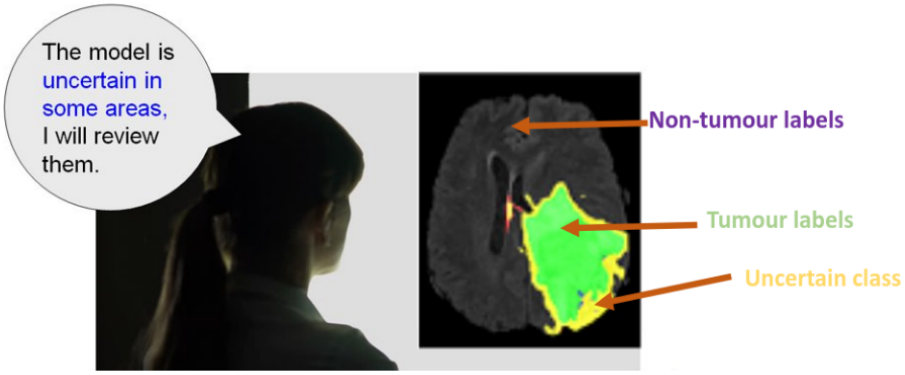
a single input instead of a distribution of possible predictions. These distributions could enable us to measure the confidence in its prediction by calculating associated uncertainties. In many critical applications, like self-driving cars, it is of paramount importance to get associated uncertainty with the output, as a wrong decision can be the difference between life and death. Take, for example, a case reported in May 2016 [6], where a fatality was reported as a self-driving car of Tesla could not successfully differentiate between the sky and the white side of a trailer. In this scenario, it would have been more appropriate if the system had provided uncertainty associated with their output. This could enable passengers to override the uncertain decisions of the self-driving car and save a human life.

Uncertainty is even more necessary in medical image analysis, where despite recent advances [253, 41, 158, 51, 109, 180, 253, 163, 167], DL models are prone to make mistakes. Reasons for this include but are not limited to noise and artifacts in medical images, a variety of shapes, sizes, and textures of pathology across populations, etc. This can create situations where the system's end-user (clinicians) will not trust the system's output while reviewing and, therefore, will hesitate to integrate them into the clinic (Figure 1.1 (a)). In this scenario, generating uncertainties associated with system output would be beneficial, as it can provide information about where a system is not confident in its predictions. This can potentially allow the end-user (clinicians) to review the system output and correct it if necessary (Figure 1.1 (b)). This brings the clinicians back into the workflow and helps in utilizing both faster computation of machine learning methods and the experience and judgment of clinicians for better diagnosis.

Classical Bayesian learning methods like Gaussian processes [30] allow us to capture and represent the model uncertainty, but they are computationally costly and often intractable. Recently, Bayesian deep learning (BDL) methods [79, 136, 250, 130, 27], especially MC-Dropout [79] which captures model uncertainty and Probabilistic U-Net [130] which captures inter-rater variability, have gained quite a bit of attention. These methods



(a) Without uncertainties a clinician would be less likely to trust automatic segmentation model output, especially when it is prone to make mistakes.



(b) With uncertainties a clinician are more likely to include automatic segmentation model output in their workflow as they can review areas where model output is not confident in its predictions, and correct it if necessary.

Figure 1.1: Example showing a clinician reviewing brain tumour segmentation output produced by a machine learning model (a) without associated uncertainties and (b) with associated uncertainties.

capture the uncertainty of the deep learning models without the drawbacks (i.e., computational cost) of classical Bayesian methods. Many metrics exist in the literature to measure the performance of model uncertainties [90, 176, 114, 87], but they don't specifically focus on medical image analysis. The quantification of uncertainty is useful if the model predictions are confident when they are correct and wrong when they are uncertain. This permits reviewing clinicians to trust the confident predictions and review other predictions. There is an unmet need to develop an uncertainty quantification metric specifically designed for medical image analysis with the above mentioned clinical goal in mind. The first part of this thesis specifically focuses on this aspect and develops a metric to measure the correlation between uncertainties and model errors for brain tumour segmentation.

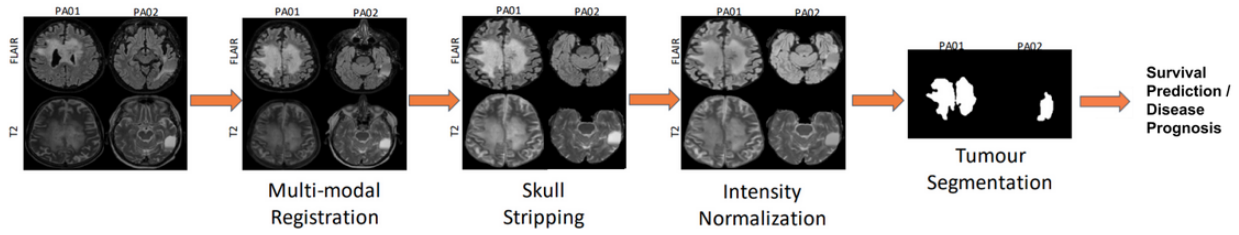


Figure 1.2: An example of a typical medical image analysis pipeline is depicted here. Here, we look at the survival prediction for patients with brain tumours [22]. Survival prediction and disease prognosis have been shown to correlate with the segmentation of brain tumours [22]. Typical sequential tasks in this pipeline include registration of multi-modal images, skull stripping for these images, intensity normalization, tumour segmentation, and survival prediction. A small mistake in any initial task (registration, skull stripping, tumour segmentation) can adversarially affect the downstream task of interest

Using uncertainties to inform the end-user (clinicians) about the confidence of a machine learning model in its prediction is indeed a great clinical use case. However, it is not necessary that we will have access to clinicians at all stages of medical image analysis pipelines. Take, for example, the survival time prediction pipeline for patients with brain tumours, shown in Figure 1.2. It consists of a sequence of inference tasks (e.g., registration [51], skull stripping [129], segmentation [92, 109, 180], etc.) before the downstream survival time prediction task. Here, we might not be able to use the domain knowledge of clinicians at all these stages and correct mistakes made by machine learning (ML) models. As these are cascaded inference tasks, small errors made by any initial task can accumulate and adversarially affect the downstream task of interest. For example, if the machine learning model makes an error in skull stripping, it would be carried forward to the survival time prediction model and can hinder its performance. In this situation, embedding uncertainty estimation across cascaded inference tasks can help build better automatic machine-learning systems for the downstream task of interest. It can inform the downstream task of any potential mistakes made by initial tasks, make the system less sensitive to these mistakes, and can lead to improved performance. In the second part of this thesis, we develop a cascaded medical image analysis pipeline where machine learning methods for each task produce uncertainty associated with its output. We show that the performance of the downstream tasks in a medical image analysis pipeline would improve if, in addition to mean output predictions, the uncertainty estimates are

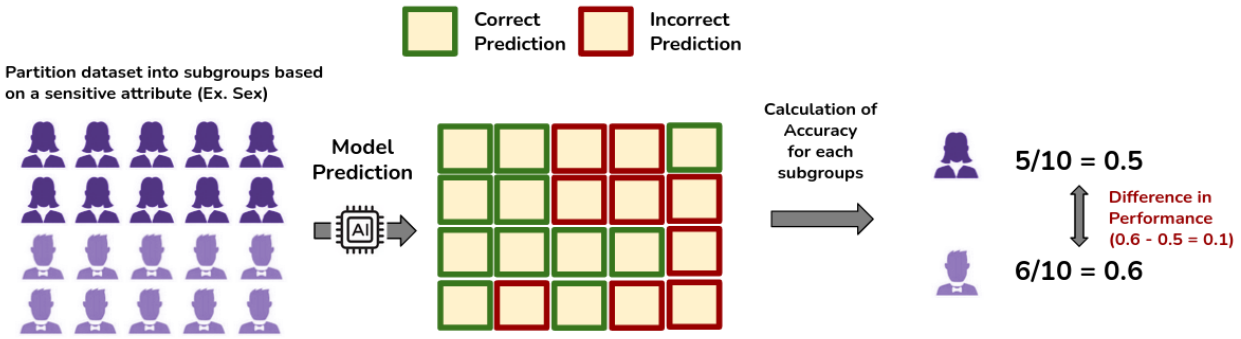


Figure 1.3: An example of a machine learning model showing different performances across sex. Here, the model is biased towards males as it shows an accuracy of 0.6 for images of males, while the same model has an accuracy of 0.5 for females. A machine learning model is considered fair for sex attributes if this difference in performance is zero.

propagated across cascaded inference tasks.

For machine learning models to be trustworthy and ready for clinical deployment, in addition to uncertainty quantification, it also requires fairness and robustness across different sensitive attributes (ex. demographic information). Fairness requires that the model's predictions and recommendations are not biased against any particular group of patients, and that the model performs equally well across different sensitive groups. In medical image analysis, bias can arise from various sources, such as differences in patient populations, imaging protocols, or variability in the task at hand (ex., size of the tumour in brain tumour segmentation). To ensure fairness in machine learning models, it is important to carefully select and prepare the training data, evaluate the model's performance across different groups, and monitor the model's outputs over time. By prioritizing fairness in medical image analysis, we can ensure that these models are accurate, reliable, and equitable for all patients. Consider a machine learning system shown in Figure 1.3. Here, analyzing the model performance, across different subgroups based on sex, reveals that it is biased toward male patients as it performs relatively better for them than for female patients. If we deploy such a model in clinical practice, it would lead to unfair recommendations for female patients. As such, it is critical to mitigate these fairness concerns before its real-world deployment.

Several methods [284] have been proposed in the machine learning literature to mitigate a lack of fairness in DL models. However, they focus entirely on the absolute performance between groups without considering their effect on uncertainty estimation. As we discussed previously, real clinical contexts would benefit from knowledge about the confidence in the model predictions, when made explicit in the form of uncertainties. A machine learning model that underperforms for a subgroup but indicates higher uncertainties associated with its output for that subgroup, could still be clinically deployed. As in this scenario, uncertainties could be useful to flag predictions from subgroup where the model is prone to underperform. Conversely, a machine learning model that achieves fairness in terms of performance across different subgroups, but produces low uncertainties for predictions where it makes mistakes, would become less trustworthy to clinicians. Considering this in mind, in this thesis, we take a look at bias mitigation models from the perspective of both fairness and uncertainty quantification. Our analysis of popular bias mitigation methods [213, 105] reveal their shortcoming in mitigating bias in terms of both absolute performance and associated uncertainties.

So far, we only talked about improving the clinical usefulness of medical image analysis systems from the perspective of their end-use. However, one of the biggest challenges in deploying a machine learning model in clinical practice is its dependency on a large labeled training dataset. The process of labeling medical images is time-consuming and often requires clinical domain expert knowledge, making it difficult to obtain large datasets for training machine learning models. Compared to that, unlabelled datasets are easier to obtain, and in many contexts, it would be feasible for an expert to provide labels for a small subset of images. Active learning can help address this challenge by identifying the most informative images to be labeled by experts, thereby reducing the amount of labeled data required for training (Figure 1.4). This can make the process of developing machine learning models for clinical practice more scalable, cost-effective, and efficient. Additionally, active learning can help address bias in medical image datasets by priori-

Figure 1.4: An example of a typical active learning framework. Here, a machine learning model is trained on initial labeled data. Following this, based on an acquisition function, many examples from a large unlabeled dataset are queried. These examples are provided to human experts for annotation, which are appended to the initially labeled dataset. This process is repeated in a loop until the machine learning model reaches a pre-defined performance criterion.

tizing the labeling of images from underrepresented patient populations, improving the overall accuracy and reliability of the model's outputs. An acquisition function is an essential part of the active learning framework determining the next data point to be labeled. Common acquisition functions include uncertainty sampling [226], which selects points with the highest uncertainty, and query-by-committee [224], which selects samples where different model versions disagree. Other acquisition functions include coreset [220] and Bayesian active learning [80].

Generally, uncertainty-based active learning approaches, particularly entropy-based methods, have been popular in medical imaging contexts where they have shown some effectiveness in addressing the issue of high-class imbalance. Entropy based methods select the samples which are the hardest for the current model to classify. The assumption for entropy based methods is that selecting these hardest samples, labeling them, and retraining the model with these samples would improve the model performance. However, just because these samples are the hardest for the current model to classify, does not nec-

essarily mean that they will lead to improvement in the performance of the model on a real-world (ex. test) dataset. In the last part of this thesis, we develop a new active learning acquisition function that explicitly measures the information gain on an unseen (evaluation) set. The hypothesis is that selecting samples based on this acquisition function should lead to better performance of the model on a real-world dataset.

Overall, we can say that uncertainty estimates should be integrated into almost all the aspects of medical image analysis systems. They would help in improving the trust of the end-user into the system, enhancing the performance of these systems, and making them ready for clinical deployment.

1.1 Contributions

In this section, a brief summary of contributions from this thesis is presented. Specifically, in this thesis following contributions are made (i) Evaluating the uncertainty produced by different methods for the task of interest (ex., tumour segmentation) by designing a task and application-specific metric, which can verify if the performance of the method indeed correlates with the uncertainty produced by them or not, (ii) Embedding the uncertainty produced by deep learning methods in cascaded medical imaging tasks for improved inference, (iii) Evaluating fairness of deep learning uncertainties for various medical imaging tasks, and (iv) Using information gain formulation in active learning framework for improved label acquisition in various medical image classification problems.

- **Developing a task-specific metric to evaluate the uncertainties produced for brain tumour segmentation**

As a first contribution of this thesis, a novel metric is developed, which can help in evaluating and comparing the uncertainties produced by different methods. This is the first metric in the literature which can help in validating the correlation between uncertainty and the task of interest in medical image analysis. Furthermore, for two consecutive

years (2019 and 2020), we organized the first MICCAI challenge on quantifying uncertainty for brain tumour segmentation (QU-BraTS) by using the developed metric. BraTS is one of the most popular challenges in the field [22, 170]. By organizing this uncertainty quantification sub-challenge with BraTS, we reached the wider medical imaging community, which will, in turn, lead to more people understanding the need to generate uncertainty and quantify it in the field. We ranked the brain tumour segmentation uncertainties generated by 14 independent participating teams of QU-BraTS 2020, all of which also participated in the main BraTS segmentation task. Overall, our findings confirmed the importance and complementary value that uncertainty estimates provide to segmentation algorithms, highlighting the need for uncertainty quantification in medical image analyses. Our evaluation code is made publicly available at https://github.com/RagMeh11/QU-BraTS. This challenge will serve as a good benchmark for the community and will also lead to more participation from the machine learning community in the field of medical image analysis, which in turn can be helpful for the development of better models for uncertainty estimation in medical image analysis.

- **Embedding uncertainties produced by deep models in cascaded medical imaging tasks for improved inference.**

As a second contribution of the thesis, we propose the first framework that embeds uncertainty estimates across cascaded inference tasks, to improve the performance of the downstream inference task. We demonstrate the effectiveness of the proposed approach in three different clinical contexts: (i) We demonstrate that by propagating T2 weighted lesion segmentation results and their associated uncertainties, from brain MRI acquired for patients with multiple sclerosis (MS), subsequent T2 lesion detection performance is improved when evaluated on a proprietary large-scale, multi-site, clinical trial dataset acquired from patients with MS. (ii) We show an improvement in brain tumour segmentation performance when the uncertainty map associated with a synthesized missing MR volume, generated using our previously published method (Appendix A), is provided as an additional input to a follow-up brain tumour segmentation network. Experiments

are performed on the publicly available BraTS-2018 dataset [22]. (iii) We show that by propagating uncertainties from a voxel-level hippocampus segmentation task, the subsequent regression of the Alzheimer's disease clinical score is improved. Experiments are performed on popular publicly available ADNI dataset [111]. The quantitative results show that uncertainty propagation improves the downstream task performance by 1-5%. However, quantitative results only demonstrate part of the gain. The qualitative results illustrate that uncertainty propagation does indeed assist in correcting clinically relevant errors even when improvement in terms of absolute numbers are small.

- **Evaluating the fairness of deep learning uncertainty estimates in medical image analysis.**

As a third contribution of this thesis, we present the first exploration of the effect of various bias mitigation methods on overcoming biases across subgroups in medical image analysis in terms of bottom-line performance and their effects on uncertainty quantification. This would allow us to analyze the trustworthiness of machine learning models for medical image analysis from the context of both fairness and uncertainty quantification. We perform extensive experiments on three different clinically relevant problems: (i) skin lesion classification, (ii) brain tumour segmentation, and (iii) Alzheimer's disease clinical score regression. Our results indicate that popular ML methods, such as data-balancing and distributionally robust optimization, succeed in mitigating fairness issues in terms of the model performances for some of the tasks. However, this can come at the cost of poor uncertainty estimates associated with the model predictions. This tradeoff must be mitigated if bias mitigation models are to be adopted in medical image analysis.

- **Improving active learning image label acquisition via information gain criterion for deep learning models utilized for medical image classification.**

As the last contribution of this thesis, we present a novel information-theoretic active learning framework that guides the optimal selection of images from the unlabelled pool to be labeled based on the expected information gain (EIG) on an evaluation dataset. We

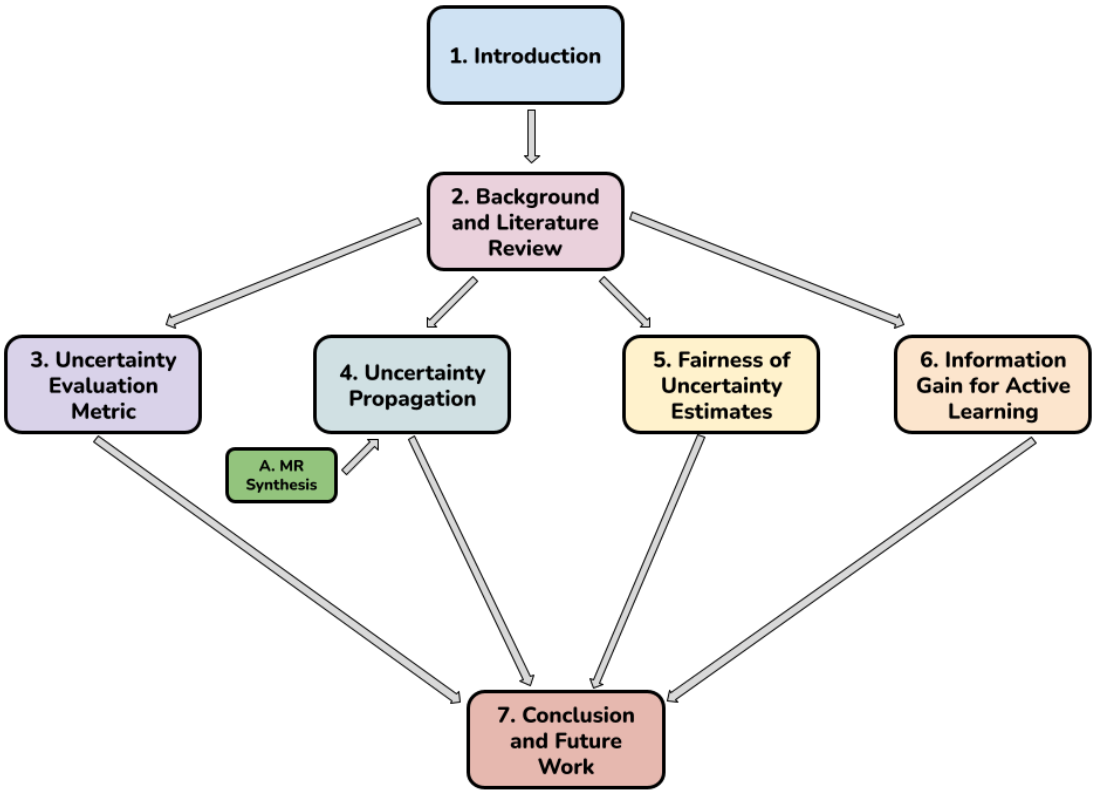


Figure 1.5: Outline of the thesis

show that, by careful design choices, our model can be easily integrated into existing deep learning classifiers. Experiments are performed on two different medical image classification datasets: multi-class diabetic retinopathy disease scale classification and multi-class skin lesion classification. Results indicate that by adapting EIG to account for class imbalances, our proposed Adapted Expected Information Gain (AEIG) outperforms several popular baselines, including the diversity-based CoreSet and uncertainty-based maximum entropy sampling. Specifically, AEIG achieves 95% of overall performance with only 19% of the training data, while other active learning approaches require around 25%. The proposed method would greatly benefit in reducing the cost of the label acquisition process by optimally selecting the data to be labeled.

1.2 Outline of the Thesis

The organization of the work reported in the thesis is described in this section (see Figure 1.5). This introductory Chapter 1 provided an overview and the context, motivation, and contribution of this thesis.

Chapter 2 presents the necessary background and the literature review of all relevant works. Specifically, Section 2.1 provides relevant clinical background, while Section 2.2 provides background on machine learning models for uncertainty estimation. Section 2.3 gives literature reviews on the application of various uncertainty estimation methods in both computer vision and medical image analysis. Section 2.4 discusses various recent attempts at designing task-specific metrics to evaluate uncertainties produced by deep neural networks and the remaining challenges for the same. Section 2.5 provides recent development in estimating uncertainty in multi-rater systems. Section 2.6 provides a literature review on the fairness of deep learning models. The chapter concludes in Section 2.7 by discussing various recent methods for active learning.

Chapter 3 introduces a metric for evaluating uncertainties produced for brain tumour segmentation. Section 3.1 gives a brief introduction to the necessity of a task-specific uncertainty evaluation metric. Section 3.2 provides the thought process behind the developed metric, how various parts for the same were designed, and a small example explaining the working of this metric. Section 3.3 lists the dataset utilized in the uncertainty quantification for the brain tumour segmentation challenge, the evaluation framework for the same, and the approaches of various participating teams from the QU-BraTS 2020 challenge. In Section 3.4 different analyses of participating teams from the challenge, an ablation study for the different components of the developed metric and their subsequent effect on the ranking of teams, and some qualitative results of participating teams are provided. At last, Section 3.5 summarizes the findings. It concludes that segmentation uncertainties provide complementary information to absolute performance and future challenges should evaluate both absolute performance and associated uncertainties for

all participating teams.

Chapter 4 discusses the propagation of uncertainties in a cascade of medical image analysis tasks and their subsequent effect on the downstream task of interest. Specifically, Section 4.1 provides a brief introduction to the chapter. Section 4.2 discusses the developed methodology of uncertainty propagation for three different contexts of brain tumour segmentation, MS T2 lesion segmentation, and Alzheimer's disease clinical score prediction. Section 4.3 lists the implementation details, datasets, and evaluation metrics deployed for the above-mentioned three different clinical contexts. Section 4.4 shows experiments and results for the effectiveness of uncertainty propagation, the effect of different uncertainty generation methods like MC-Dropout, Deep Ensemble, and Ensemble Dropout, and different uncertainty measures. Results demonstrate that by propagating uncertainties to the downstream task of interest, the performance can be improved by 2-10%. The chapter concludes by providing a brief summary in Section 4.5.

Chapter 5 evaluates the fairness of deep learning uncertainty estimates for a variety of medical image analysis tasks, ranging from image classification, and image segmentation, to clinical score regression. Section 5.1 provides an introduction to fairness in machine learning and medical image analysis models. Section 5.2 discuss the developed evaluation protocol for fairness in uncertainty estimation and how fairness and uncertainty together can provide more trustworthy models. Section 5.3 gives details of various experiments performed on three different clinical contexts (brain tumour segmentation, skin lesion classification, and Alzheimer's disease clinical score regression) for three different fairness mitigation methods. Results demonstrate that fairness can come at the cost of associated uncertainty estimates with the model predictions. At last, Section 5.4 summarizes the findings from this chapter.

Chapter 6 presents the final contribution of the thesis describing an active learning framework based on information gain. Section 6.1 provides an introduction to the chapter. Sec-

tion 6.2 describes in detail the developed information gain based active learning framework, and lists all design choices which allowed easy integration of the developed framework in modern deep neural networks. Section 6.3 describes the multi-class medical image classification context utilized for experiments performed in the following section. Section 6.4 describes the results of the proposed framework and compares them against other popular active learning methods. It shows that the proposed method outperforms previous active learning methods by 4-5%. At last, Section 6.5 gives a summary of the findings from this chapter.

Chapter 7 concludes the thesis by summarizing the contributions of the thesis and discussing its limitations and possible future extensions. Finally, Appendix A details our previously published work on synthesizing full-resolution missing brain MRI in the presence of pathologies. This work served as a basis for experiments related to the brain tumour segmentation pipeline in Chapter 4.

2

Background and Literature Review

> Human knowledge is never contained in one person. It grows from relationships we create between each other and the world, and still it is never complete.
>
> — *Paul Kalanithi, When Breath Becomes Air*

In the previous chapter, we gave brief motivation and introduction about the different problems tackled in this thesis. In the first part of this chapter, we discuss the background of various different clinical contexts and uncertainty estimations in deep learning models. The aim is to motivate the clinical necessity of different contexts, and provide a theoretical understanding of uncertainty estimations in deep learning models. During the discussion, without going into much detail, we only briefly mention how these contexts are utilized in the rest of the thesis. When discussing our proposed methods in the relevant chapters, we assume that the reader is already familiar with specific ideas.

The second part of this chapter reviews literature related to applications of uncertainty estimation, metrics to evaluate uncertainties, fairness of deep learning models, and active learning. We aim to provide a reader with a starting point and to assist in developing a basic understanding of the field.

2.1 Background: Clinical Context

Even though the proposed methods in this thesis are expected to be generalizable to any medical image analysis pipeline which uses deep learning models, the primary focus of this thesis is on various sub-problems in the context of brain tumours, multiple sclerosis (MS), and Alzheimer's disease (AD). These three contexts provide us the opportunity to test our proposed methods on heterogeneous yet clinically relevant problems. In addition to that, unlike many other medical image analysis problems, we have access to the large-scale datasets for these three contexts either through publicly available datasets [22, 111] or through our clinical collaborators (specifically for MS).

2.1.1 Brain Tumours

Worldwide, it is estimated that roughly 296,851 [76] cases of brain tumours are diagnosed every year. One of the primary diagnostic evaluation tools for brain tumours is an MRI

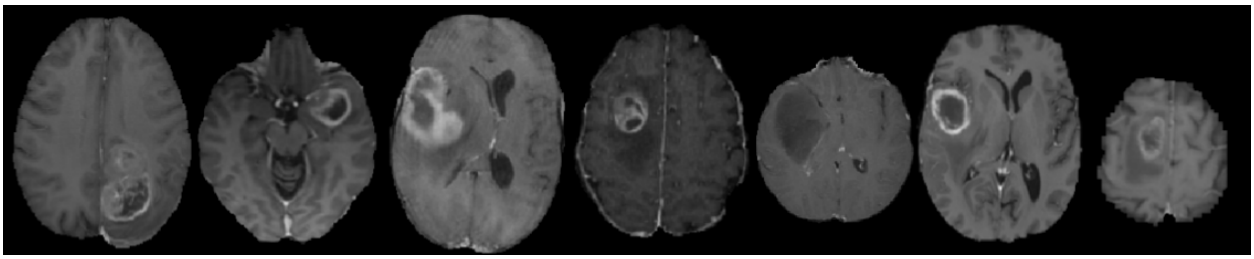


Figure 2.1: Heterogeneity in tumour shape, size, and location of a single slice across seven subjects. Here, a slice of T1ce MR for seven subjects is shown. Image Courtesy: BraTS [22]

of the brain, which helps in evaluating the size of the tumour and its proximity to critical structures of the brain. Due to its potential life-threatening characteristics and the high number of deaths, it is critical to analyze the boundaries and substructures of brain tumour accurately. This need becomes more critical for surgical planning as segmentation and boundary delineation play a crucial role in understanding the prognosis of the disease.

In clinical practice, usually, experts examine different contrast MR volumes of patients with brain tumours and delineate the boundaries of tumour manually. Manual delineation is a long and arduous process, prone to human error. However, it might also not be possible to attain a single "ground-truth" tumour boundary from the MR image alone, as tumour might have infiltrated the surrounding structure. Automatic segmentation techniques try to overcome this problem by using image analysis or machine learning techniques for delineating brain tumour and their sub-structure. Segmentation of brain tumours from healthy brain tissue in MR is a particularly challenging task given the wide variability in their shape, size, position, texture, and intensity over a population of patient images (Figure 2.1). This variability can create situations where automatic tumour segmentation techniques can make mistakes [92] and may require human intervention before its use in the downstream task of interest (e.g., surgical planning). As we discussed in the Introduction chapter of this thesis (Chapter 1), uncertainty estimates associated with automatic segmentation techniques can help in flagging predictions where these techniques are not confident. These uncertain predictions can be corrected by clinicians. Keeping this clinical use case in mind, as a part of this thesis (Chapter 3), we develop an uncertainty

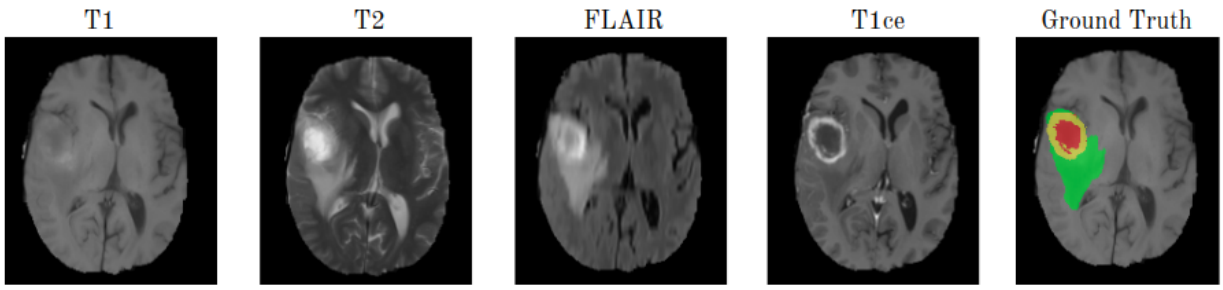


Figure 2.2: Different MR contrast used in a clinical setting to segment tumour from its surrounding healthy structures and into sub-structures. From left to right: T1-weighted MR, T2-weighted MR, FLAIR MR, T1-post contrast (T1ce) MR, Ground Truth tumour segmentation (**Edema**,**Enhancing Tumour**, **Non-enhancing core**). Image Courtesy: BraTS [22]

evaluation metric specifically designed for the brain tumour segmentation task.

In a clinical setting, different types of MR contrasts are used to delineate tumours and their sub-structures from the surrounding healthy tissues (Figure 2.2). The performance of tumour segmentation should increase if several contrasts of MRI be available, as these different contrasts assist in differentiating healthy tissues from tumours [93]. However, in real clinical practice, not all MR contrasts are always available for each patient for various reasons, including cost or time constraints or image corruption from noise, patient motion, or inappropriate acquisition parameters. Clinical practice and automatic techniques would benefit significantly from synthesizing one or more of the missing 3D MRI volumes based on the others provided [254, 106]. Many different works [158, 113, 210, 41][1] can be utilized to synthesize the missing MR contrasts. However, given the challenges presented in synthesizing high-resolution volumes with pathologies, synthesized MR contrasts may not be reliable on their own. In this case, similar to segmentation uncertainties, synthesis uncertainties would be useful in determining the confidence in synthesized contrasts. In this thesis (Chapter 4) experiments are proposed and validated which show that by propagating synthesis uncertainty to the downstream task of tumour segmentation, we can improve segmentation results.

[1]To keep the thesis focused, our previous work [158] is included in the Appendix A.

Deployment of medical image analysis systems into real-world clinical contexts, particularly maintaining clinicians' trust, requires that robustness and fairness across different sub-populations are maintained [202]. A mismatch between the distribution of demographic information of patients used to train the system and the real-world distribution can hinder the performance of automatic medical image analysis systems [284, 202]. This is especially true for patients with brain tumours, a disease well-known for its heterogeneity. In a part of this thesis (Chapter 5), we examine the fairness and robustness of deep learning model outputs for brain tumour segmentation.

2.1.2 Multiple Sclerosis

Multiple Sclerosis is a chronic, inflammatory demyelinating disease of the central nervous system with presently no known cure [84]. It affects the health of nearly 2.3 million people worldwide and one hundred thousand Canadians [264]. The presence of lesions in MRI is one of the hallmarks of MS. As a result, MRI has been used to diagnose and monitor disease progression and treatment efficacy. The number of new or enlarging T2w lesions as well as gadolinium-enhancing lesions has been used as markers of disease activity [203, 172, 240]. Disease activity is used as a clinical outcome to monitor the progression of the disease and also the efficacy of new treatments in clinical trials for patients with relapsing-remitting MS (RRMS) [118, 242].

Previous work [222] demonstrated that future disease activity from baseline MR images could be predicted with relatively good accuracy from both MRIs and T2w lesion labels at the baseline. Their results suggest that the presence of T2w lesion labels plays a vital role in disease activity prediction accuracy. This result shows the need to delineate MS lesions from their surroundings, which is a tough task due to the heterogeneity of the lesions (Figure 2.3). Manual segmentation of these lesions is a long and arduous process, prone to human error. This process also results in inter-rater variability and disagreement between them, due to inherent sources of noise like an ambiguous boundary between struc-

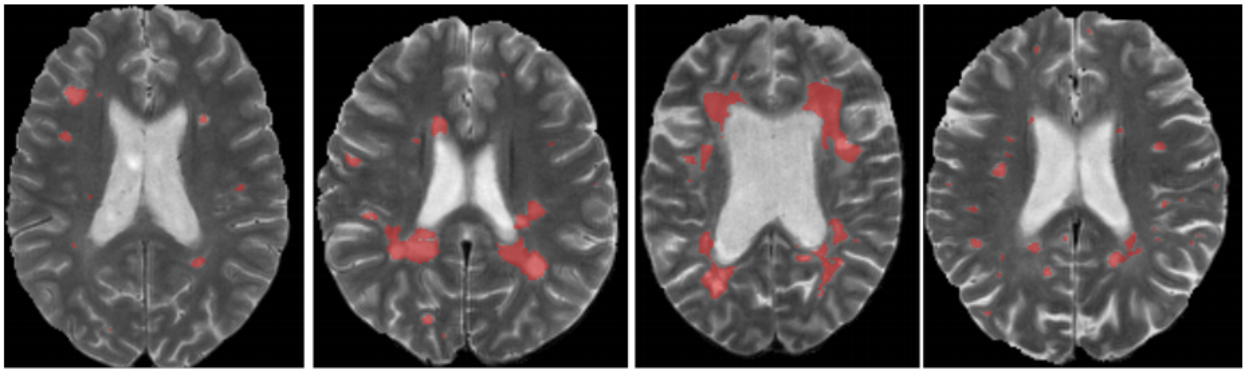


Figure 2.3: Variability of lesion size and location in a single slice across four different subjects. Manual, expert lesion labels in red are overlaid over a single slice of the T2 MRI modality. Image courtesy: NeuroRx (clinical collaborator).

tures, variations in acquisition parameters, the experience of clinicians, and variations in annotation "styles". As such automatic lesion segmentation methods would be useful for MS. However, as shown in [180], modern deep learning methods are prone to make mistakes in lesion segmentation, especially for small lesions, but uncertainty generated by these deep networks correlates with the mistakes. As a part of this thesis (Chapter 4), we show that by training a second segmentation network, which takes as input T2 lesion segmentation and associated uncertainty from [180], we could further improve the lesion segmentation/detection performance.

2.1.3 Alzheimer's Disease

Alzheimer's disease (AD) is a neurodegenerative disease known as the most common cause of dementia worldwide [234]. Older patients are more likely to suffer from AD than younger ones. While the general cause of AD is not easily understood, many symptoms are associated with AD. For example, memory loss, mood swings, disorientation, problems with language, cognitive ability, etc. Based on the symptoms, AD can be broadly classified into three different disease stages: (i) cognitive normal (CN) - patients who do not have any symptoms related to AD, (ii) mild cognitive impairment (MCI) - patients with early signs of AD, and (iii) Alzheimer's disease (AD) - patient with advance symptoms of AD (Figure 2.4).

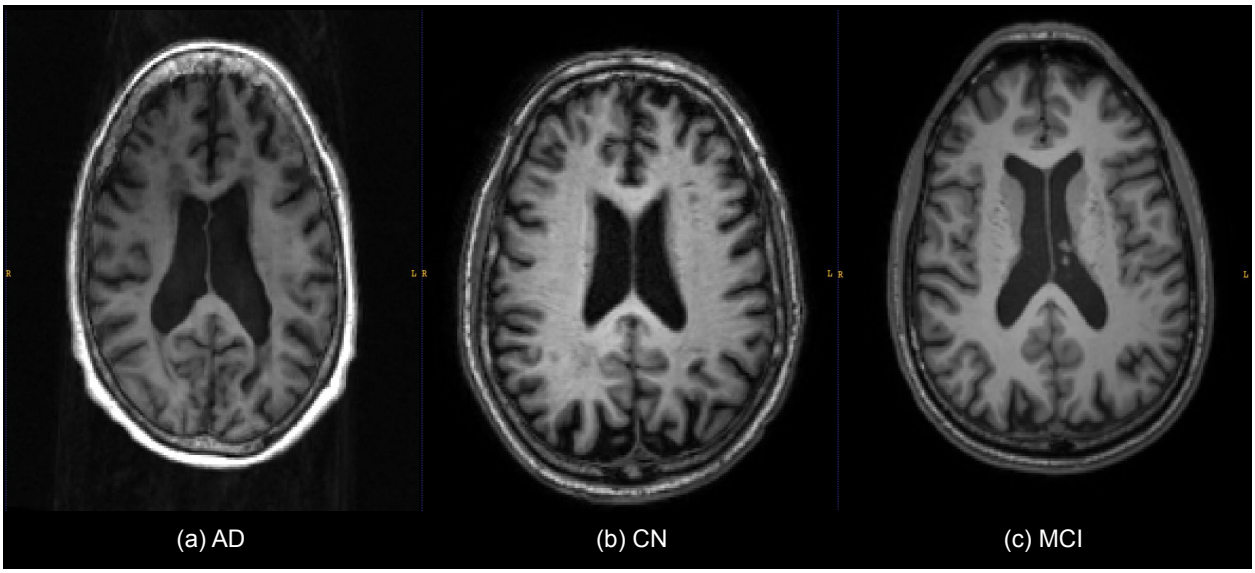
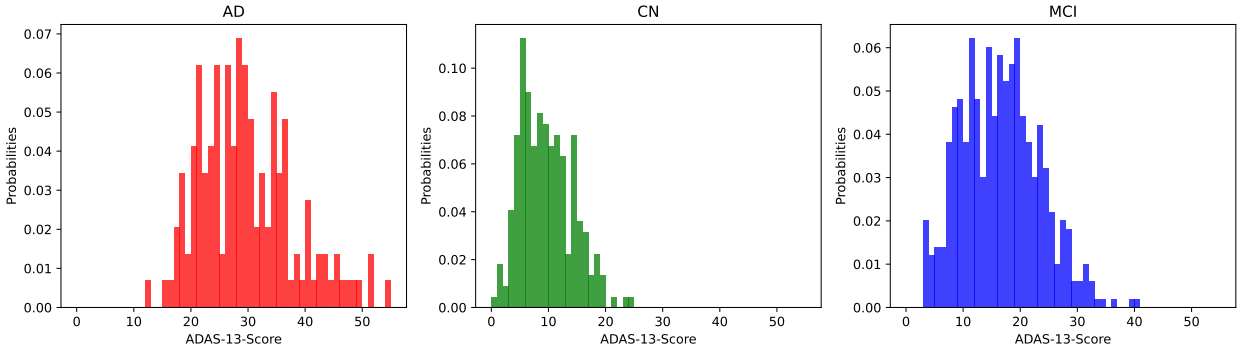
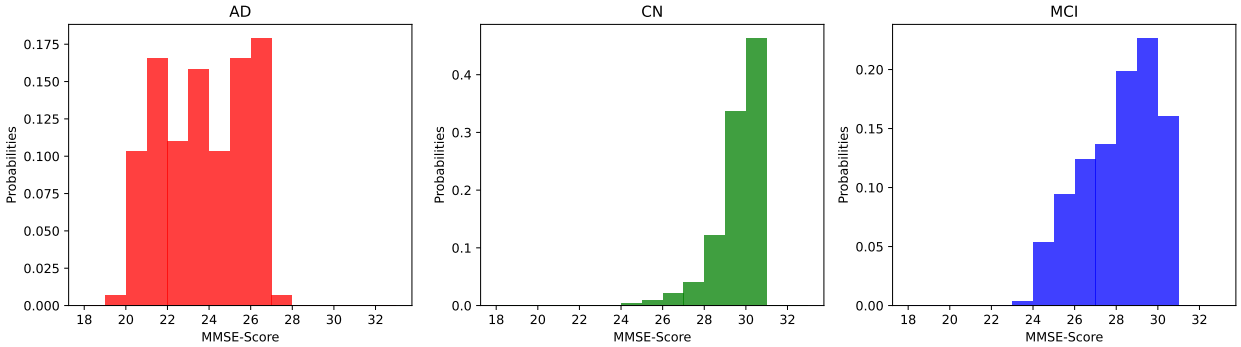


Figure 2.4: Example slice of structural T1 MR image for (a) a patient with Alzheimer's Disease (AD), (b) Cognitive normal (CN) patient, and (c) a patient with onsets of AD, also known as mild cognitive impairment (MCI). Image Courtesy: ADNI [111].

Many tests exist for the diagnosis of AD stages. For example, assessing the neuropsychological ability of patients through questions [207, 69], measuring the presence of neurological symptoms like amyloid plaques, measuring reduction in hippocampus volume associated with memory loss [73], measuring reduction cortical thickness through function and structural imaging. While neurological symptoms are better correlated with a patient's current state, it is the neuropsychological ability of a patient that can provide insights into the future progression of the disease. Clinicians are also more likely to treat symptoms based on the results of structured neuropsychological assessments. Two of the most popular assessment scores are the Alzheimer's Disease Assessment Scale (ADAS-13) [207] and Mini-Mental State Examination (MMSE) [69]. These scores vary greatly across disease stages (Figure 2.5). While a higher value of ADAS-13 is associated with AD, a lower value of MMSE indicates AD. These clinical scores can be directly predicted based on known AD biomarkers like hippocampus volume [28], which could help clinicians in making better diagnosis decisions. In this thesis (Chapter 4), we build on this principle and show that accuracy of machine learning models to predict clinical scores would greatly benefit from the propagation of uncertainty associated with machine learn-



(a) ADAS-13



(b) MMSE

Figure 2.5: Distribution of ADAS-13 (top) and MMSE (bottom) score for patients with (a) Alzheimer's Disease - AD, (b) Cognitive Normal - CN, and (c) Mild Cognitive Impairment - MCI. Image Courtesy: ADNI [111].

ing models trained for segmentation of the hippocampus. AD is a highly heterogeneous disease where its prevalence varies across demography [14]. In this thesis (Chapter 5), we also examine the ability of machine learning models to make unbiased clinical score predictions irrespective of the patient's age.

2.2 Background: Uncertainty Estimation in Deep Learning Models

In the following sections, we give a brief overview of the methods which allow us to capture uncertainty associated with the deep learning model output.

2.2.1 Multiple Sample Generation in Deep Models

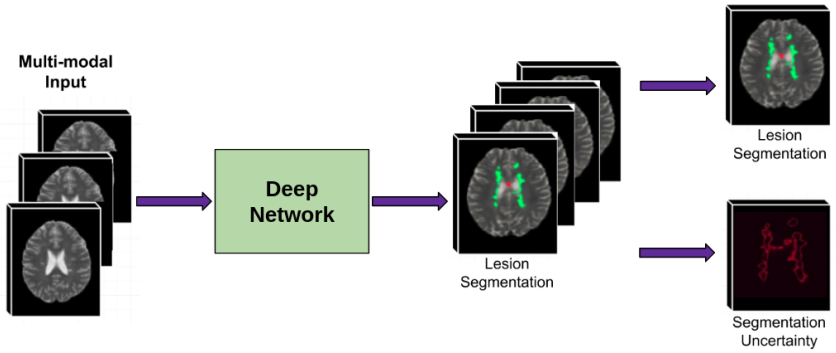
Multiple different methods exist in the literature that can generate multiple different samples for the same input in deep learning models. Broadly, they can be classified into two categories: (i) Bayesian neural networks, which includes monte-carlo dropout (MC-Dropout) [79], mean-field variational Inference [31], single-model deep uncertainty [142], radial Bayesian neural networks [66], Laplace approximation [204], etc. (ii) Ensembling methods, which provides uncertainty estimates based on ensembling. Examples for this category of methods include deep ensemble [136], stochastic-weight averaging - gaussian (SWAG) [147], batch ensemble [271], snapshot ensemble [101], hyperparameter ensemble [272], ensemble dropout [237], etc.

In this thesis, we focus on MC-Dropout, Deep Ensemble, and Ensemble Dropout. We chose these specific methods as they are easy to implement and have shown great success in the literature [79, 180, 114, 5, 11, 139, 208, 13, 71]. The following subsections provide a brief background of these methods.

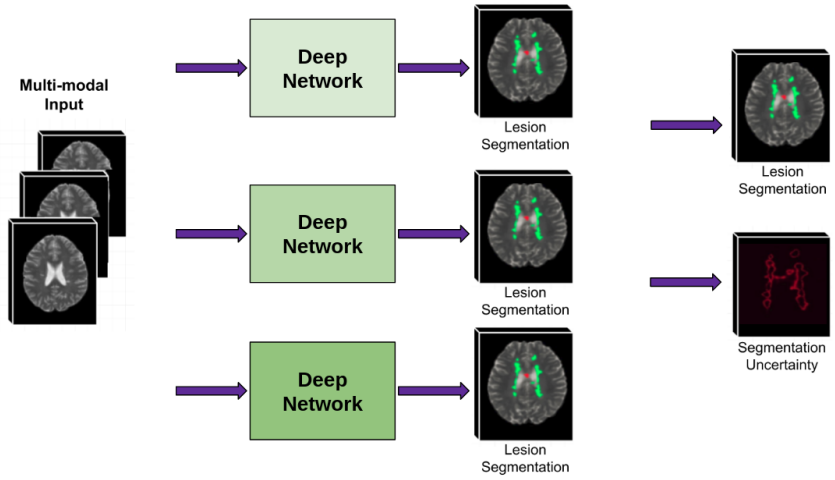
Notations: Consider a dataset $D_{train} = \{X, Y\} = \{(x_i, y_i)\}_{i=1}^{N}$ with N total samples. Here, $x_i \in \mathbb{R}^{P \times Q}$ or $x_i \in \mathbb{R}^{P \times Q \times S}$ represents 2D or 3D input image, and y_i represents corresponding ground truth labels. y_i depends on the task at hand: $y_i \in \{0, 1, .., C\}$ for image-level classification, $y_i \in \mathbb{R}$ for image-level regression, $y_i \in \{0, 1, ..C\}^{P \times Q}$ or $y_i \in \{0, 1, ..C\}^{P \times Q \times S}$ for 2D/3D voxel-level segmentation, and $y_i \in \mathbb{R}^{P \times Q}$ or $y_i \in \mathbb{R}^{P \times Q \times S}$ for 2D/3D voxel-level regression.

MC-Dropout

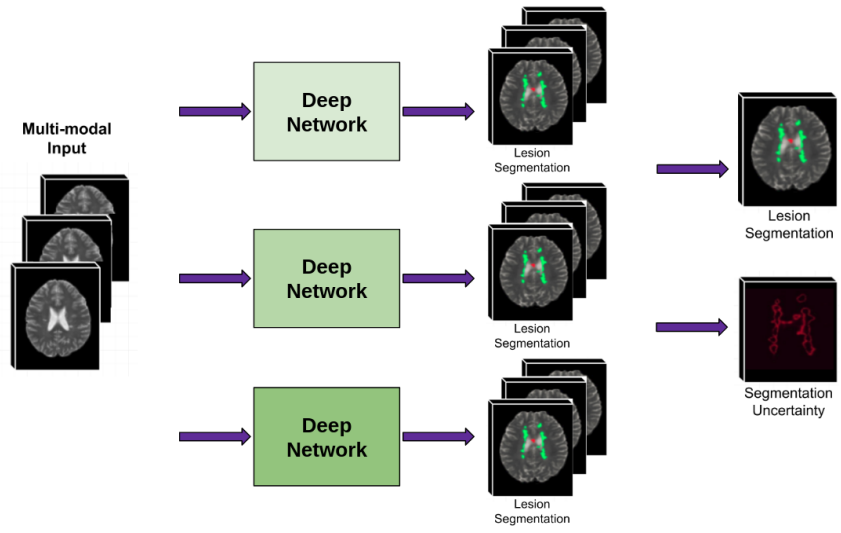
Bayesian neural networks (BNNs) are neural networks with prior probability distribution [146, 79], $p(W)$, placed on their weights, W: $W \sim p(W)$. For a training dataset D_{train}, given a likelihood, $p(y|x, W)$, the posterior distribution over the output, y, for input, x, can then be calculated as following:



(a) MC Dropout



(b) Deep Ensemble



(c) Dropout Ensemble

Figure 2.6: Illustration of sampling multiple outputs for the same input for three different Bayesian deep learning methods: (a) MC-Dropout, (b) Deep Ensemble, and (c) Dropout Ensemble. Here, we look at the MS lesion segmentation network for an illustration perspective. Mean segmentation and its corresponding segmentation uncertainty estimates are estimated based on generated multiple samples. Different shades of green colour for Deep Network represent different random initialization.

$$p(y = c|x, D_{train}) = \int p(y = c|x, W)p(W|D_{train})dW. \tag{2.1}$$

Since, in this case, exact inference often time is not tractable, various stochastic regularization techniques can be used to approximate it. The most common approximation method used in the literature for such networks is dropout [241]. In [79] (MC-Dropout) authors show that using dropout before every weight layer both at training and inference time is equivalent to performing approximate variational inference where the true posterior distribution over weights, $p(W|D_{train})$, is approximated with a tractable distribution $q(W)$. The parameters of this function are estimated by minimizing the Kullback-Leibler (KL) divergence between this function and the true posterior:

$$q^*(W) = \text{argmin}_{q(W)} KL(q(W)||p(W|D_{train})). \tag{2.2}$$

In [79], dropout is shown to be equivalent to the variational inference approximation where $q(W)$ is a mixture of two Gaussians with small variances where the mean of one of the Gaussians is set to zero. Having approximated the true posterior $p(W|D_{train})$ with $q^*(W)$, Equation 2.1 can be approximated using Monte-Carlo estimation:

$$p(y = c|x, D_{train}) \approx \int p(y = c|x, W)q^*(W)dW \approx \frac{1}{T}\sum_{t=1}^{T} p(y = c|x, \hat{W}^t). \tag{2.3}$$

where $\hat{W}^t \sim q^*(W)$. As is evident from Equation 2.3 MC-Dropout allows us to estimate the posterior probability over the output and, therefore, estimate the uncertainty associated with each prediction.

In MC-Dropout, the same input is passed through the neural network multiple times, leading to a collection of T different samples. Uncertainty is estimated using statistics computed across these samples. Interestingly enough, MC-Dropout can be seen as an ensemble method where an average of an ensemble of neural networks with shared pa-

rameters is used as the network prediction.

Deep Ensemble

In [136], the authors show how using an ensemble of independently trained deterministic deep neural networks that are randomly initialized, as known as Deep Ensembles, can reliably predict model uncertainty. This method is introduced as an alternative to the MC-Dropout method where instead of collecting predictions from passing through the same network T times, predictions are collected from T independently trained deterministic models. Deep Ensembles, if wide enough and under some assumptions, have recently been shown to exhibit similar training dynamics to a Gaussian process whose mean and variance do not always correspond to the posterior sampling of a probabilistic model [138]. Therefore, although non-Bayesian, deep ensembles can still be considered as a probabilistic method that can approximate model uncertainty.

Dropout Ensemble

Variational inference methods are prone to underestimate the posterior uncertainty as they tend to fit the approximate posterior probability to local modes. As a result, the MC-Dropout method is shown to exhibit similar behavior, at least for some regions [136]. To overcome this issue, a simple solution is to combine the MC dropout method with an ensemble method, where an ensemble of different dropout models is used to approximate the better posterior. MC-Dropout captures local variability across a single network and, in turn, captures how uncertain a single network is about its prediction. Deep Ensemble captures global variability in the prediction across different networks in an ensemble and uncertainty associated with this variability [71]. Thus, Dropout Ensemble can capture both local and global variability in model predictions [75].

Practically, Dropout Ensemble [237] combines both MC-Dropout and Deep Ensemble by training N independent networks in an ensemble and using dropout at test time for each of these networks to collect M different samples from each network. This results in a total

of $T = M * N$ sample outputs across these networks.

2.2.2 Uncertainty Measures

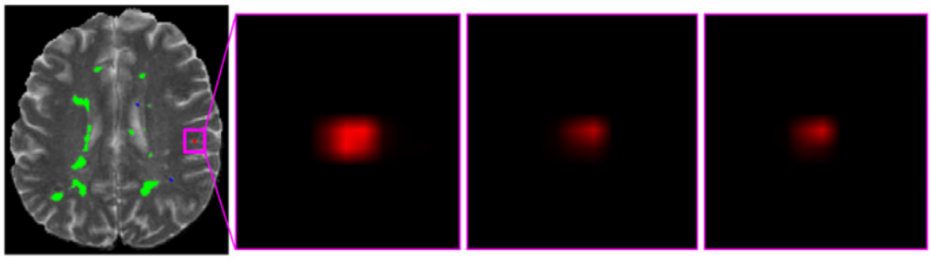


Figure 2.7: Examples of the segmentation and the uncertainty measures for MS lesion segmentation. From left to right: network segmentation output over the corresponding T2 slice, entropy uncertainty, mutual information uncertainty, and sample variance. Here, for uncertainty estimates, we look at the zoomed-in version of one particular false positive lesion. The segmentation is coloured with the following colour scheme: green true positive, pink false positive, and blue false negative. Increased intensity of red indicates greater uncertainty.

In this section, we give details about three popular uncertainty measures: sample variance, predictive entropy, and mutual information.

Sample Variance

The simplest uncertainty measure, sample variance, is estimated by computing the variance across the T samples collected using either Bayesian neural networks (e.g., [79]) or ensembles (e.g., [136]). For a regression task the variance in the output $\hat{y}_i$ for any input x_i, is defined as follows:

$$\text{Var}(\hat{y}_i) = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_{i_{(t)}}^2 - \left(\frac{1}{T}\sum_{t=1}^{T} \hat{y}_{i_{(t)}}\right)^2. \tag{2.4}$$

where $\hat{y}_{i_{(t)}}$ is a prediction for sample t.

For classification and segmentation tasks with C classes, the variance in the output $\hat{y}_i$ is defined as follows for any input x_i:

$$\text{Var}(\hat{y}_i) = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{T}\sum_{t=1}^{T} p(\hat{y}_{i_{(t)}} = c|x_i)^2 - \left(\frac{1}{T}\sum_{t=1}^{T} p(\hat{y}_{i_{(t)}} = c|x_i)\right)^2\right). \tag{2.5}$$

Here, $p(\hat{y}_{i_{(t)}} = c|x_i)$ denotes output softmax probability for class c for a sample t. Sample variance can be more simply interpreted as a measure of model output consistency across different samples.

Predictive Entropy

The predictive entropy is a measure of the informativeness of the model's predictive density function for each model output $\hat{y}_i$. It is defined as:

$$
\begin{aligned}
H[\hat{y}_i|x_i] &= -\sum_{c=1}^{C} p(\hat{y}_i = c|x_i) \log \Big(p(\hat{y}_i = c|x_i) \Big) \\
&\approx -\sum_{c=1}^{C} \Big(\frac{1}{T} \sum_{t=1}^{T} p(\hat{y}_{i_{(t)}} = c|x_i) \Big) \log \Big(\frac{1}{T} \sum_{t=1}^{T} p(\hat{y}_{i_{(t)}} = c|x_i) \Big).
\end{aligned}
\tag{2.6}
$$

where C is the total number of class labels, and $p(\hat{y}_{i_{(t)}} = c|x_i)$ denotes output softmax probability for class c for sample t. High entropy implies a flatter probability distribution across classes, while low entropy implies a more peaky probability distribution. Lower entropy shows that model is more confident in its prediction of the output class.

Mutual Information

The mutual information (MI) captures how much information we gain about the model parameters by knowing the label for input x_i. Similar to sample variance, mutual information also captures the variability in model predictions. MI is calculated as the difference between the entropy of the average model prediction ($\hat{y}_i$) and the average of the entropies of each model prediction ($\hat{y}_{i_{(t)}}$) [80]:

$$MI[\hat{y}_i, x_i] \approx H[\hat{y}_i|x_i] - \frac{1}{T}\sum_{t=1}^{T} H[\hat{y}_{i_{(t)}}|x_i]. \tag{2.7}$$

2.3 Application of Deep Learning Uncertainty Estimates in Computer Vision and Medical Imaging

In [121, 221, 97], the authors divide the uncertainty into mainly two parts: (i) epistemic uncertainty, and (ii) aleatoric uncertainty. Epistemic uncertainty captures the uncertainty associated with the model parameters. This uncertainty arises as our model parameters are not able to capture the true distribution of the model which generated our data. This uncertainty can become zero if we have an infinite amount of data, as it will allow our model parameter to learn the true distribution of the data generation model. This uncertainty is useful for capturing out-of-distribution examples. Recall that intensity values of MR images vary greatly across different scanner manufacturers [166]. Now consider an example where we train our machine learning model on MRI scans from Phillip and Siemens scanners, but during test time we use MRI scans from GE scanners. In this case, the epistemic uncertainty of the machine learning model for MRI scans from GE scanners will be higher than for Phillips or Siemens scanners, as the model has not seen the data from GE scanners during the training.

Aleatoric uncertainty captures the uncertainty associated with the inherent ambiguity in the data. Aleatoric uncertainty cannot be reduced even if more data were to be collected. Predictive entropy (Section 2.2.2) measures both epistemic and aleatoric uncertainties (which will be high whenever either epistemic is high or aleatoric is high) [78, 80]. A combination of both predictive entropy and MI (Section 2.2.2) together could be used to isolate the aleatoric uncertainty component through a simple subtraction if needed [78].

The aleatoric uncertainty is further divided into two parts: homoscedastic and heteroscedastic uncertainty. Homoscedastic uncertainty remains constant with the input, but instead,

it changes with a change in the task at hand. Heteroscedastic uncertainty changes with a change in input. Let us take an example of MRI segmentation problem. In this case, if we use an MRI scan of the same patient with different noise levels then the heteroscedastic uncertainty will change, while homoscedastic uncertainty will remain constant. Instead, if we change the task from segmentation to regression then for the same MRI scan of a patient with the same noise, it will result in the change of homoscedastic uncertainty, while heteroscedastic uncertainty will remain constant.

In [121], the authors showed how a neural network could learn to predict heteroscedastic uncertainty as an additional output for the task of classification and segmentation. They interpret this learned uncertainty as a learned loss attenuation, which makes the loss more robust to noisy data. In the end, they combined both epistemic and heteroscedastic (aleatoric) uncertainty. Through experiments on various computer vision datasets [33, 233, 216] for the task of semantic segmentation, the authors showed why it is necessary to combine both uncertainty estimates. Through these experiments, they demonstrated how epistemic uncertainty decreases when the model is trained using more data, but it does not result in a decrease of aleatoric uncertainty.

Homoscedastic uncertainty can be thought of as a task-dependent uncertainty that changes with a change in task. In [122], authors used homoscedastic uncertainty to weigh different losses in a multi-task setting. They experimented on the cityscape dataset [48] for the task of semantic segmentation, instance segmentation, and depth estimation. They showed how a single parameter associated with each task can help in learning the relative weights of these tasks in a multi-task setting. This weight is inversely proportional to the homoscedastic uncertainty of each task. Their results indicate that multi-task learning indeed helps to improve overall performance when the weights of each task are learned using homoscedastic uncertainty; it gives better performance in comparison to when weights are manually tuned.

One of the disadvantages of the model presented in [121] is that it introduces an additional output of the network to learn heteroscedastic uncertainty. This additional output increases the burden on the network. In addition to this, learned uncertainty (variance) is also not directly dependent on the model output (mean). Although this assumption of independence can be valid in some scenarios, it may not always be accurate. In [134] paper, authors derive an approximation of the uncertainty, such that, there is no need for extra parameters to learn heteroscedastic uncertainty. They show that this approximation of heteroscedastic uncertainty represents associated uncertainty with input in a better way that the approximation derived in [121], through experimentation on ischemic stroke lesion segmentation (ISLES) 2015 challenge dataset [149].

Many papers in the applied machine learning field utilize uncertainty estimation of deep learning model output [120, 180, 248, 82]. Bayesian SegNet [120] was one of the first papers to use MC-Dropout based uncertainty estimate in a classical computer vision task of semantic segmentation. In the paper, authors extended the paper on SegNet [17], which was the state-of-the-art deep network for semantic segmentation at that time, to include MC-Dropout based uncertainty estimate. It was shown that with an increase in the number of MC samples, the overall accuracy of the model prediction increases. Moreover, the performance of the system is higher than the standard weight averaging method. They also experimented with various variants of Bayesian SegNet, where they placed the dropout layers at different depths and sides (encoder and decoder). They concluded that when dropout is applied only at the layers with the lowest resolution, we get maximum performance. The experiments were done on standard computer vision datasets for semantic segmentation [65, 33, 239]. They also reported that augmenting any semantic segmentation architectures [144, 17, 102] with dropout at test-time (MC-Dropout [79]) can help in improving the performance of the networks by 2-3%.

Spurred by the success in computer vision tasks, uncertainty estimation has recently been used in a variety of medical imaging tasks [188, 208, 139, 16], ranging from MR super-

resolution [246] to nodule detection [188] and lesion detection and segmentation [180]. In [16], the authors proposed and validated that just by using augmentation at the test time, the network can produce better uncertainty estimation. They experimented on the Kaggle dataset of diabetic retinopathy (DR) for the 5-class classification task.

Papers like [208, 246] show that predicted uncertainty indeed correlates with the places where a model is prone to make mistakes. All papers experimented on the different types of tasks and different types of applications. [208] showed that when voxel-wise uncertainty estimation is converted into structure-wise uncertainty it correlates with the estimated dice score by evaluating segmentation output and uncertainty on publicly available brain sub-structure segmentation dataset. [246] performed similar type of analysis for dMRI super-resolution.

Other medical image analysis papers demonstrate improved performance when the network output is evaluated on its most certain predictions [180, 139]. In [180], the authors explored the above-mentioned uncertainty measures (Section 2.2.2) for the task of multiple sclerosis (MS) lesion segmentation and detection. Through experimentation on a private multi-cite multimodality MRI dataset, it was shown how uncertainty measurements could be useful in choosing better operating points. In [139], authors evaluated MC-Dropout uncertainty measures in diagnosing DR from fundus images. The uncertainty-informed decision referral was shown to improve diagnostic performance. A similar type of analysis was done for mammograms and chest X-ray in [248] and [82], respectively.

Other work also exploits uncertainty estimation to improve model performance [188, 277, 218]. Uncertainty estimation is also used in semi-supervised scenarios for improved segmentation of left atrium from chest MRI [277] and retinal layers from OCT images [218]. Uncertainty estimation-based active learning [227, 285] and omni-learning [260] methods try to address data scarcity problems in medical imaging.

While the above mentioned approaches illustrate how estimating uncertainty in medical imaging tasks is helpful in a clinical scenario, they do not show how uncertainty can be used to inform or improve network performance on a downstream task. Recent work in medical imaging has demonstrated how uncertainty estimates can be used to improve model performance [188, 95]. In [188], authors show that uncertainty generated from a 2D lung nodule segmentation network can be used to reduce the *false positives* in a subsequent 3D detection network centered on regions of interest. Although appropriate in the context of lung nodule detection, this is not the general case in medical imaging applications where false negative reduction is also, sometimes, more so, of interest. For example, in MS lesion detection, one false negative lesion can convert a patient from active to inactive, and in turn change the course of the patient treatment [117, 118]. Furthermore, existing works [188, 95] only explore uncertainty propagation when both inference steps are similar to each other. As we discussed in the Introduction (Chapter 1), a typical medical image analysis pipeline involves a variety of cascaded inference tasks. It is important to explore whether propagating uncertainty estimates from different but related tasks, can lead to better performance or not. In Chapter 4, we tackle this open problem, and explore and validate that the propagation of uncertainty maps from an upstream task can improve performance on a related but dissimilar task.

2.4 Evaluating Uncertainty Produced by Deep Learning Models for the Task of Interest

Some of the most popular metrics for measuring [79, 136, 250, 13, 271] model confidence output are the expected calibration error (ECE) and the maximum calibration error (MCE). These metrics measure the difference between the predicted output softmax probabilities of the model matches the actual probabilities of the correct ("ground-truth") prediction. In general, both metrics require softmax predictions ($\hat{p}_i$) to be grouped into M bins based on softmax score (ranging from 0 to 1). The most common way of group-

ing is to place all predictions into bin B_m if their confidence score falls in the interval $(\frac{m-1}{M}, \frac{m}{M}]$. Using $\hat{p}_i$ to represent the model's confidence in its predicted class, $\hat{y}_i$, and y_i to denote the ground truth class label, [87] then define the accuracy and confidence associated with each bin as: $\text{acc}(B_m) = \frac{1}{|B_m|}\sum_{i\in|B_m|}\mathbf{1}(\hat{y}_i = y_i)$ and $\text{conf}(B_m) = \frac{1}{|B_m|}\sum_{i\in|B_m|}\hat{p}_i$. ECE is calculated as the average difference between accuracy and confidence for each bin, while MCE is calculated as the worst-case difference. However, these metrics are biased estimates, as they cannot differentiate between a model that makes constant uniform predictions and another model that makes predictions with actual probabilities of the correct ("ground-truth") prediction [13, 186, 87]. Another issue with these metrics is that are based on softmax probabilities, which cannot capture epistemic or aleatoric uncertainty [78]. As we discussed previously, these uncertainties are helpful in differentiating between different sources of uncertainties. We require measures like entropy and MI to capture aleatoric and epistemic uncertainties.

In [136], authors evaluate the usefulness of the predictive uncertainty (measured by entropy) for decision-making. They evaluate the model outputs only in cases where the model's confidence is below a user-specified threshold. For example, let us consider a model that makes a total of 100 predictions, and also consider the user-specified confidence threshold as 0.8. In this case, the authors only evaluate predictions whose confidence is below 0.8. Any predictions whose confidence is above 0.8 are "referred" to the end-users. This has potential in medical image analysis, where we want clinicians to make the final decision for machine learning model prediction. Through experiments on toy datasets like MNIST, the authors of [136] showed that when the model is evaluated on its most confident prediction, the model accuracy is high compared to when the model is evaluated on all outputs. This demonstrates that when the model is confident, it is usually more correct in its predictions. Though this is encouraging and could potentially be useful for the comparison of different uncertainty generation methods, it does not consider how many model predictions were discarded at a certain threshold. Consider a model, which has really low accuracy for its predictions, but uncertainties associated with most of these predictions are really high. The above-mentioned evaluation criteria would still

consider this to be a good model, as it will produce high accuracy when discarding (referring) highly uncertain predictions. However, using such a model in a clinic would increase the burden on the reviewing clinicians, as most of the machine learning model predictions need to be reviewed by the clinicians due to their high uncertainties. As such, this type of model would not be much useful in a practical sense.

[176] designed a metric to quantify uncertainty for the task of semantic segmentation of computer vision images. They made the following assumption during the metric design: if a model is confident about its prediction, it should be accurate. This also implies that if a model is inaccurate on output, it should be uncertain. The converse of these assumptions may not hold. For instance, a model may have a high epistemic uncertainty on a class which appears infrequently in the training set but can still be accurate on its prediction. With this in mind, they calculate the following two probabilities at different uncertainty thresholds: (i) p(accurate|certain): the probability that the model is accurate on its output given that it is confident; (ii) p(uncertain|inaccurate): the probability that the model is uncertain about its output given that it has made a mistake in its prediction (i.e., is inaccurate). They used the metric to compare different BDL methods for the semantic segmentation task. Though this metric is useful for semantic segmentation, where each pixel in an image is labeled as one class, it is not useful for the task of pathology segmentation where there is a high class-imbalance problem, and the number of pixels (voxels) of interest (pathology) is low compared to the background-healthy class. For example, in the brain tumour segmentation task, 99.9% of the voxels belong to the background (healthy tissue), while only 0.1% belongs to the foreground (pathology). Due to a high-class imbalance, p(accurate|certain) would be dominated by healthy (background) voxels, most of which can be accurately classified with high certainty.

[90] developed a metric, probability-based detection quality, to evaluate the uncertainty estimate for the object detection task. The authors combine the class labeling measure (i.e., label quality) and the bounding box detection measure (i.e., spatial quality) into the

metric. Here, spatial quality measures how well the detection describes where the object is within the image. Label quality measures how effectively a detection identifies the object class. These are averaged over all possible combinations of bounding boxes and labels generated using multiple samples. The authors also organized a challenge associated with this task at the annual conference on computer vision and pattern recognition (CVPR) 2019. The paper and its associated challenge [245] illustrate the importance of developing uncertainty quantification metrics that are tailored to the task of interest.

