Imperial College of Science, Technology and Medicine
Department of Computing

Symbiotic Deep Learning for Medical Image Analysis with *Applications in Real-time Diagnosis for Fetal Ultrasound Screening*

Samuel Budd

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of Imperial College, September 2021

Copyright

Declaration

I hereby declare this work original unless otherwise stated.

Samuel Budd

September 2021

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. Bernhard Kainz for the wonderful mentorship he has given me throughout this PhD, both as an academic and as a role-model for the future. His guidance, support and enthusiasm have been invaluable and provided an unending source of motivation throughout, without which the quantity and quality of research would undoubtedly be a lot less!

I would also like to thank everyone within the BioMedIA group past and present, who have been (and will continue to be) wonderful friends and a brilliant inspiration. Special thanks to Dr. Emma C Robinson, Prof. Daniel Rueckert, Dr. Ben Glocker, Thanos Vlontzos, Dr. Benjamin Hou, Dr. Matthew Sinclair, Miguel Monteiro, Nick Pawlowski and Jeremy Tan for many insightful discussions and lucrative collaborations and much more.

I express my thanks to my friends and house-mates who have endured endless discussions of topics that span many more domains than this thesis and have kept me grounded and helped ensure this research will one day be of a great use to many.

And of course I would like to express my love and thanks to my family for providing unconditional support and reminding me to take stock and remember why we're all doing this in the first place!

Dedication

I dedicate this thesis to my late Grandmother and Grandfather, Margaret and Fred, who without which so much joy would have been kept from the world. Here's to immortalising you and everything you gave our world.

'The world is full of wonders, but they become more wonderful, not less wonderful when science looks at them.'

Sir David Attenborough

Abstract

The last hundred years have seen a monumental rise in the power and capability of machines to perform intelligent tasks in the stead of previously human operators. This rise is not expected to slow down any time soon and what this means for society and humanity as a whole remains to be seen. The overwhelming notion is that with the right goals in mind, the growing influence of machines on our every day tasks will enable humanity to give more attention to the truly groundbreaking challenges that we all face together. This will usher in a new age of human machine collaboration in which humans and machines may work side by side to achieve greater heights for all of humanity. Intelligent systems are useful in isolation, but the true benefits of intelligent systems come to the fore in complex systems where the interaction between humans and machines can be made seamless, and it is this goal of symbiosis between human and machine that may democratise complex knowledge, which motivates this thesis. In the recent past, data-driven methods have come to the fore and now represent the state-of-the-art in many different fields. Alongside the shift from rule-based towards data-driven methods we have also seen a shift in how humans interact with these technologies. Human computer interaction is changing in response to data-driven methods and new techniques must be developed to enable the same symbiosis between man and machine for data-driven methods as for previous formula-driven technology.

We address five key challenges which need to be overcome for data-driven human-in-the-loop computing to reach maturity. These are **(1)** the 'Categorisation Challenge' where we examine existing work and form a taxonomy of the different methods being utilised for data-driven human-in-the-loop computing; **(2)** the 'Confidence Challenge', where data-driven methods must communicate interpretable beliefs in how confident their predictions are; **(3)** the 'Complexity Challenge' where the aim of reasoned communication becomes increasingly important as the complexity of tasks and methods to solve also increases; **(4)** the 'Classification Challenge' in which we look at how complex methods can be separated in order to provide greater reasoning in complex classification tasks; and finally **(5)** the 'Curation Challenge' where we challenge the assumptions around bottleneck creation for the development of supervised learning methods.

Contents

List of Tables

List of Figures

Chapter 1

Introduction: The Challenge of Human Input

1.1 A Short Story of Symbiosis

The last century has seen the capabilities of machines grow exponentially. The early days of computing saw humans instructing machines with analogue punch cards to make calculations too numerous for any group of individuals to do by hand. Since then machines have been used to perform feats beyond comprehension for humans 100 years ago. From the literally out-of-this-world accomplishments of landing humans on the moon, to more down-to-earth accomplishments of computers defeating chess grandmasters, one constant has remained from day one - these achievements were not made alone: not by any individual human, and not by any individual machine, but by humans and machines co-operating in symbiosis to achieve a greater goal (although there have been and will continue to be many remarkable individuals whose contributions have been invaluable in this endeavour).