In many recent papers related to medical imaging [277, 82, 180, 248], it has been shown that uncertainty estimation indeed correlates with the places where the network is prone to make errors. These results are indeed useful and can lead to better adaptation of deep learning models in real-world scenarios. However, as we saw in the above paragraphs, in the medical image analysis field, to date, there is an unmet need to (1) systemically quantify and compare how well different uncertainty estimates properly communicate the degree of confidence in the output and (2) rank the performance of competing estimates, given the objectives of the task and the requirements during a clinical review.

[114] made the first step towards quantifying uncertainty for the brain tumor segmentation task. They compared various uncertainty generation methods such as MC-Dropout, Deep Ensemble [79, 136], and others, using standard metrics like ECE, MCE, and reliability diagrams. In addition, they proposed a new metric, uncertainty-error (U-E) overlap. The results showed that Deep Ensemble could produce more reliable uncertainty measures than other methods. One of the limitations of the above work was that they only report U-E overlap for a specific threshold where U-E overlap was highest. In a real world scenario, we prefer if end users define this threshold, and want to make sure that different uncertainty generation methods are evaluated on various ranges of these thresholds before concluding which one produces more reliable uncertainties. Similarly, U-E overlap doesn't consider the actual metric of interest (e.g., dice) and doesn't differentiate between false positives and false negatives, which changes based on the size of the pathology

(e.g., tumour). Keeping all this in mind, in Chapter 3 we propose a new evaluation metric, specifically designed for quantifying the uncertainty of brain tumour segmentation task, that overcomes this limitation. More details about how we overcome this limitation are given in the respective chapter.

2.5 Uncertainty Estimation in Multi-rater System

Several recent machine learning approaches [130, 27, 174] address an additional type of uncertainty in medical image analysis caused by the fact that a unique label cannot necessarily be attained in some regions of an image (e.g., at boundaries between tumour and healthy tissue in MRI). These papers assume that, in this case, they have access to many different annotators, and these annotators might systematically label things differently. They then model these inherent uncertainties (in various ways) using label variability as a proxy. Given the requirement of having access to multiple annotations, a context that is not common in practice due to the expenses incurred in attaining them, we do not focus on this type of ambiguity directly in this thesis. Nonetheless, for the sake of completeness, we provide details of some of the relevant work below.

One of the most straightforward ways of designing a system that can mimic multiple different annotations from different raters is to train an ensemble machine learning (deep learning) model, where each individual model is trained on a different "ground truth" annotation. These models can generate a hypothesis similar to different annotations, which allows the overall system to model inter-rater variability.

Another plausible solution explored in [211] was to train a single network with M heads. Experimentation on optical flow and human pose estimation showed that this indeed results in overall better performance and better uncertainty estimation compared to [79] and [136]. Though, these results are good and show promising results, two common dis-

advantages of both models (ensembles and M heads) are their scalability issues regards to large numbers of hypotheses and their requirement of fixing the number of allowed hypotheses at training time.

There has been quite substantial work done to counter the above-mentioned disadvantages. One of the pioneering works in this field was done in a probabilistic U-Net [130] paper, where conditional variational auto-encoder (cVAE) was combined with the famous U-Net architecture [26]. The authors showed that, given ground-truth annotations from multiple experts, the method could produce an unlimited number of realistic segmentation samples. Moreover, the method was shown to outperform various related methods, including network ensembles, M-heads [211], and the Bayesian SegNet [120]. In [100], authors extend this work and show that it is possible to model both epistemic (model) and aleatoric (data) uncertainty using this approach, and when these uncertainties are modeled separately, model performance increases.

Though the results of [130, 100] are promising, one of their drawbacks is that stochasticity is only introduced in the last stage of U-Net, which restricts its representation power and does not allow it to reconstruct multiple "ground truth" marking available. To overcome this issue, two parallel papers, hierarchical probabilistic U-Net [130] and probabilistic hierarchical segmentation (PHiSeg) [27], introduces stochasticity at all different levels of U-Net architectures, inspired by Laplacian Pyramids. The model generates image-conditional segmentation samples by generating the output at a low resolution and then continuously refining the distribution of segmentations at increasingly higher resolutions. Through experimentation on a publicly available LIDC-IDRI dataset of CT lesions [12], authors show that this approach can mimic inter-rater variations better than [130]. These results show the usefulness of these cVAE based approaches. It should be noted that these approaches are validated only on a dataset that has a single lesion or structure of interest in the image, and the size of the structure of interest is somewhat homogeneous. These approaches need to be validated in a scenario where there is more than one disconnected

pathology with varying size and shape, i.e. MS T2 lesions. One more thing that needs to be noted is that both Prob-U-Net and PHiSeg use unimodal Gaussian distribution to parameterize the network, as in the case of VAE models. Though this is useful, it is a very simplifying assumption, and many medical imaging tasks may not be able to be modeled by this. To overcome this limitation, recent papers [219, 29] utilized advanced methods like normalizing flows [127, 189], which can model more complex distributions.

2.6 Fairness of Deep Learning Methods

In addition to uncertainty estimation, robustness and fairness across different sub-populations are also required to maintain the trust of clinicians, and deploying DL models in real clinical contexts. This means a machine learning model should maintain its performance when evaluated on different sub-populations. A recent review [202] discusses the necessity to address the issue of fairness, potential sources of biases, and the remaining challenges, for machine learning models in medical imaging. According to them, there are mainly three different sources that can lead to a biased (unfair) system (1) the data being fed to the system during training, (2) design choices for the model, (3) and the people who develop those systems. This work mainly focuses on the first two sources of potential biases.

Several recent studies have indeed exposed significant biases in DL models across subpopulations (e.g., according to race, sex, age) in the context of medical image analysis [284, 137, 35, 225, 197]. In [137], it is shown that a Computer-Assisted Diagnosis system trained on a predominantly male dataset for diagnosing thoracic diseases gives lower performance when tested on female patient images. In [35], the authors show how data imbalance in the training dataset leads to a disparity in accuracies across subpopulations (dark vs. light skinned individuals) in diagnosing diabetic retinopathy from fundus images. [225] exposed a significant underperformance of X-ray pathology classification models when evaluated on groups under-represented in the training dataset (e.g., black patients). [197] found that segmentation models deployed in cardiac MR image

analysis pipelines exhibit a racial bias resulting from imbalanced training data. A similar analysis has also been performed for brain MR segmentation [107].

Several methods have been proposed in the machine learning literature to mitigate the lack of fairness [155] in the models. Popular fairness mitigation methods include data balancing [112, 105], where the training dataset is balanced across the sensitive attribute (e.g., sex - male vs. female). The hypothesis is that when a machine learning model is trained with a balanced dataset in this manner, it should not be biased toward one of the sensitive attributes (ex. male). Data balancing has shown to be successful for some medical imaging contexts [197, 107]. While several other fairness mitigation methods exist in the literature [268, 148, 282, 124, 249, 215, 40, 70], one of the most popular and effective methods is known as group distributionally robust optimization (GroupDRO) [213]. It tackles the fairness problem from an optimization perspective by minimizing the worst-case training loss across different subgroups. In [284], authors provide a framework to benchmark the fairness of several machine learning models for medical image analysis. Through several experiments on ten different medical image analysis datasets and eleven different fairness mitigation methods, they find that state-of-the-art bias mitigation algorithms do not significantly improve fairness outcomes.

Most fairness mitigation methods focus on correcting performance differences across subgroups without considering their effect on the uncertainties associated with the model output. As we briefly discussed in the Introduction (Chapter 1) and will see in the subsequent chapters of this thesis (Chapter 3, Chapter 4), real clinical contexts would benefit from knowledge about confidence in the model predictions when made explicit in the form of uncertainties [24]. Specifically, the trust would be established should uncertainties associated with the predictions be higher when the model is incorrect, and low when model outputs are correct. Various successful frameworks for quantifying models uncertainties in the context of medical image analysis have been presented for tasks such as image segmentation [180, 115], image super-resolution [247], and image classification

[173, 82]. However, these methods only analyze the output uncertainties for the entire population, without consideration of the results for population subgroups. In Chapter 5, we analyze the popular fairness mitigation methods both in terms of their absolute performance and quantification of their output uncertainties for three different medical image analysis tasks.

2.7 Active Learning

The performance of deep learning methods is largely dependent on the availability of large, labeled datasets for model training [244]. However, large, annotated datasets are not widely available in medical image analysis due to the prohibitive time, costs, and challenges of labeling large datasets. The labeling task is particularly challenging in patient images with pathological structures (e.g., lesions, tumours) and requires significant clinical and domain expertise. Various approaches have been proposed for optimally leveraging a small subset of annotated data that has been (passively) provided along with an otherwise unlabelled medical imaging dataset. These approaches range from transfer learning [43, 104], weakly supervised [193, 123], semi-supervised [190, 81] to synthetic data generation [175, 72].

Active learning (AL) frameworks [223, 280], on the other hand, have been successfully developed for "human-in-loop" computer vision [229] and medical imaging classification contexts [280]. AL tries to maximize the performance of a machine learning model while annotating the fewest training samples possible. These AL approaches work by training a model on a small, available, labeled subset, running inference on the larger unlabeled dataset, and then identifying an optimal set of samples to be labeled and added to the training pool in an iterative fashion. Sampling is optimized to attain the highest performance with the smallest number of samples. Sampling strategies can be broadly categorized as: (i) *uncertainty based*, which includes selecting samples with the least confidence in its estimated most probable class [49], the smallest margin between the first and sec-

ond most probable classes [217], the maximum predicted entropy [226], the minimum expected generalization loss [209], as well as deep Bayesian active learning approaches [80] (MCD-Entr and MCD-BALD [98]) and (ii) *representative based*, which focuses on selecting the most representative and diverse images from the unlabeled set (e.g., CoreSet [220], variational adversarial [235, 232], reinforcement learning [266]). Combinations of multiple strategies [275, 125, 276] have also been proposed.

In this thesis, we focus on uncertainty-based active learning approaches. Generally, these approaches, particularly entropy-based methods, have been popular in medical imaging contexts where they have shown some effectiveness in addressing the issue of high-class imbalance. In [80], authors showed that when the acquisition function in AL, is dependent on uncertainty approximated using MC-Dropout [79], it gives better performance than other state-of-the-art AL methods. They explored various functions like entropy, variation ratio (VR), and mean standard deviation (MSD), as an uncertainty estimate. They concluded that MI, VR, and entropy outperform MSD, while VR achieves the best performance among all others. They reported the usefulness of the AL framework in a clinically relevant task of cancer diagnosis from image data of skin segments. [270] proposed an active learning method that uses uncertainty sampling to support quality control of nucleus segmentation in pathology images. Through extensive experiments, they found uncertainty sampling for deep networks to be most useful for nucleus segmentation. [131] exploited geometric uncertainty for mitochondria segmentation from EM images and tumour segmentation from MR images. The authors of [275] introduced Suggestive Annotation, a deep active learning framework for medical picture segmentation that combines a representativeness density weighting method with an alternate formulation of uncertainty sampling. They demonstrated state-of-the-art performance using 50% of the available data on the MICCAI Gland segmentation challenge and a lymph node segmentation task. [236] suggested MedAL, an active learning framework for segmenting medical images. To obtain the most useful samples from an unlabeled data set, they suggested a sampling technique that combines uncertainty and the distance between feature descriptors.

Entropy-based methods select the samples which are the hardest for the current model to classify, however, entropy alone does not convey the particular source of the uncertainty (e.g., which classes are the source of confusion in a multi-class classification task). In addition, it does not provide information about how the addition of the sample's labels will influence downstream performance. The assumption for entropy based methods is that selecting these hardest samples, labeling them, and retraining the model with these samples would improve the model performance. However, just because these samples are the hardest for the current model to classify, does not necessarily mean that they will lead to improvement in the performance of the model on a real-world (ex. test) dataset. In Chapter 6, we develop a new active learning acquisition function that explicitly measures the information gain on an unseen (evaluation) set. The hypothesis is that selecting samples based on this acquisition function should lead to better performance of the model on a real-world dataset.

2.8 Summary

In this chapter, we provided the necessary background for three different clinical contexts, we discussed how we will tackle specific subproblems related to these contexts in the rest of the thesis. We looked at different methods for generating multiple samples from deep learning models, and how different uncertainty measures can be calculated from these samples. Following this, we provided deep dive into the related work for the application of uncertainty estimates in computer vision and medical imaging, fairness of deep learning models, and active learning framework. We set up the open problems in these fields. In the subsequent chapters, we will utilize this knowledge, and provide the details of contributions made in this thesis.

3

Evaluating Uncertainty Estimates in Brain Tumour Segmentation

> Fisher realized that the uncertain answer to the right questions is much better than a highly certain answer to the wrong question.
>
> — *Dana Mackenzie and Judea Pearl,* The book of Why

Related Paper

It should be noted that this is not a manuscript based thesis. However, considerable material from the following paper has been utilised in this chapter.

- **R. Mehta**, A. Filos, U. Baid, ..., S. Bakas, Y. Gal, T. Arbel, (95 authors), "QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain tumour Segmentation - Analysis of Ranking Scores and Benchmarking Results", *The Journal of Machine Learning for Biomedical Imaging (MELBA), August 2022* [161].

3.1 Introduction

In the previous two chapters, we provided a brief introduction and relevant background and literature review for the thesis. In this chapter, we focus on the development of an uncertainty evaluation metric for the brain tumour segmentation task, analyze various parts of it, and use it to rank uncertainty provided by participants in the popular brain tumour segmentation (BraTS) challenge. Our results indicate that tumour segmentation output and its associated uncertainties give complementary information, and these uncertainties can give information necessary for integrating these models in real clinical practice.

The main focus of this part of the thesis is three-fold: i) to develop an uncertainty evaluation score with a down-stream clinical goal in mind; ii) to benchmark the various participating teams from a recent BraTS challenge [22], using the developed evaluation score; and iii) to make the associated evaluation code publicly available for future benchmarking of uncertainty estimation methods for medical image segmentation. In particular, we focus on developing an uncertainty evaluation criterion for brain tumour segmentation.

We aim to develop a computer-aided diagnosis (CAD) system where the pathology size is smaller than the surrounding healthy tissue. In this context, the objectives are that the uncertainty estimates associated with an automatic segmentation system reflect that the system is (a) confident when correct and (b) uncertain when incorrect. These criteria would mainly permit uncertain predictions to be flagged and brought to the attention of the clinical expert, rather than overburdening the expert by having to review the entirety of the prediction. To this end, we present the resulting uncertainty evaluation score [162] and the rankings and results for 14 teams participating in the quantification of uncertainty for brain tumour segmentation (QU-BraTS) 2020 challenge. The various analyses of the methods and results produced by the different teams highlight the necessity of the different components of our developed score. The results indicate that the segmentation results and the associated uncertainties give complementary information as teams performing well on one task do not necessarily perform well on the other. Qualitative results show that the developed score measures the desired real-world properties for tumour segmentation uncertainties.

3.2 Uncertainty Evaluation Score

The objective of the uncertainty quantification task is to evaluate and rank the uncertainty estimates for the task of brain tumour segmentation. To this end, each uncertainty estimation methods provide output labels for the multi-class segmentation task and the estimated voxel-wise uncertainties for each of the associated tumour entities, namely, whole tumour (WT), tumour core (TC), and enhancing tumour (ET). These uncertainties are required to be normalized in the range of $0 - 100$ for ease of computation. For each tumour entity, the uncertain voxels are filtered at N predetermined uncertainty threshold values $\tau_{1,..,N}$, and the model performance is assessed based on the metric of interest (i.e., the Dice DSC in this case) of the remaining voxels at each of these thresholds ($\tau_{1,..,N}$). For example, $\tau = 75$ implies that all voxels with uncertainty values ≥ 75 are marked as uncertain, and the associated predictions are filtered out and not considered for the subsequent DSC

calculations. In other words, the DSC values are calculated for the remaining predictions of the unfiltered voxels. This evaluation rewards models where the confidence in the incorrect assertions (i.e., false positives, denoted FPs, and false negatives, denoted FNs) is low and high for correct assertions (i.e., true positives, denoted TPs, and true negatives, denoted TNs). For these models, it is expected that as more uncertain voxels are filtered out, the DSC score, calculated only on the remaining unfiltered voxels, increases.

Although the criterion mentioned above helps to measure performance in terms of DSC, the metric of interest, it does not keep track of the total number of filtered voxels at each threshold. In real practice, an additional penalty should be provided to a system that filters out many voxels at a low threshold to achieve high performance on the metric of interest, as it will increase the reviewing burden on clinical raters. One solution is to add a penalty based on the total number of filtered voxels at each uncertainty threshold. This strategy is also not ideal as it will also penalize methods that filter out FPs/FNs, areas where mistakes are made. Instead, the evaluation criterion chosen penalizes methods that filter out only the correctly predicted voxels (i.e., TP and TN). Given that the specific tumour segmentation task has a high-class imbalance between pathological and healthy tissue, different penalties are assigned to TPs and TNs. The ratio of filtered TPs (FTP) is estimated at different thresholds ($\tau_{1,..,N}$) and is measured relative to the unfiltered values ($\tau = 100$) such that FTP = (TP_{100} - TP_{τ}) / TP_{100}. The ratio of filtered TNs is calculated similarly. This evaluation essentially penalizes approaches that filter out a large percentage of TP or TN relative to $\tau = 100$ voxels (i.e., more uncertain about correct assertions) to attain the reported DSC value, thereby rewarding approaches with a lower percentage of uncertain TPs/TNs.

Figure 3.1 and Table 3.1 depict qualitative examples and their associated quantitative results. Here, decreasing the threshold (τ) leads to filtering out voxels with incorrect assertions. This filtering, in turn, leads to an increase in the DSC value for the remaining voxels. Example 2 indicates a marginally better DSC value than the slice in example 1 at

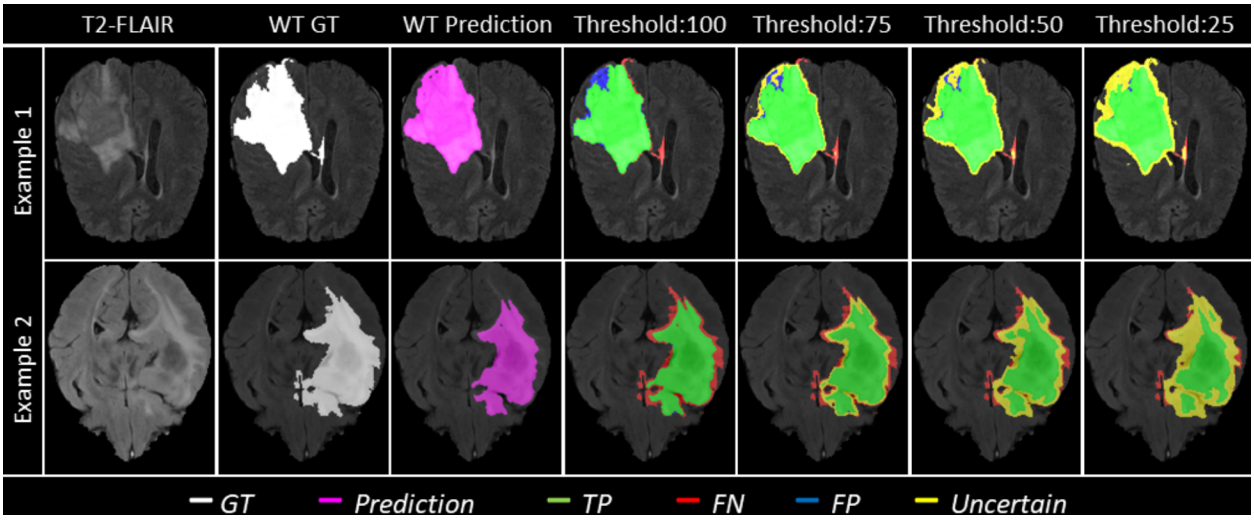


Figure 3.1: Effect of uncertainty thresholding on two different examples of patient MRI slices (Row-1 and Row-2) for whole tumour (WT) segmentation. (a) T2-FLAIR MRI. (b) WT ground truth (c) overall model prediction (d) results with no filtering, uncertainty threshold = 100. (e) uncertainty threshold = 75 (f) uncertainty threshold = 50 (g) uncertainty threshold = 25. It is desired that with decrease in the uncertainty threshold, more false positives (blue) and false negative (red) voxels are filtered out (marked as uncertain - yellow) while true positive (green) and true negative voxels remain unfiltered. ©[2022] CC-BY. Reprinted, with permission, from [161].

uncertainty thresholds (τ) 50 and 25. However, the Ratio of FTPs and FTNs indicates that this is at the expense of marking more TPs and TNs as uncertain.

To ensure that the generated output segmentations are directly associated with the BraTS challenge protocol, the generated uncertainties are expected to be produced for these "binary" tumour entities, i.e., ET, TC, and WT. The associated uncertainties are evaluated using the scores defined above for each tumour entity.

Finally, the resulting uncertainty measures for each team are ranked according to a unified score that combines the area under three curves: 1) DSC vs τ, 2) FTP vs τ, and 3) FTN vs τ, for different values of τ. The unified score is calculated as follows:

$$\text{score}_{\text{tumour\_entity}} = \frac{AUC_1 + (1 - AUC_2) + (1 - AUC_3)}{3}. \tag{3.1}$$

In the context of the BraTS uncertainty evaluation task (QU-BraTS), the score is estimated

Table 3.1: Change in DSC, filtered true positives (FTP) ratio, and filtered true negatives (FTN) ratio with change in uncertainty thresholds for two different example slices shown in Figure 3.1. ©[2022] CC-BY. Reprinted, with permission, from [161].

	DSC			
	DSC at 100 (baseline)	**DSC at 75**	**DSC at 50**	**DSC at 25**
Example-1	0.94	0.96	0.965	0.97
Example-2	0.92	0.955	0.97	0.975
	Ratio of Filtered TPs (1 - (TP_x / $TP_{baseline\ (\tau=100)}$))			
	FTP at 100	**FTP at 75**	**FTP at 50**	**FTP at 25**
Example-1	0.00	0.00	0.05	0.1
Example-2	0.00	0.00	0.15	0.25
	Ratio of Filtered TNs (1 - (TN_x / $TN_{baseline\ (\tau=100)}$))			
	FTN at 100	**FTN at 75**	**FTN at 50**	**FTN at 25**
Example-1	0.00	0.0015	0.0016	0.0019
Example-2	0.00	0.0015	0.0026	0.0096

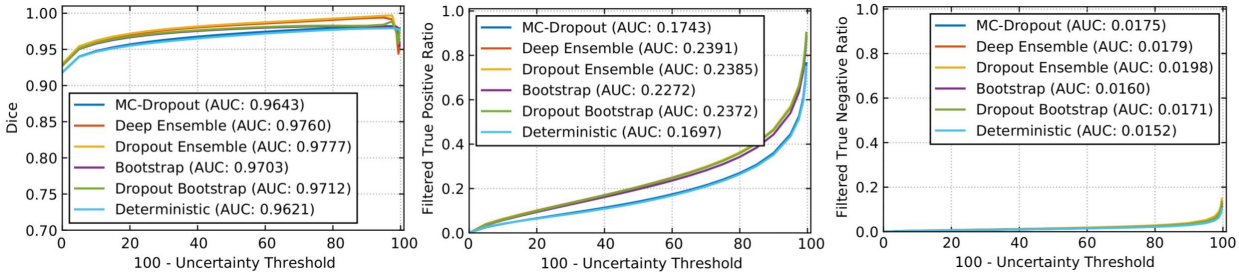


Figure 3.2: Effect of changing uncertainty threshold (τ) on WT for entropy measure. Specifically, we plot (left) DSC, (middle) filtered true positive ratio, and (right) filtered true negative ratio as a function of 100 - τ. We plot the curves for six different uncertainty generation methods, namely, MC-Dropout, Deep Ensemble, Dropout Ensemble, Bootstrap, Dropout Bootstrap, and Deterministic. All methods use entropy as a measure of uncertainty. ©[2022] CC-BY. Reprinted, with permission, from [161].

for each tumour entity separately and then used to rank the participating methods.

3.2.1 A 3D U-Net Based Experiment

Experiments were devised to show the functioning of the derived uncertainty evaluations and rankings. A modified 3D U-Net architecture [157] generates the segmentation outputs and corresponding uncertainties. The network was trained ($n = 228$), validated ($n = 57$), and tested ($n = 50$) based on the publicly available BraTS 2019 training dataset ($n = 335$) [170, 22, 21, 19, 20]. The performances of WT segmentation with the entropy uncertainty measure [80], which captures the average amount of information contained in the predictive distribution, is shown in Figure 3.2. Here uncertainties are estimated us-

ing MC-Dropout [79], Deep Ensemble [136], Dropout Ensemble [237], Bootstrap, Dropout Bootstrap, and a Deterministic softmax entropy measure. Dropout bootstrap shows the best DSC performance (highest AUC) and has the worst performance for FTP and FTN curves (highest AUC). This result shows that the higher performance in DSC is at the expense of a higher number of filtered correct voxels. Overall, the score is working in line with the objectives. However, there is no clear winner amongst these uncertainty methods in terms of rankings.

3.3 BraTS 2020 Quantification of Uncertainty (QUBraTS) challenge – Materials and Methods

3.3.1 Dataset

The BraTS 2020 challenge dataset [170, 22, 21, 19, 20] is divided into three cohorts: Training, Validation, and Testing. The Training dataset is composed of multi-parametric MRI (mpMRI) scans from 369 diffuse glioma patients. Each mpMRI set contains four different sequences: native T1-weighted (T1), post-contrast T1-weighted (T1ce), T2-weighted (T2), and T2 Fluid-Attenuated-Inversion-Recovery (FLAIR). Each MRI volume is skull-stripped (also known as brain extraction) [251], co-aligned to a standard anatomical atlas (i.e., SRI24 [205]), and resampled to $1mm^3$ voxel resolution. Expert human annotators provided ground truth (GT) tumour labels, consisting of 3 classes described previously. Note that there is no "ground-truth" uncertainty label.

The BraTS 2020 Validation cohort is composed of 125 cases of patients with diffuse gliomas. Similar to the training dataset, this also contains four different mpMRI sequences for each case. The validation dataset allows participants to obtain preliminary results in unseen data and their cross-validated results on the training data. The GT labels for the validation data are not provided to the participants.

The BraTS 2020 Testing cohort is then used for the final ranking of the participating team. It is comprised of a total of 166 cases. The exact type of glioma is not revealed to the participating teams. Each team gets a window of 48 hours to upload their results to the challenge evaluation platform (https://ipp.cbica.upenn.edu/) [53].

3.3.2 Evaluation Framework

The University of Pennsylvania image processing portal (https://ipp.cbica.upenn.edu/) is used to evaluate all BraTS participating algorithms quantitatively. This portal allows the registration of new teams to access the BraTS datasets and the framework for automatically evaluating all participating algorithms on all three (i.e., training, validation, and testing) cohorts[1]. In addition to the IPP, and in favor of reproducibility and transparency, we make the uncertainty evaluation framework publicly available through GitHub[2]. As mentioned previously, the evaluation framework expects the participants to provide multi-class brain tumour segmentation labels and their associated voxel-wise uncertainties for three tumour entities: whole tumour (WT), tumour core (TC), and enhancing tumour (ET). These uncertainties are expected to be normalized between 0-100 for ease of computation.

3.3.3 Participating Methods

In total, 14 teams participated in the QU-BraTS 2020 challenge. All teams utilized a convolutional neural network (CNN) based approach for the tumour segmentation task and the generation of associated uncertainty maps. Detailed descriptions of 12/14 proposed approaches are given below[3]. Details regarding the CNN segmentation architectures utilized by each team are not described in detail here, as this chapter focuses on uncertainty generation methods rather than the segmentation itself. Readers are requested to refer

[1] Access to the BraTS testing datasets is not possible after the conclusion of the challenge.
[2] https://github.com/RagMeh11/QU-BraTS
[3] Two teams, namely Frankenstein [61] and NSU-btr [86], withdrew after their initial participation.

to each team's individual papers (as cited below) for more details about the CNN architecture used for the segmentation task. A preliminary version of the QU-BraTS challenge was run in conjunction with the BraTS 2019 challenge. Appendix B provides details about the participating teams and their performance. We did not include the analysis results of the QU-BraTS 2019 challenge here, as the task was run as a preliminary task without employing any statistical significance analysis ranking scheme to evaluate the participating teams.

Method-1: Team SCAN [153]

The method uses the DeepSCAN [154] model. The training of the model was performed using a combination of focal loss [140] and a Kullback-Leibler divergence: for each voxel and each tumour entity, the model produces an output $p \in (0, 1)$ (corresponding to the output of a standard binary classification network) and an output $q \in (0, 0.5)$ which represents the probability that the classifier output differs from the ground truth on that tumour entity. The probability q is supervised by the label z, which is the indicator function for disagreement between the classifier (thresholded at the $p = 0.5$ level) and the ground truth. Given q, an annealed version of the ground truth is formed, $w = (1-x)\cdot q + x\cdot(1-q)$. Focal KL divergence between w and p is defined as follows:

$$\text{Focal}_{\text{KL}}(w||p) = (p - w)^2(w \cdot log(w) - w \cdot log(p)).$$

The final loss function is given by:

$$\text{Loss} = 0.1 \cdot \text{Focal}(p, x) + 0.9 \cdot \text{Focal}_{\text{KL}}(w||p) + 0.9 \cdot \text{BCE}(q, z).$$

An ensemble of the networks were utilized in the final output, where from different predictions, p and q were combined to a single probability $q \cdot I_{p\leq 0.5} + (1-q) I_{p\geq 0.5}$. The final uncertainty output (denoted q above) was normalized into the range of 0 to 100: $100 * (1-2q)$. The uncertainty in the ensemble can likewise be extracted as for any ordinary model with a sigmoid output x as: $100 \cdot (1 - 2|0.5 - x|)$

While this uncertainty measure gives a measure of uncertainty both inside and outside the provided segmentation, it was empirically observed that treating all positive predictions as certain and only assigning uncertain values to only negative predictions gives better performance on the challenge scores.

Method-2: Team Alpaca [177]

An ensemble of three different 2D segmentation networks [102, 42, 99] was used. The softmax probabilities from each of the three networks were averaged to generate the final probability maps. These probability maps were used to generate the uncertainty maps for each tumour entity. This was computed by mapping the most confident prediction value to 0 and the least confident value to 100.

Method-3: Team Uniandes [54]

A novel deep learning architecture named Cerberus was proposed. The uncertainty maps were produced by taking the complement of the final segmentation softmax probability maps, and rescaling them between 0 and 100.

Method-4: Team DSI_Med [50]

Five attention-gated U-Net models were trained. The uncertainty maps were normalized between 0 and 100 for the four nested tumour entities. For each uncertainty map, the maximum softmax probability from the five models for each voxel in each entity was taken. The voxels were either part of the given nested entity or not, judging by the segmentation maps acquired from the ensemble of five models. The probabilities of those voxels that belong to the nested entity were inverted and multiplied by 100. The results were then rounded to get into the 0-100 range.

Double thresholds were further applied to refine the uncertainty maps. Low and high probability thresholds for each nested entity were empirically defined: WT(0.1, 0.3), TC(0.2,

0.3) ET(0.3, 0.5). For each voxel that belongs to a nested entity, the uncertainty was set to 0 when the probability was higher than the corresponding high threshold. For each voxel that belongs to the background, the uncertainty was set to 0 when the maximum probability was lower than the low threshold. Such a method enabled the adjustment of the uncertainty of nested entities and the background independently.

Method-5: Team Radiomics_MIU [25]

The method used an ensemble of three different CNNs [265, 26, 58] for segmentation. Different models were trained for three different tumour entities (i.e., WT, TC, and ET segmentation). Three model ensembles were used, i.e., a total of nine models were trained for the task. Averaging various probabilities is one of the best and most effective ways to get a prediction of the ensemble model in classification. The uncertainty was estimated using the concept of entropy to represent voxel-wise variance and diversity information. The resulting uncertainty values were scaled to lie between 0 and 100.

Method-6: Team Med_vision [194]

The method proposed self-ensemble-resUNet. The output softmax probabilities (y_{pred}) were inverted and normalized between 0-100 to obtain the uncertainty maps ($U_{\textbf{pred}}$): $U_{\text{pred}} = 100 \cdot (1 - y_{\text{pred}})$

Method-7: Team Jaguars [206]

The method used an ensemble of a total of 7 U-Net type models. The output probabilities of each model were averaged for each label in each voxel to obtain a new probability for the ensemble. Since the model makes a binary classification of each voxel, the highest uncertainty corresponds with a probability of 0.5. Then the normalized entropy was used to get an uncertainty measure of the prediction for each voxel:

$$H = \sum_{c \in C} \frac{p_c \cdot \log(p_c)}{log(|C|)} \in [0, 1],$$

where p_c is the sigmoid output average probability of class c and C is the set of classes, (C = {0,1} in this case). These values were multiplied by 100 to normalize it between 0 and 100.

Method-8: Team UmU [263]

The method proposes a Multi-Decoder Cascaded Network to predict the probability of the three tumour entities. An uncertainty score, $u^r_{i,j,k}$, at voxel (i, j, k) was defined by:

$$u^r_{i,j,k} = \begin{cases} 200 \cdot (1 - p^r_{i,j,k}), & \text{if } p^r_{i,j,k} \geq 0.5 \\ 200 \cdot p^r_{i,j,k}, & \text{if } p^r_{i,j,k} < 0.5 \end{cases}$$

where $u^r_{i,j,k} \in [0, 100]^{|R|}$ and $p^r_{i,j,k} \in [0, 1]^{|R|}$ are the uncertainty score map and probability map, respectively. Here, $r \in R$, where R is the set of tumour entities, i.e. WT, TC, and ET.

Method-9: Team LMB [23]

The method used a V-net [171] architecture. A combination of test-time dropout and test-time augmentation was used for uncertainty estimation. In particular, the same input was passed through the network 20 times with random dropout and random data augmentation. The uncertainty map was estimated with the variance for each sub-region independently. Let $Y^i = y^i_1, y^i_2, ..., y^i_B$ be the vector that represents predicted labels for the i^{th} voxel. The voxel-wise uncertainty map, for each tumour entity (WT,TC,ET), was obtained as the variance:

$$\text{var} = \frac{1}{B} \sum_{b=1}^{B} (y^i_b - y^i_{\text{mean}})^2,$$

where y^i_{mean} represents the mean prediction across b samples.

Method-10: Team Matukituki [152]

A multisequence 2D Dense-UNet segmentation model was trained. The final layer of this model is a four-channel soft-max layer representing the labels 'no tumour', 'edema', 'necrosis', and 'ET'. Uncertainty values were obtained from the final layer of the segmen-

tation model for each label as follows: For WT, initial uncertainty values were obtained by adding the voxel-wise soft-max values of 'edema + necrosis + ET'. The initial uncertainty values for TC were the voxel-wise sum of 'necrosis + ET'. The initial uncertainty of the ET was the values of the voxel-wise soft-max channel representing ET. For all labels, the initial uncertainty values were clipped between 0 and 1. They were then modified according to the function: uncertainty = (1 – initial uncertainty) x 100. Finally, uncertainty values of 99 were changed to 100.

Method-11: Team QTIM [192]

The method used an ensemble of five networks to estimate voxel-wise segmentation uncertainty. Mirror axis-flipped inputs were passed through all models in the ensemble, resulting in 40 predictions per entity. These predictions were combined by directly averaging the model logits, denoted as l_x. A voxel with high predictive uncertainty will have $|l_x| \approx 0$, whereas a voxel with high predictive certainty will have $|l_x| \gg 5$. To explicitly quantify uncertainty (U) in the range 0 (maximally certain) to 100 (maximally uncertain), the following formula is used:

$$U_x = \begin{cases} 200 \cdot \sigma(l_x) & \text{if } 0 \leq \sigma(l_x) < 0.5 \\ 200 \cdot (1 - \sigma(l_x)) & \text{otherwise} \end{cases}$$

where the σ function converts the ensembled logits to probabilities.

Method-12: Team Nico@LRDE

A cascade of two 3D U-Net-type networks was employed for the task of brain tumour segmentation and its associated uncertainty estimation. The first network was trained for the brain tumour segmentation task. The second network was trained to predict where the segmentation network made wrong predictions. Here, the ground truth for training this network was generated as follows: the ground truth was considered ones (present) at voxels where the segmentation network was wrong, and it was considered as zeros (absent) at voxels where the segmentation network was correct. This way, the uncertainty

networks learn to return zeros where the segmentation network is generally accurate and values next to one where the segmentation networks will have issues correctly predicting the segmentation ground truth. The output of the uncertainty estimation network (second network) was normalized between 0-100.

3.4 Analysis

This section presents the complete analyses and evaluation of teams that participated in the QU-BraTS 2020 challenge. Section 3.4.1 provides the description of the evaluation and ranking strategy followed during the QU-BraTS 2020 challenge. Section 3.4.2 provides the overall ranking results (accounting for all tumour entities) according to which the winning teams were announced at the challenge (Figure 3.3). We also compare their ranking on the segmentation task in the same section. Then, Section 3.4.2 provides the ranked order of the participating teams according to the individual tumour entities (Figure 3.4-3.6), followed by our ablation study (in Section 3.4.2) on the scores incorporated in the general score (Equation 3.1) (Figure 3.7-3.9). Table 3.2 encapsulates a summary of the ranked order of the participating teams for this analysis. Finally, Section 3.4.3 provides qualitative results highlighting the effect of uncertainty thresholding filtering for all participating teams.

3.4.1 Ranking Scheme: BraTS 2020 challenge on Uncertainty Quantification (QU-BraTS)

The ranking scheme used during the challenge comprised the ranking of each team relative to its competitors for each testing subject, for each evaluated tumour entity (i.e., ET, TC, WT) using the overall score (Equation 3.1). This ranking scheme led to each team being ranked for 166 subjects for three regions, resulting in 498 individual rankings. For each team, first, the individual ranking for each patient was calculated by adding ranks across each region. This ranking is referred to as the cumulative ranking score (CRS). For each team, the normalized ranking score (NRS) was also calculated for each patient by

Table 3.2: Summary of team ranking for different analyses performed in this chapter. We use the ranking scheme described in Section:3.4.1 to rank different teams. The "QU-BraTS Ranking" column depicts the actual team ranking for all participating teams in QU-BraTS 2020 challenge (Section 3.4.2). In the "Segmentation Ranking" column, we also report segmentation ranking for all teams that participated in the QU-BraTS challenge. The segmentation ranking is across 78 teams that participated in the segmentation task during BraTS 2020. In three columns under "Ranking based on Individual tumour Entities" (Section 3.4.2), we provide a team ranking based only on one of the three tumour entities. Similarly, we also report the team ranking based on the ablation study of our developed score in the last three columns of "Ranking Based on Ablation Study" (Section 3.4.2). For each type of ranking, the total number of provided ranks (given in the bracket) varies, as we provide the same rank for teams that do not have a significant statistical difference between their performance (Section 3.4.1). ©[2022] CC-BY. Reprinted, with permission, from [161].

Teams	Challenge Ranking		Variations					
			Ranking Based on Individual tumour Entities			Ranking Based on Ablation Study		
	QU-BraTS Ranking (9)	Segmentation Ranking (18)	Whole tumour (13)	tumour Core (11)	Enhancing tumour (11)	DSC AUC (10)	DSC AUC and FTP AUC (9)	DSC AUC and FTN AUC (12)
SCAN	1	4	1	1	1	6	2	4
UmU	2	7	3	2	2	4	3	3
DSI_Med	2	13	2	2	3	9	3	7
QTIM	3	7	4	2	3	3	4	2
Uniandes	4	15	5	3	4	8	5	6
nsu_btr	5	13	10	8	10	1	4	9
LMB	5	20	8	4	3	10	7	8
radiomics_miu	6	13	7	5	5	2	8	3
Nico@LRDE	6	18	6	6	6	7	9	5
Jaguars	6	13	5	6	6	2	8	3
Team_Alpaca	7	10	9	7	7	2	1	1
Matukituki	8	19	11	9	9	7	4	12
Frankenstein	9	18	13	11	8	6	6	11
med_vision	9	14	12	10	11	5	7	10

dividing their CRS by the total number of participating teams and the total number of regions. The NRS is in the range of 0-1 for each patient. The final ranking score (FRS) was calculated by averaging the cumulative rank across all patients for each participating team. Other challenges, such as the ischemic stroke lesion segmentation challenge (ISLES - http://www.isles-challenge.org/) [149], use a similar ranking scheme.

Following the BraTS challenge, further permutation testing was done to determine the statistical significance of the relative rankings between each pair of teams. This permutation testing would reflect differences in performance that exceeded those that might be expected by chance. Specifically, for each team, given a list of observed patient-level cumulative ranks, i.e., the actual ranking described above, for each pair of teams, repeated random permutations (i.e., 100,000 times) of the cumulative ranks for each subject were performed. The difference in the FRS between this pair of teams was calculated for each

permutation. The proportion of times the difference in FRS calculated using randomly permuted data exceeded the observed difference in FRS (i.e., using the actual data) indicated the statistical significance of their relative rankings as a p-value. Teams that do not have a statistically significant difference in their FRS have similar respective ranks (group) on the leaderboard[4].

3.4.2 Team Ranking

This section reports the final rankings of all participating teams on BraTS 2020 test dataset.

Overall Ranking Results

Figure 3.3 (and QU-BraTS ranking column in Table 3.2) provides a relative ranking for each team[5]. We can see that *Team SCAN* comfortably outperforms all other methods and achieves the first rank in the challenge. Their normalized ranking score (NRS) across all patients was ~ 0.14, while the NRS (across all patients) for the teams which achieved rank 2 (*Team UmU* and *Team DSI_Med*) was ~ 0.28. There was no statistically significant difference between *Team UmU* and *Team DSI_Med*. Thus both teams were ranked at position 2 on the challenge leaderboard. *Team QTIM* ranked 3rd in the challenge leaderboard and achieved marginally (though statistically significant) lower performance compared to Rank-2 teams (average NRS of ~ 0.31 compared to average NRS of ~ 0.28).

We also report the relative segmentation ranking of each team participating in the uncertainty challenge. The reported segmentation task ranking is across 78 teams that participated in the segmentation task. From Figure Figure 3.3 (and Segmentation Ranking column in Table 3.2), we can observe that while the *Team SCAN* (pink colour) achieves a higher ranking (Rank-4) than other teams in the segmentation task, the segmentation

[4]Throughout the chapter, we report any p-value less than 0.05 as the threshold for statistically significant differences.

[5]Box plot depicting performance of each team for four different scores - DICE_AUC, FTP_RATIO_AUC, FTN_RATIO_AUC, SCORE, for three different tumour entities - WT, TC, ET, is given in Appendix B.

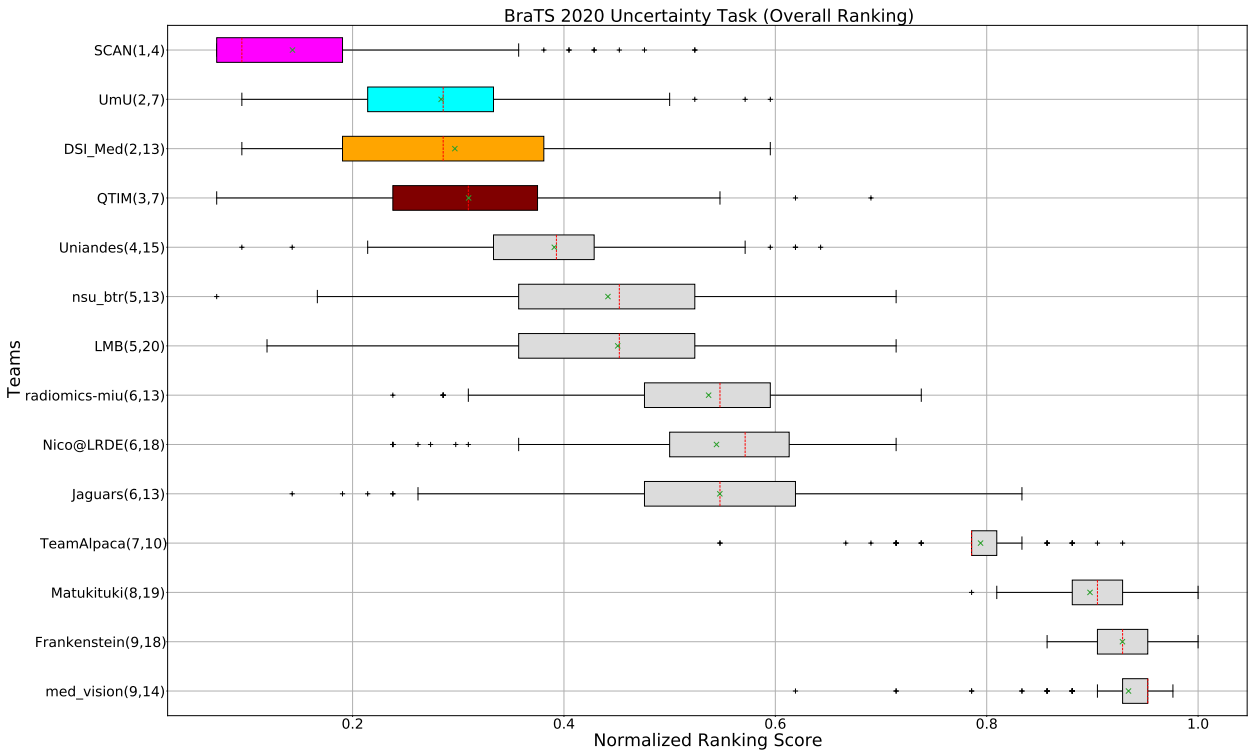


Figure 3.3: QU-BraTS 2020 boxplots of the normalized ranking score (NRS) across patients for all participants on the BraTS 2020 test set (lower is better). Boxplots for the top four performing teams are visualized using Pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for the remaining teams use gray colour. Y-axis shows the name of each team and their respective uncertainty task ranking, followed by their segmentation task ranking. There was no statistically significant difference between the per-patient ranking of teams ranked at the same position. Teams that have different ranks had statistically significant differences in their per-patient ranking. ©[2022] CC-BY. Reprinted, with permission, from [161].

task ranking and the uncertainty task (QU-BraTS challenge) ranking are not the same. This is visible for *Team UmU* and *Team QTIM*, as both achieved a similar ranking (rank-7) in the segmentation task of BraTS 2020; while *Team UmU* was ranked second in the uncertainty task, *Team QTIM* was ranked third. Similarly, we can observe that three teams that achieved Rank-13 in the segmentation task (*Team DSI_Med*, *Team nsu_btr*, and *radiomics-miu*) were ranked differently in the uncertainty evaluation task (Rank-2, Rank-5, and Rank-6, respectively). The difference in ranking across both tasks shows that performing well on the segmentation task does not guarantee good performance on the uncertainty evaluation task, and both tasks are complementary.

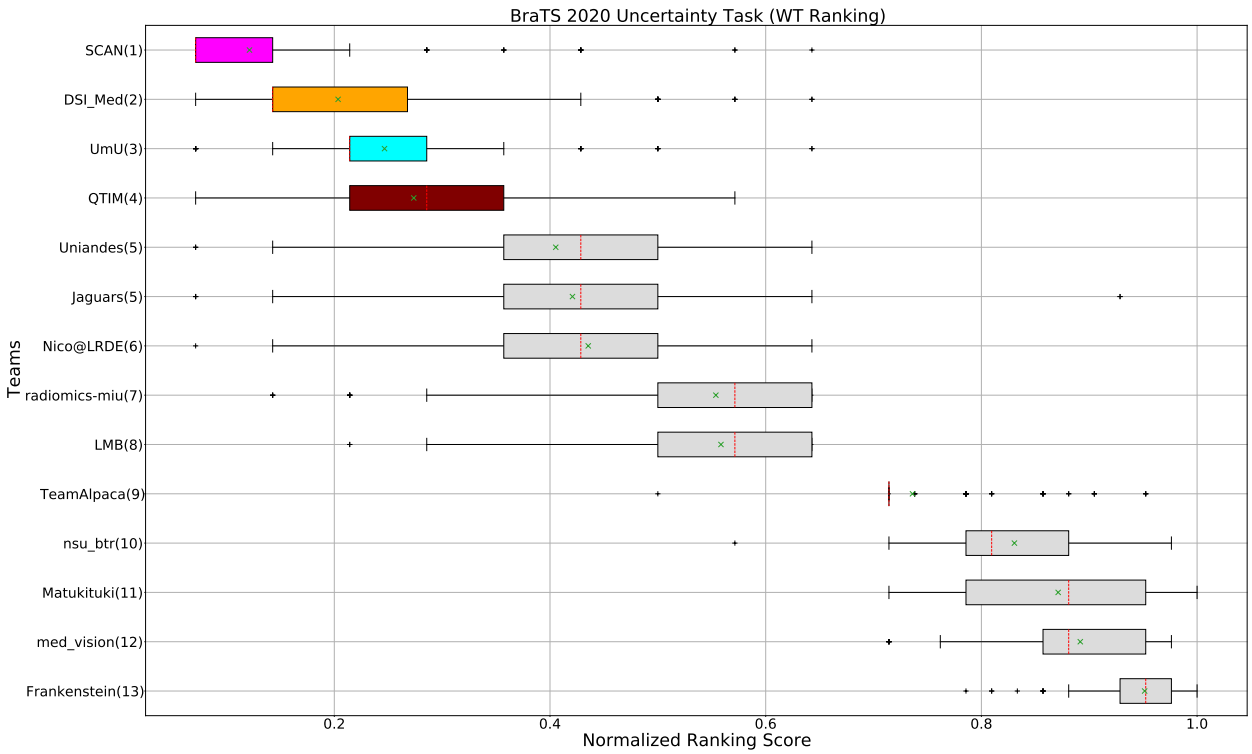


Figure 3.4: QU-BraTS 2020 boxplots of normalized ranking score (NRS) across patients for all participants on the BraTS 2020 test set only for Whole tumour (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3.3) are visualized using Pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for the remaining teams use gray colour. Y-axis shows the name of each team and their respective uncertainty task ranking, followed by their segmentation task ranking. There was no statistically significant difference between the per-patient ranking of teams ranked at the same position. Teams that have different ranks had statistically significant differences in their per-patient ranking. [2022] CC-BY. Reprinted, with permission, from [161].

Team Ranking for Individual Tumour Entities

The BraTS challenge involves three separate tumour entities (WT, TC, and ET). The segmentation performance across these entities varies, as reported in the previous BraTS challenge reports [170, 22]. Specifically, the BraTS challenge reports good DSC across different teams for the WT segmentation task, while the performance for the ET segmentation task is relatively lower. The performance gap between different tumour entities can hinder the clinical adaptation of the segmentation algorithms. The main goal for developing the uncertainty evaluation scores is to make algorithms more useful for clinical adaptation. Keeping this in mind, we further report the raking of each participating team according to the score (Equation 3.1) calculated for each tumour entity in Figure 3.4, Figure 3.5, and Figure 3.6.

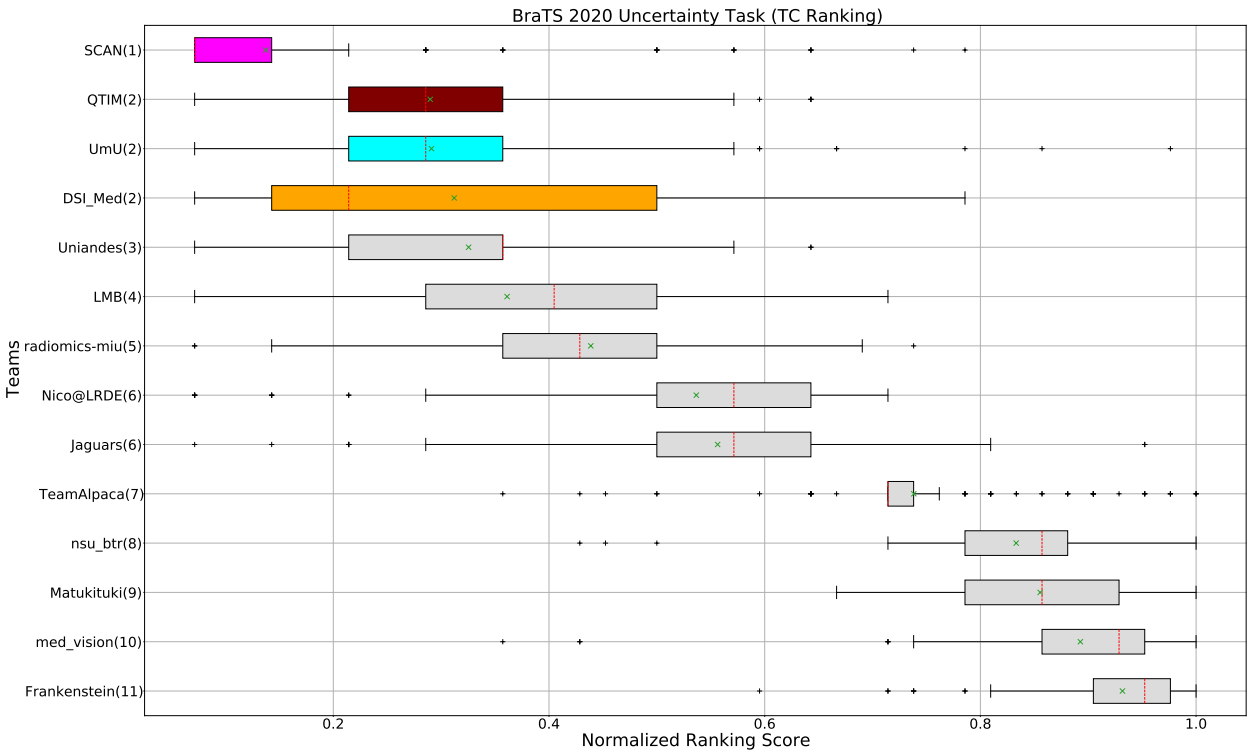


Figure 3.5: QU-BraTS 2020 boxplots of the normalized ranking score (NRS) across patients for all participants on the BraTS 2020 test set only for tumour Core (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3.3) are visualized using Pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for the remaining teams use gray colour. Y-axis shows the name of each team and their respective uncertainty task ranking, followed by their segmentation task ranking. There was no statistically significant difference between the per-patient ranking of teams ranked at the same position. Teams that have different ranks had statistically significant differences in their per-patient ranking. ©[2022] CC-BY. Reprinted, with permission, from [161].

When teams are ranked only based on their WT scores (Figure 3.4 and whole tumour column in Table 3.2), *Team SCAN* still comfortably outperforms other teams similar to the original ranking (Figure 3.3). Unlike the original ranking scheme, *Team DSI_Med* ranks statistically significantly higher compared to *Team UmU*. Similarly, from Figure 3.5 (and the tumour core column in Table 3.2), we can observe that *Team QTIM*, *Team UmU*, and *Team DSI_Med* perform similarly without any statistically significant difference when ranked only based on their TC score as all teams are ranked at the same position. In Figure 3.6 (and the enhancing tumour column in Table 3.2), *Team UmU* achieves rank-2 with statistical significance compared to *Team QTIM* and *Team DSI_Med*. We also observe no statistically significant difference between *Team QTIM*, *Team DSI_Med*, and *Team LMB*.

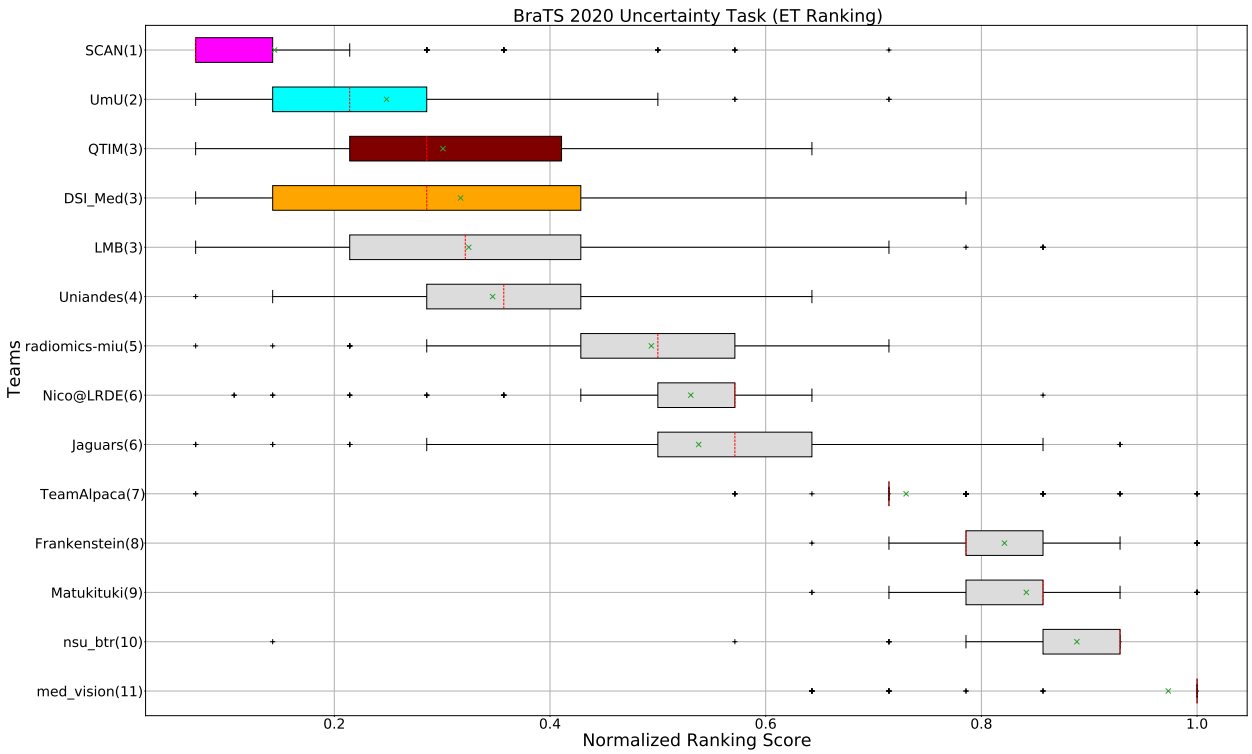


Figure 3.6: QU-BraTS 2020 boxplots of the normalized ranking score (NRS) across patients for all participants on the BraTS 2020 test set only for Enhancing tumour (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3.3) are visualized using Pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for the remaining teams use gray colour. Y-axis shows the name of each team and their respective uncertainty task ranking, followed by their segmentation task ranking. There was no statistically significant difference between the per-patient ranking of teams ranked at the same position. Teams that have different ranks had statistically significant differences in their per-patient ranking. ©[2022] CC-BY. Reprinted, with permission, from [161].

Overall, *Team SCAN* comfortably ranks first for all tumour entities. *Team UmU* ranks 3-2-2 for WT-TC-ET, while *Team DSI_Med* ranks 2-2-3 for WT-TC-ET. Both teams are ranked at position 2 when considering all tumour entities. The analysis shows that different teams achieve different ranks depending on the tumour entities, which shows that their performance differs across different tumour entities.

Ablation Study

The overall score for uncertainty evaluation is calculated as a combination of three different AUCs as described in Equation 3.1. Section 3.2 described the rationale behind the development of this score. As discussed in Section 3.2, evaluating the task-dependent

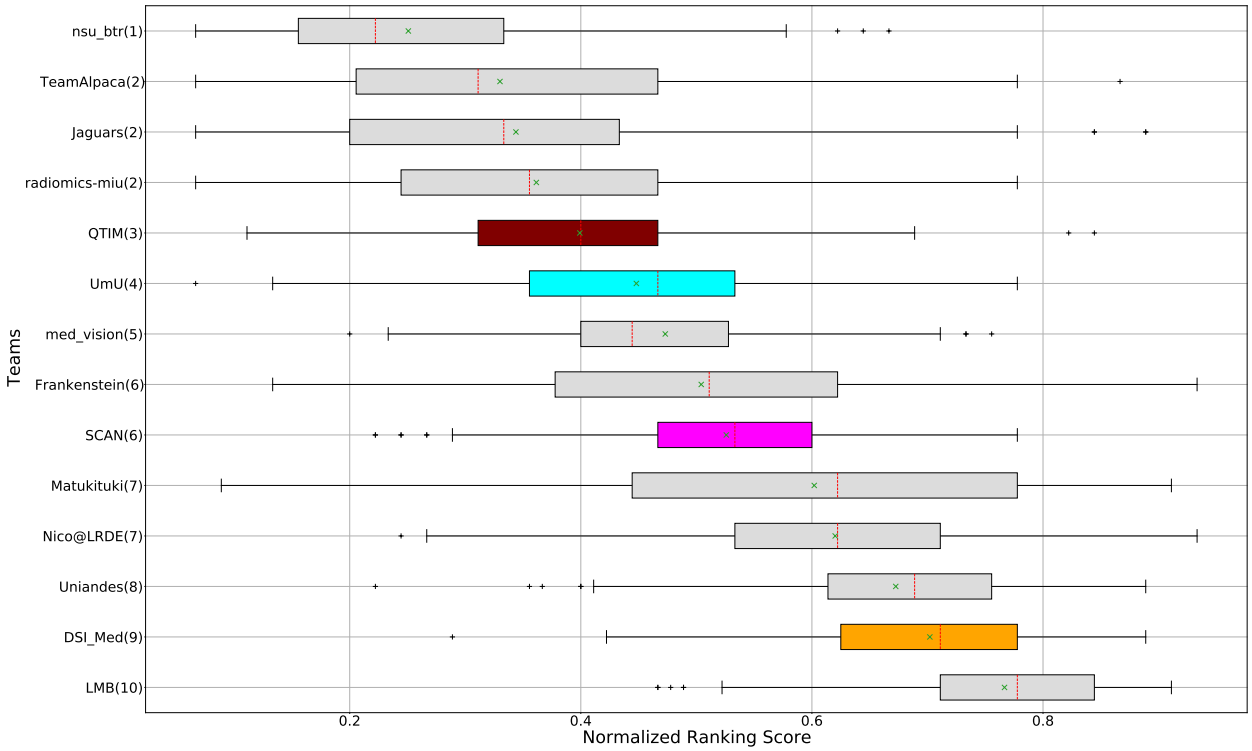


Figure 3.7: QU-BraTS 2020 boxplots of the normalized ranking score (NRS) across patients for all participants on the BraTS 2020 test set based only on DICE_AUC score (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3.3) are visualized using Pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for the remaining teams use gray colour. Y-axis shows the name of each team and their respective uncertainty task ranking, followed by their segmentation task ranking. There was no statistically significant difference between the per-patient ranking of teams ranked at the same position. Teams that have different ranks had statistically significant differences in their per-patient ranking. ©[2022] CC-BY. Reprinted, with permission, from [161].

metric (in our case, DSC) as a function of filtered samples is critical, especially in the case of pathology segmentation, where there is a high class imbalance. We expect that, by filtering voxels with a decrease in the uncertainty threshold, the performance on the remaining voxels measured using the task-dependent metric (DSC) should increase but not at the expense of filtering true positive or true negative voxels. The final score consists of the task-dependent metric and filtered true positives/negatives as a function of uncertainty thresholds. In this section, we perform an ablation study of different components of the final score (DSC, FTP, FTN). Our analysis reaffirms that only considering one or two components of the final score leads to a different ranking among participating teams.

Ranking according to DSC AUC: The main component of any uncertainty evaluation

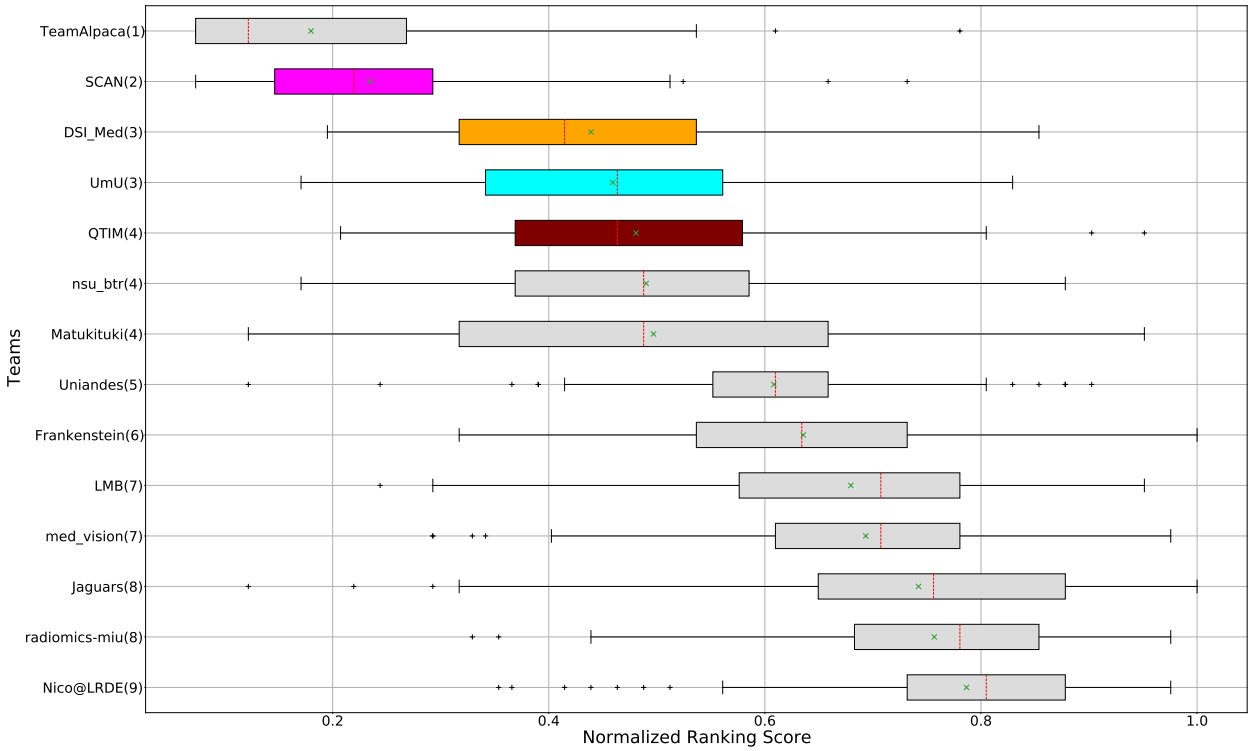


Figure 3.8: QU-BraTS 2020 boxplots of the normalized ranking score (NRS) across patients for all participants on the BraTS 2020 test set test set based on a combination of DICE_AUC score and FTP_AUC score (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3.3) are visualized using Pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for the remaining teams use gray colour. Y-axis shows the name of each team and their respective uncertainty task ranking, followed by their segmentation task ranking. There was no statistically significant difference between the per-patient ranking of teams ranked at the same position. Teams that have different ranks had statistically significant differences in their per-patient ranking. ©[2022] CC-BY. Reprinted, with permission, from [161].

score is the task-dependent metric, in our case, DSC. Many previously proposed methods for various tasks only report the value of task-dependent metrics at various uncertainty filtering thresholds – For example, the AUC score for multiple sclerosis [180]. In Figure 3.7 (and the DSC AUC column in Table 3.2), we rank participating teams according to their performance based on the AUC of DSC vs. Uncertainty threshold. The figure shows that higher ranking teams in this ranking scheme (*Team nsu_btr*, *Team Alpaca*, and *Team Jaguars*) are different from those (*Team SCAN*, *Team UmU*, and *Team DSI_Med*) in the original ranking scheme (Figure 3.3). A closer look at the higher ranking teams according to AUC of DSC (Figure 3.7) reveals that teams like *Team Alpaca* (Section 3.3.3) achieve a good score by using $100 - (100 \cdot \text{softmax\_confidence})$ as a proxy for uncertainty. Using softmax confidence in the foreground class (e.g. tumour subclass) as a direct proxy to un-

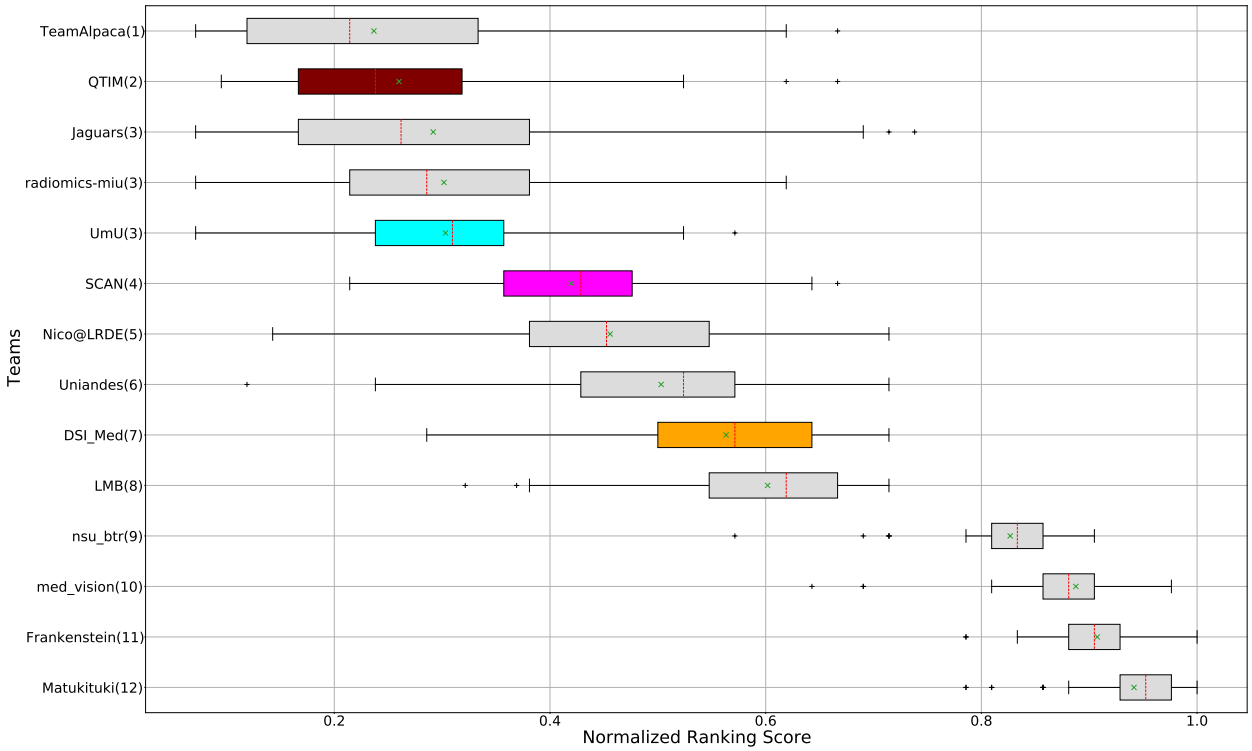


Figure 3.9: QU-BraTS 2020 boxplots of the normalized ranking score (NRS) across patients for all participants on the BraTS 2020 test set based on a combination of DICE_AUC score and FTN_AUC score (lower is better). Boxplots for the top four performing teams (in the final ranking - Figure 3.3) are visualized using Pink (*Team SCAN*), orange (*Team DSI_Med*), Cyan (*Team UmU*), and Maroon (*Team QTIM*) colour. Box plots for the remaining teams use gray colour. Y-axis shows the name of each team and their respective uncertainty task ranking, followed by their segmentation task ranking. There was no statistically significant difference between the per-patient ranking of teams ranked at the same position. Teams that have different ranks had statistically significant differences in their per-patient ranking. ©[2022] CC-BY. Reprinted, with permission, from [161].

certainty leads to all voxels belonging to the background class (i.e. healthy tissues) being marked as uncertain at a low uncertainty threshold. This would increase the burden in a system where we are asking clinicians to review all uncertain voxels (Figure 3.14). We observed that Team Alpaca used softmax confidence in the foreground class as a direct proxy for uncertainty.

Ranking according to a combination of DSC AUC and FTP or FTN AUC: In the last section, we ranked teams according to their performance on the task-dependent evaluation metrics (DSC) at different uncertainty thresholds. As mentioned in Section 3.2, ranking teams only based on their task-dependent evaluation metric rewards methods

which filter out many positive predictions at low uncertainty thresholds to attain higher performance on the metric of interest. This would increase the burden in scenarios where clinical review is needed for all uncertain predictions. To alleviate the issue, teams are ranked according to a combination of (i) AUC score for DSC and (ii) AUC for FTP or AUC for FTN. From Figure 3.8 (and DSC AUC and FTP AUC column in Table 3.2), we can conclude that a combination of both DICE_AUC and FTP_AUC alone is insufficient. It still leads to *Team Alpaca* ranked higher. As shown in Figure 3.14, *Team Alpaca* marks all healthy-tissues (True Negative) voxels as uncertain, which reflects that the segmentation method is not confident in its prediction of healthy tissue. This is problematic as it would increase the burden in scenarios where we expect clinicians to review all uncertain predictions. We see a similar problem when teams are ranked only using a combination of DICE_AUC and FTN_AUC (Figure 3.9 and DSC AUC and FTN AUC column in Table 3.2).

Analysis in the previous two sections highlights the necessity of combining all three AUCs to calculate the final score for ranking teams in the context of uncertainty quantification of the brain tumour segmentation task.

3.4.3 Qualitative Analysis

Figure 3.10 - Figure 3.14 plots the effect of uncertainty threshold based filtering on example slices from a few BraTS 2020 test cases for all participating teams. Green voxels represent True Positive predictions, while blue and red voxels represent False Positive and False Negative predictions. We filter out voxels at different thresholds (100, 75, 50, and 25). Filtered voxels are marked as yellow. According to the developed uncertainty evaluation score (Section 3.2), we want methods that filter out (marked as yellow) false positive and false negative voxels while retaining true positive and true negative voxels as we decrease the uncertainty threshold.

In Figure 3.10, we visualize the effect of uncertainty based thresholding for WT segmentation on a single slice of a BraTS 2020 test case. A closer look at some of the better per-

forming teams like *Team SCAN*, *Team UmU*, and *Team DSI_Med* reveals that these teams filter out more False Positives and False Negatives at a higher threshold than other teams like *Team QTIM* and *Team Uniandes*. We can also observe that lower-performing teams like *Team Alpaca*, *Team Matukituki*, *Team Frankenstein*, and *Team med_vision* mark all background voxels as uncertain at a low threshold. As mentioned before, marking background voxels as uncertain is problematic as it shows that the method is not confident in its healthy-tissue segmentation and requires clinicians to review the segmentation.

In Figure 3.11, we plot the effect of uncertainty based thresholding for WT segmentation on another slice of the same BraTS 2020 test case. Here we observe a similar trend where higher ranked teams can filter out False Positives and False Negatives at a higher threshold than other teams. *Team SCAN* only filters negative predictions. This results in them never filtering out their False Positive predictions of the whole tumour inside the ventricles. It is problematic in a real-world scenario as we do not want a method that is over-confident about its positive pathology segmentation predictions.

Figure 3.12 shows an example slice of a different BraTS 2020 patient and visualizes the effect of uncertainty thresholding for core tumour segmentation. The figure highlights that team ranking is different across different cases as we can see that *Team SCAN* and *Team UmU* has similar prediction at *Threshold:100*. However, *Team SCAN* starts filtering out more true negatives sooner compared to *Team UmU*, which would result in *Team SCAN* ranked lower compared to *Team UmU* for this particular BraTS test case. We can observe a similar trend when comparing *Team DSI_Med* and *Team LMB*, where *Team LMB* starts filtering out more false positives sooner than *Team DSI_Med*. Similarly, in Figure 3.13, we can observe that in scenarios where all teams are making errors by predicting a high amount of false positives, the overall uncertainty score would be more reliant on which teams can filter out these false positives sooner. For example, *Team UmU* performs better compared to *Team DSI_Med*.

Figure 3.14 depicts an example slice of uncertainty threshold based filtering for ET segmentation. Here we can see that when all teams make almost the same predictions with a high amount of true positives compared to false positives/false negatives, the overall uncertainty score is similar across teams. Except for teams that mark all background (healthy-tissue) voxels as uncertain, they perform poorly on the final score.

3.5 Summary

This chapter introduced a new score for evaluating uncertainties in the task of brain tumour segmentation during the BraTS 2020 challenge. The proposed score was used to rank different participating teams from the Uncertainty Quantification task of the BraTS 2020 challenge (QU-BraTS 2020).

The proposed evaluation score was developed with the clinical objective of enabling the clinician to review only the uncertain areas of an automatic segmentation algorithm instead of the complete segmentation. Toward this end, this score would reward algorithms that are confident when correct and uncertain when incorrect. The objective was evaluated by filtering (marking as uncertain) voxels with uncertainty higher than a specified threshold as uncertain. The task-dependent DSC is measured only on the remaining unfiltered voxels. To ensure that method does not filter out a high number of correctly predicted voxels in order to achieve a better DSC, the developed evaluation score also keeps track of the number of filtered True Positive and True Negative voxels. Keeping track of these filtered TP and TN voxels ensures that the burden on the reviewing clinicians is not increased substantially. In short, the proposed score calculates the task-dependent metric score (i.e. DSC for segmentation), the percentage of filtered true positives and true negatives at different uncertainty thresholds. It combines them to generate a single evaluation score for a single subject.

The analysis (Section 3.4.2) of algorithms developed by the participating teams from the

QU-BraTS 2020 task highlighted that the relative ranking of the participating teams for both the segmentation and uncertainty quantification tasks are different. The different ranking orders show that performing better on the segmentation task does not guarantee good performance on the uncertainty quantification task. An automatic segmentation method that provides both the segmentation and its uncertainties is more clinically relevant. Both the segmentation and the associated uncertainties provide complementary information. For example, automatic segmentation can provide accurate results with minimal clinician input. In contrast, the associated uncertainty would allow clinicians to see where to trust and review the segmentation before deploying it in clinical practice.

Results in Section 3.4.2 indicate that it is necessary to rank teams individually for each tumour entity as they rank differently across these entities. An ablation study on the proposed score (Section 3.4.2) showed the necessity of utilizing all three components (DSC, percentage of Filtered True Positive, and percentage of Filtered True Negative) for the proposed uncertainty evaluation score.

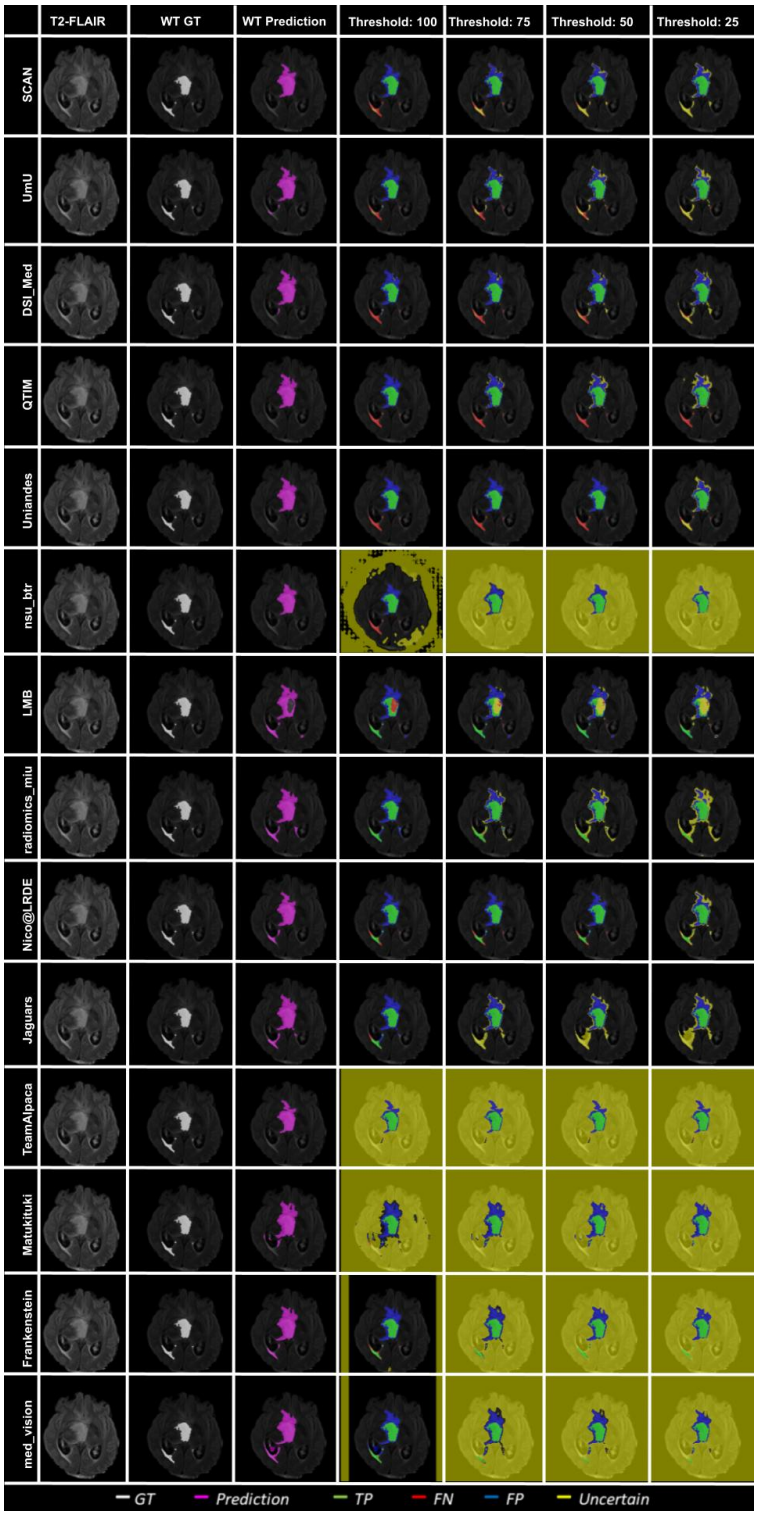


Figure 3.10: Effect of uncertainty thresholding on a BraTS 2020 test case for whole tumour segmentation across different participating teams. (a) T2-FLAIR MRI (b) Ground Truth (c) Prediction (d) No filtering. Uncertainty Threshold = 100 (e) Uncertainty Threshold = 75 (f) Uncertainty Threshold = 50 (g) Uncertainty Threshold = 25. ©[2022] CC-BY. Reprinted, with permission, from [161].

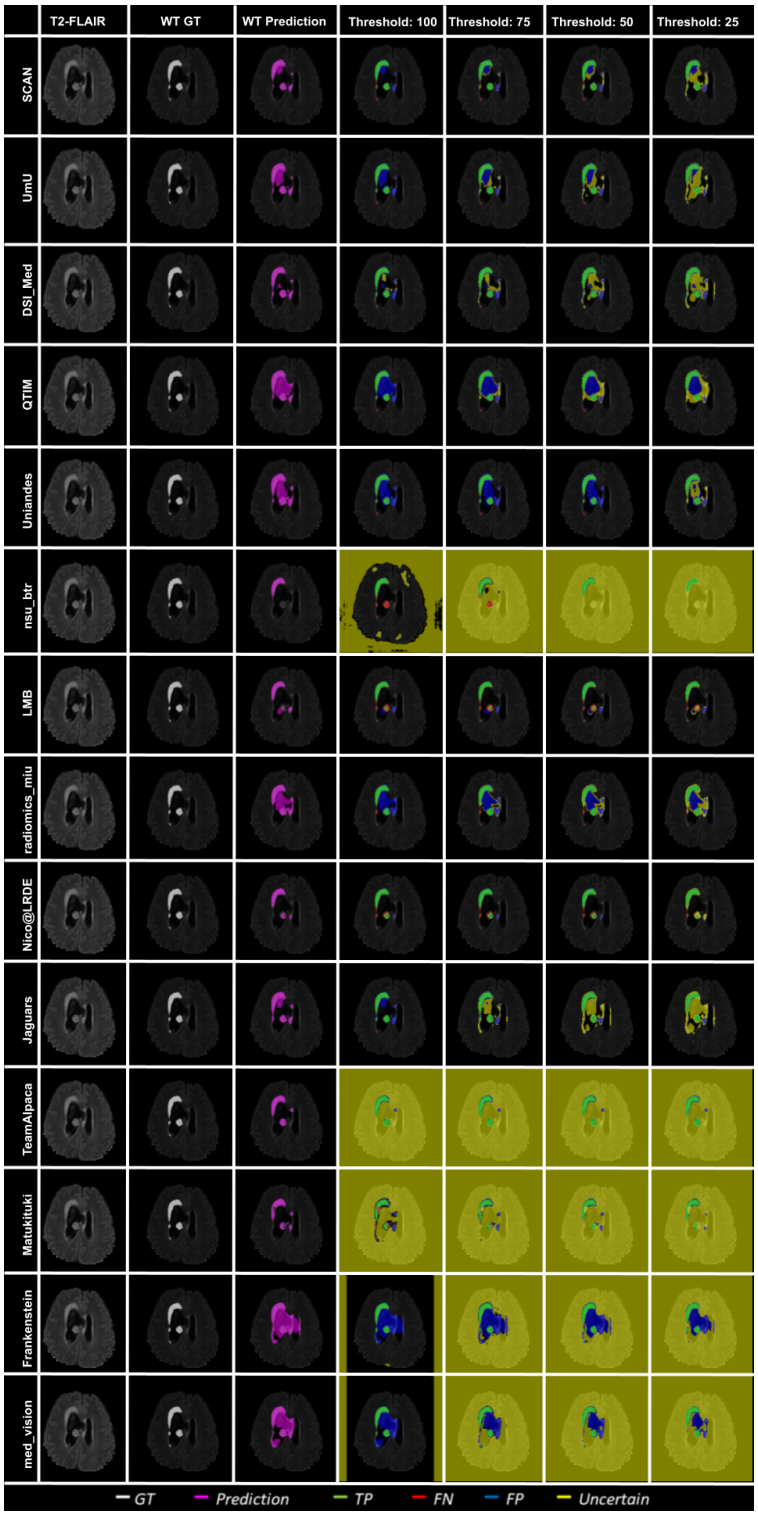


Figure 3.11: Effect of uncertainty thresholding on a BraTS 2020 test case for whole tumour segmentation across different participating teams. (a) T2-FLAIR MRI (b) Ground Truth (c) Prediction (d) No filtering. Uncertainty Threshold = 100 (e) Uncertainty Threshold = 75 (f) Uncertainty Threshold = 50 (g) Uncertainty Threshold = 25. ©[2022] CC-BY. Reprinted, with permission, from [161].

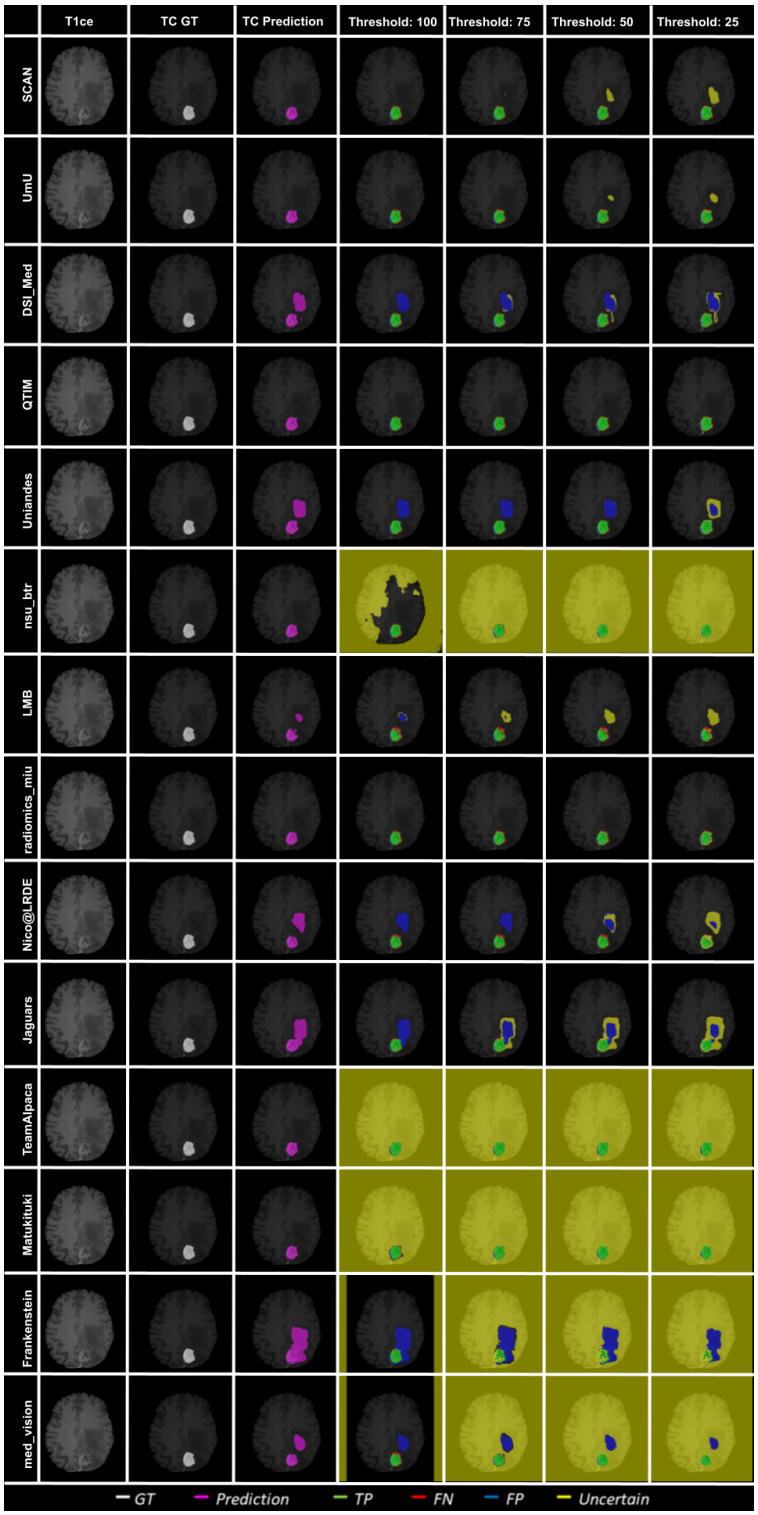


Figure 3.12: Effect of uncertainty thresholding on a BraTS 2020 test case for core tumour segmentation across different participating teams. (a) T1ce MRI (b) Ground Truth (c) Prediction (d) No filtering. Uncertainty Threshold = 100 (e) Uncertainty Threshold = 75 (f) Uncertainty Threshold = 50 (g) Uncertainty Threshold = 25. ©[2022] CC-BY. Reprinted, with permission, from [161].

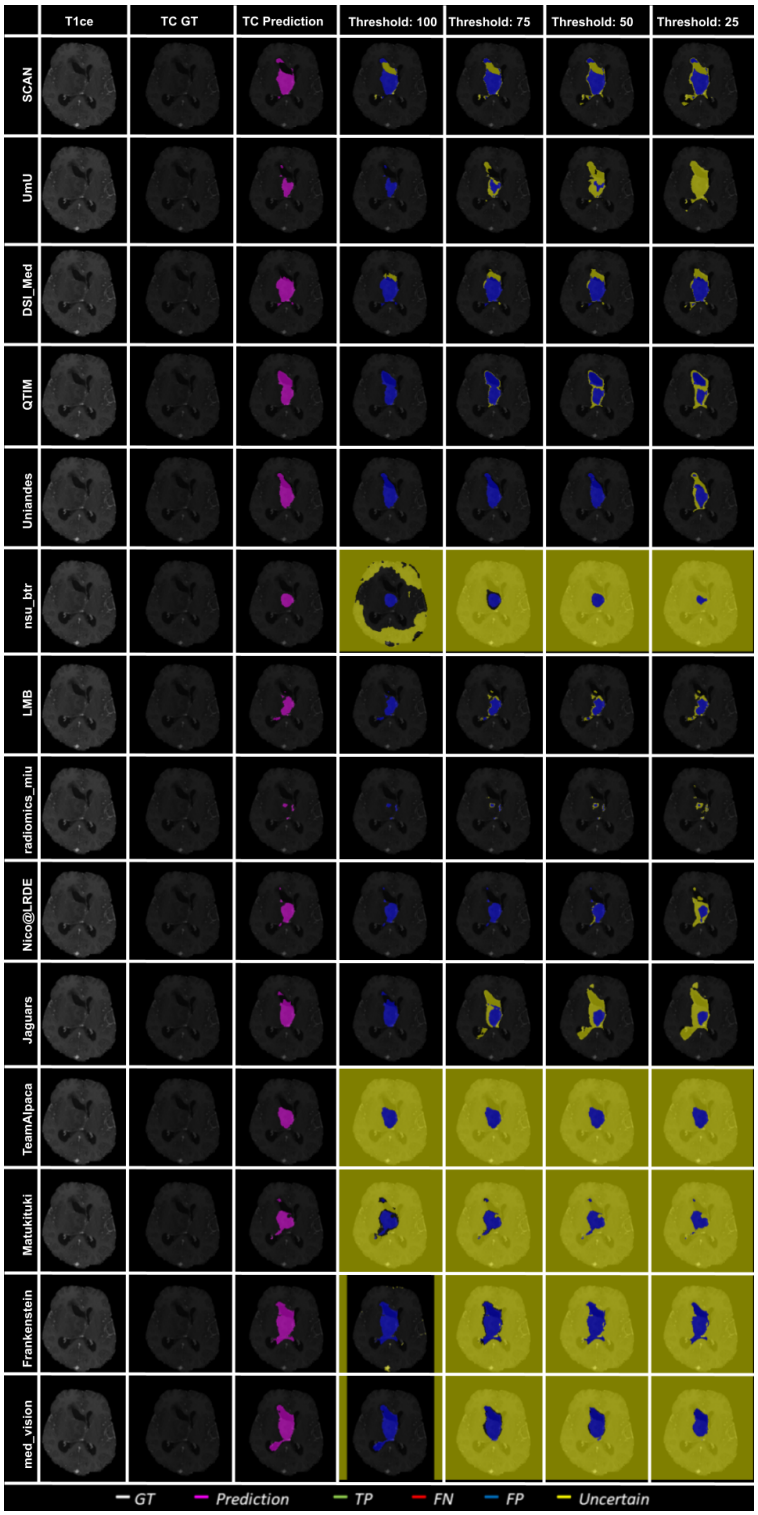


Figure 3.13: Effect of uncertainty thresholding on a BraTS 2020 test case for core tumour segmentation across different participating teams. (a) T1ce MRI (b) Ground Truth (c) Prediction (d) No filtering. Uncertainty Threshold = 100 (e) Uncertainty Threshold = 75 (f) Uncertainty Threshold = 50 (g) Uncertainty Threshold = 25. ©[2022] CC-BY. Reprinted, with permission, from [161].

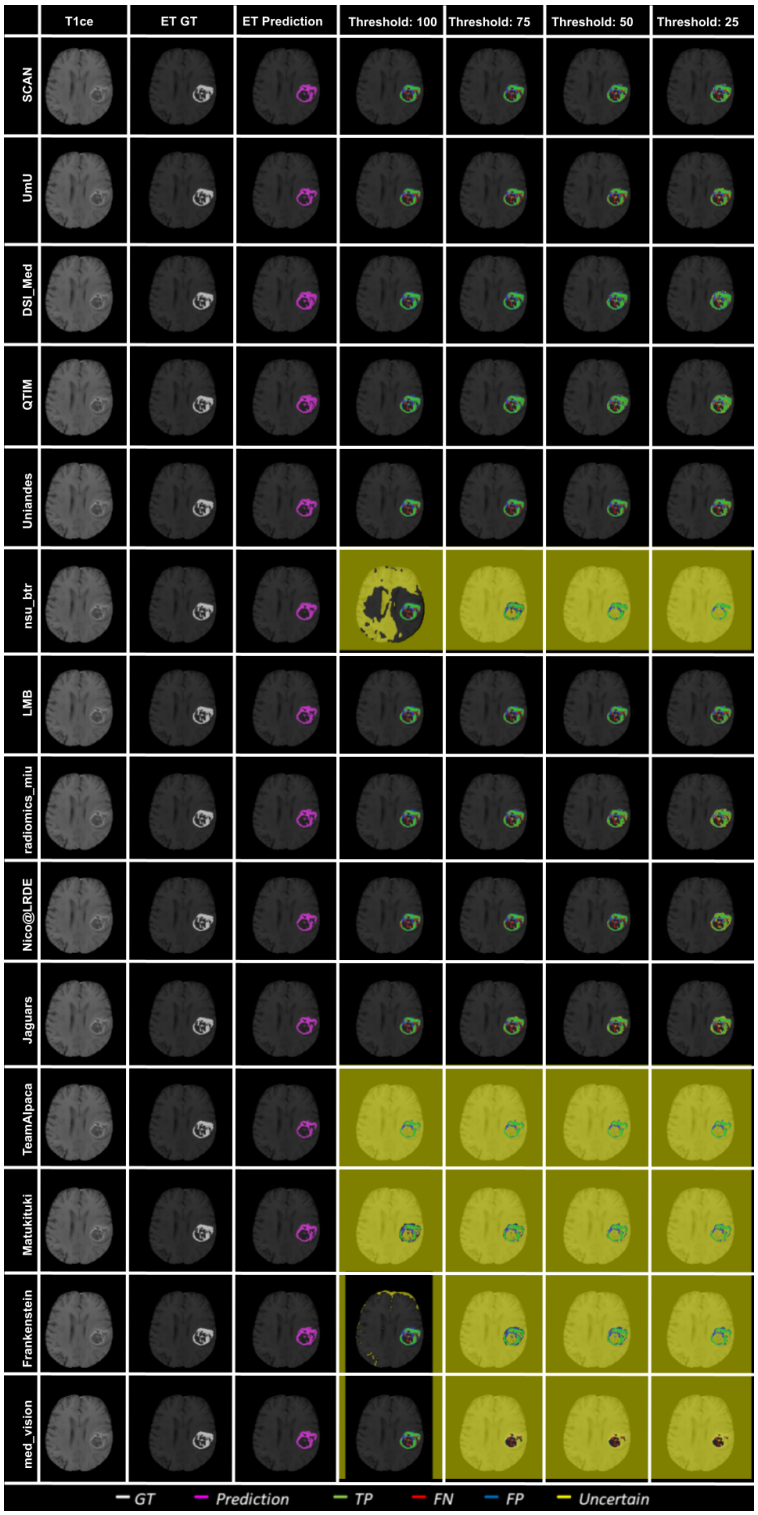


Figure 3.14: Effect of uncertainty thresholding on a BraTS 2020 test case for enhance tumour segmentation across different participating teams. (a) T1ce MRI (b) Ground Truth (c) Prediction (d) No filtering. Uncertainty Threshold = 100 (e) Uncertainty Threshold = 75 (f) Uncertainty Threshold = 50 (g) Uncertainty Threshold = 25. ©[2022] CC-BY. Reprinted, with permission, from [161].

4

Propagating Uncertainty Across Cascaded Medical Imaging Tasks

> Averages mislead by hiding a spread (a range of different numbers) in a single number.
>
> — *Anna Rosling Rönnlund, Hans Rosling, and Ola Rosling*, Factfullness

Related Paper

It should be noted that this is not a manuscript based thesis. However, considerable material from the following paper has been utilised in this chapter.

- **R. Mehta**, T. Christinck, T. Nair, A. Bussy, S. Premasiri, M. Costantino, M. Chakravarty, D. L. Arnold, Y. Gal, T. Arbel, "Propagating Uncertainty Across Cascaded Medical Imaging Tasks for Improved Deep Learning Inference", *IEEE Transactions on Medical Imaging (TMI), Volume: 41, Issue: 2, February 2022* [159].

The IEEE does not require individuals working on a thesis to obtain a formal reuse license. However, it requires that the thesis author cite the source and include IEEE copyright notice for all figures and tables [1].

4.1 Introduction

The previous chapter proposed an uncertainty evaluation measure for the brain tumour segmentation task. Results indicated that the evaluation measure could indeed quantify if the generated task-specific uncertainties are clinically relevant or not. In this chapter, we take a further step and integrate the generated task-specific uncertainties in a cascade of inference tasks, and with extensive experiments show that uncertainty propagation can improve the downstream task of interest.

This part of the thesis presents a general framework for propagating uncertainties across different classes of inference steps. This work presents a unified analysis of a variety of popular uncertainty generation methods (MC-Dropout, Deep Ensembles, Dropout Ensemble), uncertainty measures (e.g., entropy, sample variance, mutual information), and propagation techniques (summary statistics, random sampling) across three distinct contexts: (i) voxel-level binary MS T2 lesion segmentation to lesion detection, (ii) voxel-level MR modality synthesis to voxel-level multi-class brain tumour segmentation, and (iii) voxel-level hippocampus binary segmentation to volume-level Alzheimer's Disease clin-

ical score regression.

Extensive experimentation shows that uncertainty propagation from a previous task to a downstream task of interest results in performance improvements in all three contexts (1-5%) and for all three model sampling methods, with Deep Ensemble and Dropout Ensemble achieving significant performance improvements over MC-Dropout (1-5%). The maximum increase in performance gain with uncertainty propagation (2-5%) is achieved when the entire set of different uncertainty measures are propagated together to the downstream task of interest, indicating that they provide helpful complementary information. However, the quantitative results only tell part of the story. The qualitative results illustrate that uncertainty propagation does indeed assist in correcting clinically relevant errors even when improvements in terms of absolute numbers are small. Finally, experiments indicate that, should the clinical context permit that the multiple samples resulting from the first inference task themselves be available to the downstream task, rather than just the uncertainty information in the form of summary statistics (e.g., entropy, variance), comparable performance improvements on the downstream task of interest result. This might be helpful for other tasks where more complex distributions prevail.

4.2 Methodology: Propagating Uncertainty Across Inference Tasks

We consider a general medical imaging pipeline (see Figure 4.1), where input images, $\mathbf{x}_i$, are passed through a sequence of inference tasks (Task-1, Task-2, ..., Task-K) before producing the downstream output of interest (see Freesurfer [52] or ANTs [15].) The model is general, but here the context explored is one where the images may reflect some patient pathology (e.g., tumour, lesion), leading to additional challenges. The framework follows a protocol where each task is performed by a separate deep learning model sequentially. This is typical for most clinical contexts, where access to the individual training label sets

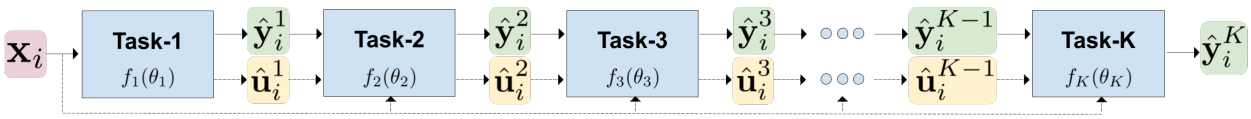


Figure 4.1: An example of a medical image analysis pipeline. During inference, the input image $\mathbf{x}_i$ (and output of previous task, $\hat{\mathbf{y}}_i^k$) is passed through a cascade of inference tasks (1,2,..,K). The neural network for any task, Task-k, is parameterized by θ_k. The output for Task-k is defined as $\hat{\mathbf{y}}_i^k = f_k(\theta_k; \mathbf{x}_i, \hat{\mathbf{y}}_i^{k-1})$. In the proposed framework, we also estimate uncertainties ($\hat{\mathbf{u}}_i^k$) associated with output ($\hat{\mathbf{y}}_i^k$) for each task. These uncertainties are used as an additional input to the subsequent task ($\hat{\mathbf{y}}_i^k = f_k(\theta_k; \mathbf{x}_i, \hat{\mathbf{y}}_i^{k-1}, \hat{\mathbf{u}}_i^{k-1})$). Here, Task-K represents the final downstream task of interest. ©[2022] IEEE. Reprinted, with permission, from [159].

for each of the tasks (e.g., reconstruction, segmentation), ($\mathbf{y}_i^1$, $\mathbf{y}_i^2$, ..., $\mathbf{y}_i^K$), is not typically available for the same input images, $\mathbf{x}_i$. This hinders end-to-end training of the whole medical image analysis pipeline. Each task model is parameterized by its corresponding parameters (θ_1, θ_2, ..., θ_K) such that $\hat{\mathbf{y}}_i^k = f_k(\theta_k; \mathbf{x}_i, \hat{\mathbf{y}}_i^{k-1})$.

We adopt a Bayesian deep learning [79, 136, 237] framework, whereby model predictions ($\hat{\mathbf{y}}_i^k$), as well as uncertainties ($\hat{\mathbf{u}}_i^k$) associated with these predictions can be generated for each task. These uncertainties are estimated by acquiring multiple output samples ($\hat{\mathbf{y}}_{i_{(t)}}^k$) for the same input images. The model prediction becomes the mean of the samples ($\hat{\mathbf{y}}_i^k$), and the uncertainties ($\hat{\mathbf{u}}_i^k$) are derived from statistics across the samples (Section 2.2.2).

In the proposed framework, depicted in Figure 4.1, in addition to passing the model predictions ($\hat{\mathbf{y}}_i^k$) from each preceding task to its subsequent task, uncertainties ($\hat{\mathbf{u}}_i^k$) are also passed onto the subsequent tasks. The hypothesis is that this would lead to better performance for the downstream task of interest. We also explore a premise where instead of passing the mean prediction and its associated uncertainties from the previous task to the subsequent task, the samples ($\hat{\mathbf{y}}_{i_{(t)}}^k$) themselves (should they be available) are passed individually to the next task. Direct sample propagation would help in scenarios where the output distribution might be multi-modal, for example, and not well represented by a single statistic (e.g., variance). It should be noted that this comes at the cost of increased storage requirements and substantial increases in inference time.

In order to prove the generality of the proposed framework, experiments are performed for three different clinical contexts with diverse inference steps: (i) T2 weighted MS lesion segmentation and detection, (ii) brain tumour segmentation, and (iii) Alzheimer's disease (AD) clinical score prediction. Here, pipelines include two different sequential inference tasks, as depicted in Figure 4.2. Note that the uncertainties produced on training cases would not properly reflect the uncertainties on unseen test cases [136, 13, 79]. In the proposed framework, the Task-1 network and the Task-2 network are trained separately to provide the Task-2 network with meaningful Task-1 uncertainties as input.

4.2.1 MS T2 Lesion Segmentation

One of the hallmarks of multiple sclerosis (MS) is the presence of multiple hyperintense lesions visible on T2-weighted MRI (i.e., T2 lesions). The detection and segmentation of T2 lesions in MRI are therefore important to monitor disease activity and treatment efficacy. However, T2 lesions can be very small (3-10 voxels) and difficult to detect. Popular neural networks, including U-Nets, have not yet proven to be effective at the detection and segmentation of small MS lesions in MRI when deployed with commonly used settings [180]. However, uncertainties based on MC-Dropout have been shown to correlate well with network errors in the context of MS lesion segmentation [180]. In this work, we propose to first segment T2 lesions from multi-sequence MRI ($\mathbf{x}_i$) acquired from patients with MS using a Bayesian U-Net [180] (Task-1). The resulting mean T2 lesion segmentation map ($\hat{\mathbf{y}}_i^1$) and its associated voxel-level uncertainties ($\hat{\mathbf{u}}_i^1$), along with the original MRI patient sequences ($\mathbf{x}_i$), are then provided as inputs to a second T2 lesion segmentation U-Net (Task-2). The conjecture is that the second network will learn to improve the lesion segmentation/detection ($\hat{\mathbf{y}}_i^2$) performance by learning to interpret the predictions and associated uncertainties from the first network (see Figure 4.2(A)). This includes learning, for example, which regions with high uncertainties should indeed be labeled as lesions and which should not, thus assisting in detecting and segmenting subtle lesions.

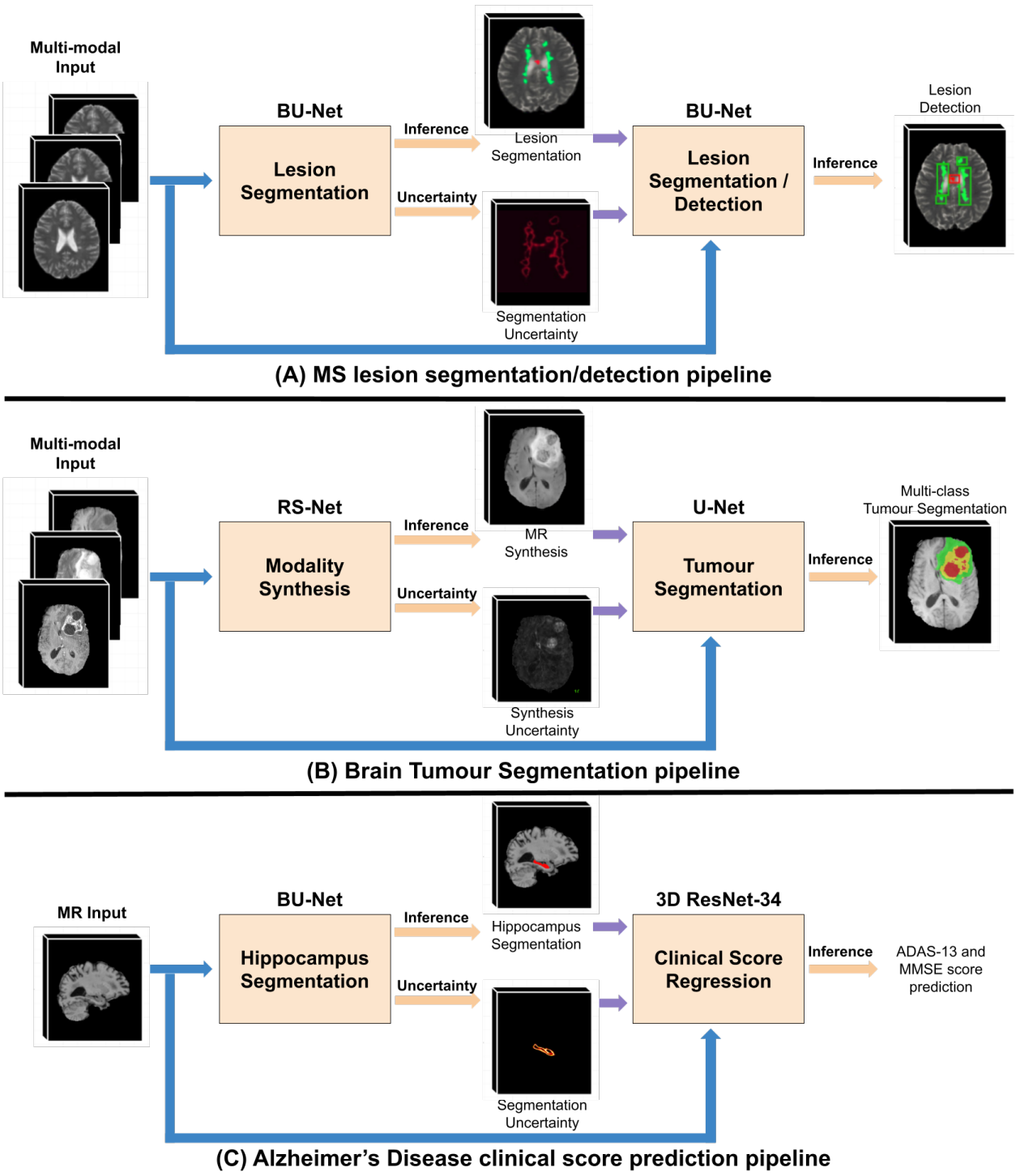


Figure 4.2: Overview of the proposed general framework for propagating inference results and their associated uncertainties across sequential tasks in medical image analysis. (A) MS T2 lesion segmentation, (B) MR synthesis - brain tumour segmentation, and (C) Alzheimer's disease clinical score prediction. ©[2022] IEEE. Reprinted, with permission, from [159].

4.2.2 Brain Tumour Segmentation

The accuracy of detecting and segmenting brain tumours increases significantly should several MRI channels be available. Different contrasts generally assist in differentiating healthy tissues from focal pathologies (e.g., T1, T1c, T2, FLAIR) [22, 93]. However, in real clinical practice, the availability of all sequences is not guaranteed for each patient for various reasons, including cost or time constraints, and corruption from noise or patient motion. As such, accurate synthesis of one or more of the missing 3D MRI volumes based on those acquired would benefit both clinical practices [196] and automatic downstream segmentation techniques [254, 113, 106]. Synthesizing high-resolution volumes in the presence of pathological structures presents significant challenges to current machine learning methods. As a result, any resulting synthesized MR volumes may not be reliable on their own. In this context, voxel-level uncertainties associated with the synthesized volume can be helpful to guide a clinician towards regions of lower confidence where further inspection is needed [158] or towards detecting an anomaly in a synthesized volume [199].

In this work, we suggest that by propagating the uncertainties associated with the synthesized missing MRI sequence provided by the synthesis network (Task-1) to a downstream tumour segmentation network (Task-2), the final results should improve. Details are shown in Figure 4.2(B). The Task-1 network is a synthesis network, which takes multimodal MR sequences acquired from a brain tumour patient as inputs. It regresses a full, synthesized image volume for the mean missing MR sequence ($\hat{\mathbf{y}}_i^1$) as well as the uncertainties ($\hat{\mathbf{u}}_i^1$) associated with the synthesis at each voxel. The synthesis network chosen here is the multi-task regression-segmentation Network (RS-Net) proposed in [158] (Appendix A). The Task-2 network is a multi-class tumour segmentation network that takes the original MRI sequences ($\mathbf{x}_i$), and the synthesized (mean) missing sequence volume ($\hat{\mathbf{y}}_i^1$) and associated uncertainties ($\hat{\mathbf{u}}_i^1$) produced from Task-1 as inputs, and produces multi-class tumour labels ($\hat{\mathbf{y}}_i^2$) at each voxel. The network is a U-Net [45] with instance normalization [256] added in order to improve performance on small batch sizes.

4.2.3 Alzheimer's Disease Clinical Score Prediction

Alzheimer's disease (AD) is the most common form of neurodegenerative disorder in elderly people [83]. Machine learning methods have performed well in providing an AD diagnosis (i.e., a classification task) [279, 85]. However, clinicians are more likely to treat symptoms based on structured clinical assessments (e.g., Alzheimer's disease assessment scale – ADAS-13, mini-mental state examination – MMSE) than on a specific diagnosis [243]. In this work, the objective is to develop an accurate model to estimate clinical disease severity scores, specifically the commonly used ADAS13 [207] and MMSE [69], directly from neuroimaging data (i.e., T1 MR image) [28]. A recognized biomarker for AD is the presence of reduced hippocampal volume as measured from a single time point, high-resolution T1-weighted MR image [73]. As such, automatic hippocampal segmentation has previously been shown to effectively diagnose AD [44, 143].

In this work, we hypothesize that a downstream clinical score prediction network's accuracy can be increased by propagating the estimated uncertainty maps from a preceding hippocampus segmentation network. Details are shown in Figure 4.2 (C). The hippocampal segmentation network (Task-1) is a BU-Net, which takes a T1 MR image ($\mathbf{x}_i$) as input and produces a mean segmentation of the hippocampus ($\hat{\mathbf{y}}_i^1$), as well as an estimate of its associated segmentation uncertainty map ($\hat{\mathbf{u}}_i^1$). The two outputs ($\hat{\mathbf{y}}_i^1$ and $\hat{\mathbf{u}}_i^1$), along with the original T1 MR image ($\mathbf{x}_i$), are then provided to a downstream deep network (3D ResNet-34 [91]) which regresses two clinical scores, ADAS-13 and MMSE ($\hat{\mathbf{y}}_i^2$).

4.3 Implementation Details, Datasets, and Evaluation Metrics

In this section, we provide both task-specific implementation details[1], as well as details about sampling for uncertainty estimation.

4.3.1 Task Specific Details

MS T2 Lesion Segmentation

As depicted in Figure 4.2(A), both the MS T2 lesion labels and their associated uncertainties produced from a Bayesian U-Net are propagated to a second T2 lesion segmentation U-Net. A large proprietary dataset of multi-modal MRI sequences acquired from a total of 1073 patients with relapsing-remitting MS (RRMS) at different stages of the disease was used for training and testing. The dataset consists of over 2700 multi-modal MRI sequences (T1, T2, fluid attenuated inverse recovery – FLAIR, and proton density – PD) federated from three different multi-site, multi-scanner clinical trials. The majority of the patients were scanned annually or bi-annually over 24 months. MRI sequences were acquired at 1mm x 1mm x 3mm resolution. T2 lesion labels were provided with the dataset and were produced through an external process where trained expert human annotators manually corrected a proprietary automated segmentation method. The dataset was split as follows: 40% of the available data was used for training/validating the first network, with a 90/10 training/validation split. Another 40% was used for training/validating the second network, again with a 90/10 training/validation split. The final 20% of the available data was used for testing the second network. The dataset was carefully divided this way to provide the second network with consistent and meaningful uncertainties reflective of unseen test cases.

The downstream outcome of interest is the accurate detection of T2 lesions. Therefore, the performance is evaluated based on lesion-level detection metrics. A connected compo-

[1]Network architecture and training details specific to each pipeline are provided in Appendix C.1

nent analysis is performed on the voxel-based segmentation provided by the network to group lesion voxels in an 18-connected neighbourhood [180]. The detection level metrics, namely true positive rate (TPR) vs. false detection rate (FDR), are calculated at the lesion level and are used to plot receiver operating characteristic (ROC)-like curves. Given that MS lesions vary significantly in size, lesions are grouped into three sized bins for performance evaluation: small (3-10 vox), medium (11-50 vox), and large (51+ vox). Given that the detection of small lesions is particularly challenging and 40% of the lesions in the dataset are small, we mainly focus on the overall detection performance for all the lesions and show the performance on only the small lesions separately. We calculate the area under the curve (AUC) for ROC-like curves and use it as a quantitative measure of the network performance.

Brain Tumour Segmentation

RS-Net (Task-1 network) [158] was developed to take in 3 real MRI sequences and synthesize the missing fourth sequence. This thesis focuses on the synthesis of T1 post-contrast (T1ce) and FLAIR MRIs as previous work [254, 158] has shown that their absence significantly decreases brain tumour segmentation performance compared to either T1 or T2 sequences. T1ce is the most challenging sequence to synthesize, as it is the only MR sequence that indicates enhancement within the tumour post-injection with a contrast agent, providing a signal of new disease activity. T1, T2, and FLAIR sequences are presented to RS-Net to synthesize the T1ce MRI, and T1, T1ce, and T2 MRI sequences are used as inputs to synthesize the FLAIR MRI.

This pipeline is evaluated using the 2018 MICCAI BraTS [22] dataset. The BraTS training dataset comprises 210 HGG and 75 LGG patients with T1, T1ce, T2, and FLAIR MRI sequences. Ground truth tumour labels were provided by expert human annotators and consist of 3 classes: edema, necrotic/non-enhancing core, and enhancing tumor core. 228 patients were randomly selected for training the network and another remaining 57 for

network validation. A separate BraTS 2018 validation dataset was used to test the segmentation performance. This dataset contains 66 patients with multi-channel MRI. The BraTS challenge provides pre-processed volumes that were skull-stripped, co-aligned, and resampled to isotropic (1mm x 1mm x 1mm) resolution. As we mentioned before, uncertainties on a training dataset would not reflect uncertainties on an unseen dataset. The RS-Net was trained in two folds, with each fold comprised of 114 volumes. This training strategy allows us to generate uncertainties on the whole training dataset in two folds, and should reflect uncertainties on an unseen dataset. The downstream segmentation U-Net was trained using all 228 volumes in a single fold.

In line with the BraTS challenge [22], the brain tumour segmentation performance is evaluated by calculating Dice scores for three different tumour sub-types: enhancing tumor, whole tumor, and tumour core. Quantitative assessment was generated by uploading the segmentation results on the challenge portal as there are no ground-truth labels available for the validation set.

Alzheimer's Disease Clinical Score Prediction

As depicted in Figure 4.2, a BU-Net [180] is used for hippocampus segmentation with T1 MRI as the input (Task-1). The segmentation maps and their associated voxel-wise uncertainties are propagated to a volume-level clinical score regression network (Task-2), which produces values for MMSE and ADAS-13 scores. A 3D ResNet-34 [91] network was used for clinical score regression. MMSE is one of the most widely used cognitive assessments for diagnosing Alzheimer's disease and related dementias. The scores range from 0 to 30, with lower scores indicating greater cognitive impairment. The ADAS-13 is a modified version of the ADAS-cog assessment, with a maximum score of 85. In contrast to MMSE, higher scores on the ADAS-13 indicate greater cognitive impairment.

The EADC-ADNI/HARP dataset [74] is used for training the hippocampus segmentation

network. This dataset consists of a subset of 135 volumes selected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, with expert manual 3D segmentations of the hippocampus. There are 45 AD, 46 Mild Cognitive Impairment (MCI), and 44 Cognitive Normal (CN) patients in this dataset. All volumes were in isotropic resolution, brain-extracted, and linearly registered to MNI152 space. We divide this dataset into an 80/20 training/validation split. The clinical score regression network (3D ResNet-34) is trained and tested using the ADNI [111] dataset. Specifically, we used baseline data from participants in the ADNIGO (n=69), ADNI1 (n=442) and ADNI2 (n=354) databases. We divide this dataset into a training/validation/testing (70/10/20) split such that the ratio of AD/MCI/CN is maintained across the split. We perform 5-fold cross-validation on this dataset. Performance evaluation for both ADAS-13 and MMSE scores is based on the Pearson correlation (r), and root mean square error (RMSE) between true and predicted clinical scores.

4.3.2 Sampling For Uncertainty Estimation

The proposed framework requires producing uncertainties at the outputs of the Task-1 network along with the estimated predictions (e.g., voxel-based segmentation, regression). This is achieved by calculating various statistics across multiple samples generated using different uncertainty estimation methods. Specifically, we use MC-Dropout [79], Deep Ensemble [136], and Dropout Ensemble [237] as uncertainty estimation methods. Also, we use predictive entropy, sample variance, and mutual information as uncertainty measures. Readers are referred back to Chapter 2 for more details about these methods.

In this work, we also explore propagating the samples from the Task-1 network directly (if available) as inputs to the Task-2 networks. Details about sampling are now provided:

MC-Dropout

For all three clinical contexts, 20 samples are generated for the Task-1 network using dropout (dropout rate=0.2) at test time. We chose this as previous studies have shown that there is a marginal improvement in performance with more samples [120].

Deep Ensemble

5 different Task-1 networks are trained with different weight initializations on the same training set to get an ensemble of size 5 for each clinical pipeline. This choice is based on previous studies [136, 13] which showed that only marginal improvement was attained with ensembles with sizes larger than 5. During test time, the 5 networks provide 5 different samples for the same input.

Dropout Ensemble

Each of the 5 trained networks developed for the Deep Ensemble model generates 20 samples using dropout at test time. This results in a total of 100 samples for Dropout Ensembles.

4.4 Experiments and Results

Several experiments were performed for each of the clinical pipelines. The goal was to evaluate the effectiveness of propagating the uncertainties from Task-1 to Task-2 in improving the final downstream results. Evaluations and comparisons were made based on (a) different uncertainty estimation methods: MC-Dropout [79], Deep Ensemble [136], and Dropout Ensemble [237], (b) different uncertainty measures: sample variance, entropy, MI, and finally (c) propagating the uncertainties derived from the samples (e.g., sample variance) against propagating the samples themselves.

Table 4.1: Comparing overall MS T2 lesion detection performance using Area Under Curve (AUC) of ROC-like curves, illustrating TPR (true positive rate) vs. FDR (false detection rate) across (A) all lesions, and (B) small lesions (3-10 voxels) with several input combinations. The inclusion of the associated uncertainties with outputs from Task-1, in addition to Task-1 outputs, as inputs to the Task-2 network results in improved detection performance. **Bold** values indicate the best performance for each method, while underlined values indicate the overall best performance across different methods. The performance of the MS T2 lesion detection for medium and large lesions is provided in Table 4.2 in Appendix C.1. ©[2022] IEEE. Reprinted, with permission, from [159].

	Method	Input						AUC all lesions (↑)	AUC small lesions (↑)
		MR sequences (T1, T2, FLR, T1ce, PDw)	Mean Segm.	Uncertainties			Segm. Samples		
				Var.	Entr.	MI			
1	**Baseline-1**	✓						0.8425	0.6704
2	**MC-Dropout**	✓	✓					0.8465	0.6837
3		✓	✓	✓				0.8643	**0.7197**
4		✓	✓		✓			0.8479	0.6876
5		✓	✓			✓		0.8419	0.6853
6		✓	✓	✓	✓	✓		**0.8652**	0.7170
7		✓					✓	0.8591	0.7019
8	**Dropout Ensemble**	✓	✓					0.8613	0.7116
9		✓	✓	✓				0.8739	0.7312
10		✓	✓		✓			0.8650	0.7235
11		✓	✓			✓		0.8654	0.7131
12		✓	✓	✓	✓	✓		**0.8781**	**0.7409**
13		✓					✓	0.8771	0.7341
14	**Deep Ensemble**	✓	✓					0.8603	0.7113
15		✓	✓	✓				0.8735	0.7349
16		✓	✓		✓			0.8697	0.7225
17		✓	✓			✓		0.8649	0.7159
18		✓	✓	✓	✓	✓		**0.8792**	**0.7410**
19		✓					✓	0.8767	0.7369

4.4.1 Effectiveness of Uncertainty Propagation

The first set of experiments was designed to evaluate the effectiveness of propagating uncertainties from the Task-1 network to the Task-2 network. To this end, we first examine the results of the proposed framework for all three clinical pipelines (Figure 4.2) based on a set of fixed experimental parameters: using MC-Dropout [79] during inference to provide 20 samples from the Task-1 network, and estimating and propagating the sample mean and variance across these samples to the Task-2 network along with the original MRI [2]. Sample variance was chosen as it is the simplest and the most commonly used uncertainty measure [120, 260, 139, 188, 253, 199]. Results were compared against

[2]Figure C.5 in the Appendix C.1 shows the effect of varying the number MC-Dropout of sample for uncertainty estimation on a downstream task of interest for MS lesion detection.

Table 4.2: Comparing overall MS T2 lesion detection performance using area under curve (AUC) of ROC-like curves, illustrating TPR (true positive rate) vs. FDR (false detection rate) across (A) large lesions (51+ voxels), and (B) medium lesions (10-50 voxels) with several input combinations. The inclusion of the associated uncertainties with outputs from Task-1, in addition to Task-1 outputs, as inputs to the Task-2 network results in improved detection performance. **Bold** values indicate the best performance for each method, while underlined values indicate the overall best performance across different methods. ©[2022] IEEE. Reprinted, with permission, from [159].

	Method	Input						AUC Med. lesions (↑)	AUC large lesions (↑)
		MR sequences (T1, T2, FLR, T1ce, PDw)	Mean Segm.	Uncertainties Var.	Uncertainties Entr.	Uncertainties MI	Segm. Samples		
1	**Baseline-1**	✓						0.9768	0.9977
2	**MC-Dropout**	✓	✓					0.9775	0.9979
3		✓	✓	✓				0.9834	**0.9992**
4		✓	✓		✓			0.9768	0.9982
5		✓	✓			✓		0.9782	0.9979
6		✓	✓	✓	✓	✓		**0.9852**	0.9989
7		✓					✓	0.9831	0.9981
8	**Dropout Ensemble**	✓	✓					0.9801	0.9988
9		✓	✓	✓				0.9858	0.9992
10		✓	✓		✓			0.9807	0.9992
11		✓	✓			✓		0.9851	0.9992
12		✓	✓	✓	✓	✓		**0.9861**	**0.9993**
13		✓					✓	0.9844	0.9989
14	**Deep Ensemble**	✓	✓					0.9806	0.9988
15		✓	✓	✓				0.9855	0.9992
16		✓	✓		✓			0.9812	0.9991
17		✓	✓			✓		0.9856	0.9993
18		✓	✓	✓	✓	✓		**0.9867**	**0.9993**
19		✓					✓	0.9854	0.9990

(1) *Baseline-1*: only passing the MR sequences to Task-2 and (2) *Baseline-2*: passing the MRIs and the sample mean outputs from the Task-1 network to Task-2. Comparisons between Baseline-1 and Baseline-2 indicate the effectiveness of cascading inference results in general. A comparison of the proposed method with Baseline-2 should reflect the effectiveness of additionally propagating uncertainties.

Tables 4.1, 4.3, and 4.4 illustrate the results for the MS lesion segmentation/detection, brain tumour segmentation, and AD clinical score prediction pipelines, respectively. We perform a two-sided paired sample t-test to find a statistically significant difference between methods that propagates uncertainty and the baseline method which doesn't consider uncertainty propagation [3].

[3]We do not report the statistical significance test result for Table I as it would require us to run multiple different runs for the large MS dataset, where each training setup takes approximately four days to run, which is practically infeasible. In this case, we have kept the folds constant across different experiments throughout the experiments (and even the random seeds for the neural network initialization), which gives

Table 4.3: Comparison of multi-class brain tumour segmentation performance on the BraTS Validation dataset. The inclusion of the associated uncertainties from the synthesis network, in addition to the synthesis output, as input to the segmentation network results in improved performance. Quantitative results are based on percentage Dice coefficients for enhancing tumor (DE), whole tumor (DT), and tumor core (DC). * indicates statistically significant ($p \leq 0.05$) differences between including and excluding uncertainty using a two-sided paired sample t-test. **Bold** values indicate the best performance for each method, while underlines indicate the overall best performance across different methods. ©[2022] IEEE. Reprinted, with permission, from [159].

		Method	Input									Dice Coefficients (%)		
			Real MR sequ.				synth. MR sequ.		Var.	synth. samples		DT	DC	DE
			T1	T2	T1ce	FLR	T1ce	FLR	Uncer.	T1ce	FLR	(↑)	(↑)	(↑)
FLR Synthesis	1	**Baseline-1**	✓	✓	✓							83.27	73.91	71.07
	2	**MC-Dropout**	✓	✓	✓			✓				84.56	76.72	72.89
	3		✓	✓	✓			✓	✓			**85.84** *	**79.25** *	**74.51** *
	4		✓	✓	✓						✓	84.83	78.43 *	73.84
	5	**Dropout Ensemble**	✓	✓	✓			✓				84.76	77.65	74.09
	6		✓	✓	✓			✓	✓			**86.45** *	**78.98**	**75.43**
	7		✓	✓	✓						✓	86.33	79.11 *	74.99
	8	**Deep Ensemble**	✓	✓	✓			✓				84.86	77.52	74.01
	9		✓	✓	✓			✓	✓			**86.51** *	**79.98** *	**75.24**
	10		✓	✓	✓						✓	86.03	79.19 *	74.99
T1ce Synthesis	1	**Baseline-1**	✓	✓		✓						87.17	50.25	26.89
	2	**MC-Dropout**	✓	✓		✓	✓					86.72	52.80	27.35
	3		✓	✓		✓	✓		✓			**88.20**	**57.29** *	**32.86** *
	4		✓	✓		✓				✓		87.91	56.71 *	31.95
	5	**Dropout Ensemble**	✓	✓		✓	✓					87.54	55.41	29.62
	6		✓	✓		✓	✓		✓			**88.38**	**58.99** *	**34.02** *
	7		✓	✓		✓				✓		88.01	58.09 *	32.91
	8	**Deep Ensemble**	✓	✓		✓	✓					87.45	55.68	29.62
	9		✓	✓		✓	✓		✓			**88.63**	**58.84** *	**33.91** *
	10		✓	✓		✓				✓		88.28	57.76 *	32.56

Row-1 to Row-3 in each of these tables illustrate that, for all three pipelines, the network for the downstream task of interest (Task-2) shows performance improvements of 0.5-4% when the Task-1 sample mean output is passed to the Task-2 network, relative to only passing MR sequences (*Baseline-1*). Propagating uncertainties leads to a further 2-12% performance improvement over only passing the Task-1 sample mean output to the Task-2 network (*Baseline-2*).

Although quantitative improvements are important, they do not tell the entire story. In some cases, the overall numerical improvements based on the standard performance metrics seem relatively small, however, there still can be significant clinically relevant improvements. For example, Figure 4.3 depicts qualitative results for three MS patient

a fair comparison without repeated runs

Table 4.4: ADAS-13 and MMSE score prediction performance comparison on the ADNI test dataset. The inclusion of the associated uncertainties from the hippocampus segmentation network, in addition to the hippocampus segmentation output, as input to the clinical score prediction network improves both ADAS-13 and MMSE. Quantitative prediction performance is based on root mean squared error (RMSE) and Pearson correlation coefficient (r). (*) indicates statistically significant ($p \leq 0.05$) differences between including and excluding uncertainty using a two-sided paired sample t-test. **Bold** values indicate the best performance for each method, while underlined values indicate the overall best performance across different methods. ©[2022] IEEE. Reprinted, with permission, from [159].

	Method	Input						ADAS-13		MMSE	
		T1 MR sequence	Mean seg.	Uncertainties			Segm. samples	RMSE (↓)	r (↑)	RMSE (↓)	r (↑)
				Var.	Entr.	MI					
1	**Baseline-1**	✓						7.87 ± 0.92	0.47 ± 0.09	2.28 ± 0.17	0.46 ± 0.11
2	**MC-Dropout**	✓	✓					7.77 ± 0.76	0.48 ± 0.06	2.28 ± 0.12	0.47 ± 0.08
3		✓	✓	✓				7.47 ± 0.76 *	0.54 ± 0.05 *	2.23 ± 0.10 *	0.51 ± 0.05 *
4		✓	✓		✓			7.71 ± 0.77	0.47 ± 0.06	2.28 ± 0.15	0.48 ± 0.08
5		✓	✓			✓		7.72 ± 0.78	0.46 ± 0.05	2.27 ± 0.14	0.47 ± 0.07
6		✓	✓	✓	✓	✓		**7.45 ± 0.72** *	**0.54 ± 0.04** *	**2.22 ± 0.11** *	**0.51 ± 0.06** *
7		✓					✓	7.51 ± 0.71 *	0.51 ± 0.06 *	2.24 ± 0.15 *	0.49 ± 0.07 *
8	**Dropout Ensemble**	✓	✓					7.67 ± 0.74	0.50 ± 0.04	2.26 ± 0.12	0.48 ± 0.09
9		✓	✓	✓				7.38 ± 0.71 *	0.57 ± 0.05 *	2.17 ± 0.13 *	0.51 ± 0.05 *
10		✓	✓		✓			7.59 ± 0.72	0.50 ± 0.04	2.26 ± 0.11	0.47 ± 0.08
11		✓	✓			✓		7.68 ± 0.68	0.50 ± 0.04	2.25 ± 0.13	0.48 ± 0.07
12		✓	✓	✓	✓	✓		**7.36 ± 0.73** *	**0.57 ± 0.06** *	**2.15 ± 0.16** *	**0.53 ± 0.06** *
13		✓					✓	7.37 ± 0.61 *	0.54 ± 0.04 *	2.19 ± 0.17 *	0.52 ± 0.06 *
14	**Deep Ensemble**	✓	✓					7.69 ± 0.74	0.50 ± 0.05	2.27 ± 0.11	0.47 ± 0.08
15		✓	✓	✓				7.38 ± 0.71 *	0.57 ± 0.05 *	2.19 ± 0.14 *	0.52 ± 0.06 *
16		✓	✓		✓			7.60 ± 0.70	0.51 ± 0.05	2.27 ± 0.14	0.47 ± 0.09
17		✓	✓			✓		7.69 ± 0.67	0.49 ± 0.05	2.25 ± 0.14	0.48 ± 0.06
18		✓	✓	✓	✓	✓		**7.34 ± 0.73** *	**0.57 ± 0.04** *	**2.15 ± 0.17** *	**0.54 ± 0.07** *
19		✓					✓	7.38 ± 0.63 *	0.55 ± 0.05 *	2.18 ± 0.18 *	0.52 ± 0.06 *

cases (top to bottom), where the propagation of uncertainties enabled the correction of both false positive (bottom case) and false negative (top two cases) lesions. The system learned how to interpret the uncertainties in the (incorrect) inferences made in those areas, and corrected the errors.