It is not only machine's growing computational power that has enabled greater and greater feats of wonder, but human's growing understanding of how best to leverage that power. Humanities greatest developments have been enabled by humans, behind the scenes, squeezing every last drop of potential from our resources, and those in the public eye, interfacing with cutting edge technology, feeding, reacting and responding to rapidly updating information. It is this dual symbiosis between human and machine that all technologies seek to acquire - depth of understanding, and ease of use. The dominant efforts of human computer interaction have been enabled by rigorously understood technology, based on formulae driven computation through which every behaviour of a system can be known and verified before use. However, in the recent past, data-driven technology, developed through learning from data and observations of the real world, has become common place in society. This data-driven approach has enabled technology to uncover and perform tasks requiring greater levels of machine intelligence, but this has come at a cost. Technology driven by models developed from observed data is no longer rigorously understood, the behaviour of these systems can no longer be verified in all unseen scenarios, and the ability to reason about the underlying mechanisms driving those technologies has been greatly reduced. This has led to a wide spanning effort to open these 'black-box' technologies to lead to a greater understanding of their inner workings and to enable a new symbiosis between humans and data-driven machines. This effort is two fold, requiring data-driven models to communicate their outputs and reasoning effectively to users, and for users to be able to provide feedback to these systems and guide their future learning and improvement.

Alongside (and intertwined with) technology's great advancements, remarkable improvements in healthcare across the globe have allowed people to live longer, healthier lives. As technology enables us to achieve more in our lifetimes, improving healthcare enables us to extend our lifetimes. An integral part of many clinical pathways in modern medicine is the use of medical image analysis. Advances in medical imaging have enabled us to develop a clearer view on the inner workings of the human body and the biological processes that govern healthy development. The ability to identify, track and predict health outcomes from medical images across all modalities has become a vital tool in any healthcare workers' tool-belt. As such, almost every one of us has experienced the benefits of improved healthcare, aided by medical image analysis, before we're even born.

Even faster than new technology emerges, new people are born, and for many (myself included) it is vital that every new life be given the best chance of survival, and not just survival, but the ability to thrive[1]. It is at the interface of technology and pre-term healthcare that I seek to establish a symbiosis, to enable the greatest care possible to be given to all. This aim is grounded in a desire for global equality and an equalness of opportunity for all, so why not start with an equal chance of a healthy birth, as that is where we all start.

1.2 Introduction to Medical Imaging Sciences and Medical Image Analysis

Ever since the first X-ray image conducted by Wilhelm Röntgen (Figure 1.2) was shown to the world, humanity has been captivated by images giving a deeper view into the human body. Throughout history, clinicians, biologists and scientists from many different disciplines have used drawings and diagrams to convey information (Figure 1.1). This has arguably had as great an impact on medical learning as writing, and inventions such as the printing press. Exquisite and detailed hand-crafted drawings and diagrams enabled the dissemination of complex information to a wider audience, and these were of great benefit to improving biological and anatomical understanding. The invention of the X-ray enabled something different. For the first time, an anatomical image was created not by hand, but by technology directly, in near real-time. This opened an avenue to a vast array of clinical applications, and jump-started the field of medical imaging towards what we know today[20].

Fast forward 100 years from the first X-Rays, and presenting that image in any clinical setting would be met with contempt, such have been the vast improvements of imaging quality since then[20]. New modalities and improved acquisition quality have enabled medical practitioners and researchers to peer ever deeper, with ever finer detail, into the inner workings of the human body, to produce the both beautiful, and life-saving imagery that we are familiar with seeing today (Figure 1.3).

[1] This is not a pro-life argument. Every parent should be able to make the most informed decision possible before deciding to bring a new life into this world.