Figure 4.4 shows example cases for three patients (top to bottom), where the downstream brain tumour segmentation network makes use of synthesized MRI sequences (here T1ce and FLAIR). The first example (top row) shows that propagating the synthesized T1ce image to the downstream tumour segmentation network results in confusion between enhancing tumour and core tumour, as the enhancing portion is not well synthesized in the generated T1ce. This result is not unsurprising as T1ce is the post-contrast injection T1 MRI, and accurate synthesis of the enhanced tumour without injection remains an open problem. Importantly, the system produces an uncertainty map that indicates that the synthesis uncertainty is higher in this region, and conveys the uncertainty information

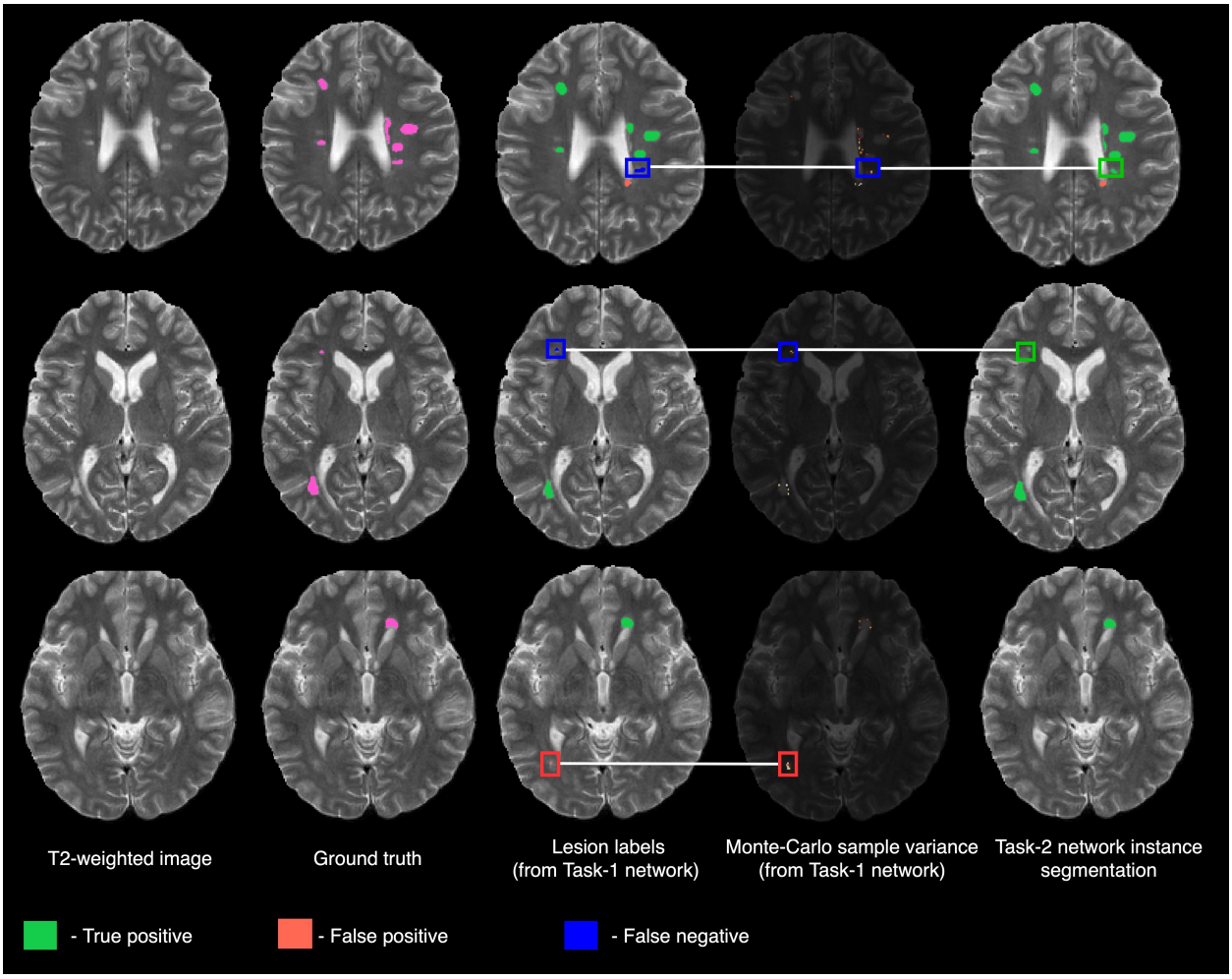


Figure 4.3: Examples demonstrating the corrective effect of uncertainty propagation for MS lesion detection for three patient cases (Rows 1-3). From left to right: T2 weighted MRI input, expert T2 lesion labels (in magenta), T2 lesion labels produced by the Task-1 network, sample variance uncertainty estimates for the Task-1 network output, and the T2 lesion labels produced by the Task-2 network. ©[2022] IEEE. Reprinted, with permission, from [159].

to the segmentation network. This enables the segmentation network to learn to correct these errors and leads to an improvement in the results. This can also be seen in the example in the second row, where the uncertainty allows the network to fix errors and correctly identify enhancing and non-enhancing cores. The third example shows the results of FLAIR synthesis, where an erroneous bright spot appears within the ventricle. This leads to the segmentation network erroneously predicting edema within the ventricle (which is clinically impossible) when the uncertainty is not propagated. However, the uncertainty maps indicate that the network is not confident in its synthesis prediction in this region. As such, cascading the uncertainty maps permits the network to learn to

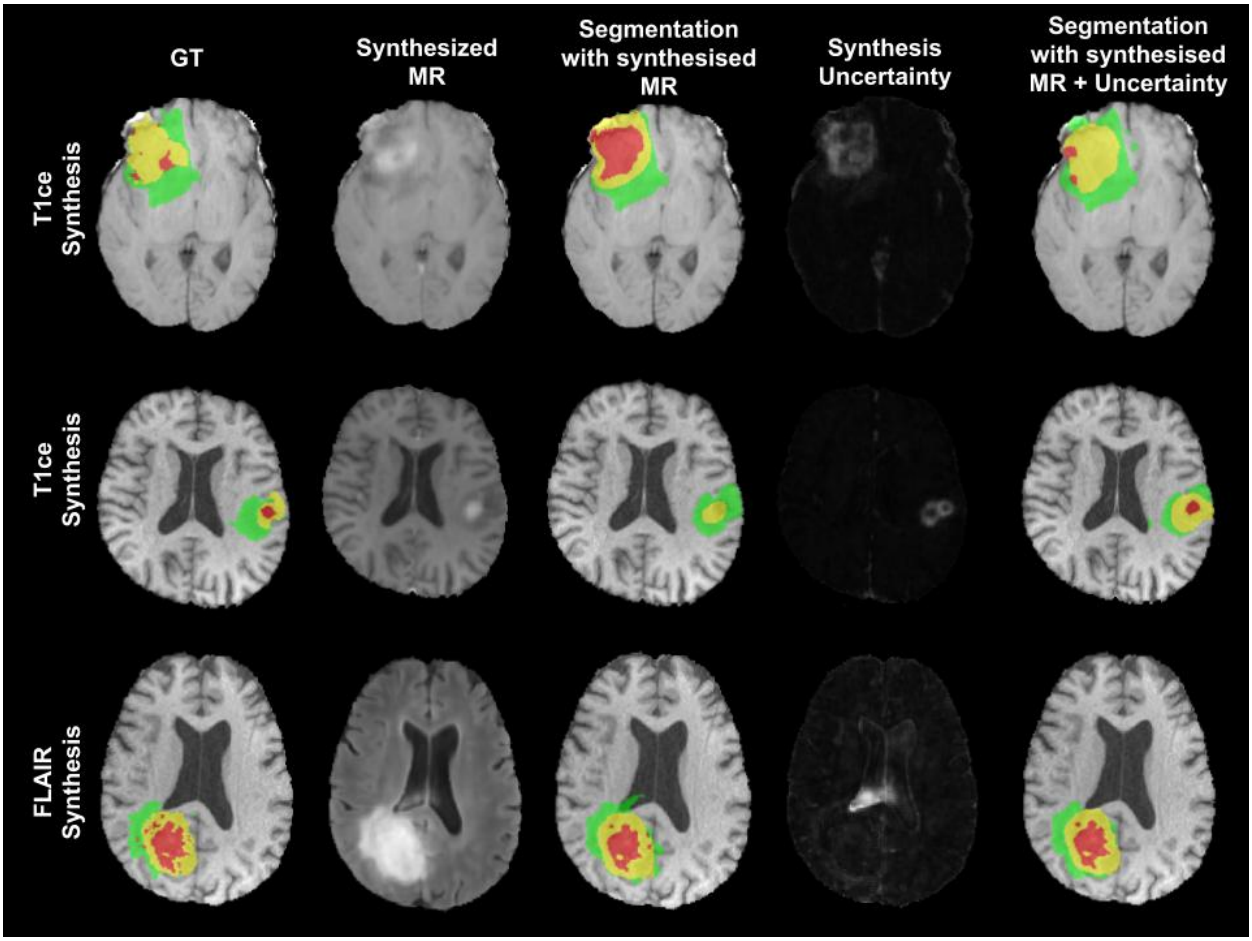


Figure 4.4: Examples of three patient cases (top to bottom) demonstrating the 3D U-Net performance on the multi-class brain tumour segmentation task [22] based on synthesized MRI sequences. From Left to Right: Expert manual segmentation, synthesized MR sequence, segmentation using real MRI (3 sequences) + synthesized MRI, synthesis uncertainty, segmentation using real MRI (3 sequences) + synthesized MRI + synthesis uncertainty. First two rows: T1ce synthesis. Last row: FLAIR synthesis. Labels: edema (green), non-enhancing or necrotic tumour core (red), enhancing tumour (yellow). ©[2022] IEEE. Reprinted, with permission, from [159].

correct its error.

The results for all 3 clinical pipelines demonstrate that multi-step medical image processing pipelines, that would otherwise accumulate errors can benefit from including the network uncertainty for each task as input to subsequent tasks.

4.4.2 MC-Dropout vs. Deep Ensemble vs. Dropout Ensemble

The next set of experiments compares the performance of uncertainty propagation using different methods for estimating sample variance uncertainties: MC-Dropout [79], Deep Ensemble [136], and Dropout Ensemble [237]. Tables 4.1, 4.3, and 4.4, Row-2 and Row-3, Row-8, and Row-9, and Row-14 and Row-15 report results for MC-Dropout, Dropout Ensemble, and Deep Ensemble, respectively. These results indicate that ensemble methods, Deep Ensemble and Dropout Ensemble, achieve 1-5% higher performance over MC-Dropout when only mean predictions are propagated across tasks. The performance gains improve by a further 1-4% when the sample variance uncertainties are additionally propagated to the downstream task of interest. A marginal performance gain of Dropout Ensemble over Deep Ensemble can be seen, both with and without uncertainty propagation.

4.4.3 Effect of Different Uncertainty Measures

Experiments were devised in order to compare the effects of propagating each of the different uncertainty measures: sample variance, entropy, and MI (Section 2.2.2), as well as the effectiveness of cascading all three measures at once for all three uncertainty estimation strategies. Experiments were performed for the clinical pipelines of MS lesion segmentation/detection and AD clinical score prediction, but not for the brain tumour segmentation pipeline as estimating entropy or MI in the context of image regression (synthesis) in this context is an open research problem [80]. Tables 4.1 and 4.4 show that the sample variance gives better performance gains over entropy and MI for both the MS T2 lesion detection task and AD clinical score prediction task. However, passing all three uncertainty measures simultaneously shows the best improvement in the performance of downstream tasks (Row-7, Row-13, and Row-19), indicating that each provides different yet relevant summary statistics [180].

4.4.4 Statistics vs Samples

Finally, the effectiveness of passing summary statistics calculated across samples is examined against propagating the samples themselves for all three uncertainty estimation strategies. Multiple samples are generated (Section 4.3.2) from the Task-1 network for these uncertainty estimation strategies. During Task-2 network training, one random sample from the available Task-1 output samples is provided as input. During inference, all Task-1 samples are independently passed to the Task-2 network. The output samples from the Task-2 network are then used to estimate the sample mean, which serves as the final Task-2 output. Table 4.1, Table 4.3, and 4.4 indicate that passing samples instead of statistics across samples results in similar performance in the contexts explored in this chapter.

4.5 Summary

This work proposes a general deep learning framework for propagating uncertainties across a sequence of inference tasks within medical image analysis pipelines. It demonstrates that cascading uncertainties (e.g., based on MC dropout, Deep Ensemble) along with the outputs from the previous inference module can lead to improvements in the performance of the downstream task. The framework was applied to three different contexts. First, we showed that by propagating voxel-based lesion segmentation uncertainties to a second segmentation network, lesion-level detection performance could be improved by reducing both FPs and FNs. Experiments were performed on a large-scale, multi-site MS patient brain MRI dataset acquired during different clinical trials. Next, using the publicly available BraTS dataset, we demonstrated that by propagating regression uncertainties from an MRI synthesis network, the performance of a downstream multi-class tumour segmentation task could be improved. In the last context, we demonstrated that uncertainty propagation from a voxel-level hippocampus segmentation network to a scan-level clinical score regression task in the context of images acquired from AD pa-

tients leads to improved predictions. These results are encouraging and suggest that uncertainties can be propagated to a downstream task of interest to improve performance in cascaded medical image processing pipelines where the upstream task is related to the downstream task of interest[4]. The expectation is that the results are generalizable to other clinical pipelines. Our experiments also showed that by propagating Task-1 samples to the Task-2 network as a proxy to the uncertainty associated with the Task-1 output, we could achieve similar performance. This is important as samples could better represent the Task-1 output distribution when it is multi-modal, compared to a single statistic like sample variance. It should be noted that the performance improvements resulting from uncertainty propagation are dependent on the number of samples taken to estimate the uncertainties (as we show in Appendix C.1 - Figure C.5), as well as sample generation method. As a result, it would be important to tune these hyper-parameters for optimal performance in the particular application of interest.

[4]Propagating uncertainties from a skull stripping task to a hippocampus segmentation task might not lead to performance improvement, as the two tasks are not directly related.

5

Evaluating the Fairness of Deep Learning Uncertainty Estimates in Medical Image Analysis

– *Bill Watterson*, Calvin and Hobbs

Related Paper

It should be noted that this is not a manuscript based thesis. However, considerable material from the following paper has been utilised in this chapter.

- **R. Mehta**, C. Shui, T. Arbel, "Evaluating the Fairness of Deep Learning Uncertainty Estimates in Medical Image Analysis", *Medical Imaging with Deep Learning (MIDL) 2023* [164].

The MIDL conference papers are published in the Proceedings of Machine Learning Research (PMLR), which follows the CC-BY license and does not require individuals working on a thesis to obtain a formal reuse license. This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use [3].

5.1 Introduction

In the previous chapter of this thesis, we looked at uncertainty propagation for medical image analysis pipelines and showed how uncertainty propagation could help in improving the performance of the downstream task of interest. While this indeed shows the effectiveness of generating uncertainty associated with medical image analysis models, uncertainty itself cannot lead to more trustworthy and clinically applicable models. In this chapter, we take the next step towards trustworthy models by looking at it from the lens of fairness and uncertainty quantification together. Related literature review for the same is given in Chapter 2.

We present the first analysis of the effect of popular fairness models at overcoming biases of Deep learning (DL) models across subgroups for various medical image analysis tasks, and investigate and quantify their effects on the estimated output uncertainties. We conjecture that uncertainty quantification can help mitigate some potential risks in clinical deployment related to a lack of robustness and fairness for under-represented

populations. However, the uncertainties will only help clinicians make more informed decisions if they are accurate. Specifically, a machine learning model that underperforms for an under-represented subgroup should indicate high uncertainties associated with its output for that subgroup. Conversely, a machine learning model that achieves fairness in terms of performance across different subgroups, but produces low uncertainties for predictions where it makes mistakes, would become less trustworthy to clinicians.

We perform extensive experiments on three different clinically relevant tasks: (i) multi-class skin lesion classification [46], (ii) multi-class brain tumour segmentation [22], and (iii) Alzheimer's disease clinical score [111] regression. Our results indicate a lack of fairness in model performance for under-represented groups. The uncertainties associated with the outputs behave differently across different groups. We show that popular methods designed to mitigate the lack of fairness, specifically data balancing [197, 107, 105, 284] and robust optimization [213, 284] do indeed improve fairness for some tasks. However, this comes at the expense of poor performance of the estimated uncertainties in some cases. This tradeoff must be mitigated if fairness models are to be adopted in medical image analysis.

5.2 Methodology: Fairness in Uncertainty Estimation

This work aims to evaluate the effectiveness of various popular machine learning fairness models at mitigating biases across subgroups in various medical image analysis contexts in terms of (a) the absolute performance of the models and (b) the uncertainty estimates across the subgroups. Although general, the framework and associated notations focus on binary sensitive attributes (e.g., sex, binarized ages, disease stages).

Consider a dataset $D = \{X, Y, A\} = \{(x_i, y_i, a_i)\}_{i=1}^{N}$ with N total samples. Here, $x_i \in \mathbb{R}^{P \times Q}$ or $x_i \in \mathbb{R}^{P \times Q \times S}$ represents 2D or 3D input image, where P, Q, and S represent the number of pixels/voxels in each dimension. Also, y_i represents corresponding ground truth

labels, and $a_i = \{0, 1\}$ represents the sensitive binary group-attribute. y_i depends on the task at hand: $y_i \in \{0, 1, .., C\}$ for image-level classification, $y_i \in \mathbb{R}$ for image-level regression, and $y_i \in \{0, 1, ..C\}^{P \times Q}$ or $y_i \in \{0, 1, ..C\}^{P \times Q \times S}$ for 2D/3D voxel-level segmentation. The dataset can be further divided into subgroups, $A = \{0, 1\}$, based on the value of the sensitive attribute: (i) $D^0 = \{X^0, Y^0, A = 0\} = \{(x_i^0, y_i^0, a_i = 0)\}_{i=1}^{M}$ and (ii) $D^1 = \{X^1, Y^1, A = 1\} = \{(x_i^1, y_i^1, a_i = 0)\}_{i=1}^{L}$, where $M + L = N$.

5.2.1 Fairness

Let us consider a DL model $f(., \theta)$ that produces a set of outputs $\hat{Y} = f(X, \theta)$ for a set of input multi-dimensional images, X. The literature on fairness metrics in medical imaging is fairly sparse as this is a relatively new area of research. The goal here is to define a global fairness metric that is applicable and consistent across a wide variety of tasks (e.g. classification, segmentation, regression). Fairness can be defined as follows: A machine learning model is considered to be fair if the difference in the task-specific performance metric between different subgroups is low. To that end, a general fairness gap (FG) metric (Equation 5.1) calculates the differences in the task-specific evaluation metric (EM) values between $\hat{Y}$ and Y when conditioned on a binary sensitive attribute A. The majority of the fairness metrics [96] are only defined for the classification task. There has been some recent work related to the fairness of segmentation models [197, 107], where fairness gap metrics are aligned with the one presented in this work. To our knowledge, fairness in regression in medical imaging has not yet been explored in the medical image analysis literature.

$$\text{FG}(A = 0, A = 1) = |\text{EM}(Y^0, \hat{Y}^0) - \text{EM}(Y^1, \hat{Y}^1)|. \tag{5.1}$$

A machine learning model is fair for the sensitive attribute A if $\text{FG}(A = 0, A = 1) = 0$. EM differs depending on the task at hand, for example, accuracy for image classification, dice value for segmentation, and mean squared error for image-level regression. EM is cal-

culated for each image separately and then averaged across the dataset for a voxel-level segmentation task. For image classification or regression tasks, EM is calculated directly at a dataset level.

5.2.2 Uncertainty and Fairness

In the machine learning and computer vision fairness literature, the objective is to bridge the performance gap across subgroups with different attributes. It is well established in the literature [283, 59], however, that reduced fairness gap across different subgroups can come at the cost of poor overall performance. For example, increasing the model performance on the underrepresented group and decreasing the model performance on the overrepresented group can lead to a smaller fairness gap, but it can come at the cost of overall worse performance of the model, which is undesirable. In [283, 59], they do not consider the effect of the bias mitigation methods on the uncertainties associated with the model output. As we saw in the previous chapters (Chapter 3, and Chapter 4), in medical image analysis, however, model output uncertainties can play an important role in gaining clinician trust. A machine learning model that is fair in terms of performance across subgroups but underperforms overall (e.g., accuracy, precision) could still be clinically useful if it indicates high uncertainties associated with its output when it is incorrect, and if it is confident when correct. With this objective in mind, this work analyzes the effectiveness of fairness mitigation methods not only in terms of absolute performance but also in terms of the quantification of uncertainties associated with their output.

There are multiple different existing works in the literature that could allow the deep learning networks to produce uncertainties associated with model output, for example, Bayesian deep learning methods [182, 79, 237], Ensembling methods [136, 272, 101], conformal prediction [9], etc. The focus of this work is not to compare these different uncertainty quantification methods, but rather to analyze the majority of the current popular fairness mitigation methods through the perspective of quantification of uncertainties as-

sociated with model output. To this end, we specifically focus on Bayesian deep learning (BDL) models and Ensembling models [182, 79, 136, 237] (Section 2.2.1), which are widely adopted within the medical image analysis community [24, 159] given their ability to produce uncertainty estimates, $\hat{u}_i$, associated with the model output $\hat{y}_i$. Popular uncertainty estimates include (Section 2.2.2) sample variance, predicted variance, entropy, and mutual information [121, 80]. Uncertainties $\hat{u}_i$ are typically normalized between 0 (low uncertainty) and 100 (high uncertainty) across the dataset.

We propose an uncertainty fairness evaluation metric (Equation 5.2), which evaluates the fairness gap metric at different uncertainty thresholds. This metric follows the popular convention employed by various papers in the literature for evaluating uncertainties in the contexts of classification [24, 82] and segmentation (Chapter 3). The rationale behind these approaches is to design an uncertainty evaluation method, where performance is measured based on the following criteria: The correct predictions should have low uncertainties, and the uncertainties for the incorrect predictions are high. Following those papers, all predictions whose output uncertainties ($\hat{u}_i$) are above a threshold (τ) are labeled as uncertain. The task-specific evaluation metric (EM) and fairness pap (FG) are then calculated on the remaining certain predictions ($\hat{Y}_\tau$ and Y_τ) (below the threshold). In this work, we perform this evaluation for a range of different uncertainty thresholds, and plot EM (and FG) vs. (100-uncertainty threshold).

$$\text{FG}_\tau(A=0, A=1) = |\text{EM}_\tau(Y_\tau^0, \hat{Y}_\tau^0) - \text{EM}_\tau(Y_\tau^1, \hat{Y}_\tau^1)|. \tag{5.2}$$

At $\tau = 100$, equations 1 and 2 become equivalent. A higher degree of fairness in uncertainty estimation is established through a reduced fairness gap ($\text{FG}_{\tau 1} \leq \text{FG}_{\tau 2}$) when the number of filtered predictions increases. In other words, when the uncertainty threshold is reduced from $\tau 1$ to $\tau 2$ (where, $\tau 1 < \tau 2$), thereby increasing the number of filtered uncertain predictions, the differences in the performances on the remaining confident predictions across the subgroups should be reduced. However, this decrease should not lead

to a reduction in overall performance. In other words, it is desirable that $EM_{\tau 1} \geq EM_{\tau 2}$. Conversely, an increase in the fairness gap ($\text{FG}_{\tau 1} > \text{FG}_{\tau 2}$) indicates the undesirable effect of having a higher degree of confidence in incorrect predictions for one of the subgroups.

5.3 Experiments and Results

Extensive experimentation involves comparisons of two established fairness models against a baseline: (i) A **Baseline-Model**: trained on a dataset without consideration of any subgroup information. This model is trained with standard ERM loss (cross entropy for classification and segmentation, and mean squared error for regression) and does not consider any subgroup information, and thus can act as a baseline method. (ii) A **Balanced-Model**: trained on a balanced dataset. Here, the training dataset is constructed to contain an equal number of images from each subgroup. This might lead to a smaller overall dataset as the underrepresented group has a lower number of total images compared to the overrepresented group, and balancing across subgroups would require sampling a smaller number of images from the overrepresented group. This is an established fairness model that focuses on mitigating biases due to data imbalance [197, 107, 105, 284]. (iii) A **GroupDRO-Model**: trained with GroupDRO loss [213]. Compared to the **Balanced-Model**, in this mitigation method, the dataset is not balanced across different subgroups, but instead, loss for different subgroups is re-weighed differently, thereby mitigating the lack of fairness through the optimization procedure. This model might give better performance compared to the **Balanced-Model** as it doesn't require subsampling of the overrepresented group, and thus can have access to an overall higher number of training data points. The **Balanced-Model** addresses fairness from the data balancing perspective, and the **GroupDRO-Model** tackles it from the optimization perspective. The number of images in the test set is the same across all subgroups for fair comparisons.

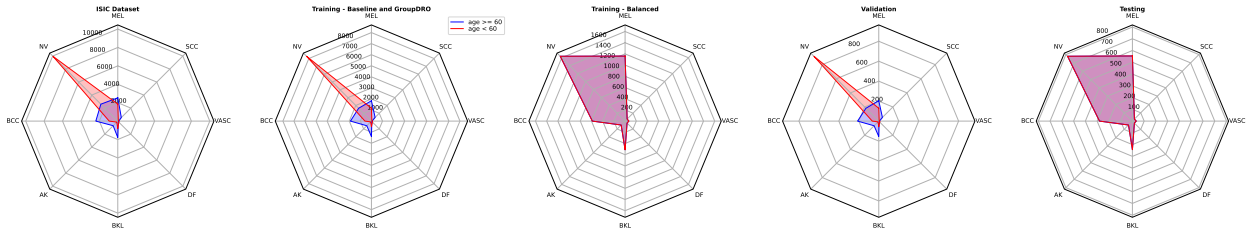


Figure 5.1: Number of images for each class and each subgroup for 5 different splits. (a) The ISIC dataset: From this, we can see a high-class imbalance across different classes. Similarly, distribution across both subgroups for a particular class is also different. For example, while melanoma (MEL), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), and squamous cell carcinoma (SCC), D^0 have a higher number of samples compared to D^1, for the rest of the classes (melanocytic nevus - NV, dermatofibroma - DF, and vascular lesion - VASC) D^1 (age < 60) has a higher number of samples compared to D^0 (age ≥ 60). (b) Training Set for the **Baseline-Model** and the **GroupDRO Model**: Similar to the ISIC dataset, we see high-class imbalance across different classes, and different distributions across both subgroups for a particular class. (c) The training set for the **Balanced-Model**: Compared to the training dataset used for the **Baseline-Model** and the **GroupDRO-Model**, we balance the number of samples across both subgroups, but we do not balance across different classes. (d) Validation set: The distribution of samples across both subgroups and across different classes is similar to the ISIC dataset. (e) Testing set: The distribution of samples across both subgroups is kept similar, but it is not similar across different classes. We kept similar distribution across both subgroups for a fair comparison of their performance, while the distribution across different classes was not kept similar to reflect real-world scenarios where some classes can be more frequent compared to others. ©[2023] PMLR. Reprinted, with permission, from [164].

5.3.1 Multi-class Skin Lesion Classification

Skin cancer is the most prevalent type of cancer in the United States [89], which can be diagnosed by classifying skin lesions into different classes. Due to the heterogeneity of skin cancer, classifying skin lesions into different classes can play an important role in disease diagnosis. As the heterogeneity and the risk of skin cancer can vary across different demographic differently [10], it is necessary to analyze the fairness of the machine learning model for skin cancer classification.

Dataset: We use the publicly available international skin imaging collaboration (ISIC) 2019 dataset [46] for multi-class skin lesion classification. A dataset of 24947 dermoscopic images is provided, with 8 associated disease scale labels, and with high class imbalance. Demographic patient information (e.g. age, sex) is provided. We consider age as the sensitive attribute (a_i). The entire dataset is divided into two subsets: patient images with age ≥ 60 in subgroup D^0 with a total of 10805 images, and patients with age < 60 in subgroup D^1 with a total of 14045 images. We chose age (and respective threshold) such that

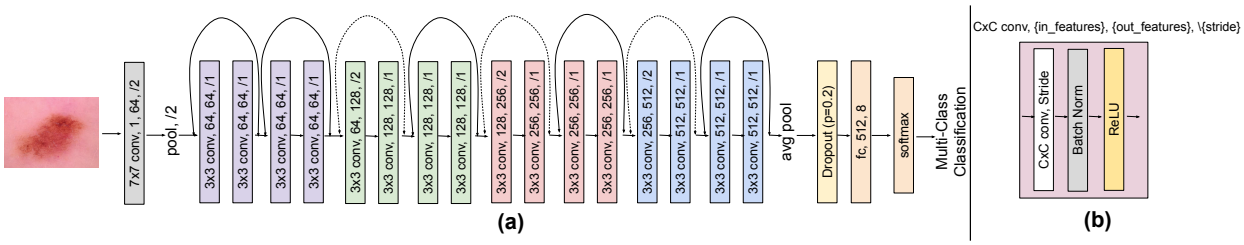


Figure 5.2: (a) A 2D ResNet-18 architecture consists of a 7x7 convolutional unit, followed by 16 3x3 convolutional units, one dropout layer (p=0.2), and one fully connected layers. The dotted shortcuts increase dimensions. (b) Each convolutional unit consists of one CxC convolutional layer with stride S, followed by Batch Normalization layer [108], and a ReLU layer. ©[2023] PMLR. Reprinted, with permission, from [164].

it led to large differences in images across subgroups and could potentially show a large fairness gap between different subgroups. This would permit the evaluation of the effects of different fairness mitigation methods on model performance and on uncertainty estimation. We did indeed run experiments with sex as a sensitive attribute, which showed similar results. To keep the experiment section brief and easy to follow, these results are included in Appendix D.1.

Figure 5.1 shows that the **Baseline-Model** and the **GroupDRO-Model** are trained on a training dataset where subgroup D^0 contains 8260 images, while subgroup D^1 contains 10892 images. While it appears that subgroup D^1 contains approximately 32% more images, it is not strictly the case for all eight classes. A **Balanced-Model** is trained on a training dataset where both subgroup D^0 and subgroup D^1 contain 7251 images. Both subgroups are balanced for each of the eight classes of the dataset (but not the same across the eight classes).

Implementation Details: An ImageNet pre-trained 2D ResNet18 [94] architecture was used for the ISIC 8-class disease scale classification task. The network architecture is depicted in Figure 5.2. A Dropout layer [241] with p=0.2 is introduced before the fully connected (fc) layer. The network was trained to reduce the categorical cross-entropy loss. An Adam optimizer [126] with a learning rate of 0.0005 and a weight decay of 0.00001 is used to train the network for a total of 100 epochs, and a batch size of 64. The learning rate is decayed with a factor of 0.995 after each epoch. All ISIC images are resized to

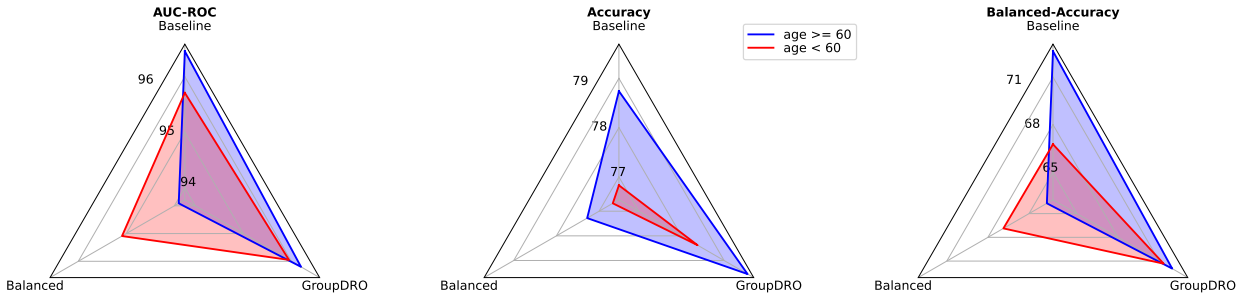


Figure 5.3: Overall AUC, Accuracy, and Balanced Accuracy for each subgroup (D^0 - age $>= 60$ and D^1 - age < 60) for all three models (**Baseline-Model**, **Balanced-Model**, and **GroupDRO-Model**). ©[2023] PMLR. Reprinted, with permission, from [164].

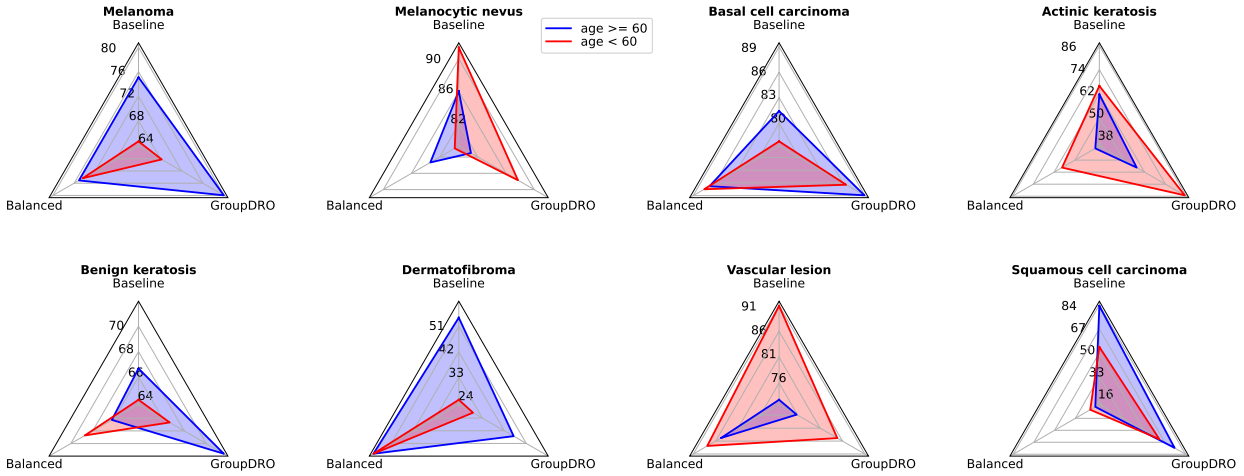


Figure 5.4: Classwise accuracy for each subgroup (D^0 - age $>= 60$ and D^1 - age < 60) for all three models (**Baseline-Model**, **Balanced-Model**, and **GroupDRO-Model**). ©[2023] PMLR. Reprinted, with permission, from [164].

600x450 size and normalized with mean subtraction and divided by std. Random Horizontal Flip, Random Vertical Flip, and Random rotation in the range of 0-30, are applied as data augmentation on each image. The code is written in PyTorch [191] and ran on Nvidia GeForce RTX 3090 GPU with 24GB memory. For generating EnsembleDropout [237], we train three different networks with different random initialization of network weights and take 20 MC-Dropout samples [79] from each. This results in a total of 60 Monte-Carlo samples for each image. The evaluation metrics (EM) are overall accuracy, overall macro-averaged AUC-ROC, and class-level accuracy.

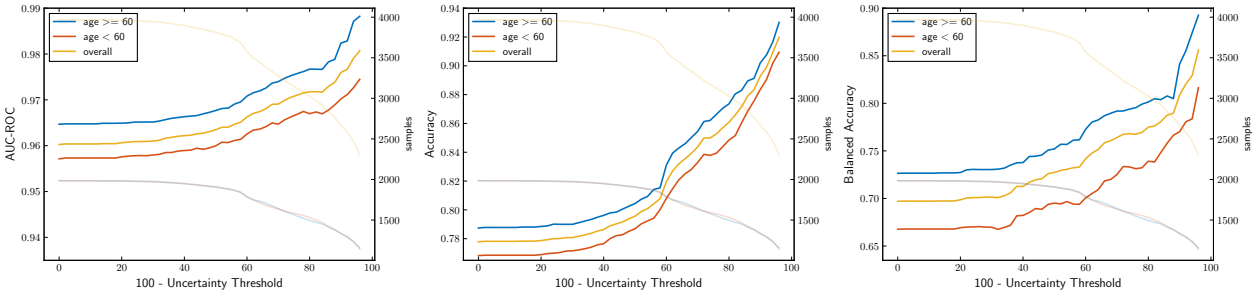
Results: First let us compare absolute performance and fairness gap without considering uncertainty filtering. Overall performance results are provided in Figure 5.3. From the figure, it can be observed that the **Balanced-Model** gives lower absolute fairness gap compared to the **Baseline-Model** for all three metrics (AUC-ROC, Accuracy, and Balanced Accuracy), but at the cost of lower overall performance. Compared to this, the **GroupDRO-Model** provides a lower fairness gap compared to the **Baseline-Model** with only a marginal decrease in absolute performance. This is clearly evident in both AUC-ROC plots and Balanced Accuracy plots. Braking performance down at a class level in Figure 5.4, we can see that neither the **Balanced-Model** nor the **GroupDRO-Model** consistently across eight classes provides a low fairness gap without a decrease in the absolute performance.

In Figure 5.5, the overall performance of all three models is plotted as a function of different uncertainty thresholds. From this, we can conclude that all three models show either an increase in the fairness gap or a similar fairness gap when more predictions are filtered based on uncertainty value (moving towards the right side of the curves). This shows that achieving better fairness in absolute performance can come at the cost of poor uncertainties.

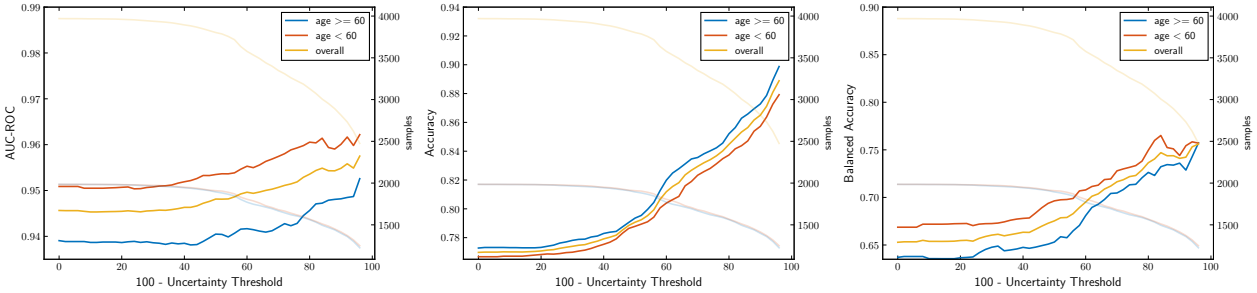
Next, we take look at the performance of these three models at a class level in Figure 5.6. For the **Baseline-Model**, almost all plots show a high fairness gap between the two subgroups when fewer predictions are filtered based on uncertainties (left side of the graph). When filtering more predictions (moving towards the right side of the curve), an increase in the accuracy for each subgroup and a reduction in the fairness gap can be observed for classes with a high number of total samples (ex. melanocytic nevus, basal cell carcinoma, and benign keratosis). This demonstrates that the model might be incorrect for more images in one of the subgroups, but it usually has higher uncertainty in those predictions compared to the other subgroup. For classes with a lower number of images (ex. dermatofibroma, vascular lesion, and squamous cell carcinoma), we do not see a similar

decrease in the fairness gap. For the **Balanced-Model**, we see a decrease in the fairness gap for absolute performance compared to the **Baseline-Model** for the majority of the classes, but it also comes at the cost of lower absolute performance. In contrast to the **Baseline-Model**, the fairness gap of the **Balanced-Model** increases with uncertainty filtering irrespective of the number of images in the classes. This behavior can be attributed to the overall less number of images used to train the **Balanced-Model** compared to the **Baseline-Model**. Figure 5.6(c) shows that the **GroupDRO-Model** might give better classwise accuracy compared to the **Baseline-Model** for classes with a high number of total samples (e.g., Melanoma, Basal cell carcinoma). But it also shows a high fairness gap when a low number of predictions are filtered (left side of the graph). The fairness gap reduces by filtering more predictions. However, it is not completely mitigated for all of the classes. Classwise accuracy for classes with a lower number of samples (ex. Dermatofibroma) sees a marginal increase in the fairness gap with uncertainty-based filtering. Results indicate that the **GroupDRO-Model** might give marginally better absolute performance than the **Baseline-Model**, but it does not produce fair uncertainty estimates across subgroups.

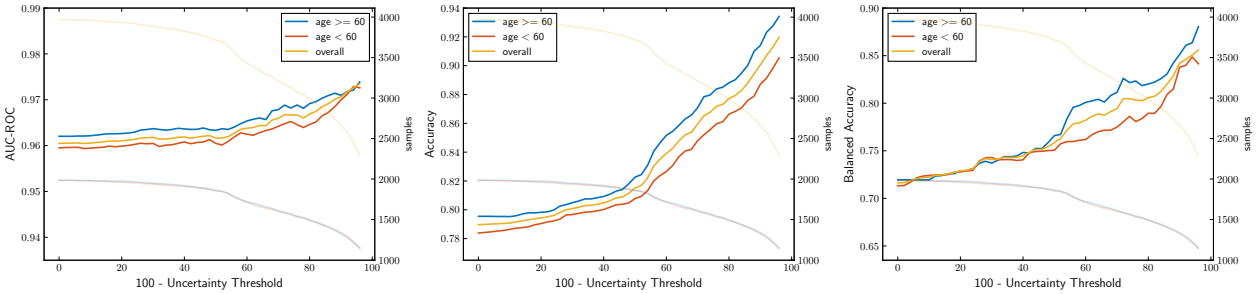
The discrepancy in performance between different models shows that performance is highly dependent not only on the task at hand, but also on the number of images in different classes. Similarly, it can be concluded that different models do not behave consistently across different classes, both in terms of fairness gap and uncertainty evaluation. It indicates that a single model cannot reduce the fairness gap and also provide good uncertainty estimation.



(a) Baseline Model

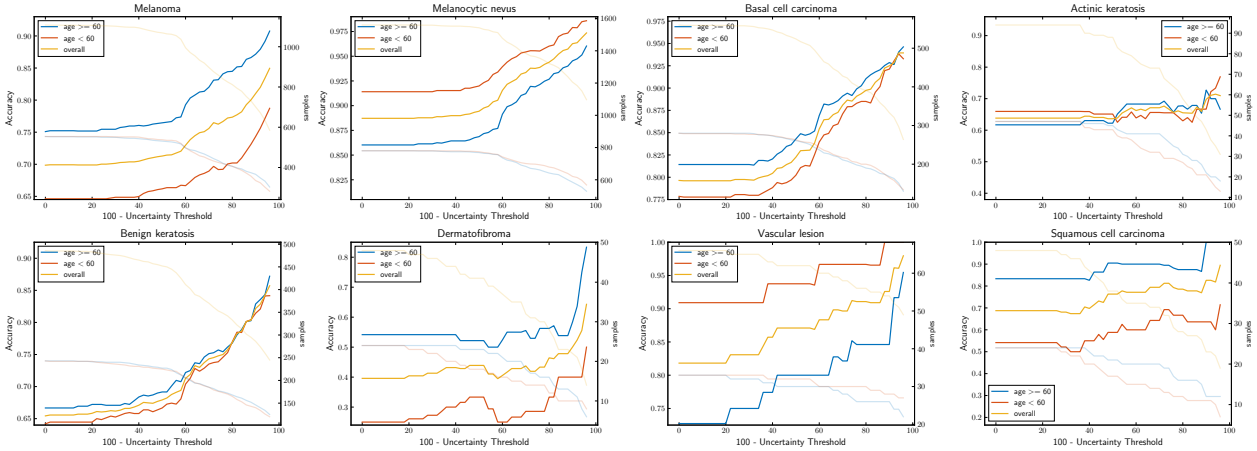


(b) Balanced Model

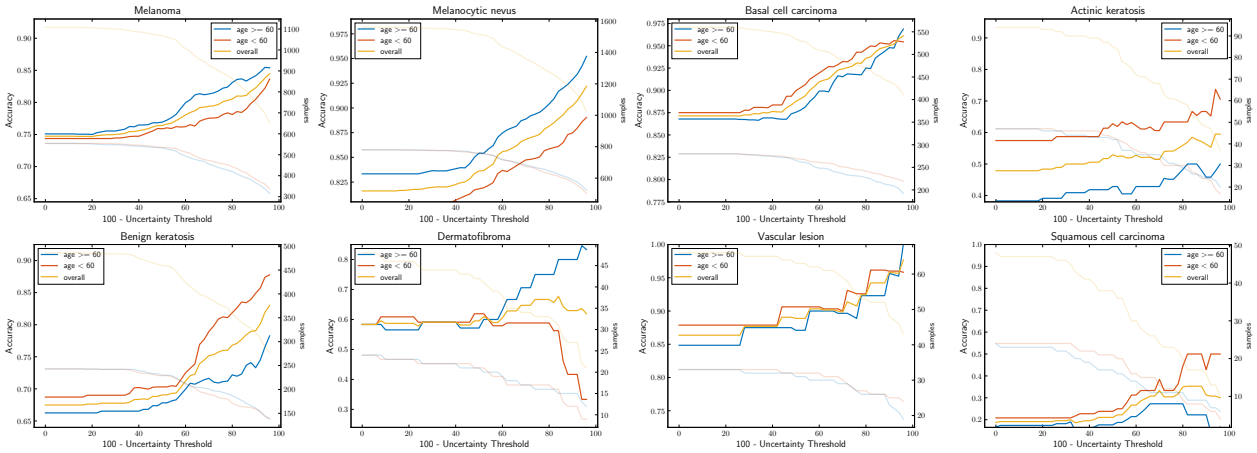


(c) GroupDRO Model

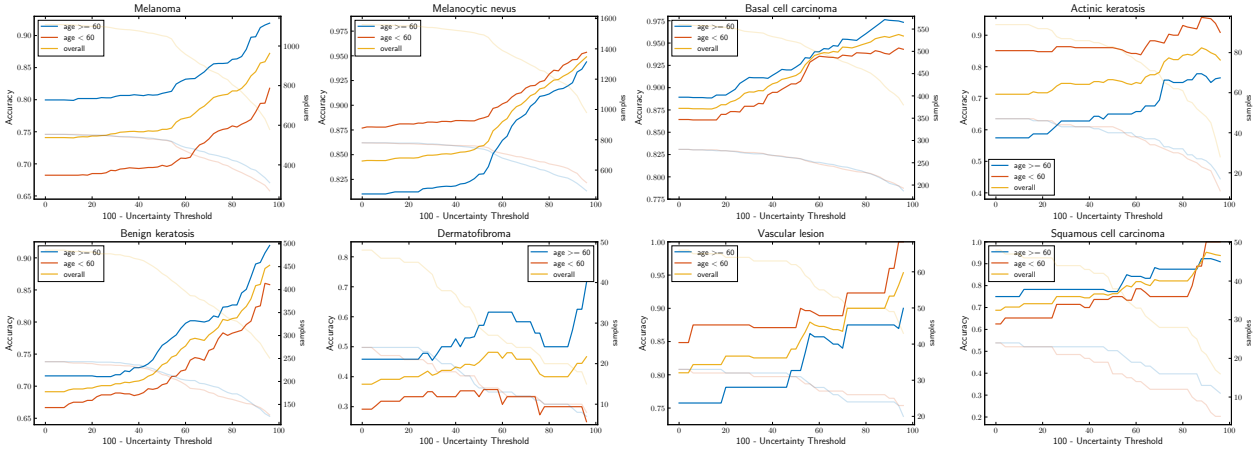
Figure 5.5: **ISIC:** Overall AUC, accuracy, and balanced accuracy (left y-axis) as a function of uncertainty threshold (x-axis) for (a) **Baseline-Model**, (b) **Balanced-Model**, and (c) **GroupDRO-Model** on the ISIC dataset. In addition to metrics, the total number of testing images for each subgroup (D^0 - age $>= 60$ and D^1 - age < 60) are shown as light colours (see y-axis labels on the right). ©[2023] PMLR. Reprinted, with permission, from [164].



(a) Baseline Model



(b) Balanced Model



(c) GroupDRO Model

Figure 5.6: **ISIC:** Class-level accuracy as a function of uncertainty threshold for (a) **Baseline-Model**, (b) **Balanced-Model**, and (c) **GroupDRO-Model** on the ISIC dataset. In addition to the accuracy, the total number of testing images for each subgroup (D^0 - age $>= 60$ and D^1 - age < 60) are shown as light colours (see axis labels on the right). ©[2023] PMLR. Reprinted, with permission, from [164].

Table 5.1: Number of samples in both D^0 and D^1 subgroups for five different splits: (i) Training Dataset used to train the **Baseline-Model** and the **GroupDRO-Model**, (ii) Training Dataset used to the train the **Balanced-Model**, (iii) Validation set for all three models, (iv) Testing set for all three models, and (v) for the whole BraTS dataset. We can observe that for the BraTS dataset, there is a high disparity between the number of samples for both subgroups. ©[2023] PMLR. Reprinted, with permission, from [164].

	Training Set		Validation Set	Testing Set	BraTS Dataset
	Baseline-Model and GroupDRO-Model	Balanced Model			
D^0	168	30	18	20	206
D^1	30	30	4	20	54
Overall	198	60	22	40	260

5.3.2 Brain Tumour Segmentation

In the previous section, we analyzed the performance of different fairness mitigation methods for the image-level classification task. In this section, the same methods are evaluated for the voxel segmentation task.

Dataset: We use the 260 high-grade glioma (HGG) images from the publicly available Brain Tumour Segmentation (BraTS) 2019 challenge dataset [22]. The BraTS dataset consists of MR images from two disease stages: high-grade glioma - HGG (260 images) and low-grade glioma - LGG (75 images). In this work, only HGG samples are considered, as the appearance of tumours across both disease stages is different and lower performance on LGG cases has been reported in the literature when a single model is trained on both HGG and LGG samples [198]. The choice for how to split the dataset is based on finding a subgroup where a performance gap is clearly present based on the provided metrics. There can be a number of such subgroups. We initially ran experiments whereby the dataset was split based on imaging centers (i.e. binary subgroups: TCIA vs non-TCIA). Our results, included in the Appendix-D.2.1, indicated that there is no bias across the resulting groups. It is well established that there is a significant bias in the BraTS dataset, whereby the performance of small tumour segmentation is significantly worse than that of large tumour segmentation [183]. This is an important bias to overcome. The image dataset is therefore divided into two subsets based on the volume of the enhancing tumour: 206 images with volumes $> 7000\text{ml}^3$ in subgroup D^0 and 54 images with volumes $\leq 7000\text{ml}^3$ in subgroup D^1. **Baseline-Model** and **GroupDRO-Model** are trained on a dataset of 168 samples from D^0 and 30 samples from D^1. While a **Balanced-Model** is

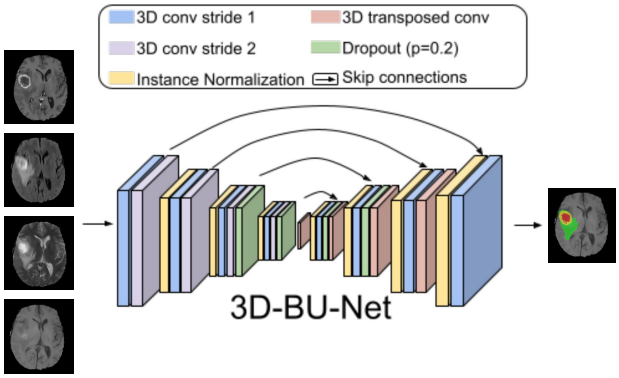


Figure 5.7: Network architecture diagram of the modified 3D-BU-Net [180], used for the multi-class brain tumour segmentation. It takes multi-modal MR images as input and produces multi-class brain tumour segmentation on the BraTS dataset. ©[2023] PMLR. Reprinted, with permission, from [164].

trained on a balanced training set with 30 samples from each subgroup.

Implementation Details: A 3D BU-Net [180] architecture is used for brain tumour segmentation on the BraTS dataset. Similar to the original 3D BU-Net, the network consists of encoder and decoder paths that embed convolution and deconvolution operations. High-resolution features from the encoder path are combined with the up-sampled output of the decoder to preserve high-resolution features. Each convolution is followed by rectified linear unit activation (ReLU). Instead of using the batch normalization layer used in the original U-Net, we use instance normalization [256]. Instance normalization typically improves performance for small batch sizes. The network is trained using Adam [126] optimizer with a learning rate of 0.0002 and weight decay of 0.00001 for a total of 240 epochs to minimize weighted cross-entropy loss. Here, the weights are defined such that the weight increases whenever there are fewer voxels in a particular class. After every epoch, class weights are decayed with a factor of 0.95, which results in equally weighted binary cross-entropy after around 50 epochs. The code is written in PyTorch [191] and ran on Nvidia GeForce RTX 3090 GPU with 24GB memory. For generating EnsembleDropout [237], we train three different networks with different random initialization of network weights and take 20 MC-Dropout samples [79] from each. This results in a total of 60 Monte-Carlo samples for each image.

Following the BraTS dataset convention, tumour segmentation performance is evaluated

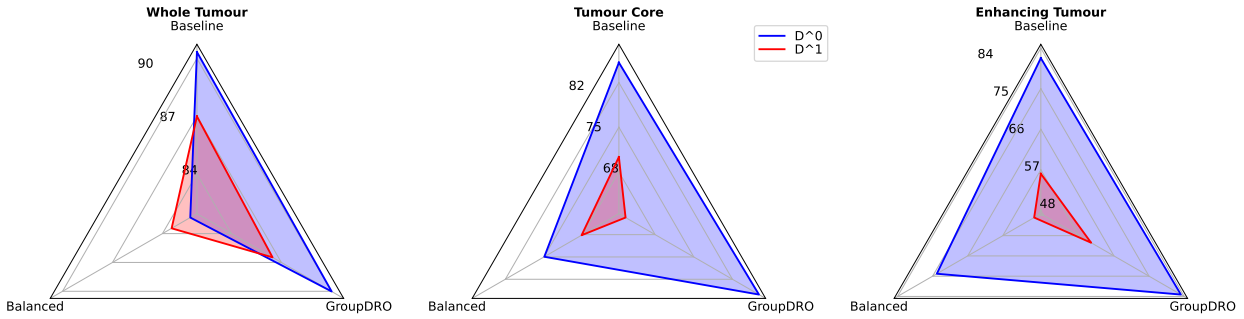


Figure 5.8: Dice results for whole tumour (WT), tumour core (TC), and enhancing tumour (ET). All three tumour Dice values are plotted for both the subgroups (D^0 and D^1) and for all three models (**Baseline-Model**, **Balanced-Model**, and **GroupDRO-Model**). ©[2023] PMLR. Reprinted, with permission, from [164].

by calculating Dice scores for three different tumour sub-types: enhancing tumor, whole tumor, and tumour core. The predictions' uncertainty is measured through the entropy of an Ensemble Dropout model [237].

Results: Figure 5.8 compares absolute performance and fairness gap for all three models without considering the uncertainty quantification. From this figure, it can be observed that similar to the classification experiments, the **Balanced-Model** provides a reduced fairness gap between two subgroups at the cost of a decrease in overall performance for all three tumour subtypes. Unlike, the classification experiments, the **GroupDRO-Model** doesn't lead to a reduced fairness gap compared to the **Baseline-Model**. In fact, for Tumour Core, it leads to an increase in the fairness gap between two subgroups.

Figure 5.10 provides the behaviour of all three models with uncertainty-based filtering. It shows that both the **Baseline-Model** and the **GroupDRO-Model** perform similarly for whole tumour (WT) across both subgroups, as an increase in Dice and decrease in the fairness gap is observed with filtering of more voxels in the images (going from left to right in the graph). For the **Balanced-Model** though initially (left most at an uncertainty threshold of 100) the fairness gap is lower compared to the other two models, it increases with the filtering of more voxels in the images. Tumour core (TC) and enhancing tumour (ET) follow a similar trend, where both the **Baseline-Model** and the **GroupDRO-Model** perform

similarly. Although for both TC and ET, the **Balanced-Model** doesn't show an increase in the fairness gap between the two subgroups with a decrease in uncertainty threshold (moving from left to right), a decrease in overall performance for both subgroups is observed. This shows that mitigating the fairness gap by filtering out more voxels is insufficient and may lead to a drop in performance in both subgroups. It can be concluded that for a challenging dataset like brain tumour segmentation, the **Balanced-Model** or the **GroupDRO-Model** do not produce fair uncertainty estimates across different subgroups.

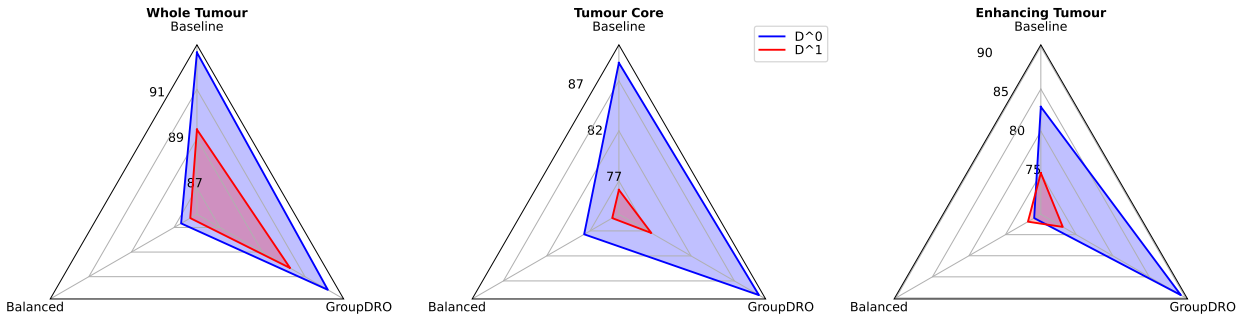
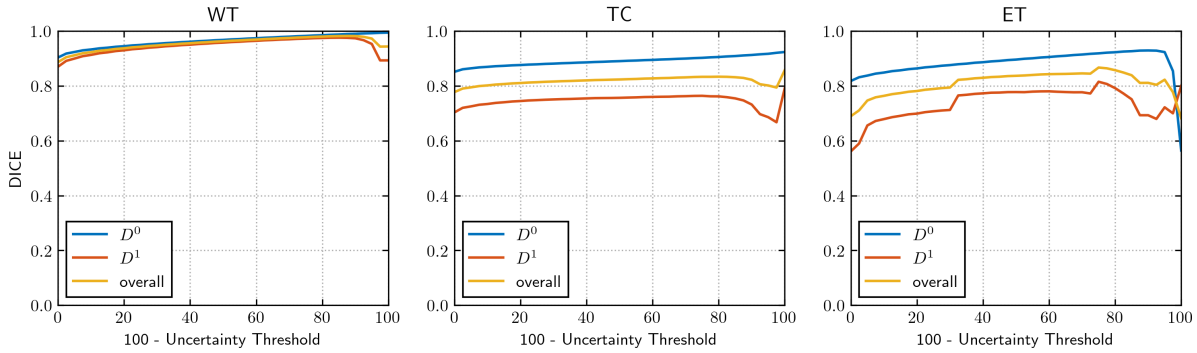
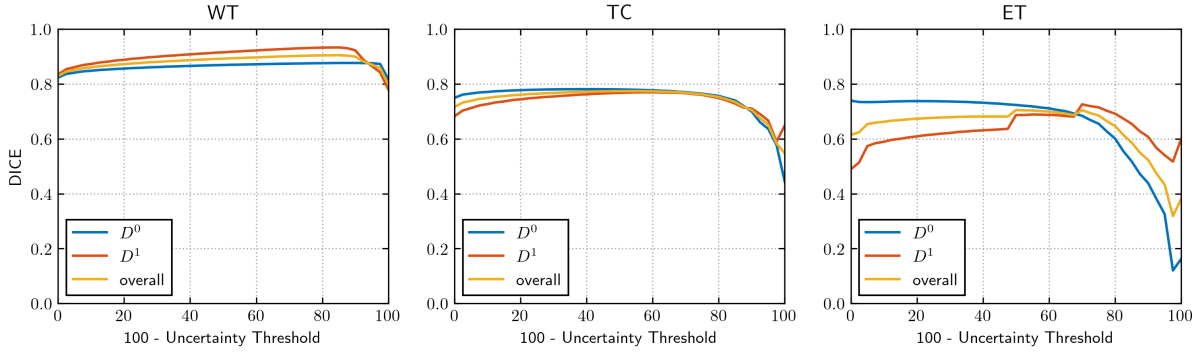


Figure 5.9: QUBraTS metric [162] for whole tumour (WT), tumour core (TC), and enhancing tumour (ET), for both the D^0 and D^1, and for all three models (**Baseline-Model**, **Balanced-Model**, and **GroupDRO-Model**). ©[2023] PMLR. Reprinted, with permission, from [164].

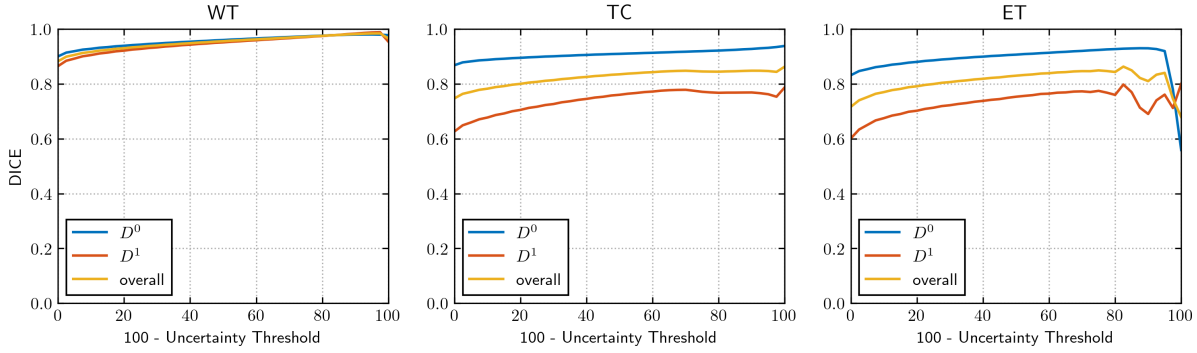
As we argued in our previous Chapter 3 that when the goal is to develop a Computer-Aided diagnosis (CAD) system for the brain tumour segmentation task, it is not sufficient to only increase the Dice values with uncertainty-based voxel filtering. In real practice, an additional penalty should be provided to a model that filters out many voxels at a low uncertainty threshold to achieve high Dice values, as it will increase the reviewing burden on clinicians. Results on our metric (Equation 3.1) is provided in Figure 5.9. From this, we can observe that similar to the Dice performance, the **Balanced-Model** provides a reduced fairness gap at the cost of reduced absolute performance, and the **GroupDRO-Model** provides a marginally greater fairness gap compared to the **Baseline-Model**. These results are consistent with Figure 5.10. Further plots for all three components of QU-BraTS metrics are provided in the Appendix D for all three tumour types and models.



(a) Baseline Model



(b) Balanced Model



(c) GroupDRO Model

Figure 5.10: Averaged sample Dice as a function of (100 - uncertainty threshold) for (a) **Baseline-Model**, (b) **Balanced-Model**, and (c) **GroupDRO-Model** on the BraTS dataset. Dice results for whole tumour (WT), tumour core (TC), and enhancing tumour (ET), for both the D^0 and D^1, set are shown in each column. ©[2023] PMLR. Reprinted, with permission, from [164].

5.3.3 Alzheimer's Disease Clinical Score Regression

As the prevalence of Alzheimer's disease (AD) varies across demography [14], in this section, we specifically analyze the effect of different fairness mitigation methods for the task of clinical assessment score regression from brain MR images.

Dataset: Experiments are based on the MRIs of a subset (865 patients) of the Alzheimer's disease neuroimaging initiative (ADNI) dataset [111] at different stages of diagnosis: Alzheimer's disease (145), mild cognitive impairment (498), and cognitive normal (222). The dataset also provides demographic patient information such as age and gender. Here, we consider age as a sensitive attribute (a_i). As can be seen from Figure 5.11, the dataset is divided such that patients with age < 70 are grouped into D^0 (259 patient images), and patients with age ≥ 70 are grouped into D^1 (606 patient images). Based on the clinical relevance and different prevalence of AD across patients with different ages [14], we chose the threshold for the sensitive attribute. This also showed a clear performance gap between these subgroups. A **Baseline-Model** and a **GroupDRO-Model** are trained on a dataset that contains 163 samples from D^0 and 440 samples from D^1. While a **Balanced-Model** is trained on a balanced training set with 163 samples from each subgroup.

Implementation Details: A 3D ResNet34 [91] architecture is designed for the task of clinical score prediction [1]. The network is modified to be a multi-task network, such that it predicts both ADAS-13 and MMSE scores simultaneously. The network is trained to reduce the combined mean squared error losses for both ADAS-13 and MMSE. An Adam optimizer with a learning rate of 0.0002 and a weight decay of 0.00001 is used to train the network for a total of 200 epochs. The learning rate is decayed with a factor of 0.995 after each epoch. The code is written in PyTorch [191] and ran on Nvidia GeForce RTX 3090 GPU with 24GB memory. For generating EnsembleDropout [237], we train three different networks with different random initialization of network weights and take 20 MC-Dropout samples [79] from each. This results in a total of 60 Monte-Carlo samples

[1] https://github.com/kenshohara/3D-ResNets-PyTorch/blob/master/models/resnet.py

Figure 5.11: Number of images for each disease stage (AD, MCI, and CN) and each subgroup for five different sets. (a) ADNI dataset: We can see a high disparity between the total number of samples in each disease stage. Similarly, distribution across subgroups for a particular disease stage is also different. (b) Training Set - **Baseline Model** and **GroupDRO Model**: Similar to the ADNI dataset, a high disparity between the total number of samples in each disease stage is visible. Similarly, distribution across subgroups for a particular disease stage is also different. (c) Training Set - **Balanced Model**: Compared to the training dataset used for the **Baseline-Model** and the **GroupDRO-Model**, we balance the number of samples across both subgroups for each disease stage, but not across disease stages. (d) Validation Set: The distribution of samples across both subgroups and across different disease stages is similar to the ADNI dataset, (e) Testing Set: The distribution of samples across both subgroups is kept similar, but it is not similar across different disease stages. We kept similar distribution across both subgroups for a fair comparison of their performance, while the distribution across different disease stages was not kept similar to reflect real-world scenarios where some disease stages can occur more frequently compared to others. ©[2023] PMLR. Reprinted, with permission, from [164].

for each image. A combination of Sample Variance and Predicted Variance, known as total variance [121], is used to measure uncertainty associated with the model output. We choose total variance as an uncertainty measure as it is computationally more feasible compared to the entropy for the regression task, and similar to the entropy, it also measures both aleatoric and epistemic uncertainties.

Root mean squared error (RMSE) is used as an evaluation metric (EM), where a lower value of RMSE represents better performance. As the total number of images is low in this dataset, we run the same experiments on five different folds and aggregate their results.

Results: Figure 5.13 shows a reduced fairness gap and better overall performance (lower RMSE) for the **GroupDRO-Model** compared to the **Baseline-Model** for both ADAS-13 and MMSE. In contrast to this, the **Balanced-Model** shows only marginal improvement in the fairness gap at the expense of poor overall performance (high RMSE) compared to the **Baseline-Model** for both ADAS-13 and MMSE. This trend is consistent with what we observed for the classification and the segmentation experiments where the **GroupDRO-**

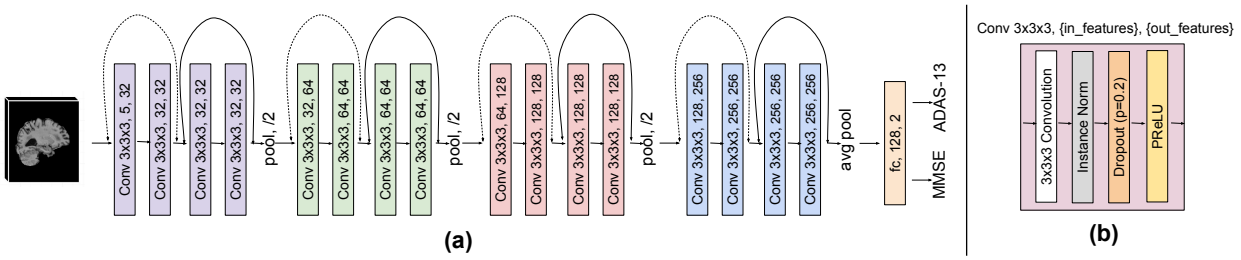


Figure 5.12: Network architecture diagram of modified 3D-ResNet-18 [91] for the Alzheimer's Disease clinical regression pipeline for predicting both ADAS-13 and MMSE scores. The network takes 3D T1-weighted MR image as input. ©[2023] PMLR. Reprinted, with permission, from [164].

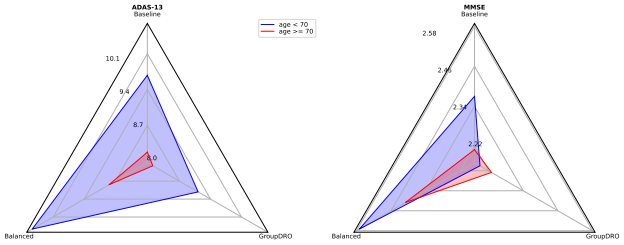


Figure 5.13: Overall Root Mean Squared Error (RMSE) of ADAS-13 (Left) and MMSE (Right) score prediction tasks for each subgroup (D^0 - age $<$ 70 and D^1 - age $\geq$ 70) for all three models (**Baseline-Model**, **Balanced-Model**, and **GroupDRO-Model**). ©[2023] PMLR. Reprinted, with permission, from [164].

Model performed better than the **Balanced-Model** for overall metrics. Further breaking down the performance at a disease type level in Figure 5.14, inconsistency of both models for both ADAS-13 and MMSE can be observed. For example, while the **GroupDRO-Model** decreased the fairness gap and reduced RMSE for ADAS-13 of CN patients, it is inverse for MMSE, where it increases the fairness gap with increased RMSE. Similar observation can be made for the **Balanced-Model** and AD patients.

Figure 5.15 column 1 shows that compared to the **Baseline-Model**, the **Balanced-Model** only marginally decreases the fairness gap in the initial performance between two subgroups, that too at the cost of poor (higher RMSE) absolute performance for each of the subgroups. The **GroupDRO-Model** shows better absolute performance (lower RMSE) and also a lower fairness gap between each subgroup compared to either of the other two models. The **Baseline-Model** shows a decrease in the fairness gap between subgroups with a decrease in uncertainty threshold (moving from left to right) for MMSE, but it is

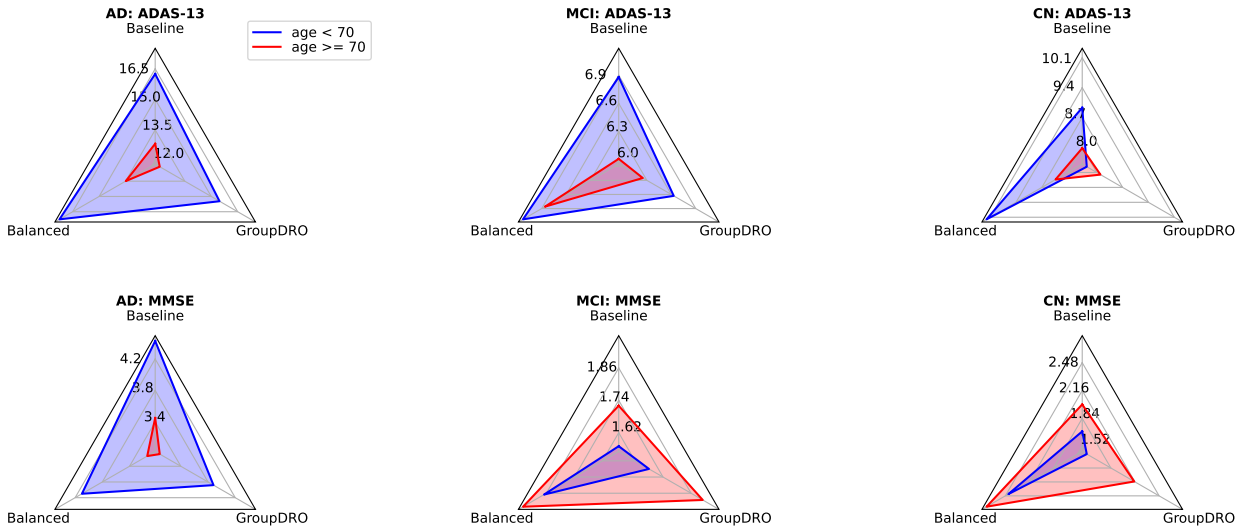


Figure 5.14: Per disease stage (AD, MCI, and CN) Root Mean Squared Error (RMSE) of ADAS-13 (Top) and MMSE (Bottom) score prediction tasks for each subgroup (D^0 - age $<$ 70 and D^1 - age $\geq$ 70) for all three models (**Baseline-Model**, **Balanced-Model**, and **GroupDRO-Model**). ©[2023] PMLR. Reprinted, with permission, from [164].

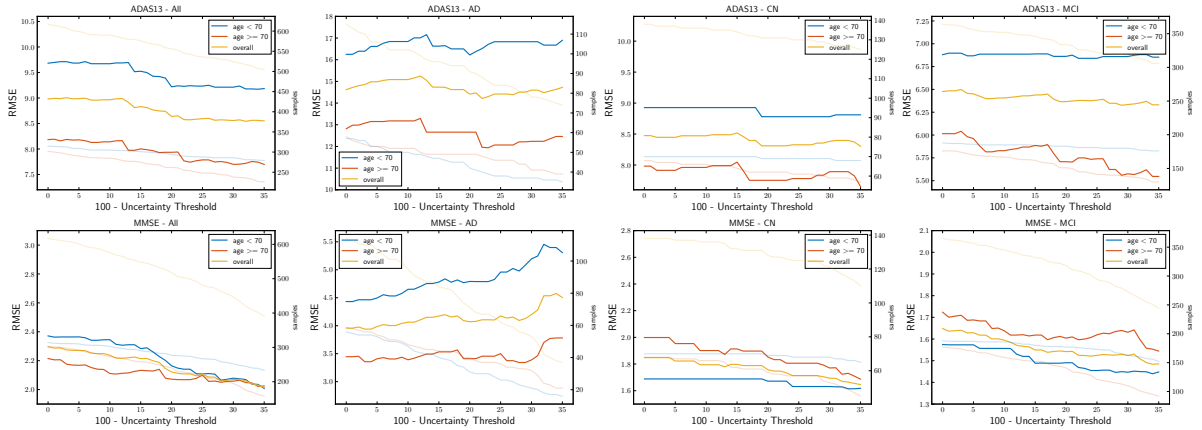
not true for ADAS-13. On the contrary, the **Balanced-Model** shows an increase in the fairness gap with a decreased uncertainty threshold for both ADAS-13 and MMSE. The **GroupDRO-Model** gives the best performance as the fairness gap decreases with a decrease in uncertainty threshold for both ADAS-13 and MMSE.

Further breaking down the performance for each of the different disease types in Figure 5.15 column 2 - column 4, shows that each model shows different performance across different disease types. For example, while at a dataset level, the **GroupDRO-Model** shows both good absolute performance, lower fairness gap, and decrease in fairness gap without an increase RMSE; it doesn't hold true for images belonging to AD patients (column 2), as both fairness gap and RMSE increases with decrease in uncertainty threshold for MMSE. Similarly, we see that the **Balanced-Model** shows a decrease in RMSE for MMSE with a decrease in uncertainty threshold (moving from left to right), which is in contrast to its behaviour for all images and AD images. This further reinforces our observation from skin lesion classification experiments that the performance of the fairness mitigation methods not only varies across subgroups and across different tasks, but also

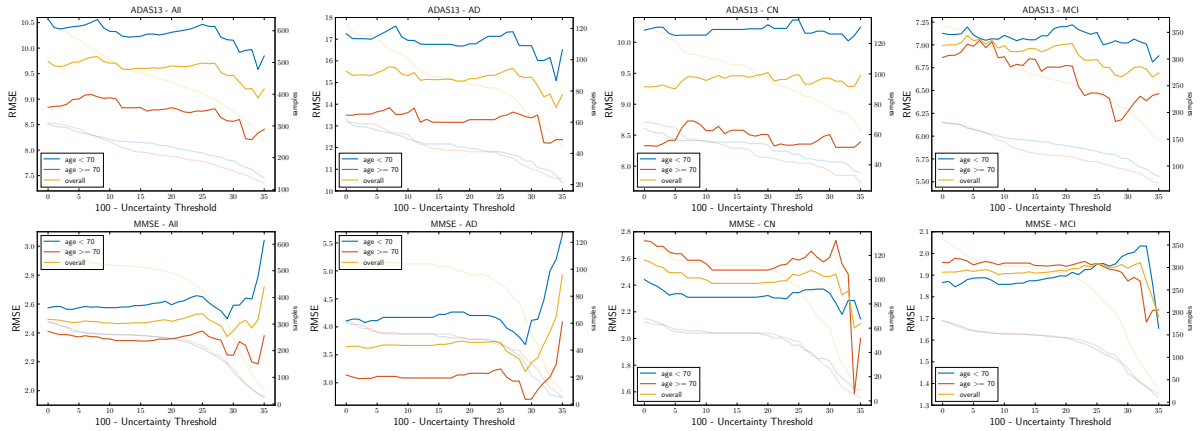
changes for different classes for the same task. And these types of models should be chosen with care when applying them to real-world medical image analysis applications.

5.4 Summary

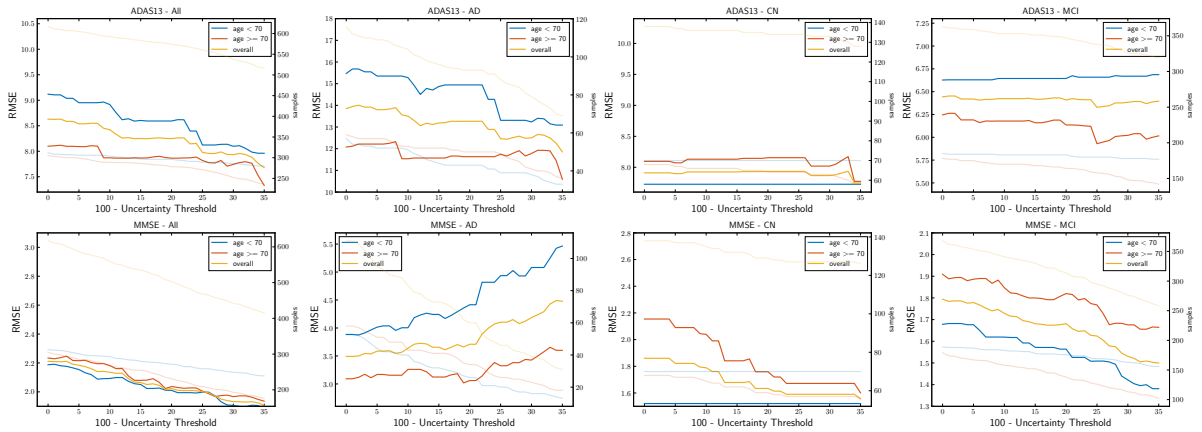
Accurate uncertainty estimation of deep learning predictions in medical image analysis is necessary for their safe clinical deployment. In this chapter, we presented the first exploration of fairness models in mitigating biases across subgroups, and their resulting effect on uncertainty quantification accuracy. Results on a wide range of experiments for three different tasks (classification, regression, and segmentation) indicate that popular fairness methods, such as data balancing and robust optimization, do not always work for different tasks. Furthermore, mitigating fairness in terms of performance can come at the cost of poor uncertainty estimation associated with output. Future work on overcoming these additional fairness issues is required prior to the clinical deployment of these models.



(a) Baseline Model



(b) Balanced Model



(c) GroupDRO Model

Figure 5.15: **ADNI:** Root mean squared error (RMSE) of ADAS-13 (Top) and MMSE (Bottom) score prediction tasks as a function of uncertainty threshold for (a) **Baseline-Model**, (b) **Balanced-Model**, and (c) **GroupDRO-Model** on the ADNI dataset. Specifically, we plot RMSE for all samples as well as samples for each of the disease stages (AD, MCI, and CN) in each subgroup (D^0 - age < 70 and D^1 - age ≥ 70). The total number of samples as a function of uncertainty thresholds are depicted with light colours in these plots (see the scale on the right vertical axis). ©[2023] PMLR. Reprinted, with permission, from [164].

6

Information Gain Sampling for Active Learning in Medical Image Classification

When stupid ideas work, they become genius ideas.

— *Andy Weir*, Project Hail Mary

Related Paper

It should be noted that this is not a manuscript based thesis. However, considerable material from the following paper has been utilised in this chapter.

- **R. Mehta**, C. Shui, B. Nichyporuk, T. Arbel, "Information Gain Sampling for Active Learning in Medical Image Classification", *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE) Workshop held in conjunction with 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2022* [165].

The Springer does not require individuals working on a thesis to obtain a formal reuse license. However, it requires that the thesis author cite the source and include Springer copyright notice for all figures and tables [4].

6.1 Introduction

The previous chapter of this thesis built towards building trustworthy medical image analysis models by analyzing the model performance for both fairness across demography and the quantification of uncertainty associated with them. While trustworthy models can indeed help in the clinical deployment of these models, they do not mitigate the need for a large amount of data required to build these models. In this chapter, we tackle the necessity of a large labeled dataset by building an active learning framework that selects the optimal images to label from an unlabeled pool of images based on information gain. Related work for active learning is provided in Chapter 2.

This part of the thesis proposes an information-theoretic active learning framework that drives the selection of new image samples to label based on maximal information gain. An active learning framework that selects samples based on expected information gain (EIG) has been previously used [209] for structure prediction tasks using Support Vector Machines (SVM). As the first contribution of this part of the thesis, we first adapt an ef-

ficient EIG computation to deep networks with careful design choices. To alleviate the high-class imbalance issue in medical imaging, we further improve the original EIG by proposing a novel adapted expected information gain (AEIG) method. In AEIG, the predicted softmax probability of the trained model is adjusted with the class frequencies of the validation distribution. The hypothesis is that AEIG based sampling strategy will lead to higher performance with a lower number of labeled samples, as different labeled samples provide different information about inter-class ambiguity.

Experiments are performed on two different challenging medical image classification tasks: (1) multi-class diabetic retinopathy (DR) classification into disease scales from colour fundus images, (2) multi-class skin lesion classification from dermoscopic images. Our experiments indicate that for the DR dataset, AEIG achieves 95% of overall performance with only 19% of the training data. In comparison, other active learning methods require around 25% (random: 27%, maximum entropy: 21%, CoreSet: 27%, MCD-Entropy: 24%, MCD-BALD: 21%). AEIG achieves higher performance than competing methods due to its ability to sample more images from the minority classes compared to other methods on highly imbalanced datasets.

6.2 Active Learning Framework with Information Gain Sampling

Consider a labeled training dataset $D^L : \{(X_L, Y_L)\}$. Here, $(X_L, Y_L) = \{(x_i, y_i = c)\}_{i=1}^{M}$, represents that there are a total of M samples (x_i) in the dataset; and $y_i = c$ represents its corresponding classification label, where there are a total of C classes ($c \in \{0, 1, ..., C-1\}$). Now, consider an unlabeled dataset $D^U : \{(X_U)\}$, with N samples. Similarly, an evaluation dataset $D^{\text{eval}} : \{(X_{\text{eval}}, Y_{\text{eval}})\}$ with K samples. Here, X^{eval} represents the set of all samples in the evaluation set, and Y^{eval} its corresponding labels. $\hat{Y}^{\text{eval}}$ would represent the predicted classification label for each sample in the evaluation set using a machine learning model. Note that $M \ll N$ and $K < N$.

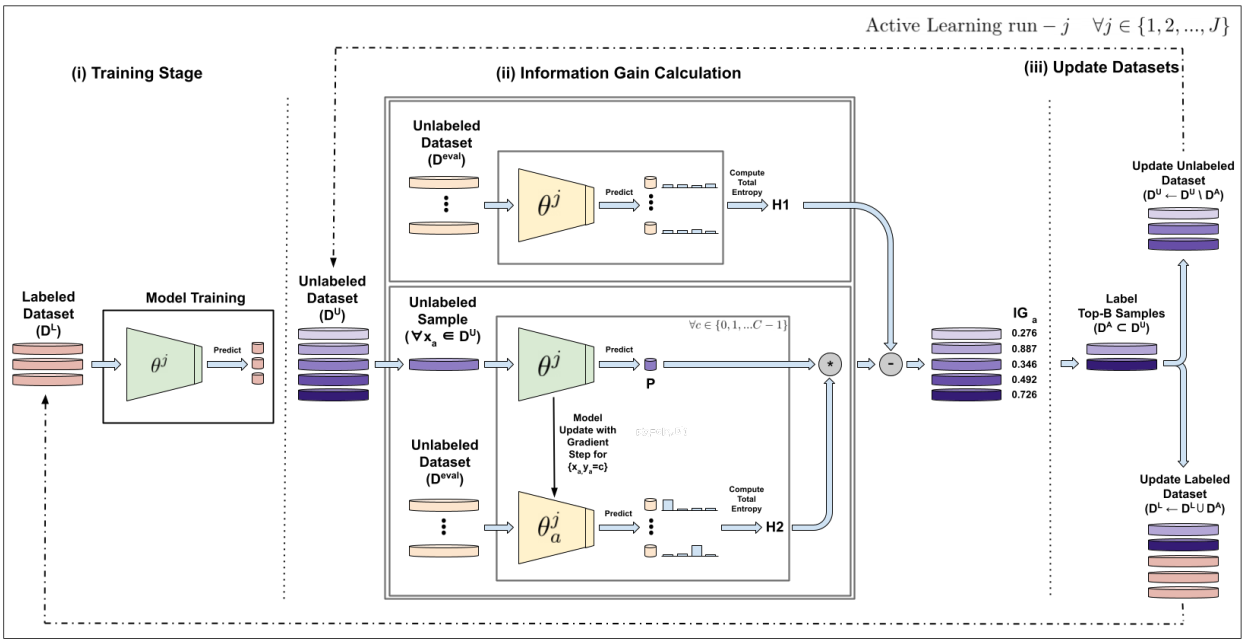


Figure 6.1: Active learning via information gain framework. Each active learning run consists of three different phases: **(i) Training Stage** - Model ($\theta^{j-1} \rightarrow \theta^j$) is trained using the labeled set D^L, **(ii) Information gain calculation** - AEIG$_a$ (Equation(6.2)), EIG$_a$ (Equation(6.1)), or its variants are calculated for each image in the unlabeled dataset ($\forall x_a \in D^U$). The entropy H1 of the evaluation set (D^{eval}) is calculated using the trained model (θ^j). For each image x_a, The conditional entropy (H2) of the evaluation set is calculated after updating the trained model (θ^j) using a single gradient step ($\theta^j \rightarrow \theta^j_a$) for all possible labels $y_a = c$, $\forall c \in \{0, 1, ..., C-1\}$. **(iii) Update Datasets** - Finally, the top-B images (D^A) from the unlabeled set are selected, and both the labeled ($D^L \leftarrow D^L \cup D^A$) and unlabeled datasets ($D^U \leftarrow D^U \setminus D^A$) are updated. The framework is executed for a total of J runs. ©[2022] Springer. Reprinted, with permission, from [165].

The general active learning framework starts by training a supervised machine learning model (θ^0) on a small labeled dataset (D^L). It then selects the B most informative subset of images to label ($D^A : \{x_a\}_{a=0}^B$, $D^A \subset D^U$) from a larger unlabeled dataset (D^U). A human annotator provides the labels for the selected subset of data ($D^{A*} : \{(X_A, Y_A\} = \{x_a, y_a\}_{a=0}^B$). Both the labeled ($D^L \leftarrow D^L \cup D^{A*}$) and the unlabeled datasets ($D^U \leftarrow D^U \setminus D^A$) are then updated. The model is retrained using the updated labeled dataset ($\theta^0 \rightarrow \theta^1$). The process is repeated for a total of J runs.

6.2.1 Information Gain (IG) for Active Learning

An active learning framework can select the subset of data from the unlabeled set based on the information gain.

Expected Information Gain (EIG): Let us consider the case of expected information gain (EIG). In an active learning context, EIG($\hat{Y}^{\text{eval}}; y_a$) measure the reduction in the entropy of the predicted labels $\hat{Y}^{\text{eval}}$ of the evaluation set, if we have access to the true state (label - y_a) of an image (x_a) in the unlabeled set. In short, EIG($\hat{Y}^{\text{eval}}; y_a$) measures difference in the entropy of $\hat{Y}^{\text{eval}}$ for two models. (i) **H1**: the entropy of the $\hat{Y}^{\text{eval}}$ for a model trained on D^L. (ii) **H2**: the conditional entropy of the $\hat{Y}^{\text{eval}}$ for a model trained on $\{D^L \cup (x_a, y_a)\}$.

$$
\begin{aligned}
&\text{EIG}(\hat{Y}^{\text{eval}}; y_a) = \text{EIG}(\hat{Y}^{\text{eval}}; y_a | X^{\text{eval}}, x_a, D^L) \\
&= \mathbf{H}[\hat{Y}^{\text{eval}} | X^{\text{eval}}, D^L] - \mathbf{H}[\hat{Y}^{\text{eval}} | X^{\text{eval}}, y_a, x_a, D^L] \\
&= \underbrace{\mathbf{H}[\hat{Y}^{\text{eval}} | X^{\text{eval}}, D^L]}_{\textbf{H1}} - \sum_{c=0}^{C-1} \underbrace{p(y_a = c | x_a, D^L)}_{\textbf{P}} \underbrace{\mathbf{H}[\hat{Y}^{\text{eval}} | X^{\text{eval}}, y_a = c, x_a, D^L]}_{\textbf{H2}}
\end{aligned}
\tag{6.1}
$$

$\mathbf{P} = p(y_a = c | x_a, D^L)$ denotes the predicted softmax probability of output having class label $y_a = c$ for an image x_a using a model trained on D^L.

Adapted Expected Information Gain (AEIG): The predicted softmax probability **P** can be quite erroneous due to the limited observations and poor calibration [87]. Thus, other alternatives can be considered to improve the reliability of **P**, such as injecting prior information about the class distribution. In the natural image classification literature, several methods [212, 252, 269] have been proposed that adapt the softmax probabilities in the context of highly imbalanced datasets. As such, a variant of the EIG method is considered here, where the predicted softmax probability (**P**) of the training model is adjusted with the class frequencies of the validation distribution. The adapted version of EIG, denoted adapted expected information gain (AEIG), provides a modification for **P** to be-

come $\mathrm{p}(y_a = c|x_a, D^L) * \frac{|y_{\text{eval}}=c|}{\sum_{j=0}^{C-1}|y_{\text{eval}}=j|}$, where $|y_{\text{eval}} = c|$ denotes the total number of samples with class-label c in the evaluation dataset:

$$\text{AEIG}(\hat{Y}^{\text{eval}}; y_a) = \mathbf{H1} - \underbrace{p(y_a = c|x_a, D^L)\frac{|y_{\text{eval}} = c|}{\sum_{j=0}^{C-1}|y_{\text{eval}} = j|}}_{\mathbf{P}}\mathbf{H2}. \tag{6.2}$$

6.2.2 Efficient IG computation in Deep Networks

As we saw in the previous section, computing both EIG (Equation 6.1) and AEIG (Equation 6.2) involves estimating the conditional entropy (**H2**) by retraining the models for each possible label for an image (i.e., a total of C classes) in the unlabeled set. In the active learning framework, this calculation is repeated for each image in the unlabeled set (i.e., total N images). Although this process might be feasible in the context of SVMs [209], it would be very computationally expensive (almost infeasible) in the context of a deep learning model (i.e., a total N*C model retraining). Following design simplifications are made to reduce the associated computation load.

Choice of Evaluation Set: In the first design simplification, we consider the validation set as our evaluation dataset ($D^{\text{eval}} = D^{\text{valid}}$).

Model Parameters: The second design simplification is based on the observation [60] that any machine learning model, including deep learning, consists of two components: representation and classification. In the context of modern convolutional neural network architectures, initial convolutional layers can be considered as feature representation learning layers, while the last MLP layers can be considered as a classification layers. While updating the model parameters during the IG calculation, only the classification layer parameters are updated. The convolutional layer's parameters are not updated. Given that most of the computation cost comes from the convolutional layers, this design permits computing IG scores (EIG or AEIG) with minimal computational overhead.

Model Updates: In the third design simplification, instead of retraining the whole model with the labeled dataset and each sample in the unlabeled dataset, the already trained model on the labeled set is only updated once through a single gradient step for one sample in the unlabeled set. This design simplification is based on the assumption that the size of the labeled dataset is greater than a single sample, and including just one more sample would not lead to a drastic change in the model parameters.

Algorithm 1 Information Gain Based Active Learning

Input: Labeled training dataset $D^L : \{(X^L, Y^L)\}$, an unlabeled dataset $D^U : \{(X^U)\}$, and an evaluation dataset D^{eval}

Require: initial machine model (with parameters θ^0) trained on labeled dataset D^L, total active learning iterations J, and active learning batch acquisition size B

$j \leftarrow 1$
while active learning iteration $j < J$ **do**

  Calculate $\mathbf{H}[\hat{Y}^{\text{eval}}|X^{\text{eval}}, D^L]$ based on the model parameters θ^{j-1}

  for each image $x_a \in D^U$ **do**
    Calculate $\text{p}(y_a = c|x_a, D^L)$ based on the model parameters θ^{j-1}
    $\theta_a^{j-1} \leftarrow \theta^{j-1}$

    for each possible class label $c \in \{0, 1, .., C\}$ **do**
      Using a single gradient step update model parameters (θ_a^{j-1}) with x_a and $y_a = c$
      Calculate $\mathbf{H}[\hat{Y}^{\text{eval}}|X^{\text{eval}}, y_a = c, x_a, D^L]$
    end for

    Compute Score based on AEIG (or EIG) according to Equation [2] (or [1])
  end for

  Select subset of top-B images (D^A) from D^U according to their score S
  Acquire ground-truth labels for D^A ((D^{A*}))
  Update Unlabeled dataset $D^U \leftarrow D^U \setminus D^A$
  Update Labeled dataset $D^L \leftarrow D^L \cup D^{A*}$
  Retrain the model (θ^j) with the updated labeled training dataset D^L
  $j \leftarrow j + 1$

end while

6.3 Multi-Class Medical Image Disease Classification

The active learning framework is applied to two different medical imaging contexts. The first context involves multi-class disease classification of diabetic retinopathy (DR) patients from colour fundus images. Fundus images are classified into five disease scales representing disease severity: No DR, mild DR, moderate DR, severe DR, and proliferative DR. A publicly available DR disease scale classification dataset is used for this task. Experiments in this part of the thesis use a subset of 8408 retinal fundus images provided by the Kaggle challenge organizers. A label with one of the five disease scales is provided with each retinal fundus image. For each of the five disease scales, there are 6150/588/1283/221/166 images, respectively. The differences in the total number of images per class highlight a high-class imbalance for the task. We randomly divide the whole dataset into 5000/1000/2408 images for training/validation/testing sets.

The second context involves multi-class classification of skin lesions from dermoscopic images. We use the publicly available international skin imaging collaboration (ISIC) 2018 dataset [46]. In this dataset, dermoscopic images are classified into 7 different classes: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesion. The challenge organizers provide a subset of 10015 dermoscopic images. A label with one of the seven disease scales is provided with each dermoscopic image. For each of the seven classes, there are 1113/6705/514/327/1099/115/142 images, respectively. The differences in the total number of images per class highlight a high-class imbalance for the task. We randomly divide the whole dataset into 6000/1500/2515 images for training/validation/testing sets.

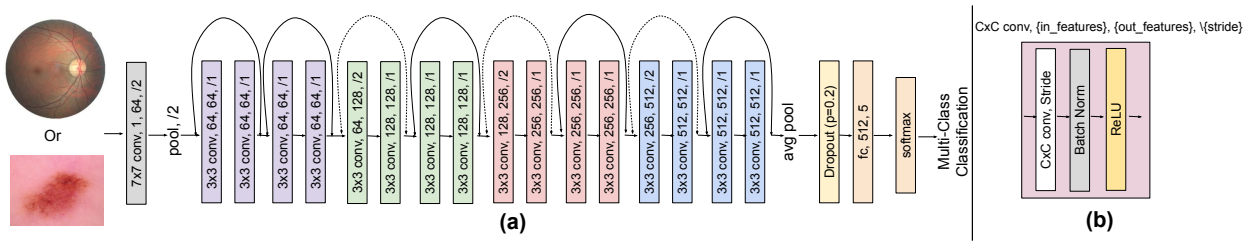


Figure 6.2: (a) A 2D ResNet-18 architecture consists of a 7x7 convolutional unit, followed by 16 3x3 convolutional units, one dropout layer (p=0.2), and one fully connected layers. The dotted shortcuts increase dimensions. Colour fundus images (or dermoscopic images) were given as input to the network. (b) Each convolutional unit consists of one CxC convolutional layer with stride S, followed by Batch Normalization layer, and a ReLU layer. ©[2022] Springer. Reprinted, with permission, from [165].

6.4 Experiments and Results

6.4.1 Implementation Details

Network Architectures: An ImageNet pre-trained 2D ResNet18 [94] architecture was used for the DR and the ISIC multi-class disease scale classification task. The network architecture is depicted in Figure 6.2. A dropout layer with p=0.2 is introduced before the fully connected (fc) layer. The network was trained to reduce the categorical cross-entropy loss. An Adam optimizer with a learning rate of 0.0005 and a weight decay of 0.00001 was used to train the network for a total of 100 epochs and a batch size of 64. The learning rate was decayed with a factor of 0.995 after each epoch. All fundus images (DR) were resized to 512x512 size, normalized with mean subtraction, and divided by std. All dermographic images (ISIC) were resized to 600x450 size, normalized with mean subtraction, and divided by std. Random horizontal flip, random vertical flip, and random rotation in the range of 0-30, were applied as data augmentation on each image. The code was written in PyTorch and ran on Nvidia GeForce RTX 3090 GPU with 24GB memory.

The 'macro' area under the receiver operating characteristic curve (ROC AUC) was used as a metric for both classification tasks. For both tasks, a macro average (unweighted) of one-vs-rest (ovr) classifier ROC AUC [67] was performed.

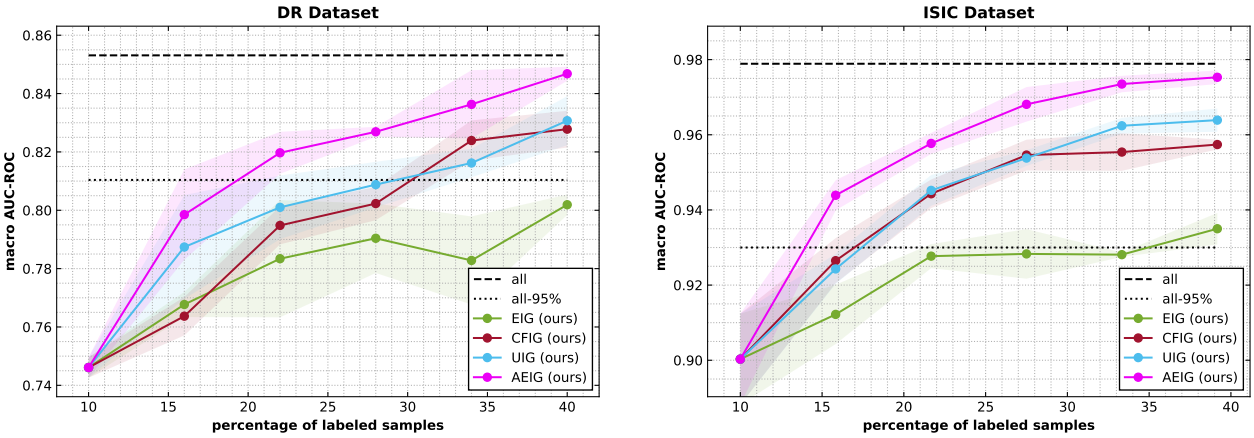


Figure 6.3: Comparison of the EIG, AEIG, UIG, and CFIG based active learning sampling methods for both the DR dataset (left) and the ISIC dataset (right). The horizontal solid dashed line ('all') at the top represents model performance when the entire training set is labeled. The dotted line ('all-95%') represents 95% of that performance. We report the mean and std of evaluation metrics across five different runs (See Table-1 and Table-2 in the appendix for exact values). ©[2022] Springer. Reprinted, with permission, from [165].

AL framework: The active learning framework was initialized by randomly selecting 10% of the training dataset (i.e., 500 for DR, 600 for ISIC) as the labeled dataset and the rest as the unlabeled dataset. It was deployed for a total of $J = 6$ active learning runs in both cases. Based on previous studies [266, 125], in each run, we select a total of $\approx 6\%$ of the dataset (B = 300 for the DR, and B = 350 for the ISIC) from the unlabeled dataset (D^U). We acquire an oracle label, and then, once labeled, these are used to update the labeled dataset ($D^L \leftarrow D^L \cup D^{A*}$) and the unlabeled dataset ($D^U \leftarrow D^U \setminus D^A$). Active learning experiments were repeated five times with different initial randomly selected images. The means and variances of the evaluation metrics were then recorded across the five repetitions.

6.4.2 Information Gain Performance

In this section, we compare the proposed AEIG and EIG based active learning sample selection against two different baseline alternatives for IG computation. Equation 6.1 describes the estimation of EIG, which includes weighing H2 with the predicted softmax probability **P**. Instead of relying on the predicted probabilities, we can compute two different baseline alternatives based on the prior information of the class distributions: (i) Uniform Information Gain (UIG) assumes a uniform distribution such that $\mathbf{P} = \frac{1}{C}$,

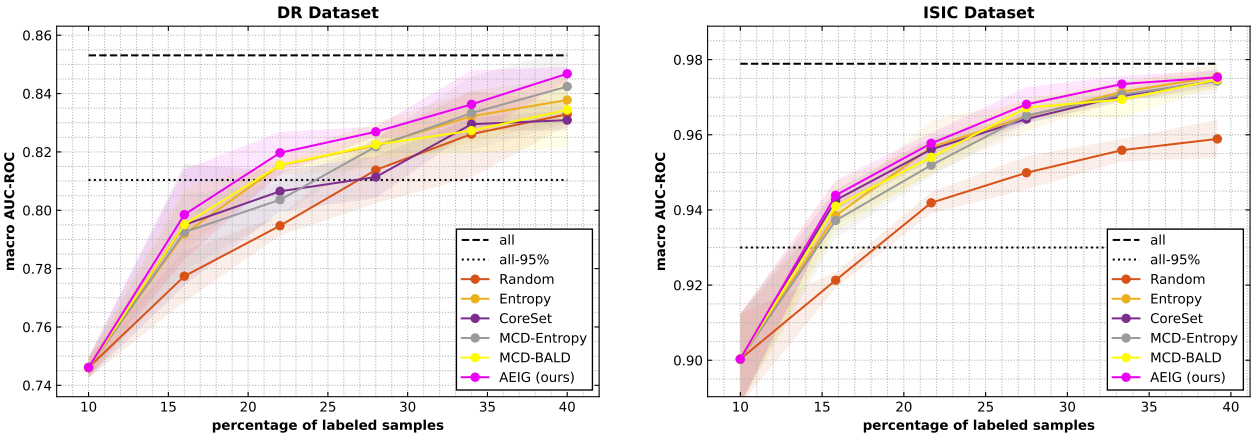


Figure 6.4: Comparison of the AEIG based active learning sampling method with Random, Entropy, MCD-Entropy, MCD-BALD, and CoreSet based sampling methods for both the DR dataset (left) and the ISIC dataset (right). The horizontal solid dashed line ('all') at the top represents model performance when the entire training set is labeled. The dotted line ('all-95%') represents 95% of that performance. We report the mean and std of evaluation metrics across five different runs. (See Table-3 and Table-3 in the appendix for exact values.)

$\forall c \in \{0, 1, ..., C-1\}$. (ii) Class-Frequency Information Gain (CFIG) assumes a distribution based on the class frequency such that $\mathbf{P} = \frac{|y_{\text{eval}}=c|}{\sum_{j=0}^{C-1}|y_{\text{eval}}=j|}$, where $|y_{\text{eval}} = c|$ denotes the total number of samples with class-label c in the evaluation dataset.

In Figure 6.3, we compare EIG, UIG, CFIG, and AEIG by experimenting on both datasets. Experiments indicate that the AEIG achieves 95% of the overall performance ('all-95%') with only 19% (for DR) and 14% (for ISIC) of the training dataset. CFIG, UIG, and EIG require 29%, 30% and >40% of the training dataset for DR; and 17%, 17.5%, and 35% of the training dataset for ISIC. We hypothesize that the better performance of AEIG is due to its ability to sample more images from minority classes.

6.4.3 Comparisons Against Active Learning Baselines

In this section, the proposed AEIG based sampling active learning framework was compared against five different baseline methods: Random, Entropy-based sampling [226], MC-Dropout with Entropy [80], MC-Dropout with BALD [98], and CoreSet [220]. The macro AUC ROC curve for experiments on the DR and ISIC datasets can be found in Figure 6.4. Overall, the proposed method gives better (or, in some cases, similar) per-

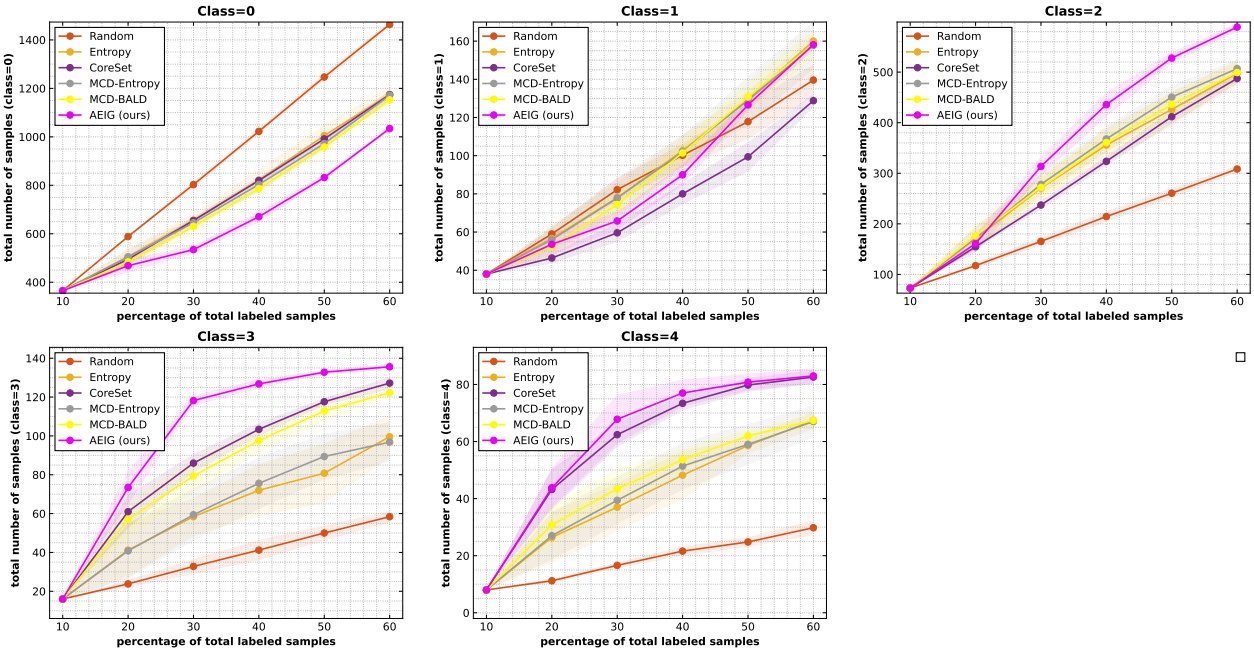


Figure 6.5: Plots depicting the total number of samples labelled per class against the percentage of labeled samples for the DR dataset for the competing active learning sampling methods. Classes 1, 3, and 4 are the minority classes. ©[2022] Springer. Reprinted, with permission, from [165].

formance compared to the other methods for both datasets and all six active learning iterations. Applying standard methods for comparison, the proposed method (AEIG) achieves 95% of the overall performance ('all-95%') with only 19% of the labeled training dataset for the DR dataset. MCD-Entropy, MCD-BALD, Entropy, CoreSet, and Random require approximately 24%, 21%, 21%, 27%, and 27% of the labeled training dataset to achieve similar performances. For the ISIC dataset, the proposed method (AEIG) achieves 95% of the overall performance ('all-95%') with only 14% of the labeled training dataset for the DR dataset. MCD-Entropy, MCD-BALD, Entropy, CoreSet, and Random require approximately 14.8%, 14.2%, 14.7%, 14.1%, and 18.2% of the labeled training dataset to achieve similar performances. It is worth pointing out that although all methods are giving a somewhat similar performance at 'all-95%' cutoff, the trend is consistent for all 6 AL acquisitions. The total active learning score computational time for each image in the unlabeled set was around 1 ms, 6 ms, 10 ms, 10 ms, 16 ms, and 28 ms for Random, Entropy, MCD-Entropy, MCD-BALD, CoreSet, and AEIG based methods. The computation times highlight that although the proposed method can achieve better performance compared to other methods, it is a bit slower. Compared to the time taken by a human annotator for

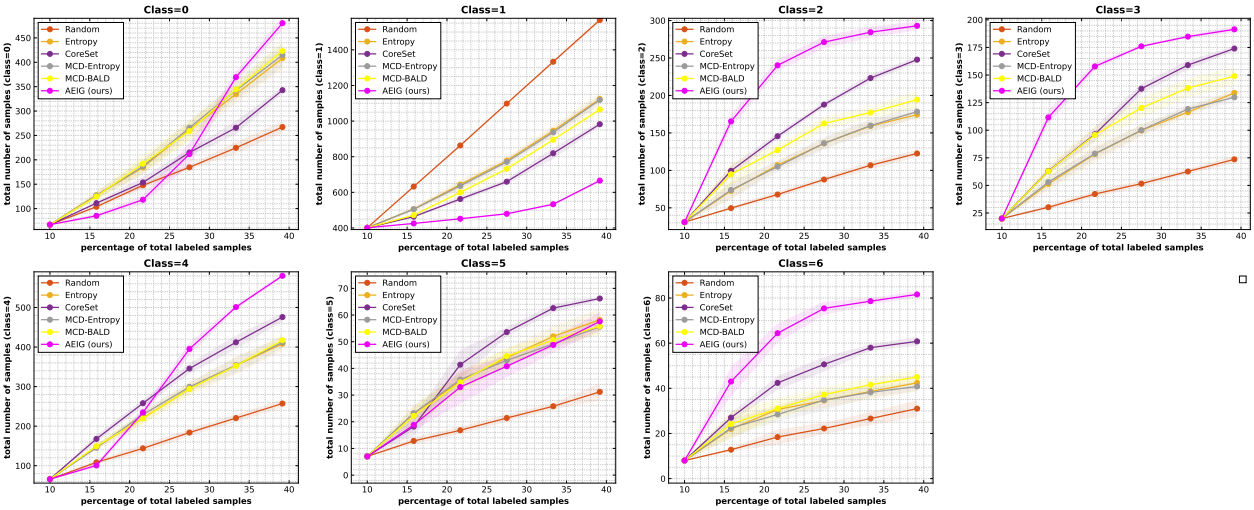


Figure 6.6: Plots depicting the total number of samples labeled per class against the percentage of labeled samples for the ISIC dataset for the competing active learning sampling methods. Classes 0, 2,3,4,5, and 6 are the minority classes. ©[2022] Springer. Reprinted, with permission, from [165].

additional labeling, this difference in computational time will not be significant.

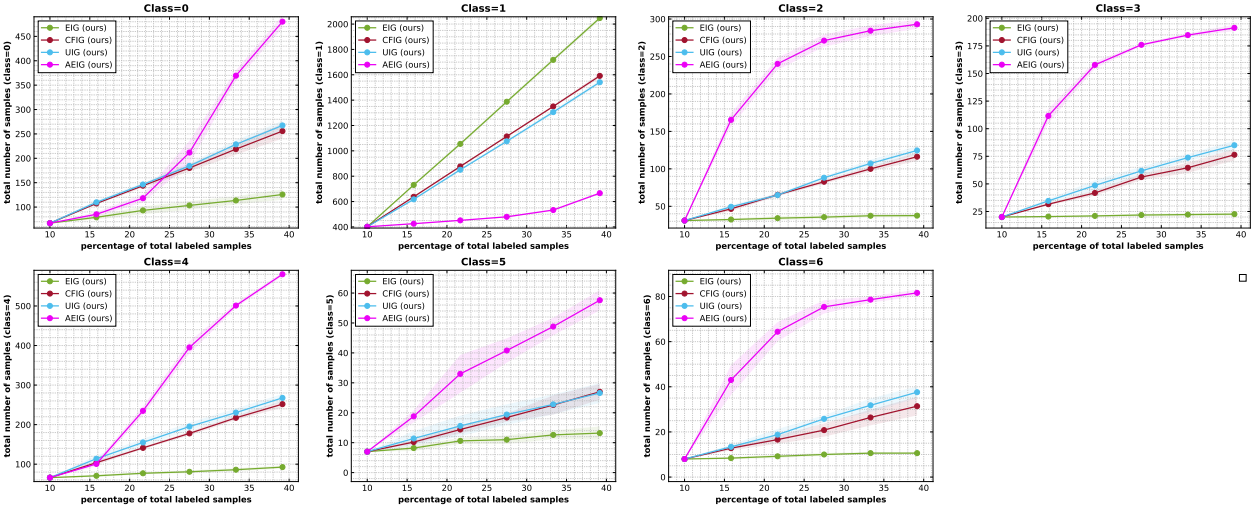


Figure 6.7: Plots depicting the total number of samples labeled per class against the percentage of labeled samples for the ISIC dataset for EIG, CFIG, UIG, and AEIG sampling methods. Classes 0, 2,3,4,5, and 6 are the minority classes. ©[2022] Springer. Reprinted, with permission, from [165].

Figure 6.5 illustrates the different number of acquired images per class on the DR dataset at each of the active learning acquisition steps for all six acquisition methods (Random, Entropy, CoreSet, MCD-Entropy, MCD-BALD, and AEIG). The results indicate that the AEIG based active learning sampling policy results in sampling and labeling of a higher

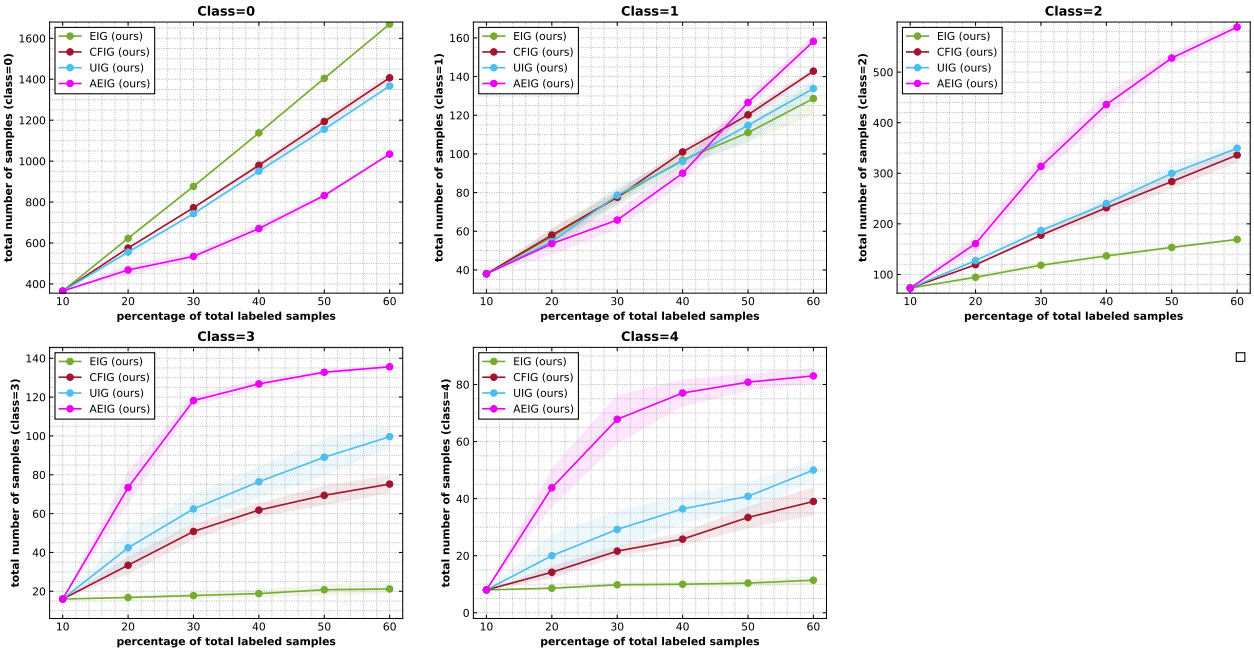


Figure 6.8: Plots depicting the total number of samples labeled per class against the percentage of labeled samples for the DR dataset for EIG, CFIG, UIG, and AEIG sampling methods. Classes 1, 3, and 4 are the minority classes. ©[2022] Springer. Reprinted, with permission, from [165].

number of images from the minority classes (e.g., classes 1, 3, and 4) compared to other sampling methods. This, in turn, leads to better overall performance for contexts with highly class-imbalance datasets, as is the case with the DR dataset.

6.5 Summary

In this chapter, we proposed an active learning framework that drives the selection of new image samples to label based on maximal Adapted expected information gain on an unseen evaluation dataset. Experiments were performed on two different medical image classification datasets, and results showed that the AEIG method performs better than Random, maximum Entropy, MCD-Entropy. MCD-BALD, and CoreSet based sampling strategies. The AEIG samples minority classes at a greater rate than competing strategies, improving performance on highly imbalanced datasets, although with a small computational overhead.

7

Conclusion and Future work

> So perhaps at a certain perspective what we leave behind is often wonderland, always different from what it was and generally more beautiful.
>
> — *Norman Maclean*, A river runs through it and other short stories

In this thesis, we developed multiple different methods for integrating Bayesian deep

learning uncertainties in medical image analysis to improve its applicability in clinical deployment.

As a first contribution, we developed an evaluation metric for quantifying uncertainty in brain tumour segmentation. The metric was designed with the clinical end goal in mind, where we can expect an end-user (clinician) to correct the most uncertain predictions made by a machine learning model. As such, it rewards uncertainty estimates that produce high confidence in correct assertions and those that assign low confidence levels at incorrect assertions; in addition to that, it penalizes uncertainty estimates that lead to a higher percentage of under-confident correct assertions. To evaluate its usefulness, the developed metric was used to rank 14 participating teams from QU-BraTS 2020 challenge. The ranking and analysis of uncertainties generated by these participating teams showed the complementary information provided by both the segmentation and their associated uncertainties, highlighting the need for uncertainty quantification in medical image analyses. The developed metric is general and can be applicable to a wide variety of other medical image segmentation problems, as was shown by a recent paper [37].

Next, we developed the first framework in medical image analysis where uncertainties are propagated from one task to another. This is an important direction as cascaded inference tasks are prevalent in medical image analysis, where mistakes made by an upstream task can lead to poor performance on the downstream task of interest. In this case, as we saw in the previous paragraph, if uncertainties are correlated with places where a model is prone to make mistakes, then we can use them as a proxy to inform the downstream task of interest about mistakes made by upstream tasks. Our experiments on a wide variety of clinically important tasks showed that propagating uncertainties from upstream tasks to the downstream task of interest improves performance.

Following this, we focused our attention on the fairness of machine learning models. Fairness of ML models across different sensitive attributes (e.g., age, race, sex) is important as

it shows that these models are not biased towards any particular subgroups, which would allow deployment of these models into the real world. However, till now the majority of fairness analysis was done with only absolute performance in mind. In this thesis, for the first time, we also analyzed the fairness of machine learning models from both absolute performance and uncertainty quantification perspectives. Our results on a wide range of clinical problems demonstrated that popular bias mitigation ML models do not always work on these types of problems, and when they work, they come at the cost of poor uncertainties associated with them. It would create situations where despite mitigating fairness from an absolute performance perspective, these models can not be deployed as poor uncertainties could lead to distrust by clinicians. As such future bias (fairness) mitigation methods should be designed by considering fairness for both the output and its associated uncertainties.

At last, we tackled an important problem of data scarcity in medical image analysis through the active learning (AL) framework. We developed a new active learning framework based on information gain. Unlike previous approaches for AL, which select samples that are hardest for the current model to classify, without considering its impact on the performance gain; our developed framework was specifically designed to consider expected information gained by the machine learning model based on selected new data points. Our results indicated that the proposed framework could achieve better model performance with a lower number of acquired labels compared to other competing methods. This has a real clinical impact as the proposed method should lead to overall lower labeling costs.

Overall, we can say that contributions made in this thesis show the necessity of quantifying uncertainties in all aspects of machine learning models for medical image analysis, and how it can be useful in making them ready for clinical deployment.

7.1 Future Work

7.1.1 Uncertainty Evaluation Metric

The uncertainty evaluation metric developed in Chapter 3 employed Dice score (DSC) for task-specific evaluation. The DSC is a good segmentation metric when the interest structure contains a high number of voxels. However, it is not a stable metric when calculated on a low number of voxels [200]. In the developed evaluation score, instability of the DSC leads to low performance at a lower threshold (more filtered voxels), as DSC calculation considers only a few remaining unfiltered voxels (Figure 3.2). The poor stability of DSC is a well-known challenge in the literature [200]. As such, future work could explore other task-dependent metrics.

Our analysis also revealed that *Team SCAN* performed better on the overall score by not marking any positive prediction (both true positive and false positive) as uncertain. In a real-world scenario, a method that is always confident about its positive predictions leads to confident over-segmentation. This shows that the developed uncertainty evaluation score is not perfect, and we need to keep improving it. One possible future direction could be to calculate Precision ($\frac{TP}{TP+FP}$) and Recall ($\frac{TP}{TP+FN}$) at different uncertainty thresholds and calculate the AUC of these curves (Precision vs Uncertainty threshold, and Recall vs Uncertainty threshold). A high-performing team should get a high AUC for both Precision and Recall (same as AUC for DSC). To achieve a high AUC for Precision, participating teams have to reduce FP (mark them as uncertain). Similarly, to attain a high AUC for Recall, participating teams have to reduce FN (mark them as uncertain). In this way, we can penalize teams that are highly confident in their positive predictions and those that are highly confident in their false negative predictions.

The proposed evaluation framework evaluates uncertainties for each tumor entity as a single-class segmentation/uncertainty problem, while the overall tumor segmentation is a multi-class problem. Future extensions could involve developing methods to eval-

uate uncertainties in multi-class segmentation. Multi-class segmentation uncertainties and single-class segmentation uncertainties are different and can lead to different outcomes [37]. In addition, the current evaluation framework focuses on filtering individual voxels, as most of the developed uncertainty frameworks generate per-voxel uncertainties that are not spatially correlated [79, 136]. The recent development of spatially correlated uncertainty generation methods [174] indicates the necessity of developing uncertainty evaluation scores that consider the spatial correlation between pixels/voxels.

Another future direction is obtaining "ground-truth" uncertainty maps and evaluating automatic uncertainty generation methods against these maps. One recent promising direction uses inter-observer and intra-rater variation to proxy for "ground-truth" uncertainty [130, 27, 169, 55]. One limitation of this approach is that it assumes that "ground-truth" uncertainties can be estimated through multiple labels provided by different raters for the same (often small) set of images. In recent papers [278, 228], it was noted that institutional biases [151] play an essential factor in deep learning medical imaging model performance. However, variability in labeling across raters reflecting institutional biases are not direct proxies for "ground-truth" uncertainties. To expand on this point, inter-rater and intra-rater variability rely on the assumption of attaining a unique label. However, there are many situations where a unique label cannot necessarily be attained in some regions of an image. For example, at boundaries between tumor and healthy tissue in MRI due partly to partial volume effects but also because the labels cannot be seen in the MRI (and cannot be verified without a biopsy in the case of a tumour). For the latter case, each annotator is "guessing" the location of the boundary when none are confident in their annotations. The result might be measuring contextual rater biases (e.g., based on their radiology backgrounds) but not reflecting the true uncertainties in the labels themselves (e.g., whether a particular pixel is an enhancing tumour). One alternative approach could be asking annotators to mark areas they are not certain about, such as tumor boundaries in an MRI scan. These "uncertain" areas can then serve as "ground truth," and uncertainty estimates generated by algorithms can be compared to it. That being said, acquiring a "ground-truth" uncertainty is still an open area of research.

7.1.2 Uncertainty Propagation across Cascaded Inference Tasks

In Chapter 4, we proposed uncertainty propagation across a cascade of inference tasks which showed improvement in the downstream task of interest. All employed methods in that chapter compute uncertainties based on generated multiple output samples for the same input. The sample generation process is computationally costly, as it comes with the overhead of increased inference time. It would be interesting to analyze if propagating uncertainties generated based on more recent methods like sample-free uncertainty estimations [258] or learned sample-based models [150], would also lead to improved performance for the downstream task of interest. These methods do not rely on multiple samples for uncertainty estimation. They predict uncertainty directly either based on the knowledge distillation or based on the distance in the data manifold. Thus, these methods come with the added benefit of low computation overhead. Similarly, other uncertainty estimation methods like evidential deep learning [255] and conformal prediction [9] methods could also be explored.

A natural extension of the uncertainty propagation framework is to convert it into an end-to-end system. However, an end-to-end system requires access to ground-truth labels at all inference stages for the same training data. These data are generally unavailable in real clinical contexts. For example, let us look at the ADNI clinical score prediction pipeline utilized in this thesis. This pipeline involves segmentation of the hippocampus followed by clinical score regression. As we saw in the experimental section (Section 4.3.1), we did not have access to ground-truth labels for both hippocampus segmentation marking and the clinical scores for the same training data. Thus, in this case, it would not be possible to develop an end-to-end system comprised of both segmentation and regression tasks. However, should it be possible to obtain this type of data, an end-to-end system where relevant uncertainty measures for a task are learned depending on the downstream task of interest may be an exciting and essential research direction to explore. It would also be

interesting to propagate labeling uncertainties [130, 27], if multiple annotations for each patient case are available.

While in the current uncertainty propagation framework, uncertainty maps are passed to the downstream task of interest as an additional input; future work could explore methods that either embed uncertainty propagation as a part of a loss function [187] or in the design of neural networks [57].

Another future direction could explore the impact of uncertainty propagation on the uncertainties of the downstream task's outputs. One could expect better uncertainty quantification in a downstream task of interest with uncertainty propagation.

7.1.3 Fairness of Machine Learning Models

In Chapter 5, we analyzed the fairness of machine learning models from the perspective of uncertainty quantification for many different medical image analysis tasks. Our results for three different methods indicated the need to consider uncertainty quantification while analyzing fairness methods. However, the analyzed fairness mitigation methods were supervised. This means they require annotation of the sensitive attribute (ex., sex, age, etc.) during training. However, these sensitive attributes might not be available during training for reasons such as privacy. In this scenario, it would be better to deploy unsupervised fairness mitigation methods like just train twice [141], adversarial reweighed learning [135], etc. These unsupervised methods mitigate the fairness concerns of machine learning models without the need for labeled sensitive attributes. It would be interesting and necessary to analyze the uncertainty quantification of these methods for various medical image analysis tasks. In the end, it is of paramount importance to develop new unsupervised fairness mitigation methods that consider both absolute performance and uncertainties.

In this thesis, we looked at the fairness of machine learning models for segmentation and clinical score regression problems. In our work, we employed standard metrics like dice coefficient for segmentation, and mean squared error for the regression problems to measure fairness. Although these metrics are useful, they are not specifically designed to measure fairness. Many different metrics have been proposed that are specifically designed to measure the fairness of classification models with different end goals [47, 38]. However, these metrics have their limitations. Take an example of demographic parity. This metric evaluates whether the proportion of positive outcomes is equal across different subgroups. However, it is possible that the proportion of positive outcomes is dependent on the sensitive attribute (e.g., demographic) of the subgroups. For example, the prevalence of multiple sclerosis is dependent on the sex of the patient, as females are twice as likely to live with MS as males [264]. In this case, using demographic parity as a fairness evaluation metric would not be justified. Another example of a fairness evaluation metric is equalized odds, which is used to evaluate fairness with respect to the distribution of false positives and false negatives across different groups. It ensures that the true positive rate (TPR) and false positive rate (FPR) are equal across subgroups. Although, it might be difficult to achieve equalized odds in practice, especially if the groups have different base rates (i.e., the proportion of positive outcomes in each group). Additionally, equalizing FPR may not always be desirable or possible, especially in cases where false negatives are more detrimental than false positives.

The above examples show that different fairness evaluation metrics can be employed for classification based on the end goal, as no single evaluation measure can always work in all scenarios. A similar type of effort must be put into developing fairness metrics specifically designed for regression and segmentation models.

Most of the work in the fairness literature only analyzes and mitigates fairness for a single sensitive attribute. For example, fairness across different sex, different demographic, or different age. But in reality, multiple sensitive attributes could lead to bias in model per-

formance [34]. For example, a model might only underperform for a subpopulation from a particular race (e.g., Asian) and a particular range of age. In this scenario, it is necessary to mitigate bias across multiple sensitive attributes. Some recent methods [231, 56] have started to look into this aspect; however, further research is required to make them more generalizable.

7.1.4 Active Learning for Medical Image Analysis

In Chapter 6, we developed a novel active learning framework for medical image classification. However, multiple future directions could be explored to build on top of our framework. First of all, we made many design choices to ensure we could compute information gain for deep learning models without much computational overhead. These design choices could be reconsidered. For example, authors in [128] approximate expected information gain (EIG) via Gaussian approximation and generalized linear model. Based on these assumptions, EIG could be calculated without retraining the whole model, as was done in our case.

The developed framework uses the difference of entropy to measure information gain provided by the selected samples. Entropy is a global metric that captures the distribution of probabilities across different classes. Future work should consider a way to measure information gained by each class separately, as it would lead to more informative samples for lower-performing classes.

The developed framework was only designed and evaluated for medical image classification problems. One of the future possible directions could be to develop a new active learning framework for medical image segmentation, as labeling of voxel-wise ground truth segmentation marking is more time-consuming compared to image-level classification data. Some work that addresses segmentation labeling for pixel-wise annotations tends to narrow their application to segmentation of natural images [235, 103, 124]. Some

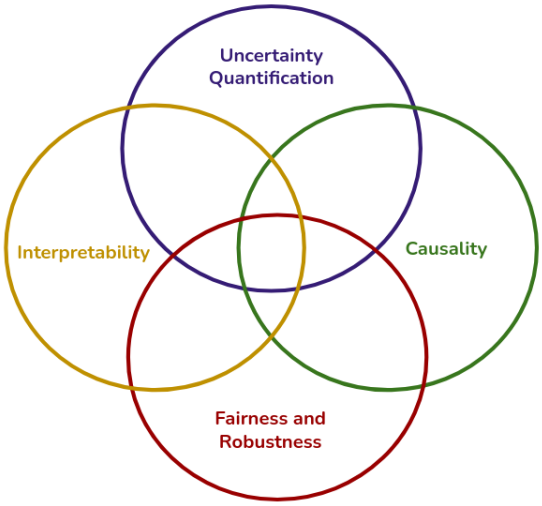


Figure 7.1: Trustworthy and safe machine learning models should exhibit four different characteristics: (i) uncertainty quantification, (ii) interpretability, (iii) robustness and fairness, and (iv) causality.

other work also addresses it from the super-pixel annotation perspective for natural image segmentation [116, 36, 7]. However, only a few works have been proposed that apply AL to medical image segmentation for deep learning models [275, 181, 77]. Future work should explore AL methods for medical image segmentation from an information gain perspective.

7.2 Towards Trustworthy and Safe Machine Learning Models

This thesis presented novel methods and applications of uncertainty quantification for medical image analysis. We also looked at the fairness of these models. However, to build a complete trustworthy, safe, and transparent system, machine learning models should also exhibit interpretability [214] and causality [39] in addition to fairness [284], robustness [274], and uncertainty quantification [286]. Interpretability allows the end users to understand how a machine learning model arrived at a particular decision [238, 145, 219]. Causality studies how changes in different input variables affect the change in the output of machine learning models. Causal machine learning models can be used to identify

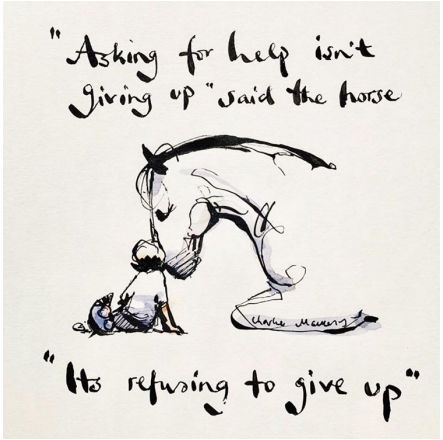
which treatment will have the greatest impact on patient outcomes [63].

As we partially saw in this thesis, all these aspects are related to each other and can help identify and alleviate issues related to one another. For example, instead of relying on the performance metrics, one could look at differences in the explanations provided by machine learning models across different subgroups to check whether these models suffer from biases [168]. Similarly, counterfactuals can be used to improve the interpretability of models [261] as well as identify and mitigate fairness concerns [133, 88]. For example, let us consider a model where the output should not change based on the skin colour of the input image, as is the case with skin lesion classification. In this scenario, if we generate a counterfactual image, where apart from the skin colour of the input image everything else remains the same, then using this counterfactual image as input should not change the output of the machine learning model. If, in this scenario, the machine learning model generates different outputs, then it is not a fair system as it is biased based on the skin colour of the input image.

Future work should focus on building models for medical image analysis that incorporates all the above-mentioned characteristics. By incorporating these aspects, we can ensure that machine learning models can provide all the necessary information to the end user (clinicians) to make more informed decisions. This will further help in the better adaptation of these models into real clinical practice.

A

Appendix: RS-Net – Synthesis of Brain MRI in the Presence of Pathologies



– *Charlie Mackesy*, The Boy, the Mole, the Fox and the Horse

Related Paper

It should be noted that this is not a manuscript based thesis. However, considerable material from the following paper has been utilised in this chapter.

- **R. Mehta**, T. Arbel, "RS-Net: Regression-Segmentation 3D CNN for Synthesis of Full Resolution Missing Brain MRI in the Presence of Tumours", *Simulation and Synthesis in Medical Imaging (SASHIMI) workshop held in conjunction with 21st International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2018, Lecture Notes in Computer Science, Springer, Vol. 11037, pp. 119-129* [158].

The Springer does not require individuals working on a thesis to obtain a formal reuse license. However, it requires that the thesis author cite the source and include Springer copyright notice for all figures and tables [4].

A.1 Introduction

The presence of a variety of different Magnetic Resonance (MR) sequences (e.g. T1, T2, Fluid Attenuated Inverse Recovery (FLAIR)) improves the analysis in the context of neurological diseases such as multiple sclerosis and brain cancers, because different sequences provide complementary information. In particular, the accuracy of detection and segmentation of lesions and tumours greatly increases should several sequences of MR be available [93], as different sequences assist in differentiating healthy tissues from focal pathologies. However, in real clinical practice, not all MR image sequences are always available for each patient for a variety of reasons, including cost or time constraints, or at times, images are available but not usable, for example due to corruption from noise or patient motion. As such, both clinical practice and automatic segmentation techniques would benefit greatly from the synthesis of one or more of the missing 3D MR image sequences based on the others provided [254, 106]. However, synthesis of full 3D brain MR image is challenging especially in the presence of pathology as different MR sequences represent pathology in a different way.

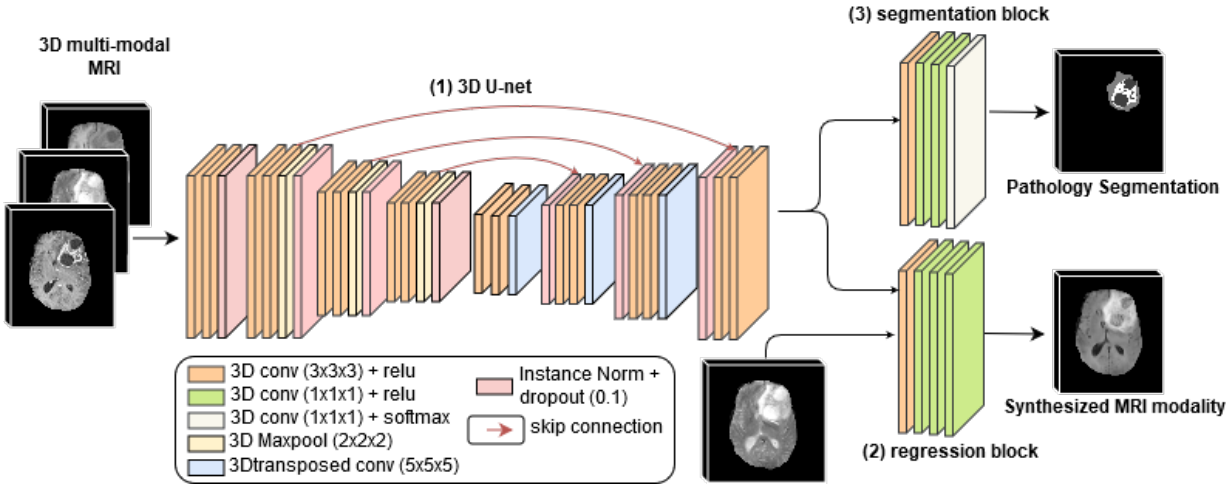


Figure A.1: Proposed Regression-Segmentation CNN architecture (RS-Net): (1) A 3D U-net, (2) Regression and (3) Segmentation convolution blocks. The model takes as input several full 3D MR image sequences, synthesizes the missing 3D MRI, while concurrently generating the multi-class segmentation of the tumour into sub-types. ©[2018] Springer. Reprinted, with permission, from [158].

Recently, modality synthesis has gained some attention from the medical image analysis community [113, 210, 259]. Several approaches have been explored, such as patch-based random forest [113] and sparse dictionary reconstruction [210]. Regression Ensembles with Patch Learning for Image Contrast Agreement (REPLICA) [113] was developed to synthesize T2-weighted MRI from T1-weighted MRI using the bagged ensemble of random forests based on nonlinear patch regression. Given the success of Convolutional Neural Networks (CNNs) [281] and Generative Adversarial Networks (GANs) [110] for image-to-image translation in the field of computer vision, several recent 2D CNN [41, 259] and 2D GANs [273] have been developed for modality synthesis in the context of medical imaging, showing promising results for synthesis of healthy subject MRI. A patch-based Location Sensitive Deep Network (LSDN) [259] was developed to combine intensity and spatial information for synthesizing T2 MRI from T1 MRI and vice versa. A 2D CNN model was developed to generate 2D synthesized images with missing input MRI [41]. Quantitative analysis showed superior performance over competing methods based on global image metrics (PSNR and SSIM). However, the performance of the method in the area of focal pathology was not examined.

In this chapter, an end-to-end 3D CNN is developed that takes as input a set of acquired

MRI sequences of patients with tumours and simultaneously performs (1) regression to generate a full resolution missing 3D MR modality and (2) segmentation of the brain tumour into subtypes. The hypothesis is that by performing regression and segmentation concurrently, the network should produce full-resolution, high quality 3D MR images, particularly the area of the tumour. The network is trained and tested on the MICCAI 2015 and 2017 BraTS datasets [22], as well as a large multi-site, multi-scanner, proprietary dataset of MS patient MRI. In the first set of experiments, the framework is evaluated against state-of-the-art synthesis methods [113, 259, 41] based on global image metrics used in previous work [41], where it is shown to give better performance compared all reported results. The second set of experiments evaluate the synthesis quality at pathological locations, by examining its performance on subsequent independent downstream tasks, namely tumour segmentation. Results show that real MR images can be swapped with the generated synthesized T1, T2, and FLAIR MR images with minimal loss in tumour segmentation performance. The network also quantifies the uncertainty of the regressed synthetic volumes through Monte Carlo dropout [79]. This permits the confidence in the synthesis results to be conveyed to radiologists and clinicians and to automatic downstream methods that would use the synthesized volumes as inputs. In the last set of experiments, we also evaluate the ability of RS-Net to synthesize missing modalities in case of Multiple Sclerosis (MS) patient MRIs. We evaluate the performance with a downstream MS T2 lesion segmentation/detection task. Results concur the findings reported for brain tumour segmentation task, and show that indeed missing modalities can be replaced by RS-Net synthesized modalities with minimal performance degradation.

A.2 Regression-Segmentation CNN Architecture

A flowchart of the proposed Regression-Segmentation CNN architecture (**RS-Net**) can be seen in Figure A.1. The network consists of three main components: (1) a modified 3D U-net [45], (2) regression convolution block for synthesizing image sequence, and (3) segmentation convolution block for multi-class tumour segmentation. RS-Net takes

as input full 3D volumes of all available sequences of a patient. The U-net generates an intermediate latent representation of the inputs which is provided to the regression and the segmentation convolution blocks. These then generate synthesis of the missing 3D MR image sequences and multi-class segmentation of tumours into sub-types, at the same resolution. The U-net learns latent representation which is common to both tumour segmentation and synthesis, with focus on high accuracy in the area containing tumour structures. In addition to the U-net output, the regression block is also provided with one of the input MRIs, which will provide necessary brain MR context to the regression block. The architecture details are now described.