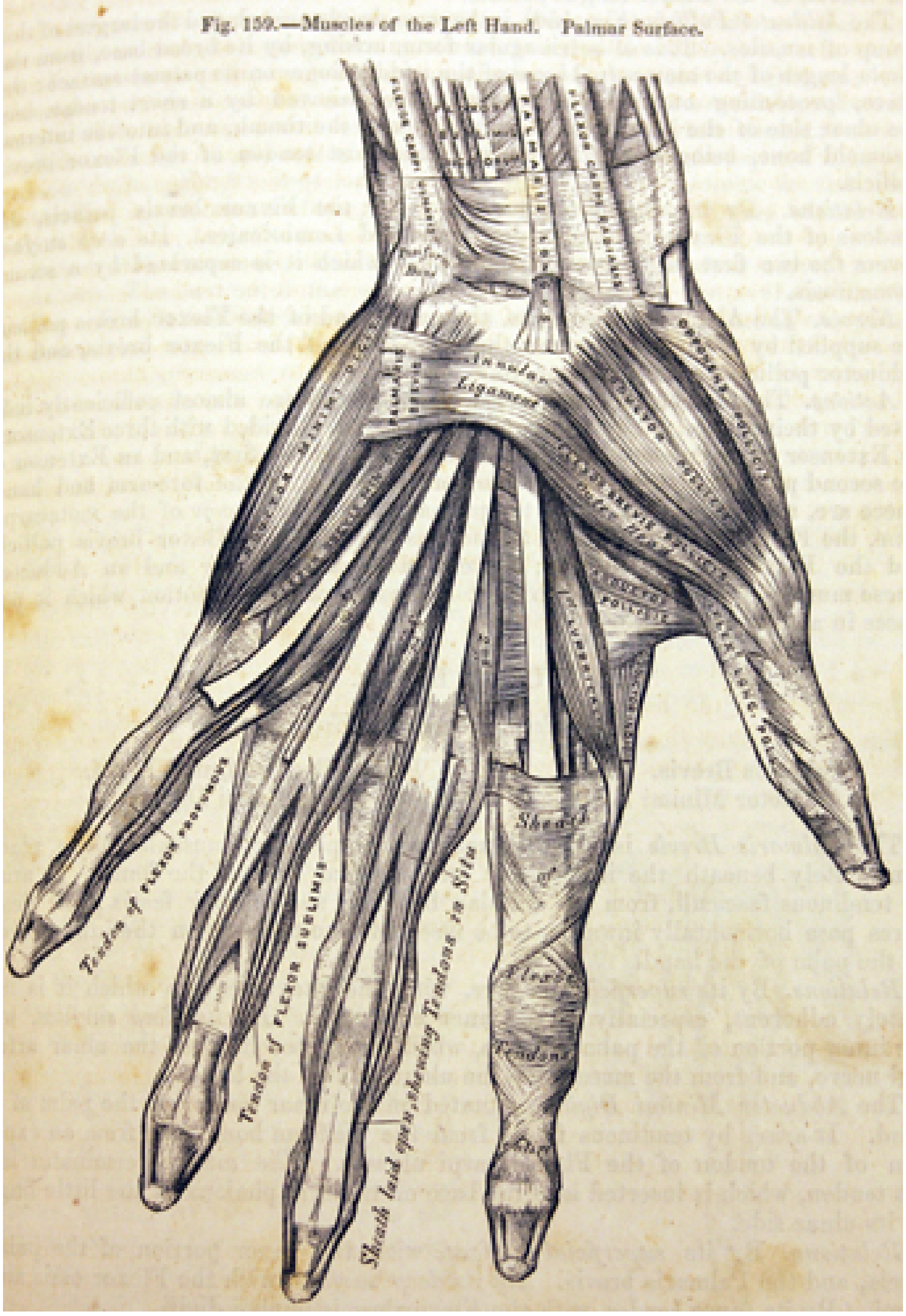


Fig. 159.—Muscles of the Left Hand. Palmar Surface.
Annular Ligament
PALMARIS LONGUS
Pisiform Bone
Tendon of FLEXOR PROFUNDUS
Tendon of FLEXOR SUBLIMIS
Sheath laid open, showing Tendons in Situ
Sheath of Flexor Tendons

Figure 1.1: Gray's Anatomy: Figure 159.- Muscles of the Left Hand, 1859[1].