The 3D U-net is similar to the one proposed in [45], with some modifications. The U-net consists of 4 resolution steps for both encoder and decoder paths. At the start, we use 2 consecutive 3D convolutions of size 3x3x3 with *k* filters, where *k* denotes the user-defined initial number of convolution filters. Each step in the encoder path consists of 2 3D convolutions of size 3x3x3 with $k * 2^n$ filters, where n denotes the U-net resolution step. This is followed by maxpooling of size 2x2x2. At the end of each encoder step, instance normalization [256] is applied, followed by dropout [241] with 0.1 probability. In the decoder path at each step, 3D transposed convolution of size 5x5x5 is applied, with 2x2x2 stride and $k * 2^n$ filters for the upsampling task. The output of the transposed convolution is concatenated with the corresponding output of the encoder path. This is, once again, followed by instance normalization and Dropout with 0.1 probability. Finally, 2 3D convolution of size 3x3x3 with $k * 2^n$ filters are applied. Rectified linear unit is chosen as a non-linearity function for every convolution layer.

Each of the segmentation and regression blocks contain 4 convolution layers. The first convolution layer is of size 3x3x3, and the rest are of size 1x1x1. The first three convolution layers have $k * 4$, $k * 2$ and k filters. In the regression block, the last layer has just 1 filter, while, for the segmentation block, there are C filters in the last layer, where C denotes the total number of classes for the segmentation task.

Weighted Mean Squared Error (W-MSE) loss is used for the synthesis task, and weighted Categorical Cross Entropy (W-CCE) loss for segmentation. Here, the weights are defined such that the weight increases whenever there are fewer voxels in a particular class.

$$\text{W-CCE}^i = -\sum_l w_l \sum_n y^i_{n,l} \log p^i_{n,l} \tag{A.1}$$

$$\text{W-MSE}^i = \sum_n w^i_n * (x^i_n - \hat{x}^i_n)^2 \tag{A.2}$$

$$w^i_n = w_l * y^i_n \qquad \text{where, } w_l = (\frac{\sum_{k=0}^{k=C} m_k}{m_l}) * r^{ep} + 1, \tag{A.3}$$

where, y^i_n, p^i_n, x^i_n, $\hat{x}^i_n$, and w^i_n denote true label, predicted label, true voxel values, predicted voxel value, and the weight for voxel n of volume i, respectively. w_l denotes the weight of class l. m_l is total number of voxels of l^{th} class in the training dataset. w_l are decayed over each epoch ep with a rate of $r \in [0, 1]$. It should be noted that w_l converges to 1 as ep becomes large.

The final loss function for the network, L^i, (for volume i) is a weighted combination of both of these loss functions:

$$L^i = \lambda_1(\text{W-MSE}^i) + \lambda_2(\text{W-CCE}^i). \tag{A.4}$$

Given the challenges associated with regressing a synthesized volume, errors are bound to exist. As such, deterministic outputs present dangers to subsequent clinical decisions as well as to downstream automatic methods that make use of the results. In this work, the network output is augmented with uncertainty estimates based on Monte Carlo dropout [79]. During testing, N Monte Carlo (MC) samples of the output are acquired by

passing each set of input volumes N times through the network to predict N different synthesized output MR volumes with probability of randomly dropping any neuron of the network equal to the dropout rate. Uncertainty in the synthesized volume, during testing, is estimated based on the variance of the MC samples at every voxel.

A.3 Experiments and Results

We now evaluate the performance of the RS-Net using two sets of experiments. In the first set of experiments, we compare the quality of the synthesized volume generated by RS-Net against other methods [41, 113, 259] using PSNR and SSIM on 2015 MICCAI BraTS dataset [22]. In the second set of experiments, we evaluate the quality of the synthesized volumes in a downstream task of tumor segmentation on 2017 MICCAI BraTS datasets [22].

RS-Net uses 4 initial convolutional filters and 4 steps for U-net encoder and decoder paths.This results in a network with a total of 674455 learnable parameters. Values of λ_1 and λ_2 in the loss function (Equation A.4), to combine CCE and MSE, were fixed to 1.0 and 0.1 respectively based on experimentation evidence. The networks were trained on a NVIDIA Titan Xp GPU for 240 epochs. Approximate training time was 3 days. The networks were trained with batch size of 1, using Adam optimizer [126] with the following hyperparameters: learning rate $= 0.0002$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-08}$. During testing time, a total of 20 samples of the output were generated to estimate the uncertainty in the synthesized volumes.

A.3.1 Comparison of RS-Net Against Other Methods

In order to compare the quality of the synthesized volumes produced by RS-Net against other state-of-the-art methods, namely REPLICA [113], LSDN [259], and 2D CNN [41], we

Table A.1: Quantitative results (mean ± std) for T1-to-T2 (top) and T1-to-FLAIR (bottom) synthesis based on PSNR and SSIM. Higher values indicate better performance. Absolute highest performing results seen in bold. ©[2018] Springer. Reprinted, with permission, from [158].

T2	REPLICA [113]	LSDN [259]	2D-CNN [41]	RS-Net (proposed)
SSMI	0.901 ± 0.01	0.909 ± 0.02	0.929 ± 0.17	$\mathbf{0.934 \pm 0.02}$
PSNR	28.62 ± 1.69	30.12 ± 1.62	30.96 ± 1.85	$\mathbf{31.13 \pm 1.78}$
FLAIR	REPLICA [113]	LSDN [259]	2D-CNN [41]	RS-Net (proposed)
SSMI	0.870 ± 0.01	0.887 ± 0.01	0.897 ± 0.01	$\mathbf{0.900 \pm 0.01}$
PSNR	28.32 ± 1.38	29.68 ± 1.56	30.32 ± 1.61	$\mathbf{30.88 \pm 1.84}$

train two different RS-Nets for T2 and FLAIR synthesis from T1 MRI, as done by Chartsias et al. [41]. We use the evaluation metrics, SSIM [267] and PSNR, defined in [41], to evaluate the quality of the synthesized volumes.

Given a ground-truth volume X and its corresponding synthesized volume $\hat{X}$, SSIM is computed as

$$SSIM(X, \hat{X}) = \frac{(2\mu_X \mu_{\hat{X}} + c_1)(2\sigma_{X\hat{X}} + c_2)}{(\mu_X^2 + \mu_{\hat{X}}^2 + c_1)(\sigma_X^2 + \sigma_{\hat{X}}^2 + c_1)} \tag{A.5}$$

where μ_X and σ_X^2 are mean and variance of volume X and $\sigma_{X\hat{X}}$ is the covariance between X and $\hat{X}$.

PSNR is computed as

$$PSNR(X, \hat{X}) = 10 \log_{10}(\frac{MAX_I^2}{MSE}) \tag{A.6}$$

where MAX_I is the maximum intensity of the volume and MSE is the mean squared error between volumes X and $\hat{X}$.

In order to compare our results to those in the paper [41], experiments were performed on the 2015 MICCAI BraTS training dataset [22]. This dataset consists of High-Grade Glioma (HGG) and Low-Grade Glioma (LGG) cases. 54 LGG cases were acquired with T1, T2,

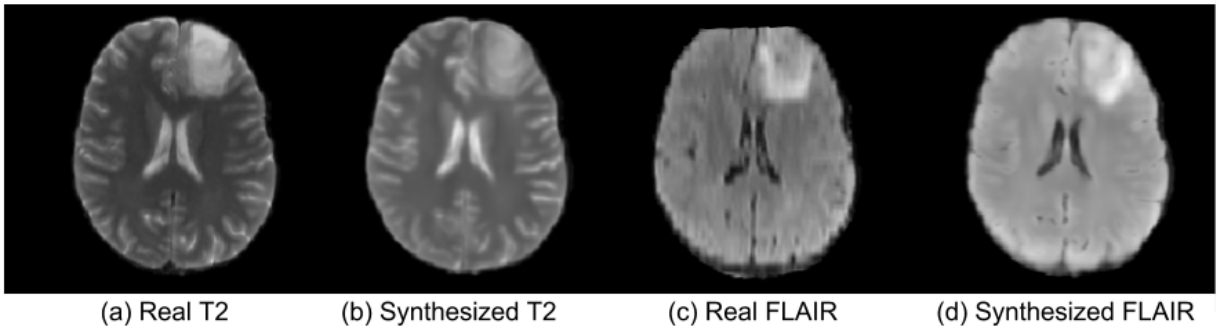


Figure A.2: Example slice from synthetic MR volumes generated by the proposed RS-Net on BraTS 2015 dataset for T1-to-T2 and T1-to-FLAIR synthesis. ©[2018] Springer. Reprinted, with permission, from [158].

T1ce, and FLAIR. Four tumour sub-classes were defined. Volumes are skull-stripped, co-registered, and interpolated to $1mm^3$ voxel dimension. Each volume is of size 240 x 240 x 155. We follow the same pre-processing steps followed in [41], where we normalize each volume by dividing by the volume's average intensity. Following [41], we perform 5-fold cross validation on the dataset (LGG cases). Here, for each cross-validation fold, the dataset is divided into three sets, namely, training, validation, and testing. Each set consists of 42, 6, and 6 volumes respectively.

Quantitative comparison of all different methods is given in Table A.1. It should be noted that we didn't reproduce the results for other methods and instead report them as listed in [41]. Results indicate that RS-Net performs slightly better than other methods based on the global metrics of PSNR and SSIM, for both T1-to-T2 and T1-to-FLAIR synthesis. The results also show the advantage of using the proposed 3D CNN over 2D CNN. An example showing qualitative results based on RS-Net for both T2 and FLAIR synthesis on a testing volume is shown in Figure A.2. Note that the resulting MR images are visually similar to the real images, particularly in the area of the tumour.

Table A.2: Comparison of multi-class brain tumour segmentation based on S-Net on the BraTS 2017 Validation dataset. The results using all 4 real MRI volumes are compared against replacing 1 real MRI volume with a synthesized MRI volume produced by RS-Net. Notation: Real MR volume (✓), and synthesized MR volume using RS-Net (⊙). Quantitative segmentation results based on Dice coefficients (mean ± std) for: enhancing tumor (DE), whole tumor (DT), and tumor core (DC). Higher values indicate better performance. ©[2018] Springer. Reprinted, with permission, from [158].

	T1	T2	FLAIR	T1ce	DE	DT	DC
Real	✓	✓	✓	✓	$\mathbf{68.2 \pm 31.0}$	$\mathbf{87.9 \pm 09.8}$	$\mathbf{75.7 \pm 23.1}$
T1 Synthesis	⊙	✓	✓	✓	67.6 ± 31.2	87.9 ± 09.8	75.5 ± 23.1
T2 Synthesis	✓	⊙	✓	✓	66.3 ± 32.1	87.3 ± 11.4	75.6 ± 23.6
FLAIR Synthesis	✓	✓	⊙	✓	66.8 ± 31.8	83.6 ± 10.7	73.1 ± 24.7
T1ce Synthesis	✓	✓	✓	⊙	24.8 ± 20.2	87.3 ± 10.0	54.0 ± 19.9

A.3.2 Evaluation of RS-Net synthesis using Tumour Segmentation

The metrics used in the previous section can be useful in assessing global synthesis quality, but in the context of volumes with pathological structures such as lesions or tumours synthesis quality assessment should focus on the pathological areas. To this end, we quantitatively evaluate the synthesis performance based on their effect on downstream method, tumour segmentation and tumour sub-class segmentation. To this end, we train a new segmentation CNN, for the specific task of multi-class tumor segmentation (referred to as **S-Net**). This network is similar to the RS-Net but modified such that the synthesis convolution block is removed. S-Net is trained using all 4 real MR volumes with weighted CCE as the loss function. To evaluate the quality of the synthesized volume, one of the real MR volumes is swapped with the synthesized one and the segmentation accuracy is measured. Note that we do not retrain the S-Net with the synthesized volume. This allows us to measure quality of the synthesized volumes in comparison to the real volumes.

Dataset and Pre-processing:

The 2017 MICCAI BraTS [22] datasets were used for all the experiments in this section. The BraTS training dataset was used to train the networks. This dataset is comprised of 210 HGG and 75 LGG patients with T1, T1 post contrast (T1ce), T2, and FLAIR MRI for

Table A.3: Comparison of multi-class brain tumour segmentation results based on S-Net on the BraTS 2017 Validation dataset, where each real MR input volume is replaced by its corresponding synthesized MR volume generated by either RS-Net or R-Net in a leave-one-out fashion. Notation: Real MR volume (✓), synthesized MR volume using RS-Net (⊙), and R-Net (•). Quantitative segmentation results based on Dice coefficients (mean ± std) for: enhancing tumor (DE), whole tumor (DT), and tumor core (DC). Higher values indicate better performance. ©[2018] Springer. Reprinted, with permission, from [158].

	T1	T2	FLAIR	T1ce	DE	DT	DC
Real	✓	✓	✓	✓	**68.2 ±31.0**	**87.9 ±09.8**	**75.7 ±23.1**
T1 Synthesis	⊙	✓	✓	✓	67.6 ± 31.2	87.9 ± 09.8	75.5 ± 23.1
	•	✓	✓	✓	67.5 ± 31.3	87.8 ± 09.9	75.3 ± 23.3
T2 Synthesis	✓	⊙	✓	✓	66.3 ± 32.1	87.3 ± 11.4	75.6 ± 23.6
	✓	•	✓	✓	66.1 ± 32.0	87.2 ± 11.9	75.4 ± 23.8
FLAIR Synthesis	✓	✓	⊙	✓	66.8 ± 31.8	83.6 ± 10.7	73.1 ± 24.7
	✓	✓	•	✓	62.9 ± 33.3	81.3 ± 17.4	71.5 ± 25.8
T1ce Synthesis	✓	✓	✓	⊙	24.8 ± 20.2	87.3 ± 10.0	54.0 ± 19.9
	✓	✓	✓	•	24.1 ± 22.1	85.9 ± 11.0	53.9 ± 23.4

each patient, along with expert tumor labels for each of 3 classes: edema, necrotic/non-enhancing core, and enhancing tumor core. 228 volumes were randomly selected for training the network and another remaining 57 for network validation. A separate BraTS 2017 validation dataset, held out during training, was used to test the synthesis and segmentation performance. This dataset contains 46 patient multi-channel MRI (with no labels provided). The BraTS challenge provided pre-processed volumes that were skull-stripped, co-aligned, and resampled to 1 mm^3 voxel volume. The intensities were additionally rescaled using mean subtraction, divided by the standard deviation, and rescaled from 0 to 1 and were cropped to 184 x 200 x 152. For this context, the additional complementary input presented to the regression block (see Figure A.1(3)) for T1, T2, T1ce, and FLAIR sequences were T1ce, FLAIR, T1, and T2 respectively. This was chosen as T1ce is the gadolinium enhanced version of T1, and FLAIR is the fluid attenuated version of T2.

Qualitative Evaluation:

Synthesis MR volumes produced in a leave-one-out approach by 4 different RS-Nets such that three real MR sequences are used to synthesize the fourth (see Figure A.3). The results indicate that the network is able to produce high-quality, high-resolution, 3D syn-

thesized MR volumes, particularly for T1 and T2 sequences, and even for FLAIR. As T1ce shows enhancement within the tumour based on injection of a contrast agent, it was not expected to be easily synthesized from other sequences and error resulted. However, the system indicates locations where the network is uncertain about the regressed output. Qualitative results indicate that errors within the tumour enhancement have associated relatively high uncertainties. This suggests that these uncertainties can be communicated to a clinician or radiologist to indicate trustworthy regions of the synthesized images, and that automatic downstream methods using the synthesized volumes can focus computations on the areas of high confidence, which should be explored in future work.

Replacing real with synthetic MRI Volumes:

In Table A.2, we compare the tumour segmentation using S-Net in two different testing scenarios, (i) all 4 real MR volumes are provided as input and (ii) 1 real MR volume is replaced with synthesized MR volume for each sequence generated by RS-Net, in turn. We train 4 different RS-Nets to synthesize 4 MR image sequences, where 3 real sequences are presented as input to RS-Net to synthesize the fourth. The synthesized MR volume, along with the 3 real corresponding MR volumes, were then presented to the S-Net previously trained on all four real MRIs. This will allow us to measure quality of the synthesized volume in comparison to the real volume. The resulting labels for BraTS 2017 validation set were uploaded to the BraTS Challenge server, where quantitative segmentation results were provided based on the Dice coefficients for: whole tumor, enhancing tumor, and tumor core. These results (Table A.2) indicate that by swapping out real MR volumes with the synthesized T1 or T2 MR volumes generated by the RS-Net leads to comparable brain tumour segmentation performance based on all three reported Dice metrics. For the slightly harder problem of FLAIR synthesis, results indicate a small degradation in tumour segmentation performance for all three Dice metrics. T1ce synthesis results in no loss of whole tumour segmentation performance, but, as predicted, led to a significant reduction in performance in terms of enhancement and necrotic core. This was expected as T1ce is a challenging MRI to synthesize due to its reliance on a contrast agent, which is

Table A.4: Comparison of multi-class brain tumour segmentation results based on S-Net against the results generated directly from the segmentation module of RS-Net for the BraTS 2017 Validation dataset. Notation: Real MR volume (✓), synthesized MR volume using RS-Net (⊙), and segmentation output of RS-Net without MR volume (×). Quantitative segmentation results based on Dice coefficients (mean ± std): enhancing tumor (DE), whole tumor (DT), and tumor core (DC). Higher values indicate better performance. ©[2018] Springer. Reprinted, with permission, from [158].

	T1	T2	FLAIR	T1ce	DE	DT	DC
Real	✓	✓	✓	✓	**68.2 ±31.0**	**87.9 ±09.8**	**75.7 ±23.1**
T1 Synthesis	⊙	✓	✓	✓	67.6 ± 31.2	87.9 ± 09.8	75.5 ± 23.1
	×	✓	✓	✓	66.4 ± 33.0	85.2 ± 15.3	71.0 ± 27.4
T2 Synthesis	✓	⊙	✓	✓	66.3 ± 32.1	87.3 ± 11.4	75.6 ± 23.6
	✓	×	✓	✓	66.5 ± 32.3	87.0 ± 10.6	71.1 ± 28.4
FLAIR Synthesis	✓	✓	⊙	✓	66.8 ± 31.8	83.6 ± 10.7	73.1 ± 24.7
	✓	✓	×	✓	69.0 ± 31.0	81.7 ± 15.1	72.4 ± 28.8
T1ce Synthesis	✓	✓	✓	⊙	24.8 ± 20.2	87.3 ± 10.0	54.0 ± 19.9
	✓	✓	✓	×	23.1 ± 19.8	86.5 ± 10.8	52.0 ± 20.8

not used by any other MR sequences.

Effectiveness of combined Regression-Segmentation task:

RS-Net has two output streams for synthesis and segmentation tasks. To check how RS-Net performs in comparison to a network which is trained only for the task of synthesis, we train a new network (**R-Net**) which is similar to RS-Net but modified such that the segmentation block is removed as well as the additional input to the regression block, and training is based only on weighted MSE. R-Net was trained for the synthesis of all 4 MR image sequences separately, in a leave-one-out approach, and tested for tumor segmentation using S-Net on the BraTS validation dataset exactly as described above. From Table A.3, we can observe that R-Net performs comparably to RS-Net, when T1 and T2 are synthesized but shows a small degradation in performance for FLAIR and T1ce synthesis on all three Dice metrics. This shows that performing synthesis and segmentation together allows the network to focus more on tumour part, and in turn gives better quality of the synthesized volume, especially for FLAIR and T1ce.

Performance of Segmentation part of RS-Net:

One of the advantages of the RS-Net is that, in addition to MRI synthesis, it also provides tumour segmentation labels. In this section, we will analyze this segmentation part of RS-Net (Figure A.1 (2)). Table A.4 indicates that the segmentation performance based on RS-Net directly is lower than the results based on using all 4 real MR volumes in S-Net, but is generally lower in comparison to the segmentation results when synthesized MR volumes generated by RS-Net is used in place of a real MR volumes. This trend is consistent across all MR image sequences for all three Dice metrics, except for FLAIR where the enhancing and core tumour Dice is higher for segmentation directly from the RS-Net over the segmentation results from S-Net with a synthesized input (for unknown reasons).

A.3.3 Evaluation of RS-Net synthesis results for Multiple Sclerosis

MS is a chronic, inflammatory demyelinating disease of the central nervous system with presently no known cure. The presence of lesions in MRI is one of the hallmarks of MS. As a result, MRI has been used for diagnosis and to monitor disease progression and treatment efficacy. Similar to brain tumours, segmentation of T2 lesion, which is useful for staging MS patients, requires availability of multiple MR sequences like FLAIR, T2, T2, PDw etc. In particular FLAIR or T2 MR images are routinely used for visualization and segmentation of T2 lesion as they appear hyperintense in FLAIR/T2 images. In this section, we validate the usefulness of RS-Net by synthesizing FLAIR or T2 images from other modalities available, and check its effectiveness by evaluating it on a downstream T2 lesion segmentation/detection task.

We train two different RS-Net to synthesize FLAIR and T2 MR sequence from the other available MR sequence (T1,T2,PDw for FLAIR synthesis and T1,FLAIR,PDw for T2 synthesis). We train a S-Net on all 4 real MR sequences and at test time replace one of them (FLAIR or T2) with the synthesized one. This allows us to measure quality of the synthesized volumes in comparison to the real volumes. We compare this against R-Net.

The method was evaluated on a proprietary, multi-site, multi-scanner, clinical trial dataset of 1064 Relapsing-Remitting MS (RRMS) patients, scanned annually over a 24-month period. T1, T2, FLAIR, and PDW MRI sequences were acquired at a 1mm x 1mm x 3mm resolution and pre-processed with brain extraction, N3 bias field inhomogeneity correction, Nyul image intensity normalization, and registration to the MNI-space. Ground truth T2 lesion segmentation masks were provided with the data. These were obtained using a proprietary approach where the result of an automated segmentation method was manually corrected by expert human annotators. All networks (RS-Net/R-Net/S-Net) were trained on 65% of the subjects, with 17.5% held out for validation and 17.5% for testing.

Since the downstream outcome of interest is the accurate detection of T2 lesions, we evaluate the performance of networks based on lesion-level True Positive Rate (TPR) and False Detection Rate (FDR). To obtain lesion-level detections from the voxel-based segmentations, a connected component analysis is performed to group lesion voxels together in an 18-connected neighbourhood. A true positive (TP) lesion is detected when the segmentation, including its 18-connected neighbourhood, overlaps with at least three, or more than 50%, of the ground truth lesion voxels. Insufficient overlap results in a false negative (FN), and candidate lesions of 3 or more voxels that do not overlap with a ground truth lesion are counted as false positives (FP). The TPR ($= \frac{TP}{TP+FN}$) and FDR ($= \frac{FP}{FP+TP}$) are then calculated at the lesion level and are used to plot receiver operating characteristic (ROC) curves. Given that MS lesions vary greatly in size, the system performance is evaluated on lesions grouped into three size bins: small (3-10 vox), medium (11-50 vox), and large (51+ vox).

Quantitative evaluation (ROC curve of TPRvsFDR) of RS-Net against R-Net for FLAIR and T2 synthesis by replacing real MR sequence with synthesized MR sequence in S-Net is given in Figure A.4 and Figure A.5. From these figures we can see that RS-Net performs better compared to R-Net for all lesions. This also holds true for all individual lesion size

Table A.5: Comparison of TPR at 0.2 FDR for different lesions sizes for RS-Net synthesized and R-Net synthesized MR sequences (FLAIR and T2) against Real sequences. ©[2018] Springer. Reprinted, with permission, from [158].

	FLAIR synthesis				T2 synthesis			
	All	**Large**	**Med.**	**Small**	**All**	**Large**	**Med.**	**Small**
All Real sequences (4)	0.740	0.999	0.970	0.360	0.740	0.999	0.970	0.360
3 Real + 1 R-Net synthesized sequences	0.695	0.998	0.952	0.300	0.705	0.990	0.925	0.350
3 Real + 1 RS-Net synthesized sequences	0.715	0.999	0.960	0.315	0.720	0.998	0.945	0.365

ROC curves. Value of TPR at 0.2 FDR (the clinical operating point of interest) is given in Table A.5. From this table, we can see that RS-Net synthesized MR sequences (FLAIR or T2) consistently gives better performance compared to R-Net synthesized MR sequence for all lesion size. This shows that performing synthesis and segmentation together gives better performance compared to only synthesizing the missing MR sequences.

A.4 Conclusions

In this chapter, a full resolution 3D end-to-end CNN was developed for the task of MR volume synthesis in the presence of brain tumours. The network was trained for the concurrent tasks of synthesizing a missing MRI sequence and tumour sub-tissue segmentation. Experimental results on BraTS 2015 challenge dataset indicated that the proposed method outperforms all previous methods in terms of traditional evaluation metrics like PSNR and SSIM. The quality of the synthesized images was further evaluated by assessing their effects on the performance in independent tumour segmentation experiments. Experiments on the BraTS 2017 challenge dataset indicated that multi-task learning helps in synthesizing high quality volumes over synthesis alone particularly in more challenging contexts (i.e. FLAIR and T1ce). Evaluation on downstream segmentation/detection task for brain tumour / Multiple Sclesions patient indicated that real MRIs can be replaced with synthesized T1, T2, and FLAIR volumes with minimum degradation in segmentation accuracy, whereas synthesizing T1ce is still too challenging. However, uncertainty measure based on Monte Carlo dropout was shown to be helpful in communicating the confidence in the synthesis results, which will be essential for their adoption by clinicians and downstream automatic methods.

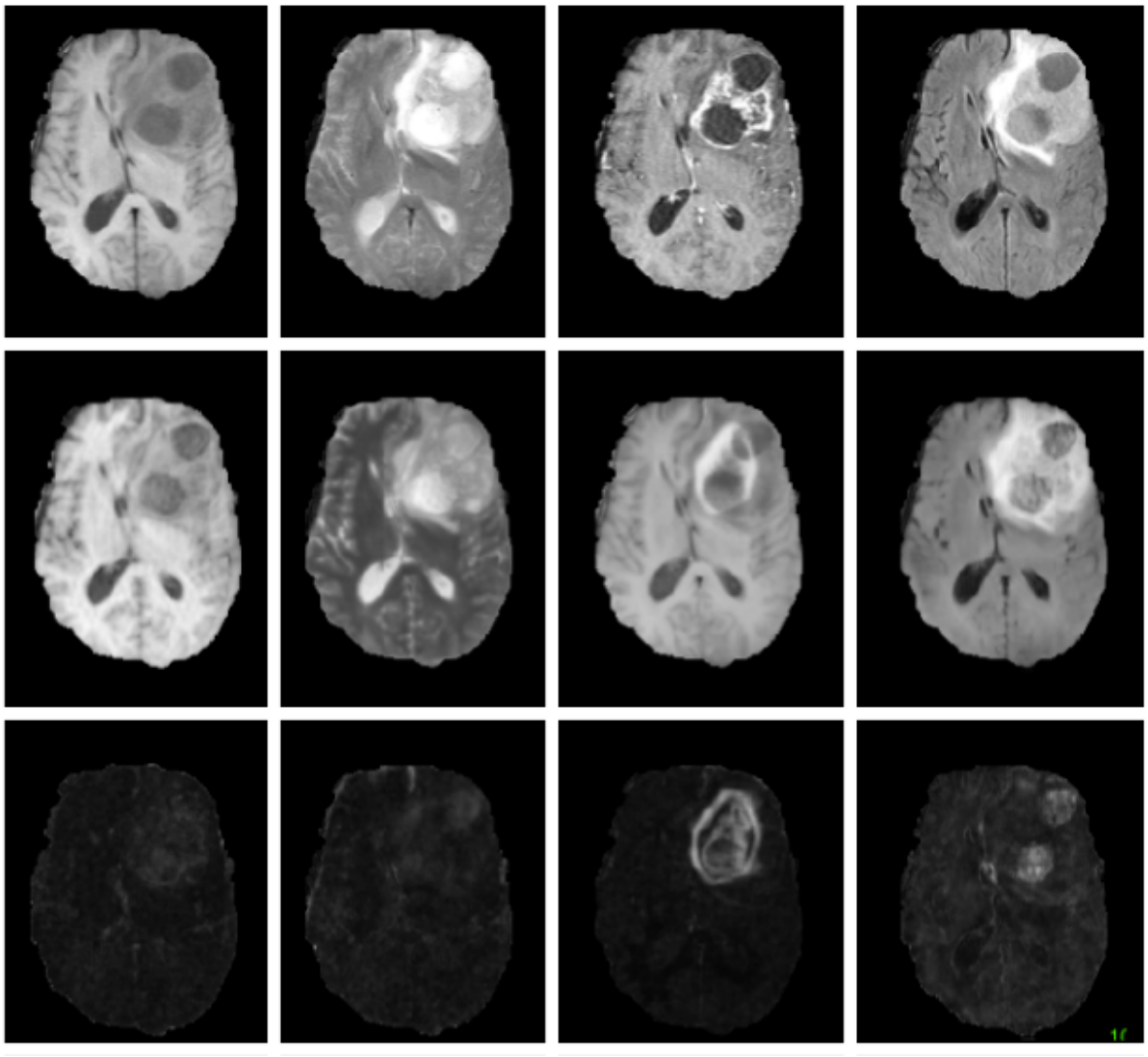


Figure A.3: Example slice from synthetic MR volumes generated using the proposed RS-Net along with its associated uncertainties. Real MRI (Row 1); synthesized volumes (Row 2) and its associated uncertainty (Row 3) produced as mean and variance across 20 MC dropout samples. Columns from left to right: T1, T2, T1ce, and FLAIR. Notice that uncertainties are highest where predicted tumour enhancements in T1ce are incorrect. ©[2018] Springer. Reprinted, with permission, from [158].

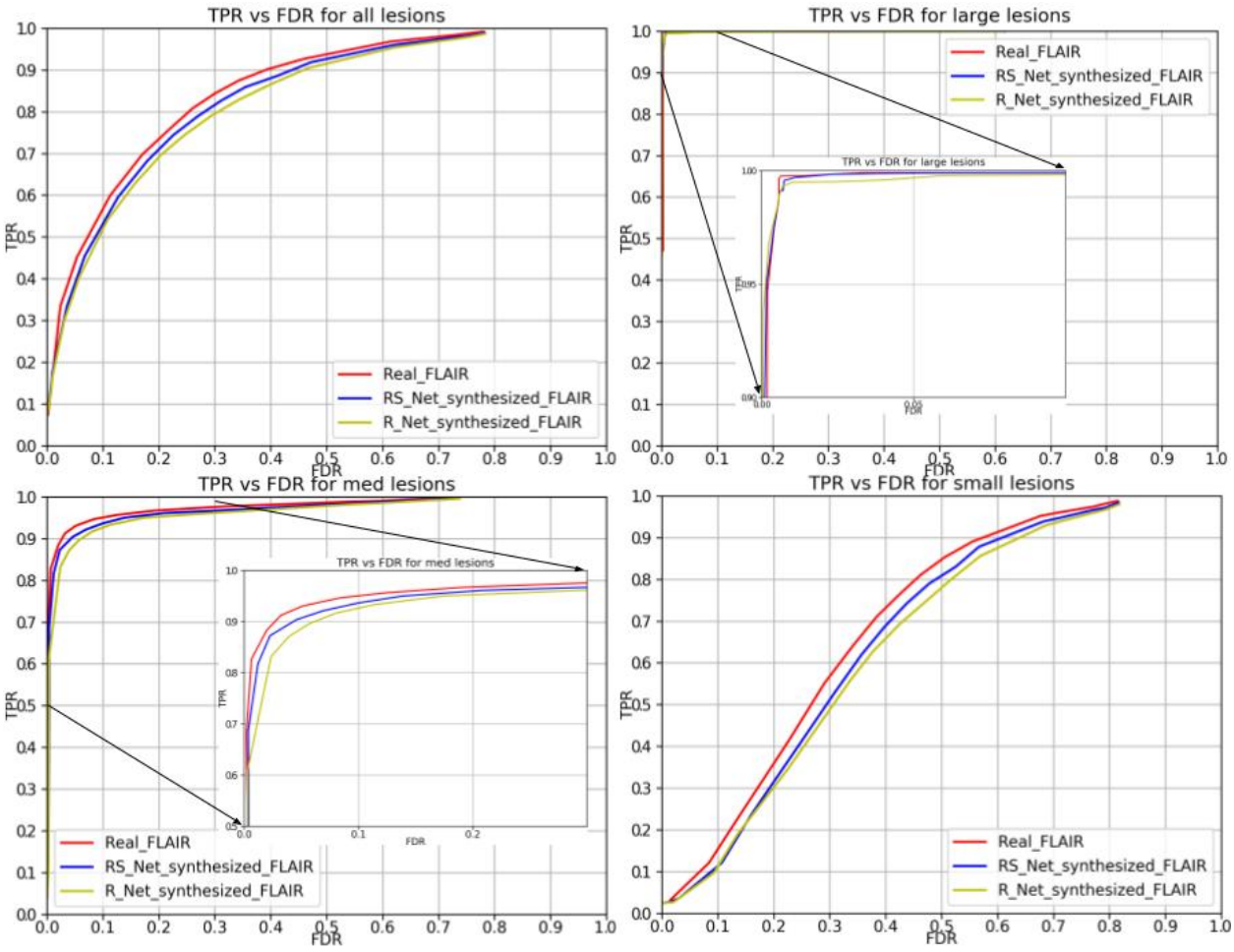


Figure A.4: Comparison of T2 lesion detection results based on S-Net (Red) for FLAIR synthesis, where FLAIR MR input image is replaced by its corresponding synthesized MR volume generated by either RS-Net (Blue) or R-Net (Yellow). Here, Receiver-operating characteristic (ROC) curves are plotted, illustrating TPR (true positive rate) vs. FDR (false detectionrate) across all lesions (Top Left), large lesions (Top Right), medium lesions (Bottom Left) and small lesions (Bottom Right). ©[2018] Springer. Reprinted, with permission, from [158].

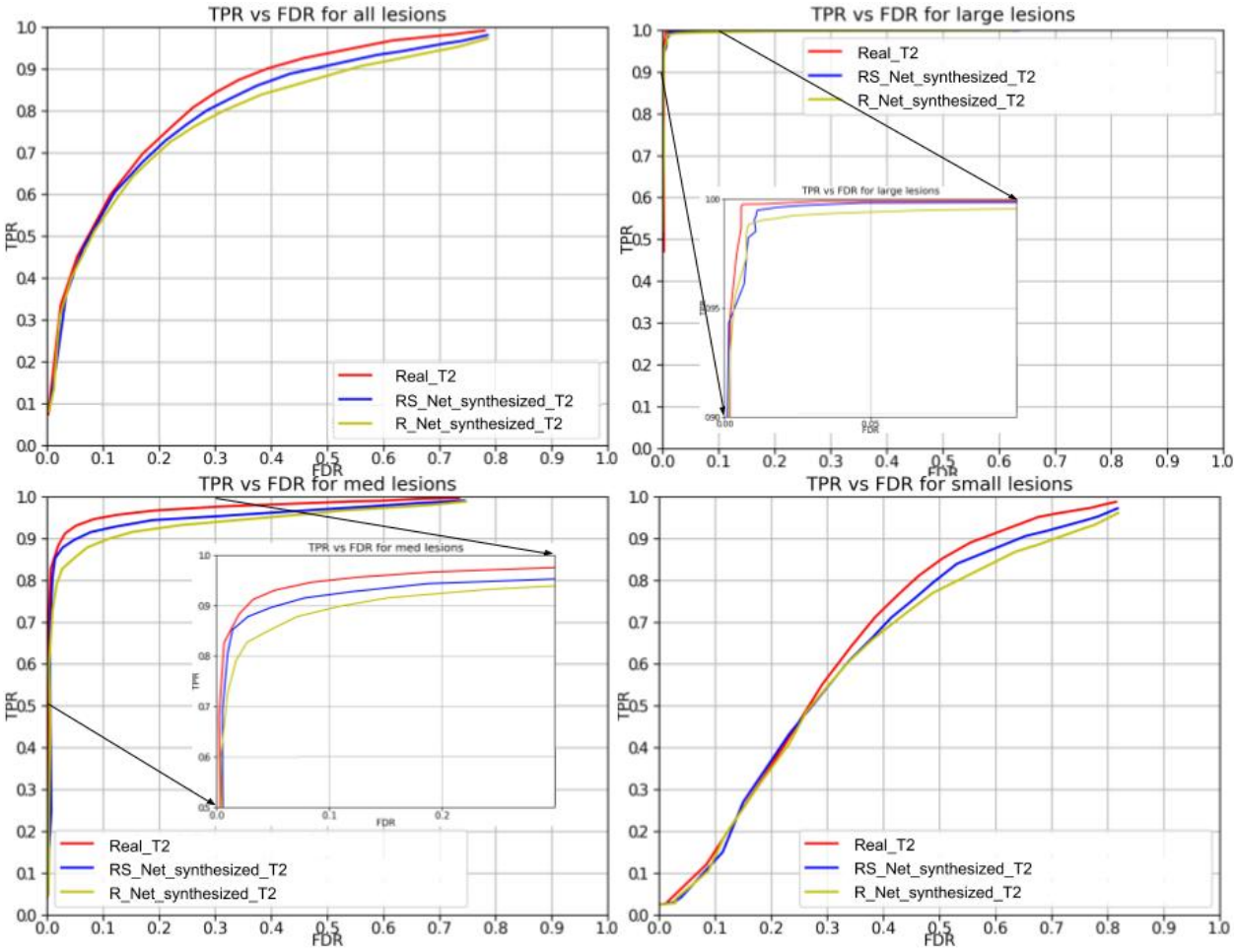


Figure A.5: Comparison of T2 lesion detection results based on S-Net (Red) for T2 synthesis, where T2 MR input image is replaced by its corresponding synthesized MR volume generated by either RS-Net (Blue) or R-Net (Yellow). Here, Receiver-operating characteristic (ROC) curves are plotted, illustrating TPR (true positive rate) vs. FDR (false detectionrate) across all lesions (Top Left), large lesions (Top Right), medium lesions (Bottom Left) and small lesions (Bottom Right). ©[2018] Springer. Reprinted, with permission, from [158].

B

Appendix: Evaluating Uncertainty Estimates in Brain Tumour Segmentation

B.1 Box Plots for Individual Scores

This appendix provides box plots for four different scores (DICE_AUC, FTP_RATIO_AUC, FTN_RATIO_AUC, and Score - Equation 3.1) for three different tumor entities (WT, TC, and ET) for each team. The teams are ranked from better to worse performance according to mean values across all patients for each score. Higher is better for DICE_AUC (Figure B.1 - Figure B.3) and Score (Figure B.10 - Figure B.12), while lower is better for FTP_RATIO_AUC (Figure B.4 - Figure B.6) and FTN_RATIO_AUC (Figure B.7 - Figure B.9).

Note that these box plots are different from ranking plots, as the ranking plots describe the overall performance across different tumor entities and different subjects as described in Section 3.4.1. From these plots, we can see that while for all three tumor entity DICE_AUC plots, *Team nsu_btr* performs better than other teams, their overall Score is comparatively lower than other teams as they do not perform well for FTP_RATIO_AUC and FTN_RATIO_AUC.

Similarly, we also observe that *Team SCAN* does not outperform other teams for DICE_AUC but comfortably outperforms other teams in FTP_RATIO_AUC. They perform relatively similar to other top-ranked teams in the FTN_RATIO_AUC score. Overall, they achieve the best performance for the Score across all three tumor entities. The main reason for them outperforming other teams for FTP_RATIO_AUC is how they developed their uncertainty generation method. They found that they achieved the best results on the given Score (Equation 3.1) by considering all positive predictions as certain (Section 3.3.3).

In terms of overall Scores, we observe that *Team SCAN* comfortably outperforms all other teams for each tumor entity. *Team QTIM* and *Team Uniandes* report better mean scores across different patients compared to *Team SCAN*. Despite this, they do not achieve an overall better ranking for each patient, which shows the usefulness of reporting ranking and statistical-significance analysis across different patients rather than just reporting mean overall Score across patients.

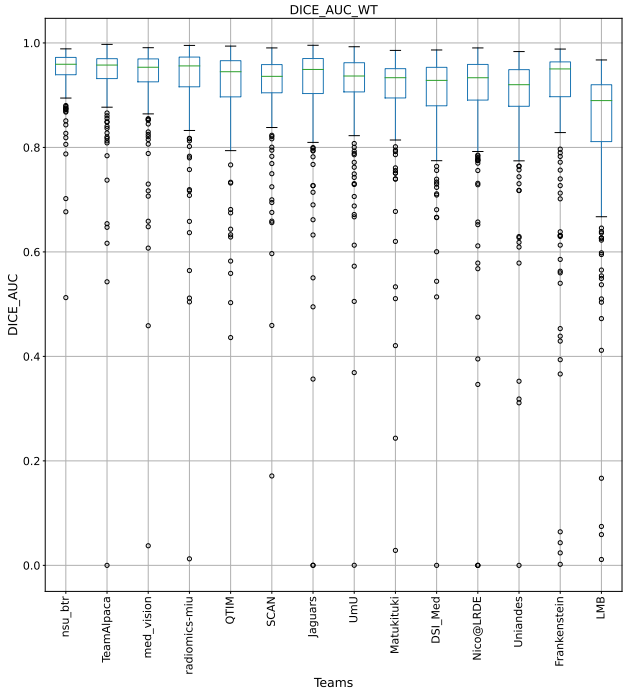


Figure B.1: QU-BraTS 2020 boxplots depicting DICE_AUC distribution for all teams across different participants for Whole Tumor on the BraTS 2020 test set (higher is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

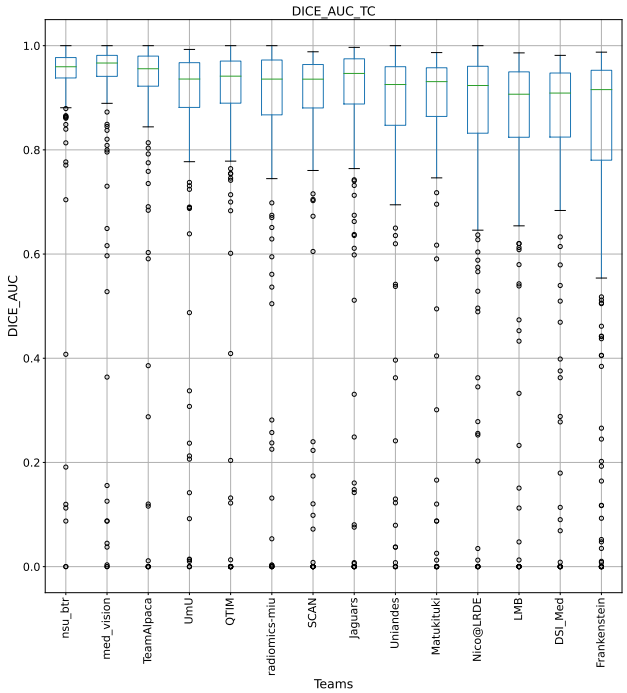


Figure B.2: QU-BraTS 2020 boxplots depicting DICE_AUC distribution for all teams across different participants for Tumor Core on the BraTS 2020 test set (higher is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

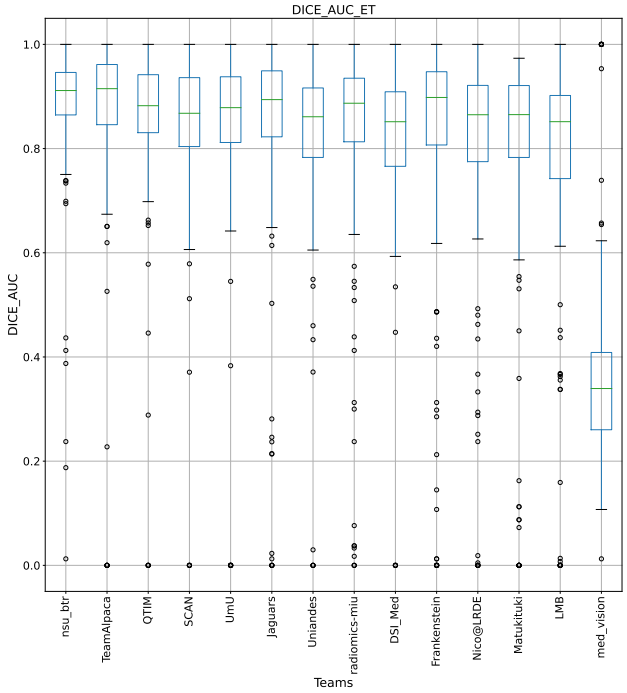


Figure B.3: QU-BraTS 2020 boxplots depicting DICE_AUC distribution for all teams across different participants for Enhancing Tumor on the BraTS 2020 test set (higher is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

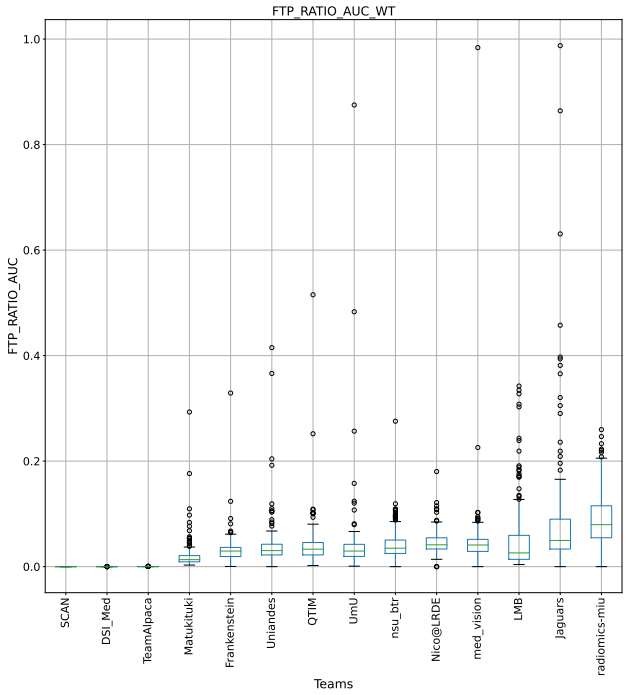


Figure B.4: QU-BraTS 2020 boxplots depicting FTP_RATIO_AUC distribution for all teams across different participants for Whole Tumor on the BraTS 2020 test set (lower is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

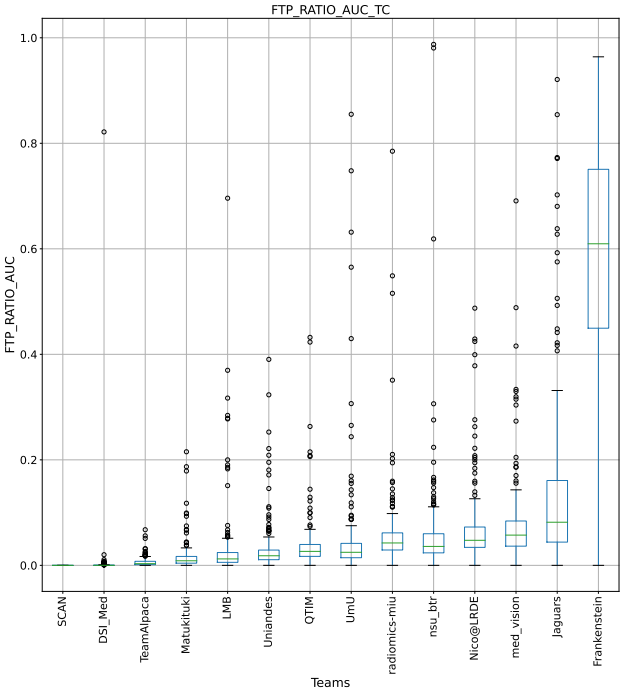


Figure B.5: QU-BraTS 2020 boxplots depicting FTP_RATIO_AUC distribution for all teams across different participants for Tumor Core on the BraTS 2020 test set (lower is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

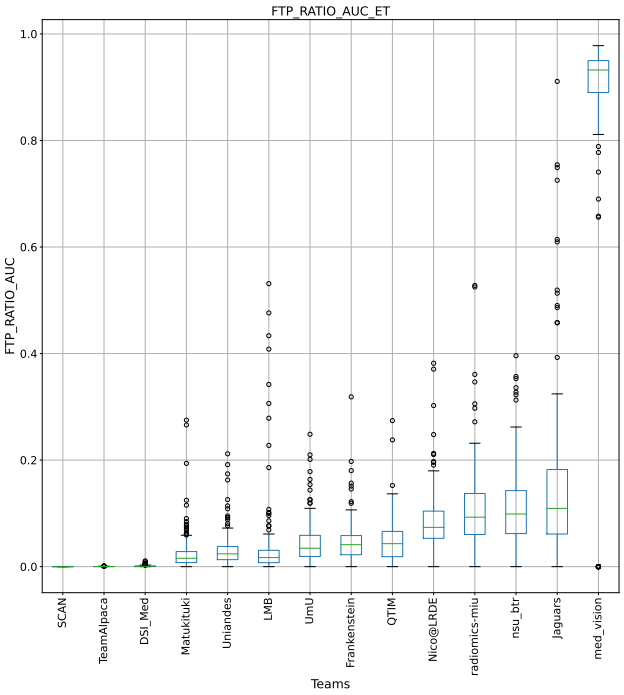


Figure B.6: QU-BraTS 2020 boxplots depicting FTP_RATIO_AUC distribution for all teams across different participants for Enhancing Tumor on the BraTS 2020 test set (lower is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

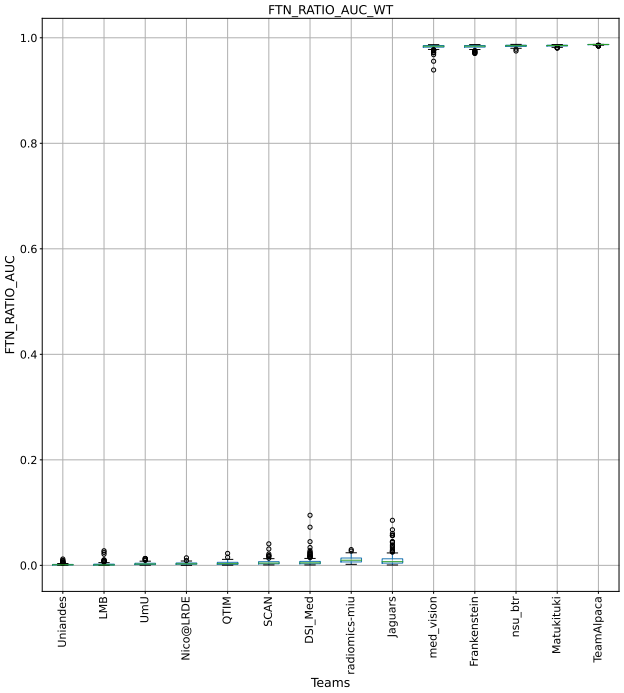


Figure B.7: QU-BraTS 2020 boxplots depicting FTN_RATIO_AUC distribution for all teams across different participants for Whole Tumor on the BraTS 2020 test set (lower is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

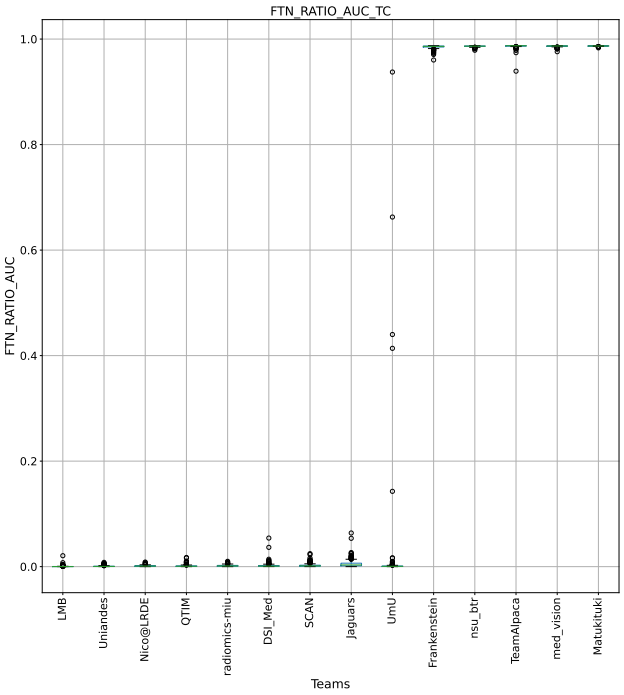


Figure B.8: QU-BraTS 2020 boxplots depicting FTN_RATIO_AUC distribution for all teams across different participants for Tumor Core on the BraTS 2020 test set (lower is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

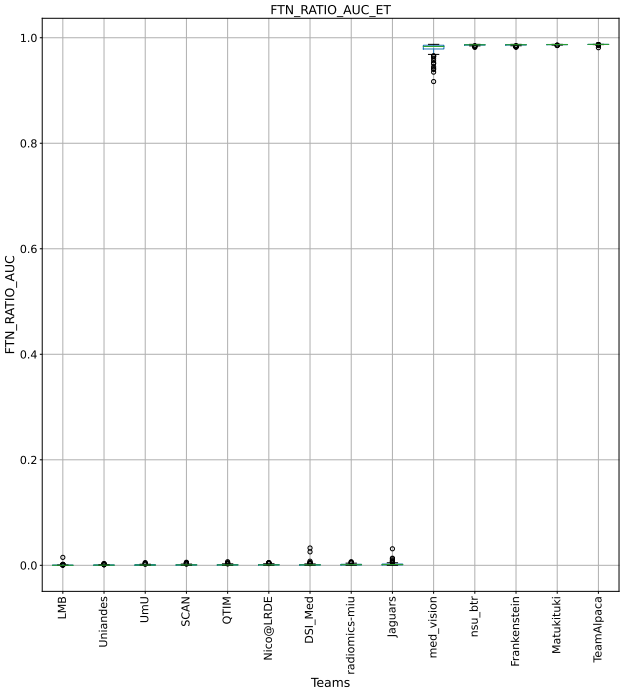


Figure B.9: QU-BraTS 2020 boxplots depicting FTN_RATIO_AUC distribution for all teams across different participants for Enhancing Tumor on the BraTS 2020 test set (lower is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

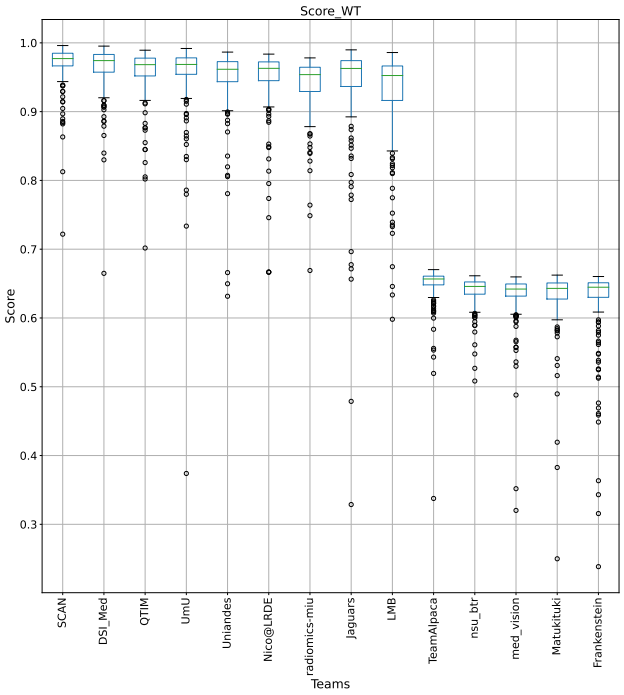


Figure B.10: QU-BraTS 2020 boxplots depicting Score distribution for all teams across different participants for Whole Tumor on the BraTS 2020 test set (higher is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

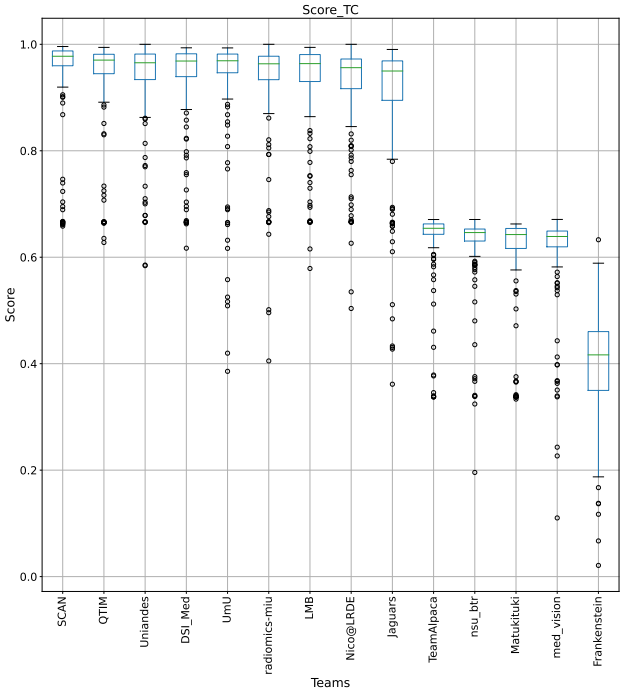


Figure B.11: QU-BraTS 2020 boxplots depicting Score distribution for all teams across different participants for Tumor Core on the BraTS 2020 test set (higher is better). [2022] CC-BY. Reprinted, with permission, from [161].

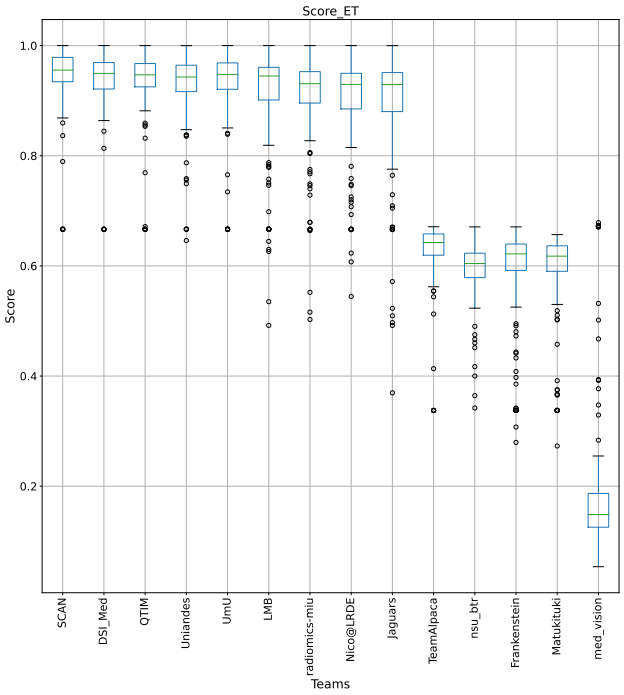


Figure B.12: QU-BraTS 2020 boxplots depicting Score distribution for all teams across different participants for Enhancing Tumor on the BraTS 2020 test set (higher is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

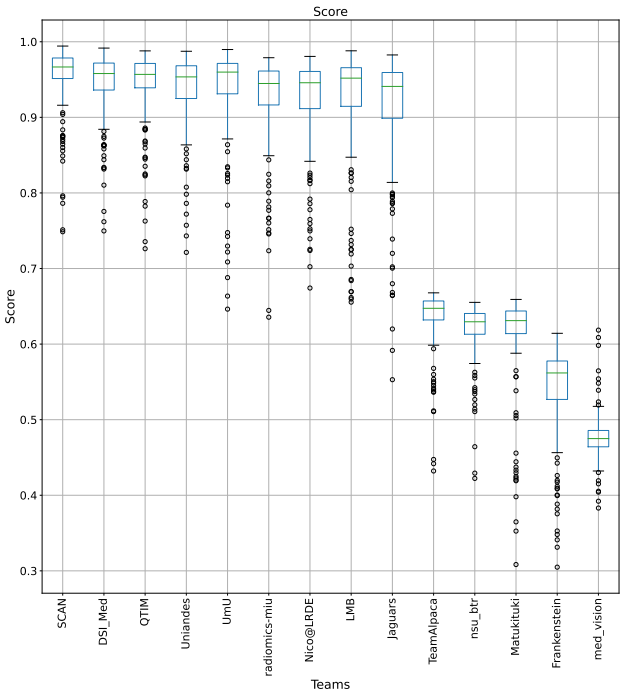


Figure B.13: QU-BraTS 2020 boxplots depicting overall Score distribution for all teams across different participants on the BraTS 2020 test set (higher is better). ©[2022] CC-BY. Reprinted, with permission, from [161].

B.2 QU-BraTS 2019

In this appendix, we analyze and briefly describe methods employed by participating teams in BraTS 2019 sub-challenge on uncertainty quantification. A total of 15 teams participated in the challenge. From these 15 teams, five teams further participated during the following QU-BraTS 2020 challenge.

BraTS 2019 dataset: As described in Section 3.2.1, BraTS 2019 dataset contains 335 patient MRIs in the training set, 125 in the validation set, and 166 in the testing set. All teams developed their method using the training set and the validation set. Ground truth segmentation for the validation set was not publicly available for the teams. The final performance of all teams was measured on the testing set, where each team had access to a 48-hour window to upload their result to the server (https://ipp.cbica.upenn.edu/).

QU-BraTS 2019 results on the test set: We ran the task of uncertainty quantification preliminary during the challenge and did not employ any ranking scheme. Also, the score used during the challenge was different from the one described in Section 3.2. Precisely, we did not calculate the AUC of Ratio of Filtered True Negatives vs. Uncertainty threshold until the validation phase was ended; and only used AUCs of DSC vs. Uncertainty Threshold and Ratio of Filtered True Positives vs. Uncertainty Threshold. After the validation phase, using qualitative inspection, we found that many teams were employing 1 - softmax confidence as an uncertainty measure, which is not helpful from a real clinical point of view as described in Section 3.2 and Section 3.4.3. Keeping this in mind, we added the AUC of Ratio of Filtered True Negatives vs. Uncertainty threshold during the final testing phase. Table B.1 lists all team names and their performance on the BraTS 2019 test phase. The table shows that teams that employed 1 - softmax_confidence as uncertainty measure performed poorly on FTN_RATIO_AUC score (Ex. *Team Alpaca*, *Team DRAG*, *Team ODU_vision_lab*, etc.). We want to point out that we did not employ the ranking strategy used in the QU-BraTS 2020 challenge during the QU-BraTS 2019 challenge. As we discussed in Appendix A, the ranking strategy and statistical significance analysis reflect the true potential of the method compared to just ranking teams according to their mean performance across testing cases.

Table B.1: Final performance on the BraTS 2019 testing dataset for teams participating in the preliminary challenge on quantification of uncertainty in brain tumor segmentation task. Here, mean values for each score across all patient in the testing dataset is listed.

Team	#cases	DICE_AUC			FTP_RATIO_AUC			FTN_RATIO_AUC			Score			overall Score
		WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	
SCAN [154]	166	0.8837	0.8253	0.8209	0.0358	0.0771	0.14958	0.01919	0.0076	0.0060	0.9429	0.9135	0.8885	0.9150
RADIOMICS-MIU [25]	166	0.8595	0.8122	0.7759	0.0421	0.0906	0.12009	0.00380	0.0012	0.0008	0.9379	0.9068	0.8850	0.9099
UmU [262]	166	0.8520	0.8077	0.7892	0.0602	0.1229	0.14089	0.00334	0.0150	0.0010	0.9295	0.8899	0.8824	0.9006
xuefeng [68]	166	0.8746	0.8432	0.8120	0.0894	0.1642	0.27216	0.00969	0.0049	0.0024	0.9252	0.8914	0.8458	0.8874
UTintelligence [8]	162	0.7800	0.6787	0.6688	0.0117	0.0528	0.12901	0.00000	0.0000	0.0000	0.9228	0.8753	0.8466	0.8816
NVDLMED [179]	166	0.8651	0.8203	0.8251	0.0213	0.0679	0.10958	0.49326	0.3883	0.2701	0.7835	0.7881	0.8151	0.7956
FightGliomas	166	0.8275	0.7783	0.4583	0.3172	0.2312	0.51028	0.00239	0.0008	0.0007	0.8360	0.8488	0.6491	0.7779
NIC-VICOROB	166	0.3077	0.6883	0.6393	0.5380	0.0458	0.08012	0.00000	0.0000	0.0000	0.5899	0.8808	0.8531	0.7746
LRDE_2 [32]	166	0.8851	0.8387	0.7725	0.5930	0.7017	0.26159	0.05312	0.0439	0.0196	0.7463	0.6977	0.8304	0.7581
LRDE_VGG [32]	166	0.8810	0.7883	0.6303	0.4930	0.7313	0.83645	0.04460	0.0280	0.0185	0.7812	0.6764	0.5918	0.6831
ANSIR	166	0.8727	0.8551	0.8349	0.0124	0.0765	0.11249	0.92500	0.9250	0.9250	0.6451	0.6179	0.5992	0.6207
med_vision [195]	166	0.8794	0.8512	0.8491	0.0203	0.0768	0.13209	0.92435	0.9253	0.9257	0.6449	0.6164	0.5971	0.6195
TEAM_ALPACA [178]	166	0.8768	0.8377	0.8116	0.0191	0.0707	0.10695	0.91639	0.9170	0.9228	0.6471	0.6167	0.5940	0.6192
ODU_vision_lab	166	0.8789	0.8517	0.8481	0.0212	0.0776	0.13283	0.92444	0.9253	0.9257	0.6444	0.6162	0.5965	0.6191
DRAG [18]	161	0.8890	0.8518	0.8105	0.0726	0.1312	0.13792	0.92280	0.9241	0.9243	0.6312	0.5989	0.5828	0.6043

C

Appendix: Propagating Uncertainty Across Cascaded Medical Imaging Tasks

C.1 Implementation Details

In section, we provide details about the network architecture, implementation details and the training process for all three pipelines explored in Chapter 4: Multiple Sclerosis lesion segmentation/detection (Section 4.2.1), brain tumour segmentation (Section 4.2.2), and Alzheimer's disease clinical score prediction (Section 4.2.3). Note that all our experiments were implemented using PyTorch, and ran on a machine equipped with an NVIDIA Titan Xp GPU with 12 GBs of memory.

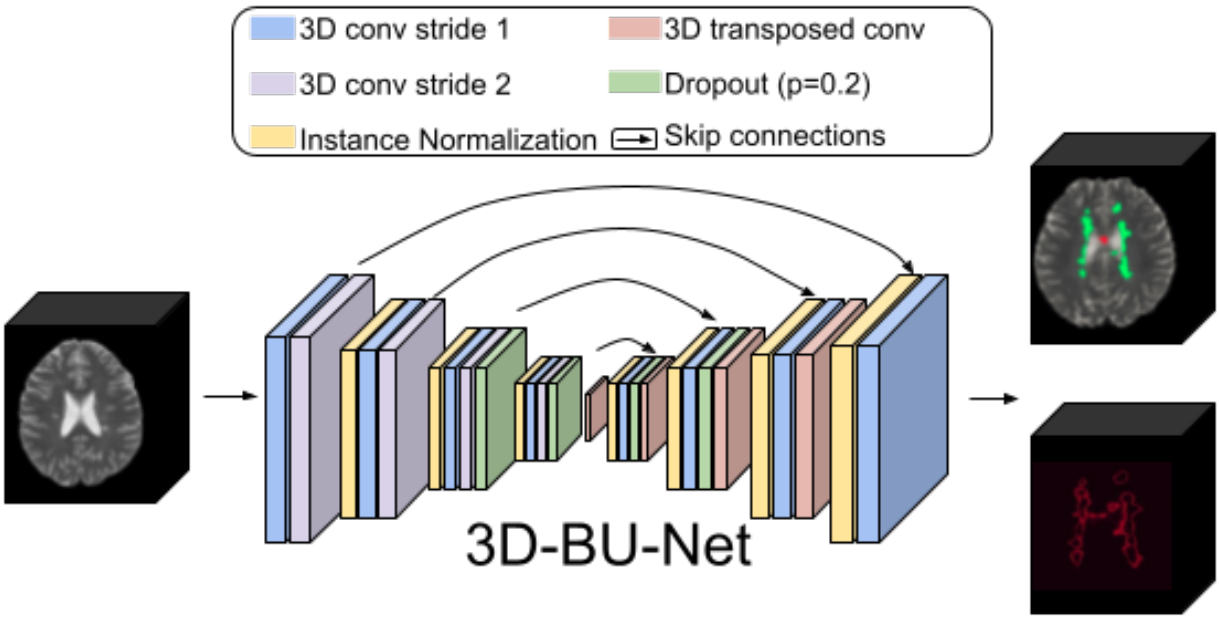


Figure C.1: Network architecture diagram for the BU-Net [180]. BU-Net provides the segmentation outputs and permits the estimation of the uncertainties associated with the outputs. BU-Net was used for both Task-1 and Task-2 in the MS lesion segmentation/detection pipeline depicted here and as a Task-1 network for hippocampus segmentation in the Alzheimer's Disease clinical score prediction pipeline. ©[2022] IEEE. Reprinted, with permission, from [159].

C.1.1 MS T2 Lesion Segmentation Detection

The pipeline (Section 4.2.1) consists of a cascade of two binary lesion segmentation tasks. We chose an off-the-shelf BU-Net [180] architecture[1] for both Task-1 and Task-2 networks, which can be seen in Figure C.1. The only differences between the two networks were their inputs. For the Task-1 network, the inputs consisted of all the MR sequences. The Task-2 network takes as input the MR sequences, the Task-1 network output, and the uncertainties associated with the Task-1 network output (in the case of the proposed framework). These additional inputs marginally increase the total number of parameters for the Task-2 network. For exact architecture details, readers can refer to the BU-Net [180] paper.

Both the Task-1 and Task-2 networks were trained to minimize a weighted binary cross-entropy loss function for the lesion segmentation task. Here, class weights were taken as an inverse of the frequency of both lesion/non-lesion voxels within the brain mask. After

[1]We reimplemented the model architecture in PyTorch following the code (link) provided by the authors.

every epoch, class weights were decayed with a factor of 0.95, which results in equally weighted binary cross-entropy after around 50 epochs. The networks were trained using an Adam optimizer with an initial learning rate of 0.0002 and a weight decay of 0.00001 for a total of 250 epochs. The learning rate was decayed with a factor of 0.995 after each epoch.

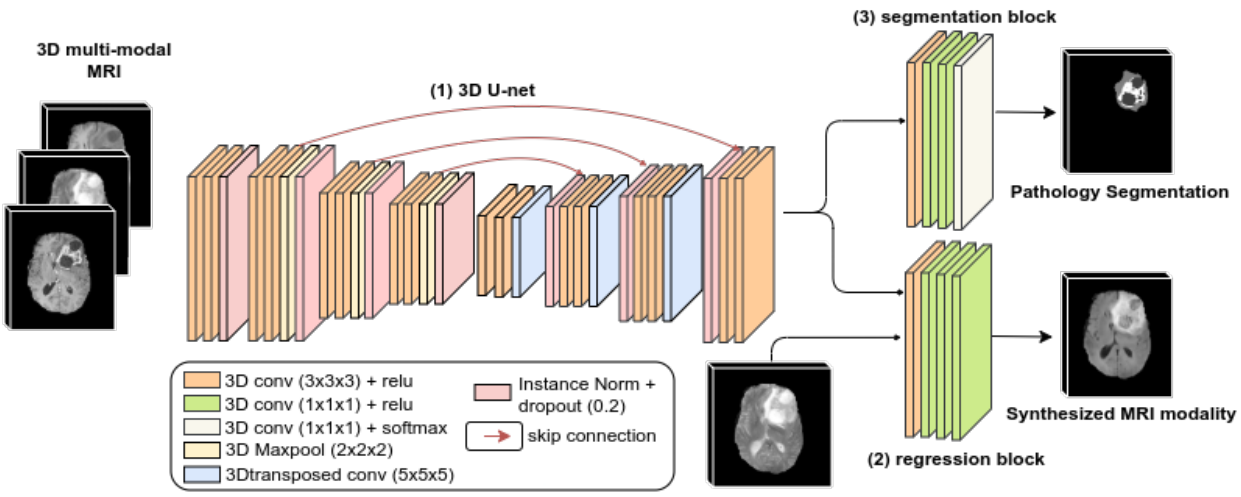


Figure C.2: Network architecture diagram of RS-Net [158]. We use RS-Net for the synthesis of the missing MRI sequence synthesis (Task-1) in the brain tumour segmentation pipeline. Note that T1, T2, and T1ce are used as inputs to the network when synthesizing FLAIR, while T1, T2, and FLAIR are used as inputs when synthesizing T1ce. ©[2022] IEEE. Reprinted, with permission, from [159].

C.1.2 Brain Tumour Segmentation

The pipeline (Section 4.2.2) consists of two different sequential inference tasks. The first network (Task-1) is designed for a 3-to-1 synthesis of a missing MRI sequence in the presence of a brain tumour. RS-Net[2] [158] was chosen for this task and can be seen in Figure C.2. RS-Net is a multi-task network designed to jointly perform the synthesis of the missing image while performing the segmentation of the tumour, with the premise that this would improve the synthesis in the tumor area. RS-Net was trained for a total of 400 epochs using an Adam optimizer with a learning rate of 0.0002 and a weight decay of 0.00001. The learning rate was decayed with a factor of 0.995 after each epoch. The network was trained to minimize a combined weighted mean squared error and weighted cross-entropy loss [158].

A modified 3D U-Net [45], depicted in Figure C.3, was developed for multi-class brain tumour segmentation (Task-2 Network). Similar to the original 3D U-Net, the network consists of encoder and decoder paths that embed convolution, pooling, and deconvolution operations. High-resolution features from the encoder path were combined with

[2]Readers are requested to refer RS-Net paper for the exact network architecture details.

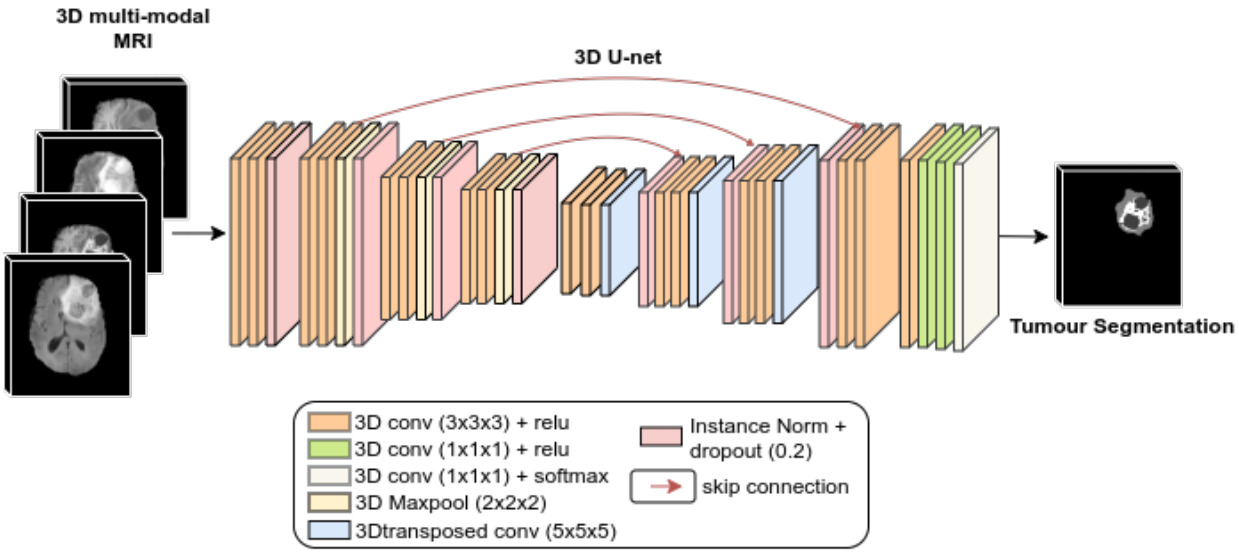


Figure C.3: Network architecture diagram of the modified 3D-U-Net [45], used for the multi-class brain tumour segmentation (Task-2) in the brain tumour segmentation pipeline. The inputs to this network vary depending on the experiment. For example, when assessing the effectiveness of uncertainty propagation, we also pass the uncertainties associated with the synthesized MR sequence as input to the network. ©[2022] IEEE. Reprinted, with permission, from [159].

the up-sampled output of the decoder to preserve high-resolution features. Each convolution was followed by rectified linear unit activation (ReLU). Instead of using the batch-normalization layer used in the original U-Net, we used instance normalization [256]. Instance normalization typically improves performance for small batch sizes. The network was trained using Adam optimizer with a learning rate of 0.0002 and weight decay of 0.00001 for a total of 240 epochs to minimize weighted cross-entropy loss. Here, the weights are defined such that the weight increases whenever there are fewer voxels in a particular class. After every epoch, class weights were decayed with a factor of 0.95, which results in equally weighted binary cross-entropy after around 50 epochs. Inputs to the network varies depending on the experiment. For example, in the proposed framework 3D U-Net takes as input the real MR sequences, the RS-Net synthesized MR sequence, and the uncertainties associated with the synthesized MR sequence. These additional inputs result in a marginal increase in the total number of parameters.

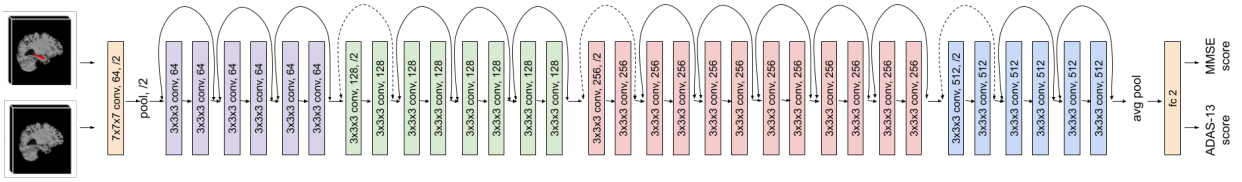


Figure C.4: Network architecture diagram of modified 3D-ResNet-34 [45] for the Alzheimer's Disease clinical regression pipeline for predicting both ADAS-13 and MMSE scores. In our framework, input to this network varies depending on the experiment. For example, when assessing the effectiveness of uncertainty propagation, uncertainties associated with the hippocampus segmentation is also provided as input to the network. ©[2022] IEEE. Reprinted, with permission, from [159].

C.1.3 Alzheimer's Disease Clinical Score Prediction

The pipeline described in Section 4.2.3 consists of two cascaded inference tasks. The BU-Net [180] was used for the binary hippocampus segmentation task (Figure C.1). The T1 weighted MRI was provided as an input to the BU-Net. The network was trained to reduce the weighted binary cross-entropy loss using an Adam optimizer with a learning rate of 0.0002 and a weight decay of 0.00001 for a total of 250 epochs. Here, class weights were taken as an inverse of the frequency of both hippocampus/background voxels within the brain mask. The learning rate was decayed with a factor of 0.995 after each epoch. After every epoch, the class weights were decayed with a factor of 0.95, which results in equally weighted binary cross-entropy after around 50 epochs.

A 3D ResNet34 [91] architecture was designed for the task of clinical score prediction (Task-2) [3]. The network (Figure C.4) was modified to be a multi-task network, such that it predicts both ADAS-13 and MMSE scores simultaneously. The network was trained to reduce the combined mean squared error losses for both ADAS-13 and MMSE. An Adam optimizer with a learning rate of 0.0002 and a weight decay of 0.00001 was used to train the network for a total of 200 epochs. The learning rate was decayed with a factor of 0.995 after each epoch.

[3] https://github.com/kenshohara/3D-ResNets-PyTorch/blob/master/models/resnet.py

C.2 Additional Results for MS Lesion Segmentation

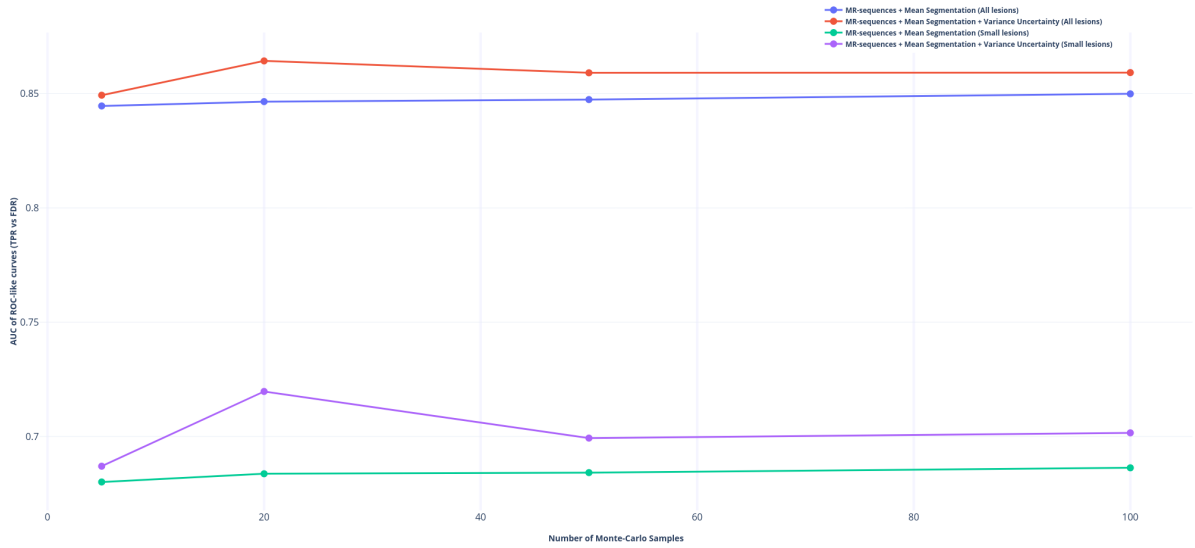


Figure C.5: Comparing overall MS T2 lesion detection performance using Area Under Curve (AUC) of ROC-like curves, illustrating TPR (true positive rate) vs. FDR (false detection rate) across all lesions, and small lesions (3-10 voxels). Here we evaluate the impact of number of samples used to estimate uncertainty (variance) measure for MC-Dropout uncertainty estimation method. From the plot we can see that for all lesion detection and small lesion detection, highest performance is achieved when 20 samples are used to estimate uncertainty. With increase in number of samples, performance saturates. ©[2022] IEEE. Reprinted, with permission, from [159].

D

Appendix: Evaluating the Fairness of Deep Learning Uncertainty Estimates in Medical Image Analysis

D.1 Multi-Class Skin Lesion Classification - Sensitive Attribute: Sex

We use sex as a sensitive attribute for experiments in this section. Specifically, we divide the ISIC dataset into two subsets based on the sex associated with each image (male vs female). The entire dataset is divided into two subsets: patient images from female patients

in subgroup D^0 with a total of 11661 images, and patient images from male patients in subgroup D^1 with a total of 13286 images.

Table D.1: Number of images for each class and each subgroup for the whole ISIC dataset. From this, we can see a high-class imbalance across different classes. Similarly, distribution across both subgroups for a particular class is also different. For example, while for Melanoma, Basal Cell Carcinoma, Actinic Keratosis, Benign Keratosis, and Squamous Cell Carcinoma, D^0 has a higher number of samples compared to D^1, for the rest of the classes (Melanocytic Nevus, Dermatofibroma, and Vascular Lesion) D^1 has a higher number of samples compared to D^0. ©[2023] PMLR. Reprinted, with permission, from [164].

	ISIC Dataset								
	Melanoma	**Melanocytic Nevus**	**Basal Cell Carcinoma**	**Actinic Keratosis**	**Benign Keratosis**	**Dermatofibroma**	**Vascular Lesion**	**Squamous Cell Carcinoma**	**Total**
D^0	1980	6379	1317	406	1134	117	125	203	11661
D^1	2461	6225	2000	458	1467	122	128	425	13286
Overall	4441	12604	3317	864	2601	239	253	628	24947

Table D.2: Number of images for each class and each subgroup for the training dataset used to train the **Baseline-Model** and the **GroupDRO-Model**. Similar to the whole ISIC dataset (Table-D.1), we see high-class imbalance across different classes, and different distributions across both subgroups for a particular class. ©[2023] PMLR. Reprinted, with permission, from [164].

	Training Dataset (Baseline-Model and GroupDRO-Model)								
	Melanoma	**Melanocytic Nevus**	**Basal Cell Carcinoma**	**Actinic Keratosis**	**Benign Keratosis**	**Dermatofibroma**	**Vascular Lesion**	**Squamous Cell Carcinoma**	**Total**
D^0	1248	4061	830	257	715	73	78	128	7390
D^1	1680	3922	1445	303	1015	78	81	328	8852
Overall	2928	7983	2275	560	1730	151	159	456	16242

Table D.3: Number of images for each class and each subgroup for the training dataset used to train the **Balanced-Model**. Compared to the training dataset used for the **Baseline-Model** and the **GroupDRO-Model** (Table-D.2), we balance the number of samples across both subgroups, but we do not balance across different classes. ©[2023] PMLR. Reprinted, with permission, from [164].

	Training Dataset (Balanced-Model)								
	Melanoma	**Melanocytic Nevus**	**Basal Cell Carcinoma**	**Actinic Keratosis**	**Benign Keratosis**	**Dermatofibroma**	**Vascular Lesion**	**Squamous Cell Carcinoma**	**Total**
D^0	1248	3922	830	257	715	73	78	128	7251
D^1	1248	3922	830	257	715	73	78	128	7251
Overall	2496	7844	1660	514	1430	146	156	256	14502

Table D.4: Number of images for each class and each subgroup in the Validation dataset for all three models (the **Baseline-Model** and the **GroupDRO-Model**, and the **Balanced-Model)**. The distribution of samples across both subgroups and across different classes is similar to the Table-D.1. ©[2023] PMLR. Reprinted, with permission, from [164].

	Validation Dataset (Baseline-Model, GroupDRO-Model, and Balanced-Model)								
	Melanoma	**Melanocytic Nevus**	**Basal Cell Carcinoma**	**Actinic Keratosis**	**Benign Keratosis**	**Dermatofibroma**	**Vascular Lesion**	**Squamous Cell Carcinoma**	**Total**
D^0	138	451	92	28	79	8	9	14	819
D^1	187	436	160	34	112	8	9	36	982
Overall	325	887	252	62	191	16	18	50	1801

Table D.5: Number of images for each class and each subgroup in the Testing dataset used to test all three models (the **Baseline-Model** and the **GroupDRO-Model**, and the **Balanced-Model)**. The distribution of samples across both subgroups is kept similar, but it is not similar across different classes. We kept similar distribution across both subgroups for a fair comparison of their performance, while the distribution across different classes was not kept similar to reflect real-world scenarios where some classes can be more frequent compared to others. ©[2023] PMLR. Reprinted, with permission, from [164].

	Testing Dataset (Baseline-Model, GroupDRO-Model, and Balanced-Model)								
	Melanoma	**Melanocytic Nevus**	**Basal Cell Carcinoma**	**Actinic Keratosis**	**Benign Keratosis**	**Dermatofibroma**	**Vascular Lesion**	**Squamous Cell Carcinoma**	**Total**
D^0	594	1867	395	121	340	36	38	61	3452
D^1	594	1867	395	121	340	36	38	61	3452
Overall	1188	3734	790	242	680	72	76	122	6904

Table D.6: Overall metrics (AUC, Accuracy, and Balanced-Accuracy) for a **Baseline-Model** trained on the ISIC dataset at $\tau = 100$. ©[2023] PMLR. Reprinted, with permission, from [164].

Baseline-Model	**AUC**	**Accuracy**	**Balanced-Accuracy**
D^0	96.24	83.02	71.77
D^1	96.83	83.02	70.23
Fairness Gap	0.59	0.00	1.54

Table D.7: Overall metrics (AUC, Accuracy, and Balanced-Accuracy) for a **Balanced-Model** trained on the ISIC dataset at $\tau = 100$. ©[2023] PMLR. Reprinted, with permission, from [164].

Balanced-Model	**AUC**	**Accuracy**	**Balanced-Accuracy**
D^0	96.26	82.24	70.26
D^1	95.92	81.66	69.42
Fairness Gap	0.34	0.58	0.74

Table D.8: Overall metrics (AUC, Accuracy, and Balanced-Accuracy) for a **GroupDRO-Model** trained on the ISIC dataset at $\tau = 100$. ©[2023] PMLR. Reprinted, with permission, from [164].

GroupDRO-Model	**AUC**	**Accuracy**	**Balanced-Accuracy**
D^0	95.76	80.56	70.25
D^1	96.31	80.59	69.90
Fairness Gap	0.55	0.03	0.35

Table D.9: Per class accuracy for a **Baseline-Model** trained on the ISIC dataset at $\tau = 100$. ©[2023] PMLR. Reprinted, with permission, from [164].

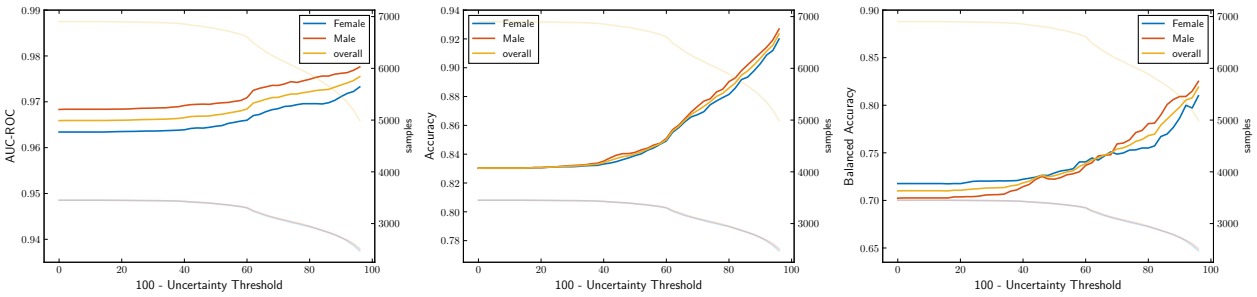
Baseline-Model	**Class-level Accuracy**							
	Melanoma	**Melanocytic Nevus**	**Basal Cell Carcinoma**	**Actinic Keratosis**	**Benign Keratosis**	**Dermatofibroma**	**Vascular Lesion**	**Squamous Cell Carcinoma**
D^0	65.32	91.64	85.06	61.16	80.29	63.88	71.05	55.74
D^1	73.91	91.27	87.09	52.07	68.53	44.44	92.11	52.46
Fairness Gap	8.59	0.37	2.03	9.09	11.76	19.44	21.06	3.28

Table D.10: Per class accuracy for a **Balanced-Model** trained on the ISIC dataset at $\tau = 100$. ©[2023] PMLR. Reprinted, with permission, from [164].

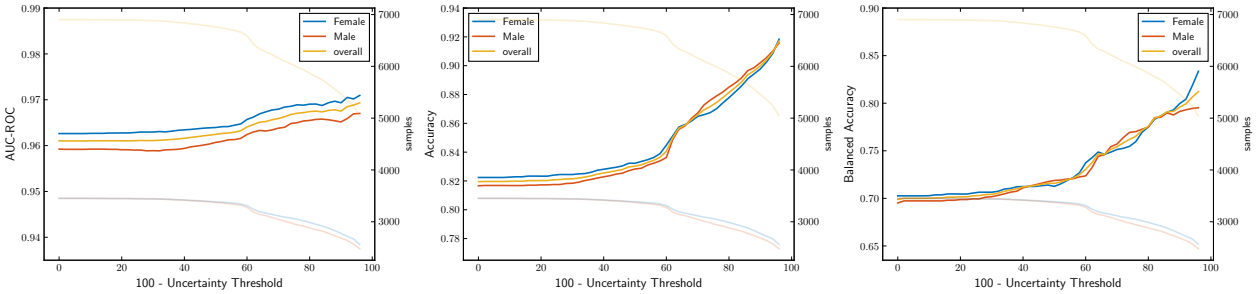
Balanced-Model	**Class-level Accuracy**							
	Melanoma	**Melanocytic Nevus**	**Basal Cell Carcinoma**	**Actinic Keratosis**	**Benign Keratosis**	**Dermatofibroma**	**Vascular Lesion**	**Squamous Cell Carcinoma**
D^0	62.79	92.34	88.35	59.50	70.88	66.67	78.95	42.62
D^1	68.52	90.95	87.59	54.55	65.88	61.11	94.74	32.79
Fairness Gap	5.73	1.39	0.76	4.95	5.00	5.56	15.79	9.83

Table D.11: Per class accuracy for a **GroupDRO-Model** trained on the ISIC dataset at $\tau = 100$. ©[2023] PMLR. Reprinted, with permission, from [164].

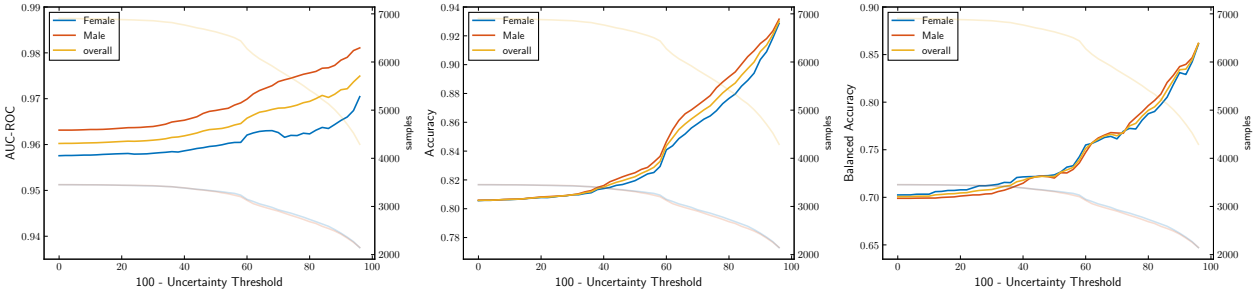
GroupDRO-Model	**Class-level Accuracy**							
	Melanoma	**Melanocytic Nevus**	**Basal Cell Carcinoma**	**Actinic Keratosis**	**Benign Keratosis**	**Dermatofibroma**	**Vascular Lesion**	**Squamous Cell Carcinoma**
D^0	55.05	92.07	83.29	66.12	70.29	55.56	78.95	60.66
D^1	63.13	90.52	82.28	59.50	68.24	50.00	81.58	63.93
Fairness Gap	8.08	1.55	1.01	6.62	2.05	5.56	2.63	3.27



(a) Baseline Model

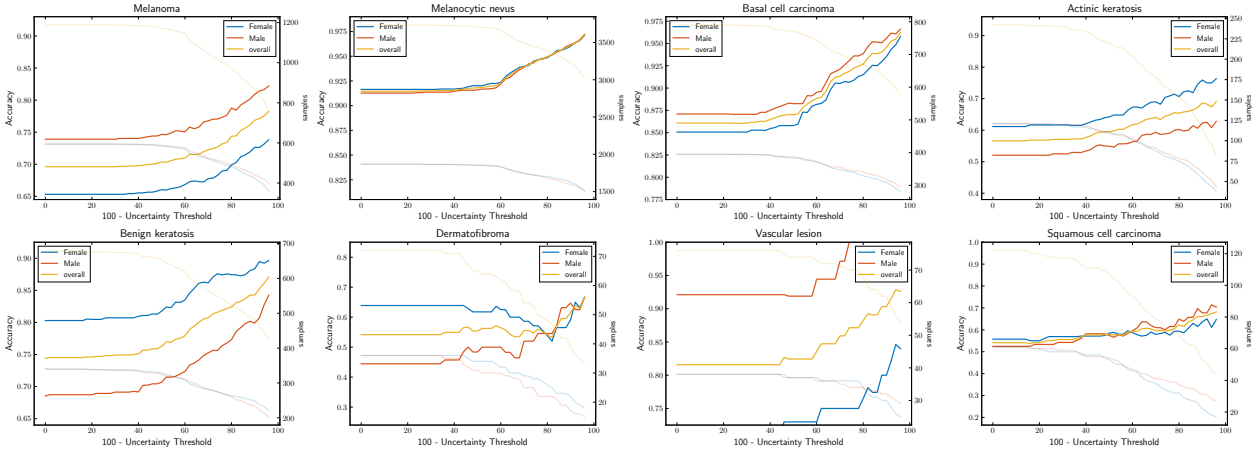


(b) Balanced Model

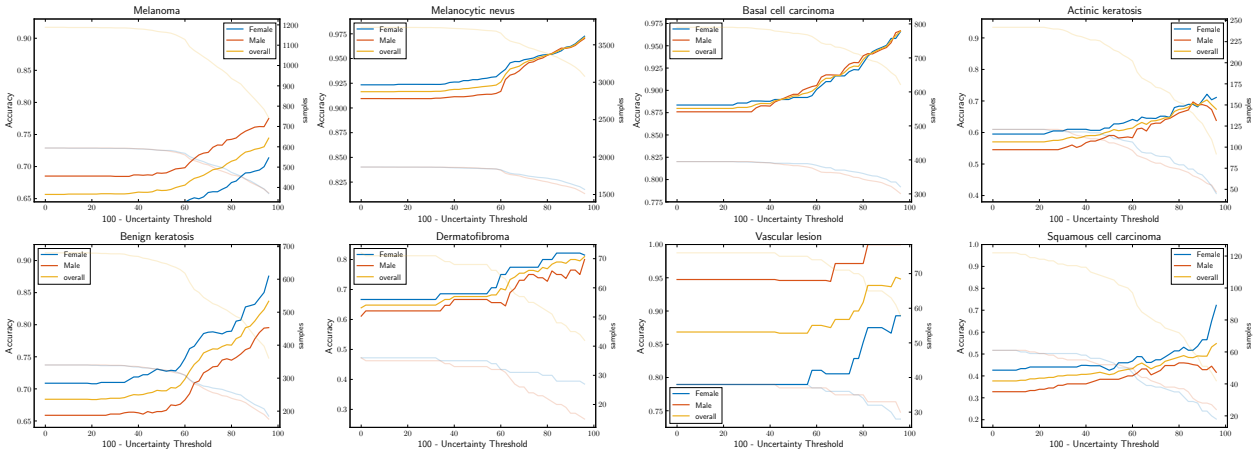


(c) GroupDRO Model

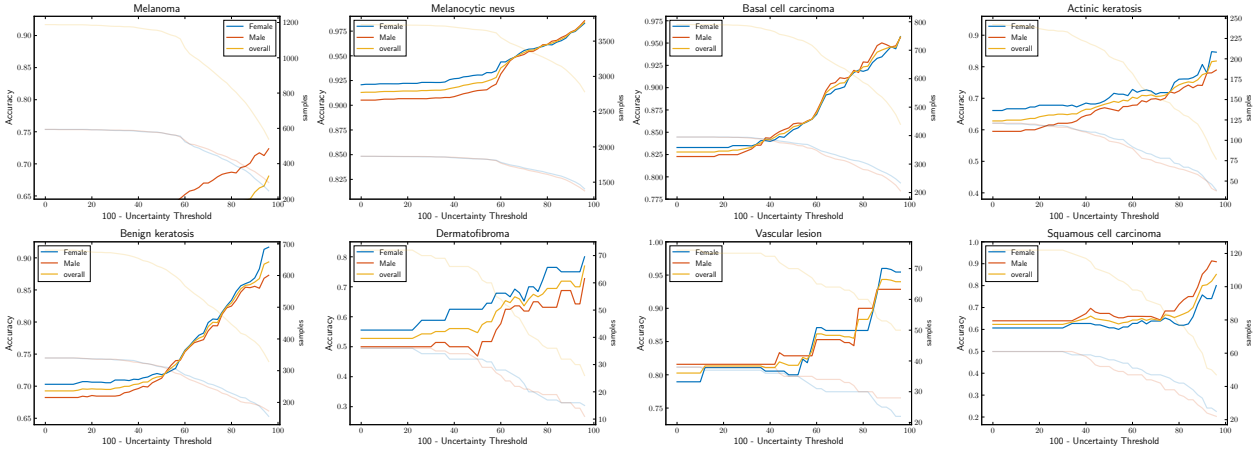
Figure D.1: **ISIC-Sex:** Overall AUC, accuracy, and Balanced Accuracy as a function of uncertainty threshold for (a) **Baseline-Model**, (b) **Balanced-Model**, and (c) **GroupDRO-Model** on the ISIC dataset. In addition to metrics, the total number of testing images for each subgroup (D^0 - Female and D^1 - Male) are shown as light colours. ©[2023] PMLR. Reprinted, with permission, from [164].



(a) Baseline Model



(b) Balanced Model



(c) GroupDRO Model

Figure D.2: **ISIC-Sex:** Class-level accuracy as a function of uncertainty threshold for (a) **Baseline-Model**, (b) **Balanced-Model**, and (c) **GroupDRO-Model** on the ISIC dataset. In addition to the accuracy, the total number of testing images for each subgroup (D^0 - Female and D^1 - Male) are shown as light colours. ©[2023] PMLR. Reprinted, with permission, from [164].

D.2 Brain Tumour Segmentation

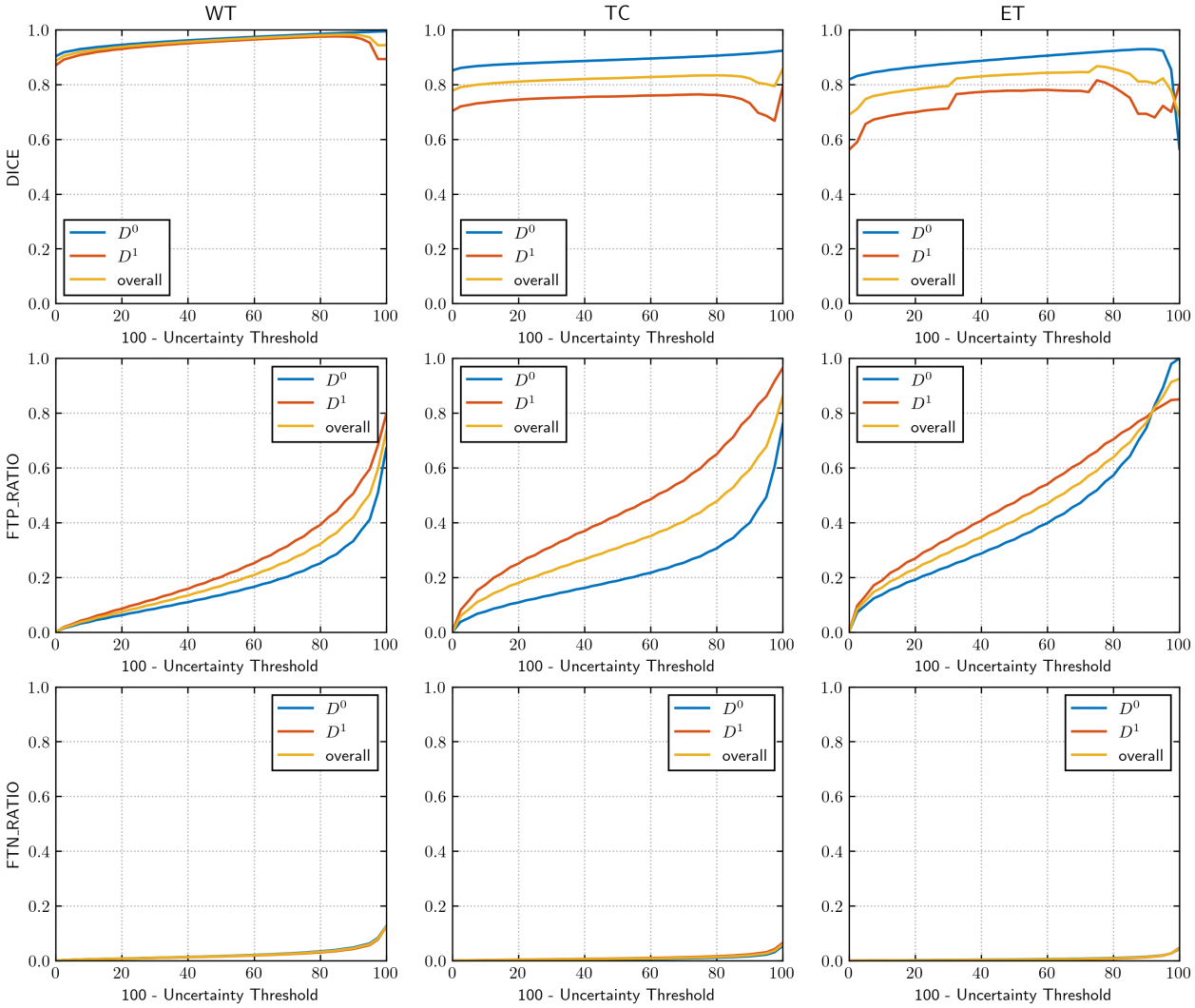


Figure D.3: **BraTS:** Dice, Filtered True Positive Ratio (FTP), and Filtered True Negative Ratio (FTN) as a function of uncertainty threshold for **Baseline-Model** on the BraTS dataset. Specifically, we plot Whole Tumour (WT), Tumour Core (TC), and Enhancing Tumour (ET) QU-BraTS [162] metrics for both the D^0 and D^1 set. ©[2023] PMLR. Reprinted, with permission, from [164].

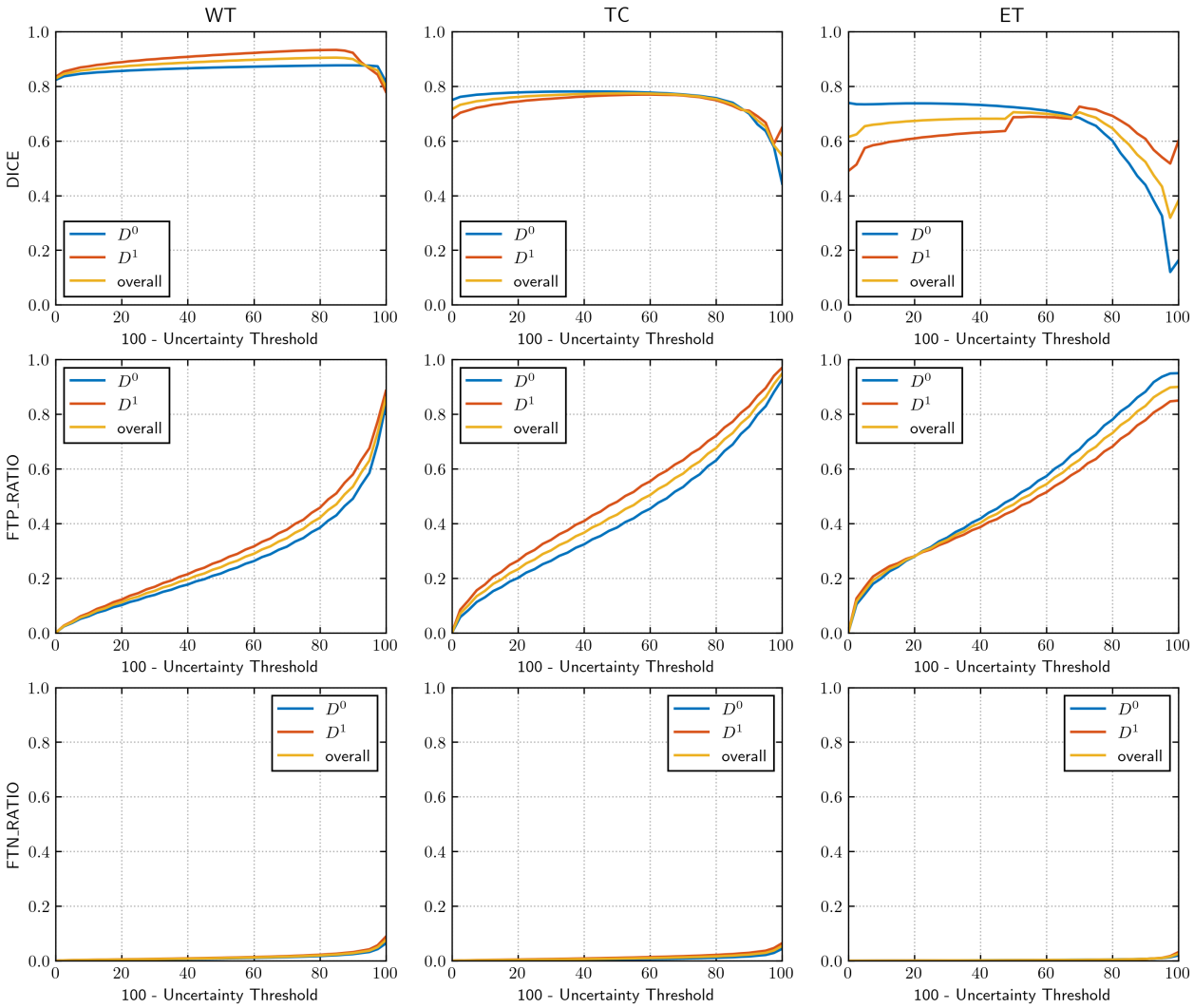


Figure D.4: **BraTS:** Dice, Filtered True Positive Ratio (FTP), and Filtered True Negative Ratio (FTN) as a function of uncertainty threshold for **Balanced-Model** on the BraTS dataset. Specifically, we plot Whole Tumour (WT), Tumour Core (TC), and Enhancing Tumour (ET) QU-BraTS [162] metrics for both the D^0 and D^1 set. ©[2023] PMLR. Reprinted, with permission, from [164].

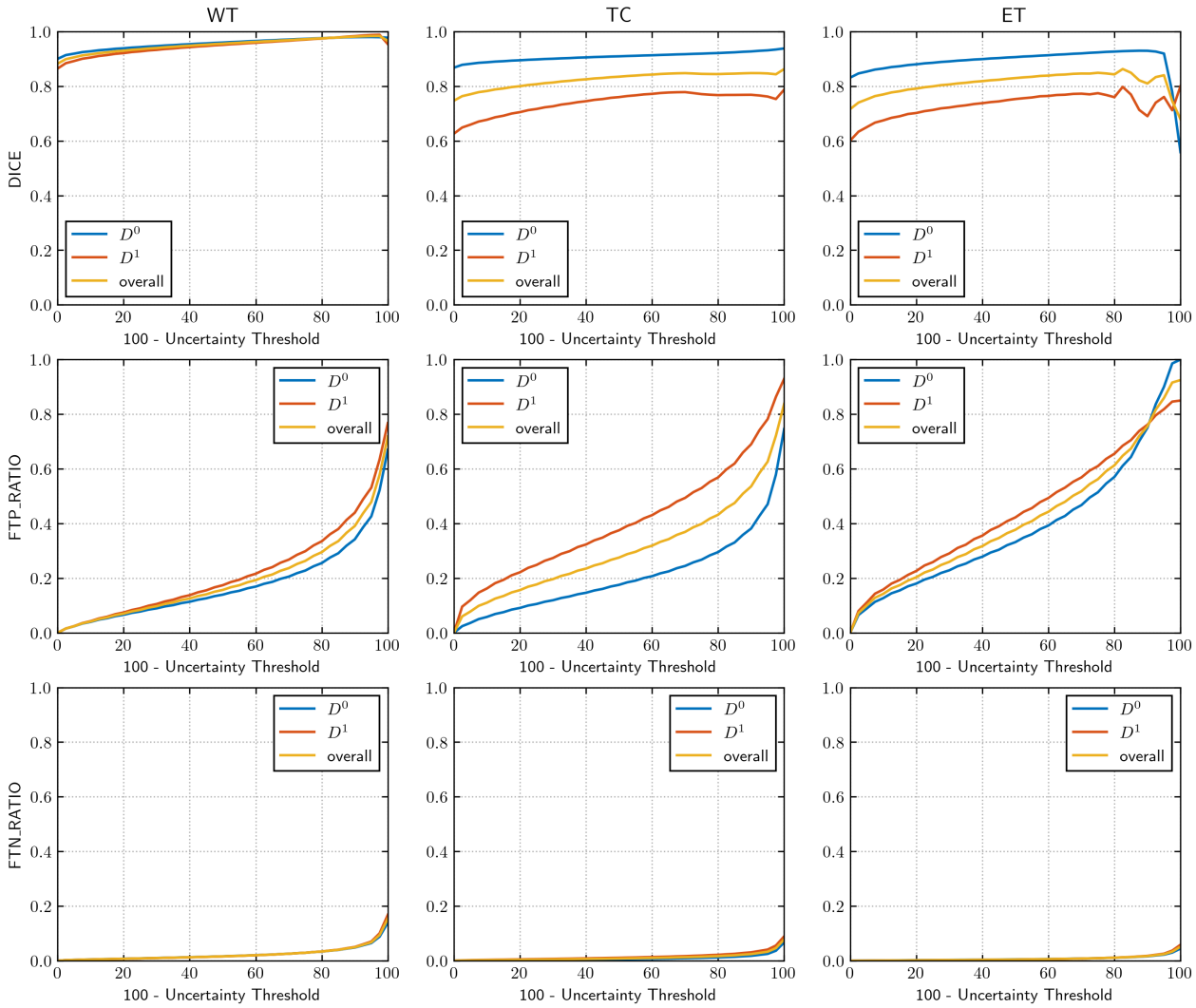


Figure D.5: **BraTS:** We plot Dice, Filtered True Positive Ratio (FTP), and Filtered True Negative Ratio (FTN) as a function of uncertainty threshold for **GroupDRO-Model** on the BraTS dataset. Specifically, we plot Whole Tumour (WT), Tumour Core (TC), and Enhancing Tumour (ET) QU-BraTS [162] metrics for both the D^0 and D^1 set. ©[2023] PMLR. Reprinted, with permission, from [164].

D.2.1 Brain Tumour Segmentation - Sensitive Attribute: Imaging Centre

In this section, We use the 260 High-Grade Glioma (HGG) images from the publicly available Brain Tumour Segmentation (BraTS) 2019 challenge dataset. The image dataset is divided into two subsets based on the imaging center. Specifically, images coming from TCIA subset were considered in subgroup D^0, while images from the rest of the imaging center were considered in subgroup D^1. A Baseline-Model and a GroupDRO-Model are trained on a dataset of 74 samples from D^0 and 124 samples from D^1. While a Balanced-Model is trained on a balanced training set with 74 samples from each subgroup.

Table D.12: Number of samples in both D^0 and D^1 subgroups for five different datasets: (i) Training Dataset used to train the **Baseline-Model** and the **GroupDRO-Model**, (ii) Training Dataset used to the train the **Balanced-Model**, (iii) Validation set for all three models, (iv) Testing set for all three models, and (v) for the whole BraTS dataset. We can observe that for the BraTS dataset, there is a high disparity between the number of samples for both subgroups. ©[2023] PMLR. Reprinted, with permission, from [164].

	Training Set		**Validation Set**	**Testing Set**	**BraTS Dataset**
	Baseline-Model and GroupDRO-Model	**Balanced Model**			
D^0	74	74	8	20	102
D^1	124	74	14	20	158
Overall	198	148	22	40	260

Table D.13: Dice (at $\tau = 100$) and QU-BraTS metric [161] values for Whole Tumour, Tumour Core, and Enhancing Tumour of a **Baseline-Model** on the BraTS dataset. ©[2023] PMLR. Reprinted, with permission, from [164].

Baseline-Model	**Dice**			**QU-BraTS Metric**		
	Whole Tumour	**Tumour Core**	**Enhancing Tumour**	**Whole Tumour**	**Tumour Core**	**Enhancing Tumour**
D^0	91.11	88.42	84.26	93.38	91.79	84.85
D^1	91.34	86.35	83.84	92.92	90.18	85.16
Fairness Gap	0.23	2.07	0.42	0.46	1.61	0.31

Table D.14: Dice (at $\tau = 100$) and QU-BraTS metric [161] values for Whole Tumour, Tumour Core, and Enhancing Tumour of a **Balanced-Model** on the BraTS dataset. ©[2023] PMLR. Reprinted, with permission, from [164].

Balanced-Model	**Dice**			**QU-BraTS Metric**		
	Whole Tumour	**Tumour Core**	**Enhancing Tumour**	**Whole Tumour**	**Tumour Core**	**Enhancing Tumour**
D^0	90.49	88.28	83.73	92.96	91.18	86.16
D^1	91.23	83.78	81.79	92.95	89.08	85.64
Fairness Gap	0.74	4.50	1.94	0.01	2.10	0.52

Table D.15: Dice (at $\tau = 100$) and QU-BraTS metric [161] values for Whole Tumour, Tumour Core, and Enhancing Tumour of a **GroupDRO-Model** on the BraTS dataset. ©[2023] PMLR. Reprinted, with permission, from [164].

GroupDRO-Model	Dice			QU-BraTS Metric		
	Whole Tumour	Tumour Core	Enhancing Tumour	Whole Tumour	Tumour Core	Enhancing Tumour
D^0	90.45	87.63	83.84	92.35	91.03	84.38
D^1	91.79	85.35	83.39	93.13	90.21	85.97
Fairness Gap	1.34	2.28	0.45	0.78	0.72	1.59

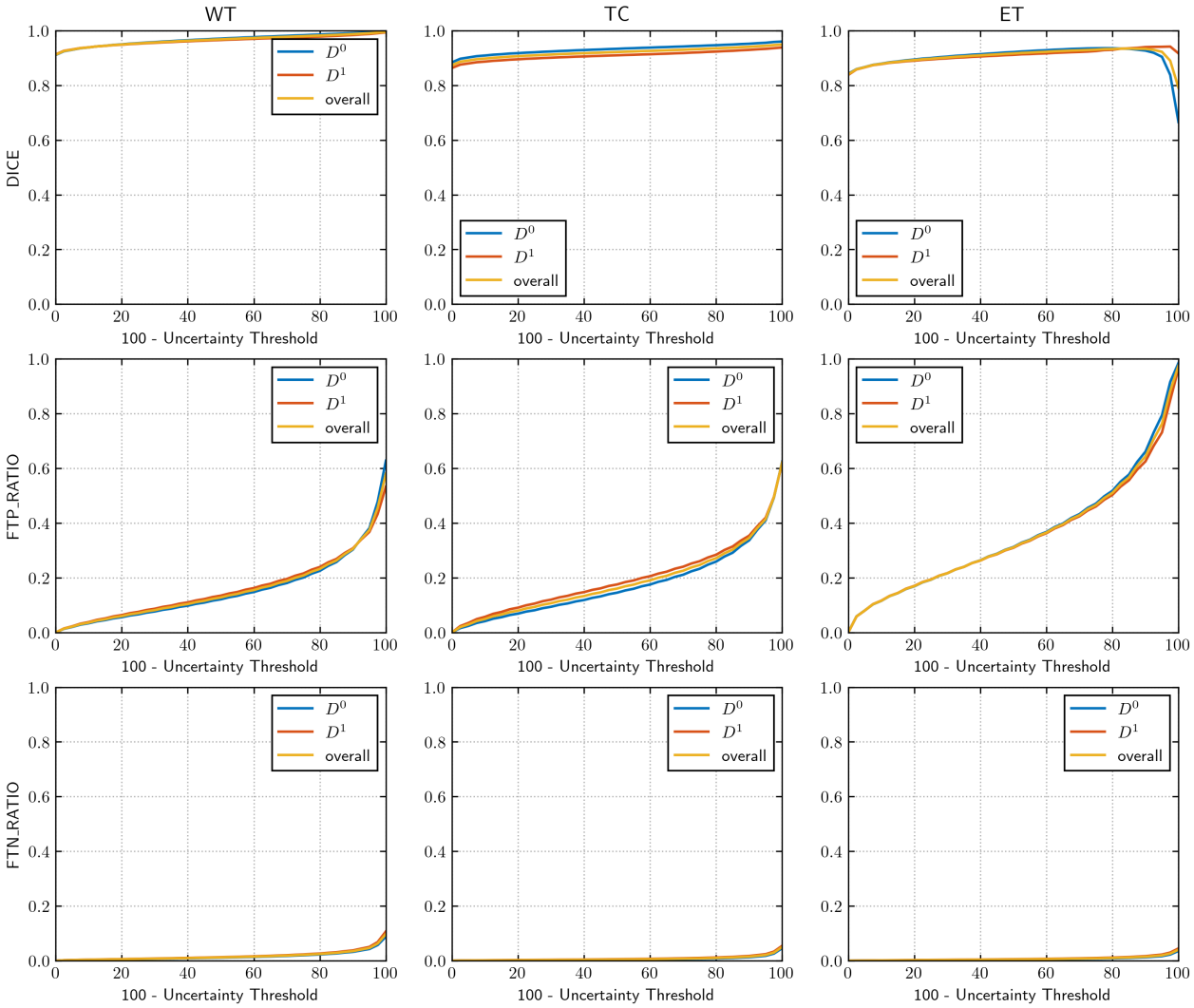


Figure D.6: **BraTS-Imaging-Centre:** Dice, Filtered True Positive Ratio (FTP), and Filtered True Negative Ratio (FTN) as a function of uncertainty threshold for **Baseline-Model** on the BraTS dataset. Specifically, we plot Whole Tumour (WT), Tumour Core (TC), and Enhancing Tumour (ET) QU-BraTS [161] metrics for both the D^0 and D^1 set. ©[2023] PMLR. Reprinted, with permission, from [164].

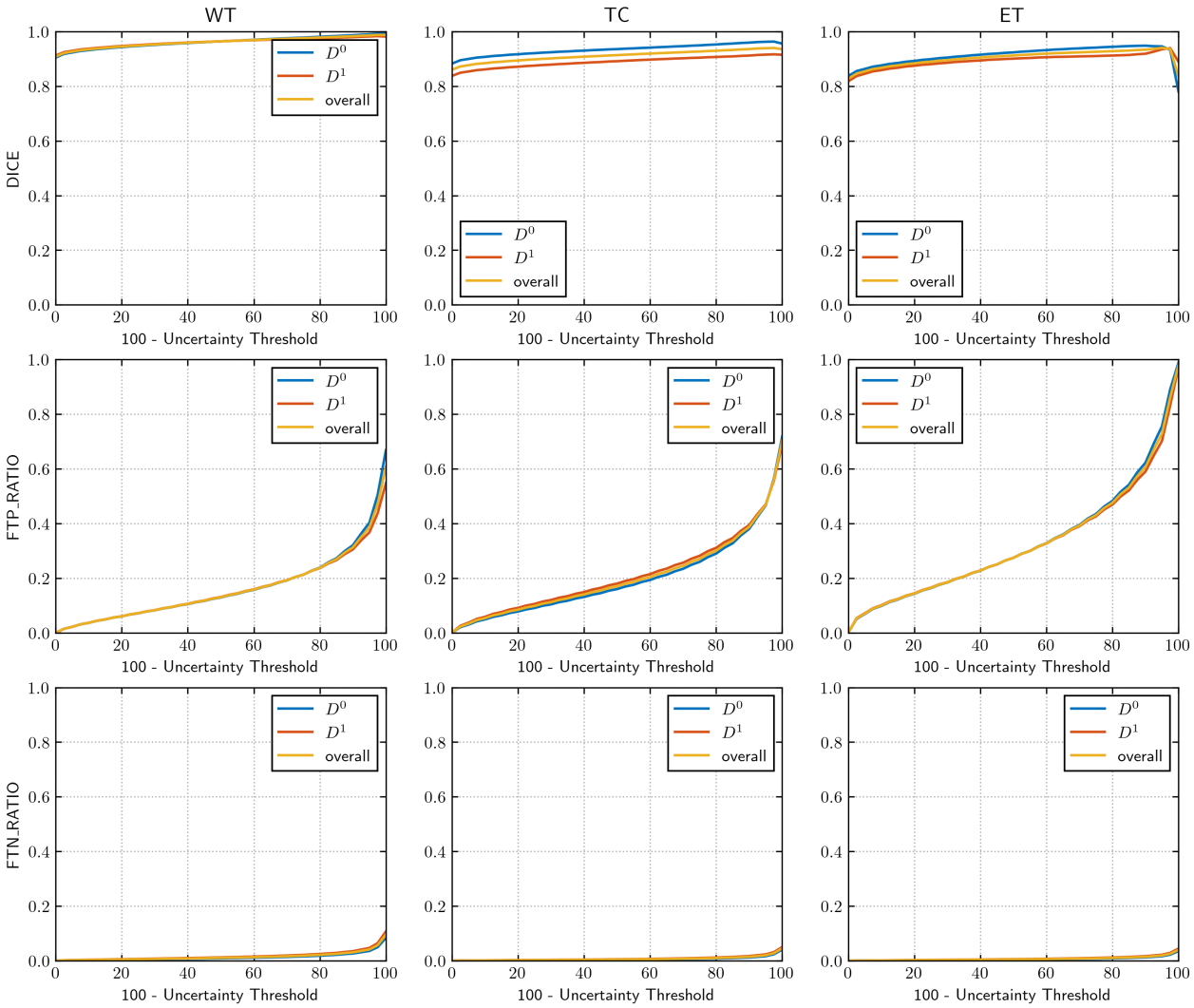


Figure D.7: **BraTS-Imaging-Centre:** Dice, Filtered True Positive Ratio (FTP), and Filtered True Negative Ratio (FTN) as a function of uncertainty threshold for **Balanced-Model** on the BraTS dataset. Specifically, we plot Whole Tumour (WT), Tumour Core (TC), and Enhancing Tumour (ET) QU-BraTS [161] metrics for both the D^0 and D^1 set. ©[2023] PMLR. Reprinted, with permission, from [164].

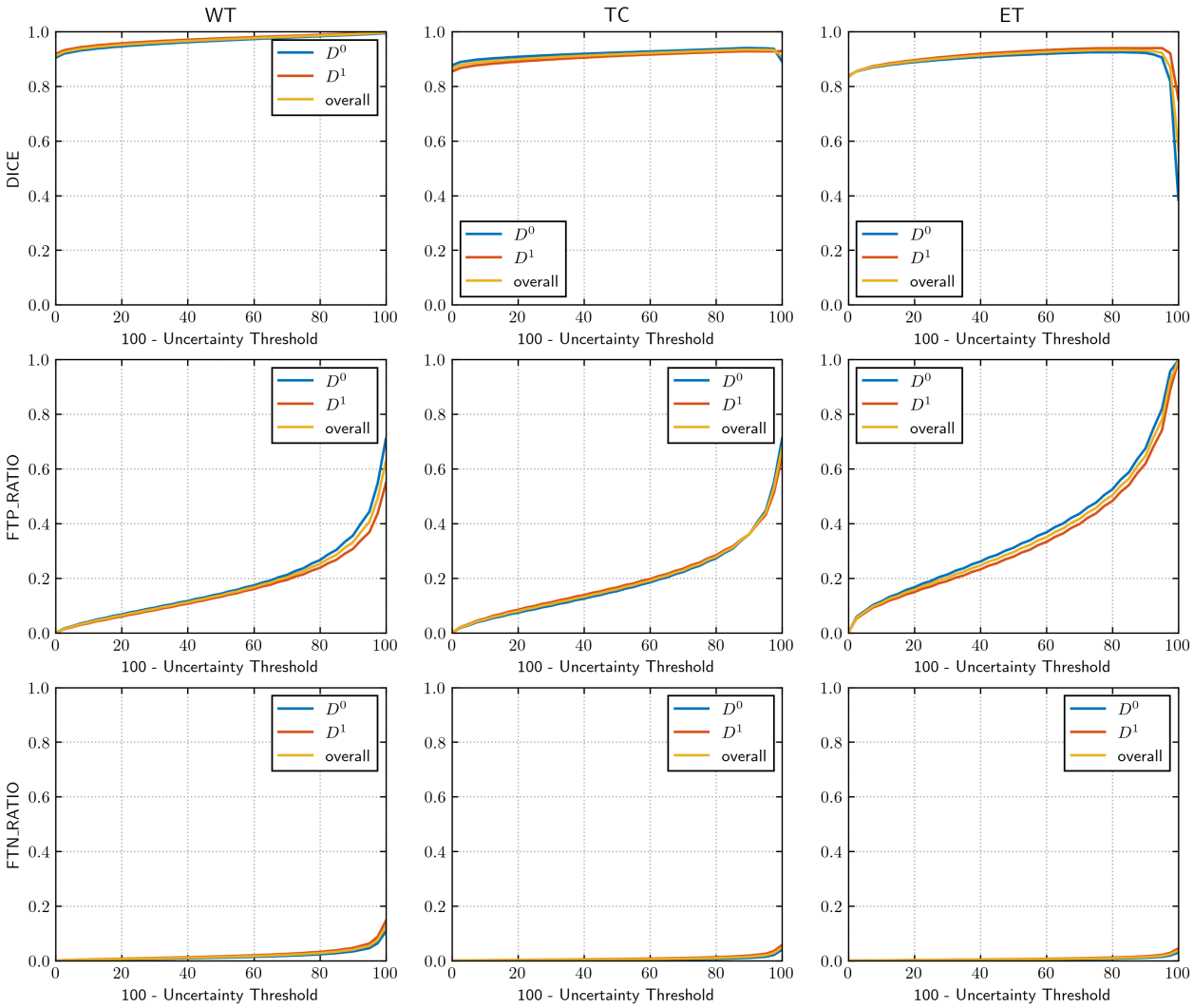
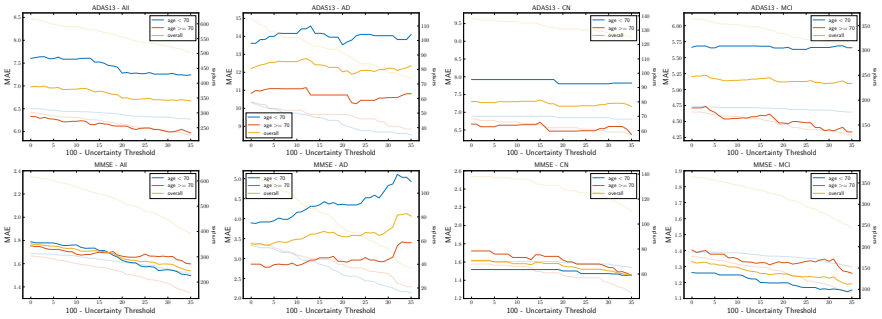
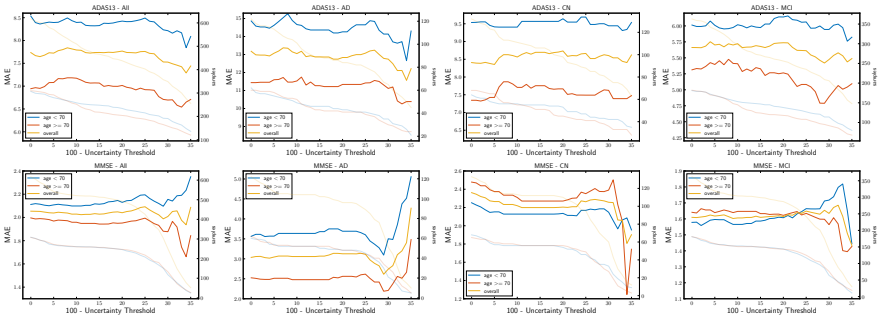


Figure D.8: **BraTS-Imaging-Centre:** We plot Dice, Filtered True Positive Ratio (FTP), and Filtered True Negative Ratio (FTN) as a function of uncertainty threshold for **GroupDRO-Model** on the BraTS dataset. Specifically, we plot Whole Tumour (WT), Tumour Core (TC), and Enhancing Tumour (ET) QU-BraTS [161] metrics for both the D^0 and D^1 set. ©[2023] PMLR. Reprinted, with permission, from [164].

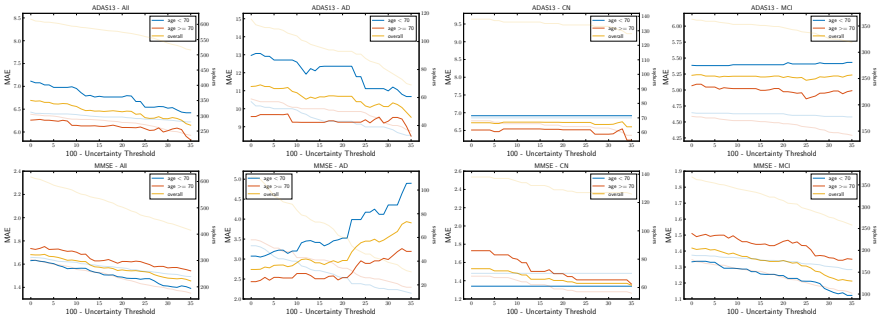
D.3 Alzheimer's Disease Clinical Score Regression



(a) Baseline Model



(b) Balanced Model



(c) GroupDRO Model

Figure D.9: **ADNI:** Mean Absolute Error (MAE) of ADAS-13 (Top) and MMSE (Bottom) score prediction tasks as a function of uncertainty threshold for (a) **Baseline-Model**, (b) **Balanced-Model**, and (c) **GroupDRO-Model** on the ADNI dataset. Specifically, we plot RMSE for all samples as well as samples for each of the disease stages (AD, MCI, and CN) in each subgroup (D^0 - age < 70 and D^1 - age ≥ 70). The total number of samples as a function of uncertainty thresholds in are depicted with light colours. ©[2023] PMLR. Reprinted, with permission, from [164].

E

Appendix: Information Gain Sampling for AL in Medical Image Classification

E.1 Tabular Results

Table E.1: Comparison of the EIG, AEIG, UIG, and CFIG based active learning sampling methods for both the DR dataset We report the mean and std of evaluation metric across five different runs. Model performance with the entire training set is 0.8561. ©[2022] Springer. Reprinted, with permission, from [165].

Method	Percentage of Labeled Sample					
	10	16	22	28	34	40
EIG	0.7461 ± 0.0035	0.7677 ± 0.0043	0.7834 ± 0.0201	0.7904 ± 0.0119	0.7828 ± 0.0151	0.8019 ± 0.0035
UIG	0.7461 ± 0.0035	0.7874 ± 0.0178	0.8010 ± 0.0109	0.8088 ± 0.0078	0.8162 ± 0.0049	0.8307 ± 0.0084
PIG	0.7461 ± 0.0035	0.7637 ± 0.0066	0.7948 ± 0.0065	0.8023 ± 0.0057	0.8239 ± 0.0069	0.8278 ± 0.0063
AEIG	0.7461 ± 0.0035	0.7985 ± 0.0155	0.8197 ± 0.0072	0.8269 ± 0.0017	0.8363 ± 0.0117	0.8468 ± 0.0024

Table E.2: Comparison of the EIG, AEIG, UIG, and CFIG based active learning sampling methods for both the ISIC dataset We report the mean and std of evaluation metric across five different runs. Model performance with the entire training set is 0.9789. ©[2022] Springer. Reprinted, with permission, from [165].

Method	**Percentage of Labeled Sample**					
	10	**15.83**	**21.67**	**27.5**	**33.33**	**39.17**
EIG	0.9033 ± 0.0121	0.9122 ± 0.0079	0.9277 ± 0.0032	0.9283 ± 0.0066	0.9281 ± 0.0006	0.9350 ± 0.0042
UIG	0.9003 ± 0.0121	0.9243 ± 0.0040	0.9452 ± 0.0041	0.9538 ± 0.0021	0.9624 ± 0.0021	0.9639 ± 0.0031
PIG	0.9003 ± 0.0121	0.9265 ± 0.0060	0.9443 ± 0.0038	0.9546 ± 0.0041	0.9554 ± 0.0050	0.9574 ± 0.0013
AEIG	0.9003 ± 0.0121	0.9439 ± 0.0040	0.9577 ± 0.0028	0.9681 ± 0.0046	0.9735 ± 0.0022	0.9753 ± 0.0018

Table E.3: Comparison of the Random, Entropy, CoreSet, MCD-Entropy, MCD-BALD, and AEIG based active learning sampling methods for both the DR dataset We report the mean and std of evaluation metric across five different runs. Model performance with the entire training set is 0.8561. ©[2022] Springer. Reprinted, with permission, from [165].

Method	**Percentage of Labeled Sample**					
	10	**16**	**22**	**28**	**34**	**40**
Random	0.7461 ± 0.0035	0.7774 ± 0.0088	0.7947 ± 0.0034	0.8138 ± 0.0114	0.8261 ± 0.0148	0.8329 ± 0.0040
Entropy	0.7461 ± 0.0035	0.7919 ± 0.0056	0.8154 ± 0.0035	0.8222 ± 0.0078	0.8322 ± 0.0088	0.8378 ± 0.0073
CoreSet	0.7461 ± 0.0035	0.7950 ± 0.0208	0.8065 ± 0.0070	0.8114 ± 0.0072	0.8295 ± 0.0022	0.8309 ± 0.0035
MCD-Entropy	0.7461 ± 0.0035	0.7925 ± 0.0132	0.8036 ± 0.0055	0.8218 ± 0.0026	0.8333 ± 0.0126	0.8424 ± 0.0127
MCD-BALD	0.7461 ± 0.0035	0.7951 ± 0.0125	0.8155 ± 0.0011	0.8225 ± 0.0044	0.8273 ± 0.0077	0.8344 ± 0.0130
AEIG	0.7461 ± 0.0035	0.7985 ± 0.0155	0.8197 ± 0.0072	0.8269 ± 0.0017	0.8363 ± 0.0117	0.8468 ± 0.0024

Table E.4: Comparison of the Random, Entropy, CoreSet, MCD-Entropy, MCD-BALD, and AEIG based active learning sampling methods for both the ISIC dataset We report the mean and std of evaluation metric across five different runs. Model performance with the entire training set is 0.9789. ©[2022] Springer. Reprinted, with permission, from [165].

Method	**Percentage of Labeled Sample**					
	10	**15.83**	**21.67**	**27.5**	**33.33**	**39.17**
Random	0.9003 ± 0.0121	0.9213 ± 0.0021	0.9419 ± 0.0028	0.9499 ± 0.0044	0.9559 ± 0.0030	0.9589 ± 0.0041
Entropy	0.9003 ± 0.0121	0.9385 ± 0.0026	0.9567 ± 0.0072	0.9649 ± 0.0039	0.9714 ± 0.0023	0.9755 ± 0.0034
CoreSet	0.9003 ± 0.0121	0.9426 ± 0.0028	0.9561 ± 0.0030	0.9642 ± 0.0011	0.9703 ± 0.0015	0.9745 ± 0.0014
MCD-Entropy	0.9003 ± 0.0121	0.9372 ± 0.0047	0.9519 ± 0.0030	0.9651 ± 0.0023	0.9707 ± 0.0010	0.9743 ± 0.0034
MCD-BALD	0.9003 ± 0.0121	0.9410 ± 0.0067	0.9540 ± 0.0044	0.9672 ± 0.0024	0.9694 ± 0.0049	0.9747 ± 0.0020
AEIG	0.9003 ± 0.0121	0.9439 ± 0.0040	0.9577 ± 0.0028	0.9681 ± 0.0046	0.9735 ± 0.0022	0.9753 ± 0.0018

Bibliography

[1] Frequently asked questions regarding ieee permissions. https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/permissions_faq.pdf, 2013. (Accessed on 01/04/2023). (↑77)

[2] arxiv license information. https://info.arxiv.org/help/license/index.html, 2023. (Accessed on 01/04/2023). (↑45)

[3] Pmlr license agreement. http://proceedings.mlr.press/pmlr-license-agreement.html, 2023. (Accessed on 01/04/2023). (↑99)

[4] Springer auther permission. https://www.springer.com/gp/rights-permissions/obtaining-permissions/882, 2023. (Accessed on 01/04/2023). (↑124), (↑149)

[5] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion (2021). (↑23)

[6] Administration, N. H. T. S. Tesla crash preliminary evaluation report. https://www.ntsb.gov/investigations/accidentreports/reports/har1702.pdf, 2017. (Accessed on 01/04/2023). (↑2)

[7] AGARWAL, S., ARORA, H., ANAND, S., AND ARORA, C. Contextual diversity for active learning. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16 (2020), Springer, pp. 137–153. (↑146)

[8] AMIAN, M., AND SOLTANINEJAD, M. Multi-resolution 3d CNN for MRI brain tumor segmentation and survival prediction. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I (2019), A. Crimi and S. Bakas, Eds., vol. 11992 of Lecture Notes in Computer Science, Springer, pp. 221–230. (↑177)

[9] ANGELOPOULOS, A. N., AND BATES, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021). (↑102), (↑142)

[10] APALLA, Z., LALLAS, A., SOTIRIOU, E., LAZARIDOU, E., AND IOANNIDES, D. Epidemiological trends in skin cancer. Dermatology practical & conceptual 7, 2 (2017), 1. (↑105)

[11] ARAÚJO, T., ARESTA, G., MENDONÇA, L., PENAS, S., MAIA, C., CARNEIRO, Â., MENDONÇA, A. M., AND CAMPILHO, A. Dr— graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Medical Image Analysis 63 (2020), 101715. (↑23)

[12] ARMATO III, S. G., MCLENNAN, G., BIDAUT, L., MCNITT-GRAY, M. F., MEYER, C. R., REEVES, A. P., ZHAO, B., ABERLE, D. R., HENSCHKE, C. I., HOFFMAN, E. A., ET AL. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics 38, 2 (2011), 915–931. (↑38)

[13] ASHUKHA, A., LYZHOV, A., MOLCHANOV, D., AND VETROV, D. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. arXiv preprint arXiv:2002.06470 (2020). (↑23), (↑33), (↑34), (↑80), (↑88)

[14] ASSOCIATION, A., ET AL. Alzheimer's disease facts and figures. Alzheimer's & Dementia 15, 3 (2019), 321–387. (↑22), (↑117)

[15] AVANTS, B. B., TUSTISON, N., AND SONG, G. Advanced normalization tools (ants). Insight j 2, 365 (2009), 1–35. (↑78)

[16] AYHAN, M. S., AND BERENS, P. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. (↑31), (↑32)

[17] BADRINARAYANAN, V., KENDALL, A., AND CIPOLLA, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39, 12 (2017), 2481–2495. (↑31)

[18] BAID, U., SHAH, N. A., AND TALBAR, S. N. Brain tumor segmentation with cascaded deep convolutional neural network. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II (2019), A. Crimi and S. Bakas, Eds., vol. 11993 of Lecture Notes in Computer Science, Springer, pp. 90–98. (↑177)

[19] BAKAS, S., AKBARI, H., SOTIRAS, A., BILELLO, M., ROZYCKI, M., KIRBY, J., FREYMANN, J., FARAHANI, K., AND DAVATZIKOS, C. Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. The cancer imaging archive 286 (2017). (↑49), (↑50)

[20] BAKAS, S., AKBARI, H., SOTIRAS, A., BILELLO, M., ROZYCKI, M., KIRBY, J., FREYMANN, J., FARAHANI, K., AND DAVATZIKOS, C. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. The cancer imaging archive 286 (2017). (↑49), (↑50)

[21] BAKAS, S., AKBARI, H., SOTIRAS, A., BILELLO, M., ROZYCKI, M., KIRBY, J. S., FREYMANN, J. B., FARAHANI, K., AND DAVATZIKOS, C. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data 4, 1 (2017), 1–13. (↑49), (↑50)

[22] BAKAS, S., REYES, M., JAKAB, A., BAUER, S., REMPFLER, M., CRIMI, A., SHINOHARA, R. T., BERGER, C., HA, S. M., ROZYCKI, M., ET AL. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018). (↑V), (↑VI), (↑XIII), (↑4), (↑9), (↑10), (↑16), (↑17), (↑18), (↑45), (↑49), (↑50), (↑61), (↑82), (↑85), (↑86), (↑94), (↑100), (↑112), (↑151), (↑154), (↑155), (↑157)

[23] BALLESTAR, L. M., AND VILAPLANA, V. MRI brain tumor segmentation and uncertainty estimation using 3d-UNet architectures. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 376–390. (↑55)

[24] BAND, N., RUDNER, T. G., FENG, Q., FILOS, A., NADO, Z., DUSENBERRY, M. W., JERFEL, G., TRAN, D., AND GAL, Y. Benchmarking bayesian deep learning on diabetic retinopathy detection tasks. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021). (↑40), (↑103)

[25] BANERJEE, S., ARORA, H. S., AND MITRA, S. Ensemble of cnns for segmentation of glioma sub-regions with survival prediction. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II (2019), A. Crimi and S. Bakas, Eds., vol. 11993 of Lecture Notes in Computer Science, Springer, pp. 37–49. (↑54), (↑177)

[26] BANERJEE, S., MITRA, S., AND SHANKAR, B. U. Multi-planar spatial-ConvNet for segmentation and survival prediction in brain cancer. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2019, pp. 94–104. (↑38), (↑54)

[27] BAUMGARTNER, C. F., TEZCAN, K. C., CHAITANYA, K., HÖTKER, A. M., MUEHLEMATTER, U. J., SCHAWKAT, K., BECKER, A. S., DONATI, O., AND KONUKOGLU, E. Phiseg: Capturing uncertainty in medical image segmentation.

In International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), Springer, pp. 119–127. (↑2), (↑37), (↑38), (↑141), (↑143)

[28] BHAGWAT, N., PIPITONE, J., VOINESKOS, A. N., CHAKRAVARTY, M. M., INITIATIVE, A. D. N., ET AL. An artificial neural network model for clinical score prediction in alzheimer disease using structural neuroimaging measures. Journal of psychiatry & neuroscience: JPN 44, 4 (2019), 246. (↑21), (↑83)

[29] BHAT, I., PLUIM, J. P., AND KUIJF, H. J. Generalized probabilistic u-net for medical image segementation. In Uncertainty for Safe Utilization of Machine Learning in Medical Imaging: 4th International Workshop, UNSURE 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings (2022), Springer, pp. 113–124. (↑39)

[30] BISHOP, C. M., AND NASRABADI, N. M. Pattern recognition and machine learning, vol. 4. Springer, 2006. (↑2)

[31] BLUNDELL, C., CORNEBISE, J., KAVUKCUOGLU, K., AND WIERSTRA, D. Weight uncertainty in neural network. In International Conference on Machine Learning (2015), PMLR, pp. 1613–1622. (↑23)

[32] BOUTRY, N., CHAZALON, J., PUYBAREAU, É., TOCHON, G., TALBOT, H., AND GÉRAUD, T. Using separated inputs for multimodal brain tumor segmentation with 3d u-net-like architectures. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I (2019), A. Crimi and S. Bakas, Eds., vol. 11992 of Lecture Notes in Computer Science, Springer, pp. 187–199. (↑177)

[33] BROSTOW, G. J., SHOTTON, J., FAUQUEUR, J., AND CIPOLLA, R. Segmentation and recognition using structure from motion point clouds. In European conference on computer vision (2008), Springer, pp. 44–57. (↑30), (↑31)

[34] BUOLAMWINI, J., AND GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability

and transparency (2018), PMLR, pp. 77–91. (↑145)

[35] BURLINA, P., JOSHI, N., PAUL, W., PACHECO, K. D., AND BRESSLER, N. M. Addressing artificial intelligence bias in retinal diagnostics. Translational Vision Science & Technology 10, 2 (2021), 13–13. (↑39)

[36] CAI, L., XU, X., LIEW, J. H., AND FOO, C. S. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021), pp. 10988–10997. (↑146)

[37] CAMARASA, R., BOS, D., HENDRIKSE, J., NEDERKOORN, P. H. J., KOOI, M. E., VAN DER LUGT, A., AND DE BRUIJNE, M. A quantitative comparison of epistemic uncertainty maps applied to multi-class segmentation. CoRR abs/2109.10702 (2021). (↑138), (↑141)

[38] CASTELNOVO, A., CRUPI, R., GRECO, G., REGOLI, D., PENCO, I. G., AND COSENTINI, A. C. A clarification of the nuances in the fairness metrics landscape. Scientific Reports 12, 1 (2022), 4209. (↑144)

[39] CASTRO, D. C., WALKER, I., AND GLOCKER, B. Causality matters in medical imaging. Nature Communications 11, 1 (2020), 3673. (↑146)

[40] CHA, J., CHUN, S., LEE, K., CHO, H.-C., PARK, S., LEE, Y., AND PARK, S. Swad: Domain generalization by seeking flat minima. Advances in Neural Information Processing Systems 34 (2021), 22405–22418. (↑40)

[41] CHARTSIAS, A., JOYCE, T., GIUFFRIDA, M. V., AND TSAFTARIS, S. A. Multimodal mr synthesis via modality-invariant latent representation. IEEE transactions on medical imaging 37, 3 (2017), 803–814. (↑2), (↑18), (↑150), (↑151), (↑154), (↑155), (↑156)

[42] CHEN, C. R., FAN, Q., MALLINAR, N., SERCU, T., AND FERIS, R. S. Big-little net: An efficient multi-scale feature representation for visual and speech recogni-

tion. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019 (2019), OpenReview.net. (↑53)

[43] CHEPLYGINA, V., DE BRUIJNE, M., AND PLUIM, J. P. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis 54 (2019), 280–296. (↑41)

[44] CHUPIN, M., GÉRARDIN, E., CUINGNET, R., BOUTET, C., LEMIEUX, L., LEHÉRICY, S., BENALI, H., GARNERO, L., AND COLLIOT, O. Fully automatic hippocampus segmentation and classification in alzheimer's disease and mild cognitive impairment applied on data from adni. Hippocampus 19, 6 (2009), 579–587. (↑83)

[45] ÇIÇEK, Ö., ABDULKADIR, A., LIENKAMP, S. S., BROX, T., AND RONNEBERGER, O. 3d u-net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (2016), Springer, pp. 424–432. (↑XXIII), (↑82), (↑151), (↑152), (↑181), (↑182), (↑183)

[46] CODELLA, N., ROTEMBERG, V., TSCHANDL, P., CELEBI, M. E., DUSZA, S., GUTMAN, D., HELBA, B., KALLOO, A., LIOPYRIS, K., MARCHETTI, M., ET AL. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019). (↑100), (↑105), (↑130)

[47] CORBETT-DAVIES, S., AND GOEL, S. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018). (↑144)

[48] CORDTS, M., OMRAN, M., RAMOS, S., REHFELD, T., ENZWEILER, M., BENENSON, R., FRANKE, U., ROTH, S., AND SCHIELE, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (2016), pp. 3213–3223. (↑30)

[49] CULOTTA, A., AND MCCALLUM, A. Reducing labeling effort for structured prediction tasks. In AAAI (2005), vol. 5, pp. 746–751. (↑41)

[50] DAI, C., WANG, S., RAYNAUD, H., MO, Y., ANGELINI, E., GUO, Y., AND BAI, W. Self-training for brain tumour segmentation with uncertainty estimation and biophysics-guided survival prediction. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 514–523. (↑53)

[51] DALCA, A. V., BALAKRISHNAN, G., GUTTAG, J., AND SABUNCU, M. R. Unsupervised learning for fast probabilistic diffeomorphic registration. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2018), Springer, pp. 729–738. (↑2), (↑4)

[52] DALE, A. M., FISCHL, B., AND SERENO, M. I. Cortical surface-based analysis: I. segmentation and surface reconstruction. Neuroimage 9, 2 (1999), 179–194. (↑78)

[53] DAVATZIKOS, C., RATHORE, S., BAKAS, S., PATI, S., BERGMAN, M., KALAROT, R., SRIDHARAN, P., GASTOUNIOTI, A., JAHANI, N., COHEN, E., ET AL. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of medical imaging 5, 1 (2018), 011018. (↑51)

[54] DAZA, L., GÓMEZ, C., AND ARBELÁEZ, P. Cerberus: A multi-headed network for brain tumor segmentation. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 342–351. (↑53)

[55] DEL TORO, O. A. J., GOKSEL, O., MENZE, B., MÜLLER, H., LANGS, G., WEBER, M.-A., EGGEL, I., GRUENBERG, K., HOLZER, M., KOTSIOS-KONTOKOTSIOS, G., ET AL. Visceral–visual concept extraction challenge in radiology: Isbi 2014 challenge organization. Proceedings of the VISCERAL Challenge at ISBI 1194 (2014), 6–15. (↑141)

[56] DENG, W., ZHONG, Y., DOU, Q., AND LI, X. On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. arXiv preprint arXiv:2301.01481 (2023). (↑145)

[57] DERA, D., BOUAYNAYA, N. C., RASOOL, G., SHTERENBERG, R., AND FATHALLAH-SHAYKH, H. M. Premium-cnn: Propagating uncertainty towards robust convolutional neural networks. IEEE Transactions on Signal Processing 69 (2021), 4669–4684. (↑143)

[58] DOSHI, J., ERUS, G., HABES, M., AND DAVATZIKOS, C. Deepmrseg: A convolutional deep neural network for anatomy and abnormality segmentation on MR images. CoRR abs/1907.02110 (2019). (↑54)

[59] DU, M., YANG, F., ZOU, N., AND HU, X. Fairness in deep learning: A computational perspective. IEEE Intelligent Systems 36, 4 (2020), 25–34. (↑102)

[60] DUDA, R. O., HART, P. E., ET AL. Pattern classification. John Wiley & Sons, 2006. (↑128)

[61] DUNCAN, C., ROXAS, F., JANI, N., MAKSIMOVIC, J., BRAMLET, M., SUTTON, B., AND KOYEJO, S. Some new tricks for deep glioma segmentation. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 320–330. (↑51)

[62] DURSO-FINLEY, J., FALET, J.-P., MEHTA, R., ARNOLD, D., PAWLOWSKI, N., AND ARBEL, T. Improving image-based precision medicine with uncertainty-aware causal models. In 26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023, Vancouver, October 2023, Proceedings (2023), Springer. (↑6)

[63] DURSO-FINLEY, J., FALET, J.-P., NICHYPORUK, B., DOUGLAS, A., AND ARBEL, T. Personalized prediction of future lesion activity and treatment effect in multiple sclerosis from baseline mri. In International Conference on Medical Imaging with Deep Learning (2022), PMLR, pp. 387–406. (↑147)

[64] EIGEN, D., PUHRSCH, C., AND FERGUS, R. Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems 27 (2014). (↑1)

[65] EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C. K., WINN, J., AND ZISSERMAN, A. The pascal visual object classes (voc) challenge. International journal of computer vision 88, 2 (2010), 303–338. (↑31)

[66] FARQUHAR, S., OSBORNE, M. A., AND GAL, Y. Radial bayesian neural networks: Beyond discrete support in large-scale bayesian deep learning. In The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy] (2020), S. Chiappa and R. Calandra, Eds., vol. 108 of Proceedings of Machine Learning Research, PMLR, pp. 1352–1362. (↑23)

[67] FAWCETT, T. An introduction to roc analysis. Pattern recognition letters 27, 8 (2006), 861–874. (↑131)

[68] FENG, X., DOU, Q., TUSTISON, N. J., AND MEYER, C. H. Brain tumor segmentation with uncertainty estimation and overall survival prediction. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I (2019), A. Crimi and S. Bakas, Eds., vol. 11992 of Lecture Notes in Computer Science, Springer, pp. 304–314. (↑177)

[69] FOLSTEIN, M. F., FOLSTEIN, S. E., AND MCHUGH, P. R. "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. Journal of psychiatric research 12, 3 (1975), 189–198. (↑21), (↑83)

[70] FORET, P., KLEINER, A., MOBAHI, H., AND NEYSHABUR, B. Sharpness-aware minimization for efficiently improving generalization. In International Conference on Learning Representations (2021). (↑40)

[71] FORT, S., HU, H., AND LAKSHMINARAYANAN, B. Deep ensembles: A loss landscape perspective. arXiv preprint arXiv:1912.02757 (2019). (↑23), (↑26)

[72] FRID-ADAR, M., KLANG, E., AMITAI, M., GOLDBERGER, J., AND GREENSPAN, H. Synthetic data augmentation using gan for improved liver lesion classification. In

2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) (2018), IEEE, pp. 289–293. (↑41)

[73] FRISONI, G. B., FOX, N. C., JACK, C. R., SCHELTENS, P., AND THOMPSON, P. M. The clinical use of structural mri in alzheimer disease. Nature Reviews Neurology 6, 2 (2010), 67–77. (↑21), (↑83)

[74] FRISONI, G. B., JACK JR, C. R., BOCCHETTA, M., BAUER, C., FREDERIKSEN, K. S., LIU, Y., PREBOSKE, G., SWIHART, T., BLAIR, M., CAVEDO, E., ET AL. The eadc-adni harmonized protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity. Alzheimer's & Dementia 11, 2 (2015), 111–125. (↑86)

[75] FUCHS, M., GONZALEZ, C., AND MUKHOPADHYAY, A. Practical uncertainty quantification for brain tumor segmentation. In Medical Imaging with Deep Learning (2021). (↑26)

[76] FUND, W. C. R. Worldwide cancer data. global cancer statistics for the most common cancers. (↑16)

[77] GAILLOCHET, M., DESROSIERS, C., AND LOMBAERT, H. Active learning for medical image segmentation with stochastic batches. arXiv preprint arXiv:2301.07670 (2023). (↑146)

[78] GAL, Y. Uncertainty in deep learning. (↑29), (↑34)

[79] GAL, Y., AND GHAHRAMANI, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning (2016), pp. 1050–1059. (↑2), (↑23), (↑25), (↑27), (↑31), (↑33), (↑36), (↑37), (↑42), (↑50), (↑79), (↑80), (↑87), (↑88), (↑89), (↑95), (↑102), (↑103), (↑107), (↑113), (↑117), (↑141), (↑151), (↑153)

[80] GAL, Y., ISLAM, R., AND GHAHRAMANI, Z. Deep bayesian active learning with image data. In International Conference on Machine Learning (2017), PMLR, pp. 1183–1192. (↑7), (↑28), (↑29), (↑42), (↑49), (↑95), (↑103), (↑133)

[81] GANAYE, P.-A., SDIKA, M., AND BENOIT-CATTIN, H. Semi-supervised learning for segmentation under semantic constraint. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2018), Springer, pp. 595–602. (↑41)

[82] GHESU, F. C., GEORGESCU, B., GIBSON, E., GUENDEL, S., KALRA, M. K., SINGH, R., DIGUMARTHY, S. R., GRBIC, S., AND COMANICIU, D. Quantifying and leveraging classification uncertainty for chest radiograph assessment. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), Springer, pp. 676–684. (↑31), (↑32), (↑36), (↑41), (↑103)

[83] GOEDERT, M., AND SPILLANTINI, M. G. A century of alzheimer's disease. science 314, 5800 (2006), 777–781. (↑83)

[84] GOLD, R., KAPPOS, L., ARNOLD, D. L., BAR-OR, A., GIOVANNONI, G., SELMAJ, K., TORNATORE, C., SWEETSER, M. T., YANG, M., SHEIKH, S. I., ET AL. Placebo-controlled phase 3 study of oral bg-12 for relapsing multiple sclerosis. New England Journal of Medicine 367, 12 (2012), 1098–1107. (↑19)

[85] GOPINATH, K., DESROSIERS, C., AND LOMBAERT, H. Learnable pooling in graph convolution networks for brain surface analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020). (↑83)

[86] GROZA, V., TUCHINOV, B., AMELINA, E., PAVLOVSKIY, E., TOLSTOKULAKOV, N., AMELIN, M., GOLUSHKO, S., AND LETYAGIN, A. Brain tumor segmentation and associated uncertainty evaluation using multi-sequences MRI mixture data preprocessing. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 148–157. (↑51)

[87] GUO, C., PLEISS, G., SUN, Y., AND WEINBERGER, K. Q. On calibration of modern neural networks. In International Conference on Machine Learning (2017), PMLR, pp. 1321–1330. (↑3), (↑34), (↑127)

[88] GUPTA, U., DHAMALA, J., KUMAR, V., VERMA, A., PRUKSACHATKUN, Y., KRISHNA, S., GUPTA, R., CHANG, K.-W., VER STEEG, G., AND GALSTYAN, A. Mit-

igating gender bias in distilled language models via counterfactual role reversal. In Findings of the Association for Computational Linguistics: ACL 2022 (2022), pp. 658–678. (↑147)

[89] GUY JR, G. P., THOMAS, C. C., THOMPSON, T., WATSON, M., MASSETTI, G. M., AND RICHARDSON, L. C. Vital signs: melanoma incidence and mortality trends and projections—united states, 1982–2030. Morbidity and mortality weekly report 64, 21 (2015), 591. (↑105)

[90] HALL, D., DAYOUB, F., SKINNER, J., ZHANG, H., MILLER, D., CORKE, P., CARNEIRO, G., ANGELOVA, A., AND SÜNDERHAUF, N. Probabilistic object detection: Definition and evaluation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2020), pp. 1031–1040. (↑3), (↑35)

[91] HARA, K., KATAOKA, H., AND SATOH, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2018), pp. 6546–6555. (↑XVII), (↑83), (↑86), (↑117), (↑119), (↑183)

[92] HAVAEI, M., DAVY, A., WARDE-FARLEY, D., BIARD, A., COURVILLE, A., BENGIO, Y., PAL, C., JODOIN, P.-M., AND LAROCHELLE, H. Brain tumor segmentation with deep neural networks. Medical image analysis 35 (2017), 18–31. (↑4), (↑17)

[93] HAVAEI, M., GUIZARD, N., CHAPADOS, N., AND BENGIO, Y. Hemis: Hetero-modal image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2016), Springer, pp. 469–477. (↑18), (↑82), (↑149)

[94] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (2016), pp. 770–778. (↑106), (↑131)

[95] HERZOG, L., MURINA, E., DÜRR, O., WEGENER, S., AND SICK, B. Integrating uncertainty in deep neural networks for mri based stroke analysis. Medical Image Analysis 65 (2020), 101790. (↑33)

[96] HINNEFELD, J. H., COOMAN, P., MAMMO, N., AND DEESE, R. Evaluating fairness metrics in the presence of dataset bias. arXiv preprint arXiv:1809.09245 (2018). (↑101)

[97] HORA, S. C. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. Reliability Engineering & System Safety 54, 2-3 (1996), 217–223. (↑29)

[98] HOULSBY, N., HUSZÁR, F., GHAHRAMANI, Z., AND LENGYEL, M. Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745 (2011). (↑42), (↑133)

[99] HU, J., SHEN, L., ALBANIE, S., SUN, G., AND WU, E. Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. 42, 8 (2020), 2011–2023. (↑53)

[100] HU, S., WORRALL, D., KNEGT, S., VEELING, B., HUISMAN, H., AND WELLING, M. Supervised uncertainty quantification for segmentation with multiple annotations. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), Springer, pp. 137–145. (↑38)

[101] HUANG, G., LI, Y., PLEISS, G., LIU, Z., HOPCROFT, J. E., AND WEINBERGER, K. Q. Snapshot ensembles: Train 1, get m for free. arXiv preprint arXiv:1704.00109 (2017). (↑23), (↑102)

[102] HUANG, G., LIU, Z., VAN DER MAATEN, L., AND WEINBERGER, K. Q. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (2017), pp. 4700–4708. (↑31), (↑53)

[103] HUANG, S., WANG, T., XIONG, H., HUAN, J., AND DOU, D. Semi-supervised active learning with temporal output discrepancy. In Proceedings of the IEEE/CVF International Conference on Computer Vision (2021), pp. 3447–3456. (↑145)

[104] HUH, M., AGRAWAL, P., AND EFROS, A. A. What makes imagenet good for transfer learning? arXiv preprint arXiv:1608.08614 (2016). (↑41)

[105] IDRISSI, B. Y., ARJOVSKY, M., PEZESHKI, M., AND LOPEZ-PAZ, D. Simple data balancing achieves competitive worst-group-accuracy. In Conference on Causal Learning and Reasoning (2022), PMLR, pp. 336–351. (↑6), (↑40), (↑100), (↑104)

[106] IGLESIAS, J. E., KONUKOGLU, E., ZIKIC, D., GLOCKER, B., VAN LEEMPUT, K., AND FISCHL, B. Is synthesizing mri contrast useful for inter-modality analysis? In International Conference on Medical Image Computing and Computer-Assisted Intervention (2013), Springer, pp. 631–638. (↑18), (↑82), (↑149)

[107] IOANNOU, S., CHOCKLER, H., HAMMERS, A., AND KING, A. P. A study of demographic bias in cnn-based brain mr segmentation. In International Workshop on Machine Learning in Clinical Neuroimaging (2022), Springer, pp. 13–22. (↑40), (↑100), (↑101), (↑104)

[108] IOFFE, S., AND SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (2015), PMLR, pp. 448–456. (↑XIV), (↑106)

[109] ISENSEE, F., KICKINGEREDER, P., WICK, W., BENDSZUS, M., AND MAIER-HEIN, K. H. No new-net. In International MICCAI Brainlesion Workshop (2018), Springer, pp. 234–244. (↑2), (↑4)

[110] ISOLA, P., ZHU, J.-Y., ZHOU, T., AND EFROS, A. A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (2017), pp. 1125–1134. (↑150)

[111] JACK JR, C. R., BERNSTEIN, M. A., FOX, N. C., THOMPSON, P., ALEXANDER, G., HARVEY, D., BOROWSKI, B., BRITSON, P. J., L. WHITWELL, J., WARD, C., ET AL. The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 27, 4 (2008), 685–691. (↑VI), (↑10), (↑16), (↑21), (↑22), (↑87), (↑100), (↑117)

[112] JAPKOWICZ, N., AND STEPHEN, S. The class imbalance problem: A systematic study. Intelligent data analysis 6, 5 (2002), 429–449. (↑40)

[113] JOG, A., CARASS, A., ROY, S., PHAM, D. L., AND PRINCE, J. L. Random forest regression for magnetic resonance image synthesis. Medical image analysis 35 (2017), 475–488. (↑18), (↑82), (↑150), (↑151), (↑154), (↑155)

[114] JUNGO, A., BALSIGER, F., AND REYES, M. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Frontiers in neuroscience 14 (2020), 282. (↑3), (↑23), (↑36)

[115] JUNGO, A., AND REYES, M. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), Springer, pp. 48–56. (↑40)

[116] KASARLA, T., NAGENDAR, G., HEGDE, G. M., BALASUBRAMANIAN, V., AND JAWAHAR, C. Region-based active learning for efficient labeling in semantic segmentation. In 2019 IEEE winter conference on applications of computer vision (WACV) (2019), IEEE, pp. 1109–1117. (↑146)

[117] KAUNZNER, U. W., AL-KAWAZ, M., AND GAUTHIER, S. A. Defining disease activity and response to therapy in ms. Current treatment options in neurology 19 (2017), 1–12. (↑33)

[118] KAUNZNER, U. W., AND GAUTHIER, S. A. Mri in the assessment and monitoring of multiple sclerosis: an update on best practice. Therapeutic advances in neurological disorders 10, 6 (2017), 247–261. (↑19), (↑33)

[119] KAUR, B., LEMAÎTRE, P., MEHTA, R., SEPAHVAND, N. M., PRECUP, D., ARNOLD, D., AND ARBEL, T. Improving pathological structure segmentation via transfer learning across diseases. In Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data. Springer, 2019, pp. 90–98. (↑7)

[120] KENDALL, A., BADRINARAYANAN, V., AND CIPOLLA, R. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680 (2015). (↑31), (↑38), (↑88), (↑89)

[121] KENDALL, A., AND GAL, Y. What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems 30 (2017). (↑29), (↑30), (↑31), (↑103), (↑118)

[122] KENDALL, A., GAL, Y., AND CIPOLLA, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE conference on computer vision and pattern recognition (2018), pp. 7482–7491. (↑30)

[123] KERVADEC, H., DOLZ, J., TANG, M., GRANGER, E., BOYKOV, Y., AND AYED, I. B. Constrained-cnn losses for weakly supervised segmentation. Medical image analysis 54 (2019), 88–99. (↑41)

[124] KIM, B., KIM, H., KIM, K., KIM, S., AND KIM, J. Learning not to learn: Training deep neural networks with biased data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 9012–9020. (↑40), (↑145)

[125] KIM, S. T., MUSHTAQ, F., AND NAVAB, N. Confident coreset for active learning in medical image analysis. arXiv preprint arXiv:2004.02200 (2020). (↑42), (↑132)

[126] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014). (↑106), (↑113), (↑154)

[127] KINGMA, D. P., AND DHARIWAL, P. Glow: Generative flow with invertible 1x1 convolutions. Advances in neural information processing systems 31 (2018). (↑39)

[128] KIRSCH, A., AND GAL, Y. Unifying approaches in data subset selection via fisher information and information-theoretic quantities. Transactions on Machine Learning Research (2022). (↑145)

[129] KLEESIEK, J., URBAN, G., HUBERT, A., SCHWARZ, D., MAIER-HEIN, K., BENDSZUS, M., AND BILLER, A. Deep mri brain extraction: A 3d convolutional neural network for skull stripping. NeuroImage 129 (2016), 460–469. (↑4)

[130] KOHL, S., ROMERA-PAREDES, B., MEYER, C., DE FAUW, J., LEDSAM, J. R., MAIER-HEIN, K., ESLAMI, S., JIMENEZ REZENDE, D., AND RONNEBERGER, O.

A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems 31 (2018). (↑2), (↑37), (↑38), (↑141), (↑143)

[131] KONYUSHKOVA, K., SZNITMAN, R., AND FUA, P. Geometry in active learning for binary and multi-class image segmentation. Computer vision and image understanding 182 (2019), 1–16. (↑42)

[132] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (2012), pp. 1097–1105. (↑1)

[133] KUSNER, M. J., LOFTUS, J., RUSSELL, C., AND SILVA, R. Counterfactual fairness. Advances in neural information processing systems 30 (2017). (↑147)

[134] KWON, Y., WON, J.-H., KIM, B. J., AND PAIK, M. C. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. Computational Statistics & Data Analysis 142 (2020), 106816. (↑31)

[135] LAHOTI, P., BEUTEL, A., CHEN, J., LEE, K., PROST, F., THAIN, N., WANG, X., AND CHI, E. Fairness without demographics through adversarially reweighted learning. Advances in neural information processing systems 33 (2020), 728–740. (↑143)

[136] LAKSHMINARAYANAN, B., PRITZEL, A., AND BLUNDELL, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (2017), vol. 30. (↑2), (↑23), (↑26), (↑27), (↑33), (↑34), (↑36), (↑37), (↑50), (↑79), (↑80), (↑87), (↑88), (↑95), (↑102), (↑103), (↑141)

[137] LARRAZABAL, A. J., NIETO, N., PETERSON, V., MILONE, D. H., AND FERRANTE, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences 117, 23 (2020), 12592–12594. (↑39)

[138] LEE, J., XIAO, L., SCHOENHOLZ, S., BAHRI, Y., NOVAK, R., SOHL-DICKSTEIN, J., AND PENNINGTON, J. Wide neural networks of any depth evolve as linear models

under gradient descent. Advances in neural information processing systems 32 (2019). (↑26)

[139] LEIBIG, C., ALLKEN, V., AYHAN, M. S., BERENS, P., AND WAHL, S. Leveraging uncertainty information from deep neural networks for disease detection. Scientific reports 7, 1 (2017), 1–14. (↑23), (↑31), (↑32), (↑89)

[140] LIN, T.-Y., GOYAL, P., GIRSHICK, R., HE, K., AND DOLLAR, P. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 2 (Feb. 2020), 318–327. (↑52)

[141] LIU, E. Z., HAGHGOO, B., CHEN, A. S., RAGHUNATHAN, A., KOH, P. W., SAGAWA, S., LIANG, P., AND FINN, C. Just train twice: Improving group robustness without training group information. In International Conference on Machine Learning (2021), PMLR, pp. 6781–6792. (↑143)

[142] LIU, J. Z., PADHY, S., REN, J., LIN, Z., WEN, Y., JERFEL, G., NADO, Z., SNOEK, J., TRAN, D., AND LAKSHMINARAYANAN, B. A simple approach to improve single-model deep uncertainty via distance-awareness. Advances in neural information processing systems 35 (2022). (↑23)

[143] LIU, M., LI, F., YAN, H., WANG, K., MA, Y., SHEN, L., XU, M., INITIATIVE, A. D. N., ET AL. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease. NeuroImage 208 (2020), 116459. (↑83)

[144] LONG, J., SHELHAMER, E., AND DARRELL, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (2015), pp. 3431–3440. (↑1), (↑31)

[145] LUNDBERG, S. M., AND LEE, S.-I. A unified approach to interpreting model predictions. Advances in neural information processing systems 30 (2017). (↑146)

[146] MACKAY, D. J. A practical bayesian framework for backpropagation networks. Neural computation 4, 3 (1992), 448–472. (↑23)

[147] MADDOX, W. J., IZMAILOV, P., GARIPOV, T., VETROV, D. P., AND WILSON, A. G. A simple baseline for bayesian uncertainty in deep learning. In Advances in Neural Information Processing Systems (2019), pp. 13153–13164. (↑23)

[148] MADRAS, D., CREAGER, E., PITASSI, T., AND ZEMEL, R. Learning adversarially fair and transferable representations. In International Conference on Machine Learning (2018), PMLR, pp. 3384–3393. (↑40)

[149] MAIER, O., MENZE, B. H., VON DER GABLENTZ, J., HÄNI, L., HEINRICH, M. P., LIEBRAND, M., WINZECK, S., BASIT, A., BENTLEY, P., CHEN, L., ET AL. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. Medical image analysis 35 (2017), 250–269. (↑31), (↑58)

[150] MALININ, A., MLODOZENIEC, B., AND GALES, M. Ensemble distribution distillation. arXiv preprint arXiv:1905.00076 (2019). (↑142)

[151] MCCARTHY, A. M., KELLER, B. M., PANTALONE, L. M., HSIEH, M.-K., SYNNESTVEDT, M., CONANT, E. F., ARMSTRONG, K., AND KONTOS, D. Racial differences in quantitative measures of area and volumetric breast density. JNCI: Journal of the National Cancer Institute 108, 10 (2016). (↑141)

[152] MCHUGH, H., TALOU, G. M., AND WANG, A. 2d dense-UNet: A clinically valid approach to automated glioma segmentation. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 69–80. (↑55)

[153] MCKINLEY, R., REBSAMEN, M., DÄTWYLER, K., MEIER, R., RADOJEWSKI, P., AND WIEST, R. Uncertainty-driven refinement of tumor-core segmentation using 3d-to-2d networks with label uncertainty. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 401–411. (↑52)

[154] MCKINLEY, R., REBSAMEN, M., MEIER, R., AND WIEST, R. Triplanar ensemble of 3d-to-2d cnns with label-uncertainty for brain tumor segmentation. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International

Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I (2019), A. Crimi and S. Bakas, Eds., vol. 11992 of Lecture Notes in Computer Science, Springer, pp. 379–387. (↑52), (↑177)

[155] MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K., AND GALSTYAN, A. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54, 6 (2021), 1–35. (↑40)

[156] MEHTA, R., ALBIERO, V., CHEN, L., EVTIMOV, I., GLASER, T., LI, Z., AND HASSNER, T. You only need a good embeddings extractor to fix spurious correlations. arXiv preprint arXiv:2212.06254 (2022). (↑6)

[157] MEHTA, R., AND ARBEL, T. 3d u-net for brain tumour segmentation. In International MICCAI Brainlesion Workshop (2018), Springer, pp. 254–266. (↑7), (↑49)

[158] MEHTA, R., AND ARBEL, T. Rs-net: Regression-segmentation 3d cnn for synthesis of full resolution missing brain mri in the presence of tumours. In International Workshop on Simulation and Synthesis in Medical Imaging (2018), Springer, pp. 119–129. (↑5), (↑XX), (↑XXI), (↑XXIII), (↑XXIX), (↑XXX), (↑2), (↑18), (↑82), (↑85), (↑149), (↑150), (↑155), (↑156), (↑157), (↑158), (↑160), (↑163), (↑164), (↑165), (↑166), (↑181)

[159] MEHTA, R., CHRISTINCK, T., NAIR, T., BUSSY, A., PREMASIRI, S., COSTANTINO, M., CHAKRAVARTHY, M. M., ARNOLD, D. L., GAL, Y., AND ARBEL, T. Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. IEEE Transactions on Medical Imaging 41, 2 (2021), 360–373. (↑2), (↑4), (↑XII), (↑XIII), (↑XXII), (↑XXIII), (↑XXVII), (↑XXVIII), (↑77), (↑79), (↑81), (↑89), (↑90), (↑91), (↑92), (↑93), (↑94), (↑103), (↑179), (↑181), (↑182), (↑183), (↑184)

[160] MEHTA, R., CHRISTINCK, T., NAIR, T., LEMAITRE, P., ARNOLD, D., AND ARBEL, T. Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. In Uncertainty for Safe Utilization of Machine Learning in

Medical Imaging and Clinical Image-Based Procedures. Springer, 2019, pp. 23–32. (↑4)

[161] MEHTA, R., FILOS, A., BAID, U., SAKO, C., MCKINLEY, R., REBSAMEN, M., DÄTWYLER, K., MEIER, R., RADOJEWSKI, P., MURUGESAN, G. K., ET AL. Qu-brats: Miccai brats 2020 challenge on quantifyinguncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results. Journal of Machine Learning for Biomedical Imaging 1 (2022). (↑1), (↑5), (↑VII), (↑VIII), (↑IX), (↑X), (↑XI), (↑XII), (↑XXI), (↑XXII), (↑XXV), (↑XXVI), (↑XXX), (↑XXXII), (↑45), (↑48), (↑49), (↑58), (↑60), (↑61), (↑62), (↑63), (↑64), (↑65), (↑66), (↑71), (↑72), (↑73), (↑74), (↑75), (↑169), (↑170), (↑171), (↑172), (↑173), (↑174), (↑175), (↑177), (↑194), (↑195), (↑196), (↑197)

[162] MEHTA, R., FILOS, A., GAL, Y., AND ARBEL, T. Uncertainty evaluation metric for brain tumour segmentation. arXiv preprint arXiv:2005.14262 (2020). (↑5), (↑XVI), (↑XXIV), (↑46), (↑115), (↑191), (↑192), (↑193)

[163] MEHTA, R., MAJUMDAR, A., AND SIVASWAMY, J. Brainsegnet: a convolutional neural network architecture for automated segmentation of human brain structures. Journal of Medical Imaging 4, 2 (2017), 024003–024003. (↑2)

[164] MEHTA, R., SHUI, C., AND ARBEL, T. Evaluating the fairness of deep learning uncertainty estimates in medical image analysis. In Medical Imaging with Deep Learning, 10-12 July 2023, Nashville, USA (2023), Proceedings of Machine Learning Research, PMLR. (↑3), (↑XIV), (↑XV), (↑XVI), (↑XVII), (↑XXIV), (↑XXV), (↑XXVIII), (↑XXX), (↑XXXI), (↑XXXII), (↑99), (↑105), (↑106), (↑107), (↑110), (↑111), (↑112), (↑113), (↑114), (↑115), (↑116), (↑118), (↑119), (↑120), (↑122), (↑186), (↑187), (↑188), (↑189), (↑190), (↑191), (↑192), (↑193), (↑194), (↑195), (↑196), (↑197), (↑198)

[165] MEHTA, R., SHUI, C., NICHYPORUK, B., AND ARBEL, T. Information gain sampling for active learning in medical image classification. In International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2022), Springer, pp. 135–145. (↑3), (↑XVIII), (↑XIX), (↑XXXII), (↑XXXIII), (↑124), (↑126),

(↑131), (↑132), (↑134), (↑135), (↑136), (↑199), (↑200), (↑201)

[166] MEHTA, R., AND SIVASWAMY, J. A hybrid approach to tissue-based intensity standardization of brain mri images. In 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI) (2016), IEEE, pp. 95–98. (↑29)

[167] MEHTA, R., AND SIVASWAMY, J. M-net: A convolutional neural network for deep brain structure segmentation. In 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017) (2017), IEEE, pp. 437–440. (↑2)

[168] MENG, C., TRINH, L., XU, N., ENOUEN, J., AND LIU, Y. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. Scientific Reports 12, 1 (2022), 7166. (↑147)

[169] MENZE, B., JOSKOWICZ, L., BAKAS, S., JAKAB, A., KONUKOGLU, E., AND BECKER, A. Qubiq - grand challenge, 2020. (↑141)

[170] 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, C., 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., AND LEEMPUT, K. V. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging 34, 10 (Oct. 2015), 1993–2024. (↑9), (↑49), (↑50), (↑61)

[171] MILLETARI, F., NAVAB, N., AND AHMADI, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 fourth international

conference on 3D vision (3DV) (2016), IEEE, pp. 565–571. (↑55)

[172] MOCCIA, M., DE STEFANO, N., AND BARKHOF, F. Imaging outcome measures for progressive multiple sclerosis trials. Multiple Sclerosis Journal 23, 12 (2017), 1614–1626. (↑19)

[173] MOLLE, P. V., VERBELEN, T., BOOM, C. D., VANKEIRSBILCK, B., VYLDER, J. D., DIRICX, B., KIMPE, T., SIMOENS, P., AND DHOEDT, B. Quantifying uncertainty of deep neural networks in skin lesion classification. In Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures. Springer, 2019, pp. 52–61. (↑41)

[174] MONTEIRO, M., FOLGOC, L. L., DE CASTRO, D. C., PAWLOWSKI, N., MARQUES, B., KAMNITSAS, K., VAN DER WILK, M., AND GLOCKER, B. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. (↑37), (↑141)

[175] MOVSHOVITZ-ATTIAS, Y., KANADE, T., AND SHEIKH, Y. How useful is photo-realistic rendering for visual learning? In European Conference on Computer Vision (2016), Springer, pp. 202–217. (↑41)

[176] MUKHOTI, J., AND GAL, Y. Evaluating bayesian deep learning methods for semantic segmentation. arXiv preprint arXiv:1811.12709 (2018). (↑3), (↑35)

[177] MURUGESAN, G. K., NALAWADE, S., GANESH, C., WAGNER, B., YU, F. F., FEI, B., MADHURANTHAKAM, A. J., AND MALDJIAN, J. A. Multidimensional and multiresolution ensemble networks for brain tumor segmentation. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 448–457. (↑53)

[178] MURUGESAN, G. K., NALAWADE, S. S., YOGANANDA, C. G. B., WAGNER, B. C., YU, F. F., FEI, B., MADHURANTHAKAM, A. J., AND MALDJIAN, J. A. Multidimen-

sional and multiresolution ensemble networks for brain tumor segmentation. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II (2019), A. Crimi and S. Bakas, Eds., vol. 11993 of Lecture Notes in Computer Science, Springer, pp. 148–157. (↑177)

[179] MYRONENKO, A., AND HATAMIZADEH, A. Robust semantic segmentation of brain tumor regions from 3d mris. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II (2019), A. Crimi and S. Bakas, Eds., vol. 11993 of Lecture Notes in Computer Science, Springer, pp. 82–89. (↑177)

[180] NAIR, T., PRECUP, D., ARNOLD, D. L., AND ARBEL, T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical image analysis 59 (2020), 101557. (↑XV), (↑XXII), (↑2), (↑4), (↑20), (↑23), (↑31), (↑32), (↑36), (↑40), (↑65), (↑80), (↑85), (↑86), (↑95), (↑113), (↑179), (↑183)

[181] NATH, V., YANG, D., LANDMAN, B. A., XU, D., AND ROTH, H. R. Diminishing uncertainty within the training pool: Active learning for medical image segmentation. IEEE Transactions on Medical Imaging 40, 10 (2020), 2534–2547. (↑146)

[182] NEAL, R. M. Bayesian learning for neural networks, vol. 118. Springer Science & Business Media, 2012. (↑102), (↑103)

[183] NGO, D.-K., TRAN, M.-T., KIM, S.-H., YANG, H.-J., AND LEE, G.-S. Multi-task learning for small brain tumor segmentation from mri. Applied Sciences 10, 21 (2020), 7790. (↑112)

[184] NICHYPORUK, B., CARDINELL, J., SZETO, J., MEHTA, R., FALET, J.-P. R., ARNOLD, D. L., TSAFTARIS, S. A., AND ARBEL, T. Rethinking generalization: The impact of annotation style on medical image segmentation. Journal of Machine Learning for Biomedical Imaging 1 (2022). (↑6)

[185] NICHYPORUK, B., CARDINELL, J., SZETO, J., MEHTA, R., TSAFTARIS, S., ARNOLD, D. L., AND ARBEL, T. Cohort bias adaptation in aggregated datasets for lesion segmentation. In Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health. Springer, 2021, pp. 101–111. (↑6)

[186] NIXON, J., DUSENBERRY, M. W., ZHANG, L., JERFEL, G., AND TRAN, D. Measuring calibration in deep learning. In CVPR workshops (2019), vol. 2. (↑34)

[187] OH, X., LIM, R., LOH, L., TAN, C. H., FOONG, S., AND TAN, U.-X. Monocular uav localisation with deep learning and uncertainty propagation. IEEE Robotics and Automation Letters 7, 3 (2022), 7998–8005. (↑143)

[188] OZDEMIR, O., WOODWARD, B., AND BERLIN, A. A. Propagating uncertainty in multi-stage bayesian convolutional neural networks with application to pulmonary nodule detection. arXiv preprint arXiv:1712.00497 (2017). (↑31), (↑32), (↑33), (↑89)

[189] PAPAMAKARIOS, G., PAVLAKOU, T., AND MURRAY, I. Masked autoregressive flow for density estimation. Advances in neural information processing systems 30 (2017). (↑39)

[190] PAPANDREOU, G., CHEN, L.-C., MURPHY, K. P., AND YUILLE, A. L. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In Proceedings of the IEEE international conference on computer vision (2015), pp. 1742–1750. (↑41)

[191] PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N., ANTIGA, L., ET AL. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32 (2019). (↑107), (↑113), (↑117)

[192] PATEL, J., CHANG, K., HOEBEL, K., GIDWANI, M., ARUN, N., GUPTA, S., AGGARWAL, M., SINGH, P., ROSEN, B. R., GERSTNER, E. R., AND KALPATHY-CRAMER, J. Segmentation, survival prediction, and uncertainty estimation of gliomas from multimodal 3d MRI using selective kernel networks. In Brainlesion: Glioma,

Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 228–240. (↑56)

[193] PATHAK, D., KRAHENBUHL, P., AND DARRELL, T. Constrained convolutional neural networks for weakly supervised segmentation. In Proceedings of the IEEE international conference on computer vision (2015), pp. 1796–1804. (↑41)

[194] PEI, L., MURAT, A. K., AND COLEN, R. Multimodal brain tumor segmentation and survival prediction using a 3d self-ensemble ResUNet. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 367–375. (↑54)

[195] PEI, L., VIDYARATNE, L., HSU, W., RAHMAN, M. M., AND IFTEKHARUDDIN, K. M. Brain tumor classification using 3d convolutional neural network. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II (2019), A. Crimi and S. Bakas, Eds., vol. 11993 of Lecture Notes in Computer Science, Springer, pp. 335–342. (↑177)

[196] PRINCE, J. L., CARASS, A., ZHAO, C., DEWEY, B. E., ROY, S., AND PHAM, D. L. Image synthesis and superresolution in medical imaging. In Handbook of Medical Image Computing and Computer Assisted Intervention. Elsevier, 2020, pp. 1–24. (↑82)

[197] PUYOL-ANTÓN, E., RUIJSINK, B., PIECHNIK, S. K., NEUBAUER, S., PETERSEN, S. E., RAZAVI, R., AND KING, A. P. Fairness in cardiac mr image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2021), Springer, pp. 413–423. (↑39), (↑40), (↑100), (↑101), (↑104)

[198] REBSAMEN, M., KNECHT, U., REYES, M., WIEST, R., MEIER, R., AND MCKINLEY, R. Divide and conquer: stratifying training data by tumor grade improves deep

learning-based brain tumor segmentation. Frontiers in neuroscience 13 (2019), 1182. (↑112)

[199] REINHOLD, J. C., HE, Y., HAN, S., CHEN, Y., GAO, D., LEE, J., PRINCE, J. L., AND CARASS, A. Validating uncertainty in medical image translation. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) (2020), IEEE, pp. 95–98. (↑82), (↑89)

[200] REINKE, A., EISENMANN, M., TIZABI, M. D., SUDRE, C. H., RÄDSCH, T., ANTONELLI, M., ARBEL, T., CARDOSO, M. J., CHEPLYGINA, V., FARAHANI, K., GLOCKER, B., HECKMANN-NÖTZEL, D., ISENSEE, F., JANNIN, P., KAHN, C. E., KLEESIEK, J., KURÇ, T. M., KOZUBEK, M., LANDMAN, B. A., LITJENS, G., MAIER-HEIN, K. H., MENZE, B. H., MÜLLER, H., PETERSEN, J., REYES, M., RIEKE, N., STIELTJES, B., SUMMERS, R. M., TSAFTARIS, S. A., VAN GINNEKEN, B., KOPP-SCHNEIDER, A., JÄGER, P., AND MAIER-HEIN, L. Common limitations of image processing metrics: A picture story. CoRR abs/2104.05642 (2021). (↑140)

[201] REN, S., HE, K., GIRSHICK, R., AND SUN, J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28 (2015). (↑1)

[202] RICCI LARA, M. A., ECHEVESTE, R., AND FERRANTE, E. Addressing fairness in artificial intelligence for medical imaging. nature communications 13, 1 (2022), 1–6. (↑19), (↑39)

[203] RÍO, J., AUGER, C., AND ROVIRA, À. Mr imaging in monitoring and predicting treatment response in multiple sclerosis. Neuroimaging Clinics 27, 2 (2017), 277–287. (↑19)

[204] RITTER, H., BOTEV, A., AND BARBER, D. A scalable laplace approximation for neural networks. In 6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings (2018), vol. 6, International Conference on Representation Learning. (↑23)

[205] ROHLFING, T., ZAHR, N. M., SULLIVAN, E. V., AND PFEFFERBAUM, A. The sri24 multichannel atlas of normal adult human brain structure. Human brain mapping 31, 5 (2010), 798–819. (↑50)

[206] ROSAS-GONZÁLEZ, S., ZEMMOURA, I., AND TAUBER, C. 3d brain tumor segmentation and survival prediction using ensembles of convolutional neural networks. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 241–254. (↑54)

[207] ROSEN, W. G., MOHS, R. C., AND DAVIS, K. L. A new rating scale for alzheimer's disease. The American journal of psychiatry (1984). (↑21), (↑83)

[208] ROY, A. G., CONJETI, S., NAVAB, N., WACHINGER, C., INITIATIVE, A. D. N., ET AL. Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage 195 (2019), 11–22. (↑23), (↑31), (↑32)

[209] ROY, N., AND MCCALLUM, A. Toward optimal active learning through monte carlo estimation of error reduction. ICML, Williamstown 2 (2001), 441–448. (↑42), (↑124), (↑128)

[210] ROY, S., CARASS, A., AND PRINCE, J. A compressed sensing approach for mr tissue contrast synthesis. In Biennial International Conference on Information Processing in Medical Imaging (2011), Springer, pp. 371–383. (↑18), (↑150)

[211] RUPPRECHT, C., LAINA, I., DIPIETRO, R., BAUST, M., TOMBARI, F., NAVAB, N., AND HAGER, G. D. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In Proceedings of the IEEE international conference on computer vision (2017), pp. 3591–3600. (↑37), (↑38)

[212] SAERENS, M., LATINNE, P., AND DECAESTECKER, C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. Neural computation 14, 1 (2002), 21–41. (↑127)

[213] SAGAWA, S., KOH, P. W., HASHIMOTO, T. B., AND LIANG, P. Distributionally robust neural networks for group shifts: On the importance of regularization for

worst-case generalization. arXiv preprint arXiv:1911.08731 (2019). (↑6), (↑40), (↑100), (↑104)

[214] SALAHUDDIN, Z., WOODRUFF, H. C., CHATTERJEE, A., AND LAMBIN, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Computers in biology and medicine 140 (2022), 105111. (↑146)

[215] SARHAN, M. H., NAVAB, N., ESLAMI, A., AND ALBARQOUNI, S. Fairness by learning orthogonal disentangled representations. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16 (2020), Springer, pp. 746–761. (↑40)

[216] SAXENA, A., SUN, M., AND NG, A. Y. Make3d: Learning 3d scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence 31, 5 (2008), 824–840. (↑30)

[217] SCHEFFER, T., AND WROBEL, S. Active learning of partially hidden markov models. In In Proceedings of the ECML/PKDD Workshop on Instance Selection (2001), Citeseer. (↑42)

[218] SEDAI, S., ANTONY, B., RAI, R., JONES, K., ISHIKAWA, H., SCHUMAN, J., GADI, W., AND GARNAVI, R. Uncertainty guided semi-supervised segmentation of retinal layers in oct images. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), Springer, pp. 282–290. (↑32)

[219] SELVAN, R., FAYE, F., MIDDLETON, J., AND PAI, A. Uncertainty quantification in medical image segmentation with normalizing flows. In Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 11 (2020), Springer, pp. 80–90. (↑39), (↑146)

[220] SENER, O., AND SAVARESE, S. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017). (↑7), (↑42), (↑133)

[221] SENGE, R., BÖSNER, S., DEMBCZYŃSKI, K., HAASENRITTER, J., HIRSCH, O., DONNER-BANZHOFF, N., AND HÜLLERMEIER, E. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. Information Sciences 255 (2014), 16–29. (↑29)

[222] SEPAHVAND, N. M., HASSNER, T., ARNOLD, D. L., AND ARBEL, T. Cnn prediction of future disease activity for multiple sclerosis patients from baseline mri and lesion labels. In International MICCAI Brainlesion Workshop (2019), Springer, pp. 57–69. (↑19)

[223] SETTLES, B. Active learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences, 2009. (↑41)

[224] SEUNG, H. S., OPPER, M., AND SOMPOLINSKY, H. Query by committee. In Proceedings of the fifth annual workshop on Computational learning theory (1992), pp. 287–294. (↑7)

[225] SEYYED-KALANTARI, L., ZHANG, H., MCDERMOTT, M., CHEN, I. Y., AND GHASSEMI, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature medicine 27, 12 (2021), 2176–2182. (↑39)

[226] SHANNON, C. E. A mathematical theory of communication. The Bell system technical journal 27, 3 (1948), 379–423. (↑7), (↑42), (↑133)

[227] SHARMA, D., SHANIS, Z., REDDY, C. K., GERBER, S., AND ENQUOBAHRIE, A. Active learning technique for multimodal brain tumor segmentation using limited labeled images. In Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data. Springer, 2019, pp. 148–156. (↑32)

[228] SHELLER, M. J., EDWARDS, B., REINA, G. A., MARTIN, J., PATI, S., KOTROTSOU, A., MILCHENKO, M., XU, W., MARCUS, D., COLEN, R. R., ET AL. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific reports 10, 1 (2020), 1–12. (↑141)

[229] SHI, X., DOU, Q., XUE, C., QIN, J., CHEN, H., AND HENG, P.-A. An active learning approach for reducing annotation cost in skin lesion analysis. In International Workshop on Machine Learning in Medical Imaging (2019), Springer, pp. 628–636. (↑41)

[230] SHUI, C., SZETO, J., MEHTA, R., ARNOLD, D., , AND ARBEL, T. Mitigating calibration bias without fixed attribute grouping for improved fairness in medical image analysis. In 26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023, Vancouver, October 2023, Proceedings (2023), Springer. (↑6)

[231] SHUI, C., XU, G., CHEN, Q., LI, J., LING, C. X., ARBEL, T., WANG, B., AND GAGNÉ, C. On learning fairness and accuracy on multiple subgroups. Advances in Neural Information Processing Systems 35 (2022), 34121–34135. (↑145)

[232] SHUI, C., ZHOU, F., GAGNÉ, C., AND WANG, B. Deep active learning: Unified and principled method for query and training. In International Conference on Artificial Intelligence and Statistics (2020), PMLR, pp. 1308–1318. (↑42)

[233] SILBERMAN, N., HOIEM, D., KOHLI, P., AND FERGUS, R. Indoor segmentation and support inference from rgbd images. In European conference on computer vision (2012), Springer, pp. 746–760. (↑30)

[234] SIMON, R. P., AMINOFF, M. J., AND GREENBERG, D. A. Clinical Neurology. Lange Medical Books/McGraw-Hill, 2009. (↑20)

[235] SINHA, S., EBRAHIMI, S., AND DARRELL, T. Variational adversarial active learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (2019), pp. 5972–5981. (↑42), (↑145)

[236] SMAILAGIC, A., NOH, H. Y., COSTA, P., WALAWALKAR, D., KHANDELWAL, K., MIRSHEKARI, M., FAGERT, J., GALDRÁN, A., AND XU, S. Medal: Deep active learning sampling method for medical image analysis. arXiv preprint arXiv:1809.09287 (2018). (↑42)

[237] SMITH, L., AND GAL, Y. Understanding measures of uncertainty for adversarial example detection. arXiv preprint arXiv:1803.08533 (2018). (↑23), (↑26), (↑50), (↑79), (↑87), (↑88), (↑95), (↑102), (↑103), (↑107), (↑113), (↑114), (↑117)

[238] SOKOL, K., HEPBURN, A., SANTOS-RODRIGUEZ, R., AND FLACH, P. blimey: surrogate prediction explanations beyond lime. arXiv preprint arXiv:1910.13016 (2019). (↑146)

[239] SONG, S., LICHTENBERG, S. P., AND XIAO, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the IEEE conference on computer vision and pattern recognition (2015), pp. 567–576. (↑31)

[240] SORMANI, M. P., AND BRUZZI, P. Mri lesions as a surrogate for relapses in multiple sclerosis: a meta-analysis of randomised trials. The Lancet Neurology 12, 7 (2013), 669–676. (↑19)

[241] SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I., AND SALAKHUTDINOV, R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15, 1 (2014), 1929–1958. (↑25), (↑106), (↑152)

[242] STANGEL, M., PENNER, I. K., KALLMANN, B. A., LUKAS, C., AND KIESEIER, B. C. Towards the implementation of 'no evidence of disease activity'in multiple sclerosis treatment: the multiple sclerosis decision model. Therapeutic advances in neurological disorders 8, 1 (2015), 3–13. (↑19)

[243] STONNINGTON, C. M., CHU, C., KLÖPPEL, S., JACK JR, C. R., ASHBURNER, J., FRACKOWIAK, R. S., INITIATIVE, A. D. N., ET AL. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. Neuroimage 51, 4 (2010), 1405–1413. (↑83)

[244] SUN, C., SHRIVASTAVA, A., SINGH, S., AND GUPTA, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision (2017), pp. 843–852. (↑41)

[245] SÜNDERHAUF, N., DAYOUB, F., HALL, D., SKINNER, J., ZHANG, H., CARNEIRO, G., AND CORKE, P. A probabilistic challenge for object detection. Nature Machine Intelligence 1, 9 (Sept. 2019), 443–443. (↑36)

[246] TANNO, R., WORRALL, D. E., GHOSH, A., KADEN, E., SOTIROPOULOS, S. N., CRIMINISI, A., AND ALEXANDER, D. C. Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2017), Springer, pp. 611–619. (↑32)

[247] TANNO, R., WORRALL, D. E., KADEN, E., GHOSH, A., GRUSSU, F., BIZZI, A., SOTIROPOULOS, S. N., CRIMINISI, A., AND ALEXANDER, D. C. Uncertainty modelling in deep learning for safer neuroimage enhancement: demonstration in diffusion mri. NeuroImage 225 (2021), 117366. (↑40)

[248] TARDY, M., SCHEFFER, B., AND MATEUS, D. Uncertainty measurements for the reliable classification of mammograms. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), Springer, pp. 495–503. (↑31), (↑32), (↑36)

[249] TARTAGLIONE, E., BARBANO, C. A., AND GRANGETTO, M. End: Entangling and disentangling deep representations for bias correction. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021), pp. 13508–13517. (↑40)

[250] TEYE, M., AZIZPOUR, H., AND SMITH, K. Bayesian uncertainty estimation for batch normalized deep networks. In International Conference on Machine Learning (2018), PMLR, pp. 4907–4916. (↑2), (↑33)

[251] THAKUR, S., DOSHI, J., PATI, S., RATHORE, S., SAKO, C., BILELLO, M., HA, S. M., SHUKLA, G., FLANDERS, A., KOTROTSOU, A., ET AL. Brain extraction on mri scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. NeuroImage 220 (2020), 117081. (↑50)

[252] TIAN, J., LIU, Y.-C., GLASER, N., HSU, Y.-C., AND KIRA, Z. Posterior re-calibration for imbalanced datasets. Advances in Neural Information Processing Systems 33 (2020), 8101–8113. (↑127)

[253] TOUSIGNANT, A., LEMAÎTRE, P., PRECUP, D., ARNOLD, D. L., AND ARBEL, T. Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data. In International Conference on Medical Imaging with Deep Learning (2019), pp. 483–492. (↑2), (↑89)

[254] TULDER, G. V., AND BRUIJNE, M. D. Why does synthesized data improve multi-sequence classification? In International Conference on Medical Image Computing and Computer-Assisted Intervention (2015), Springer, pp. 531–538. (↑18), (↑82), (↑85), (↑149)

[255] ULMER, D. A survey on evidential deep learning for single-pass uncertainty estimation. arXiv preprint arXiv:2110.03051 (2021). (↑142)

[256] ULYANOV, D., VEDALDI, A., AND LEMPITSKY, V. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016). (↑82), (↑113), (↑152), (↑182)

[257] VADACCHINO, S., MEHTA, R., SEPAHVAND, N. M., NICHYPORUK, B., CLARK, J. J., AND ARBEL, T. Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images. In Medical Imaging with Deep Learning, 7-9 July 2021, Lübeck, Germany (2021), vol. 143 of Proceedings of Machine Learning Research, PMLR, pp. 787–801. (↑7)

[258] VAN AMERSFOORT, J., SMITH, L., TEH, Y. W., AND GAL, Y. Uncertainty estimation using a single deep deterministic neural network. In International Conference on Machine Learning (2020), PMLR, pp. 9690–9700. (↑142)

[259] VAN NGUYEN, H., ZHOU, K., AND VEMULAPALLI, R. Cross-domain synthesis of medical images using efficient location-sensitive deep network. In International

Conference on Medical Image Computing and Computer-Assisted Intervention (2015), Springer, pp. 677–684. (↑150), (↑151), (↑154), (↑155)

[260] VENTURINI, L., PAPAGEORGHIOU, A. T., NOBLE, J. A., AND NAMBURETE, A. I. Uncertainty estimates as data selection criteria to boost omni-supervised learning. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2020), Springer, pp. 689–698. (↑32), (↑89)

[261] VERMA, S., DICKERSON, J., AND HINES, K. Counterfactual explanations for machine learning: Challenges revisited. arXiv preprint arXiv:2106.07756 (2021). (↑147)

[262] VU, M. H., NYHOLM, T., AND LÖFSTEDT, T. Tunet: End-to-end hierarchical brain tumor segmentation using cascaded networks. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I (2019), A. Crimi and S. Bakas, Eds., vol. 11992 of Lecture Notes in Computer Science, Springer, pp. 174–186. (↑177)

[263] VU, M. H., NYHOLM, T., AND LÖFSTEDT, T. Multi-decoder networks with multi-denoising inputs for tumor segmentation. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2021, pp. 412–423. (↑55)

[264] WALTON, C., KING, R., RECHTMAN, L., KAYE, W., LERAY, E., MARRIE, R. A., ROBERTSON, N., LA ROCCA, N., UITDEHAAG, B., VAN DER MEI, I., ET AL. Rising prevalence of multiple sclerosis worldwide: Insights from the atlas of ms. Multiple Sclerosis Journal 26, 14 (2020), 1816–1821. (↑19), (↑144)

[265] WANG, G., LI, W., ZULUAGA, M. A., PRATT, R., PATEL, P. A., AERTSEN, M., DOEL, T., DAVID, A. L., DEPREST, J., OURSELIN, S., AND VERCAUTEREN, T. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans. Medical Imaging 37, 7 (2018), 1562–1573. (↑54)

[266] WANG, J., YAN, Y., ZHANG, Y., CAO, G., YANG, M., AND NG, M. K. Deep reinforcement active learning for medical image classification. In International

Conference on Medical Image Computing and Computer-Assisted Intervention (2020), Springer, pp. 33–42. (↑42), (↑132)

[267] WANG, Z., BOVIK, A. C., SHEIKH, H. R., AND SIMONCELLI, E. P. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 4 (2004), 600–612. (↑155)

[268] WANG, Z., QINAMI, K., KARAKOZIS, I. C., GENOVA, K., NAIR, P., HATA, K., AND RUSSAKOVSKY, O. Towards fairness in visual recognition: Effective strategies for bias mitigation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020), pp. 8919–8928. (↑40)

[269] WEISS, G. M., AND PROVOST, F. Learning when training data are costly: The effect of class distribution on tree induction. Journal of artificial intelligence research 19 (2003), 315–354. (↑127)

[270] WEN, S., KURC, T. M., HOU, L., SALTZ, J. H., GUPTA, R. R., BATISTE, R., ZHAO, T., NGUYEN, V., SAMARAS, D., AND ZHU, W. Comparison of different classifiers with active learning to support quality control in nucleus segmentation in pathology images. AMIA Summits on Translational Science Proceedings 2018 (2018), 227. (↑42)

[271] WEN, Y., TRAN, D., AND JIMMY, B. Batchensemble: An alternative approach to efficient ensemble and lifelong learning. In 8th International Conference on Learning Representations, ICLR 2020-Conference Track Proceedings (2020), International Conference on Representation Learning. (↑23), (↑33)

[272] WENZEL, F., SNOEK, J., TRAN, D., AND JENATTON, R. Hyperparameter ensembles for robustness and uncertainty quantification. Advances in neural information processing systems 33 (2020). (↑23), (↑102)

[273] WOLTERINK, J. M., DINKLA, A. M., SAVENIJE, M. H., SEEVINCK, P. R., VAN DEN BERG, C. A., AND IŠGUM, I. Deep mr to ct synthesis using unpaired data. In International workshop on simulation and synthesis in medical imaging (2017), Springer, pp. 14–23. (↑150)

[274] XU, M., ZHANG, T., LI, Z., LIU, M., AND ZHANG, D. Towards evaluating the robustness of deep diagnostic models by adversarial attack. Medical Image Analysis 69 (2021), 101977. (↑146)

[275] YANG, L., ZHANG, Y., CHEN, J., ZHANG, S., AND CHEN, D. Z. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In International conference on medical image computing and computer-assisted intervention (2017), Springer, pp. 399–407. (↑42), (↑146)

[276] YOO, D., AND KWEON, I. S. Learning loss for active learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019), pp. 93–102. (↑42)

[277] YU, L., WANG, S., LI, X., FU, C.-W., AND HENG, P.-A. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), Springer, pp. 605–613. (↑32), (↑36)

[278] ZECH, J. R., BADGELEY, M. A., LIU, M., COSTA, A. B., TITANO, J. J., AND OERMANN, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine 15, 11 (2018), e1002683. (↑141)

[279] ZHANG, D., WANG, Y., ZHOU, L., YUAN, H., SHEN, D., INITIATIVE, A. D. N., ET AL. Multimodal classification of alzheimer's disease and mild cognitive impairment. Neuroimage 55, 3 (2011), 856–867. (↑83)

[280] ZHANG, J., XIE, Y., WU, Q., AND XIA, Y. Skin lesion classification in dermoscopy images using synergic deep learning. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2018), Springer, pp. 12–20. (↑41)

[281] ZHANG, R., ISOLA, P., AND EFROS, A. A. Colorful image colorization. In European conference on computer vision (2016), Springer, pp. 649–666. (↑150)

[282] ZHAO, H., COSTON, A., ADEL, T., AND GORDON, G. J. Conditional learning of fair representations. arXiv preprint arXiv:1910.07162 (2019). (↑40)

[283] ZIETLOW, D., LOHAUS, M., BALAKRISHNAN, G., KLEINDESSNER, M., LOCATELLO, F., SCHÖLKOPF, B., AND RUSSELL, C. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022), pp. 10410–10421. (↑102)

[284] ZONG, Y., YANG, Y., AND HOSPEDALES, T. Medfair: Benchmarking fairness for medical imaging. arXiv preprint arXiv:2210.01725 (2022). (↑6), (↑19), (↑39), (↑40), (↑100), (↑104), (↑146)

[285] ZOTOVA, D., LISOWSKA, A., ANDERSON, O., DILYS, V., AND O'NEIL, A. Comparison of active learning strategies applied to lung nodule segmentation in ct scans. In Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention. Springer, 2019, pp. 3–12. (↑32)

[286] ZOU, K., CHEN, Z., YUAN, X., SHEN, X., WANG, M., AND FU, H. A review of uncertainty estimation and its application in medical imaging. arXiv preprint arXiv:2302.08119 (2023). (↑146)