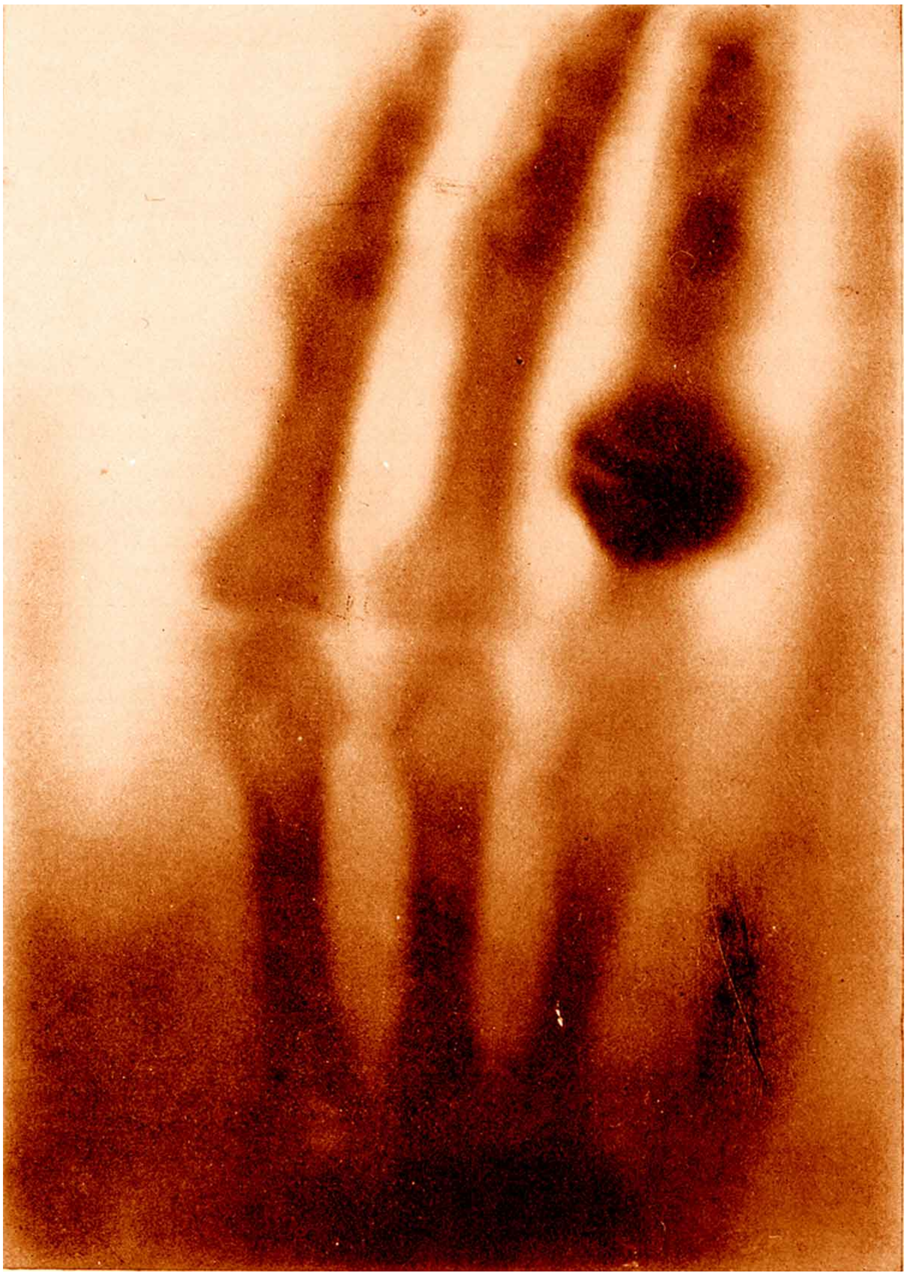


Figure 1.2: The First X-Ray: An X-Ray of Röntgen's Wife's hand, who upon seeing the image is said to have proclaimed: "I have seen my death" [2].

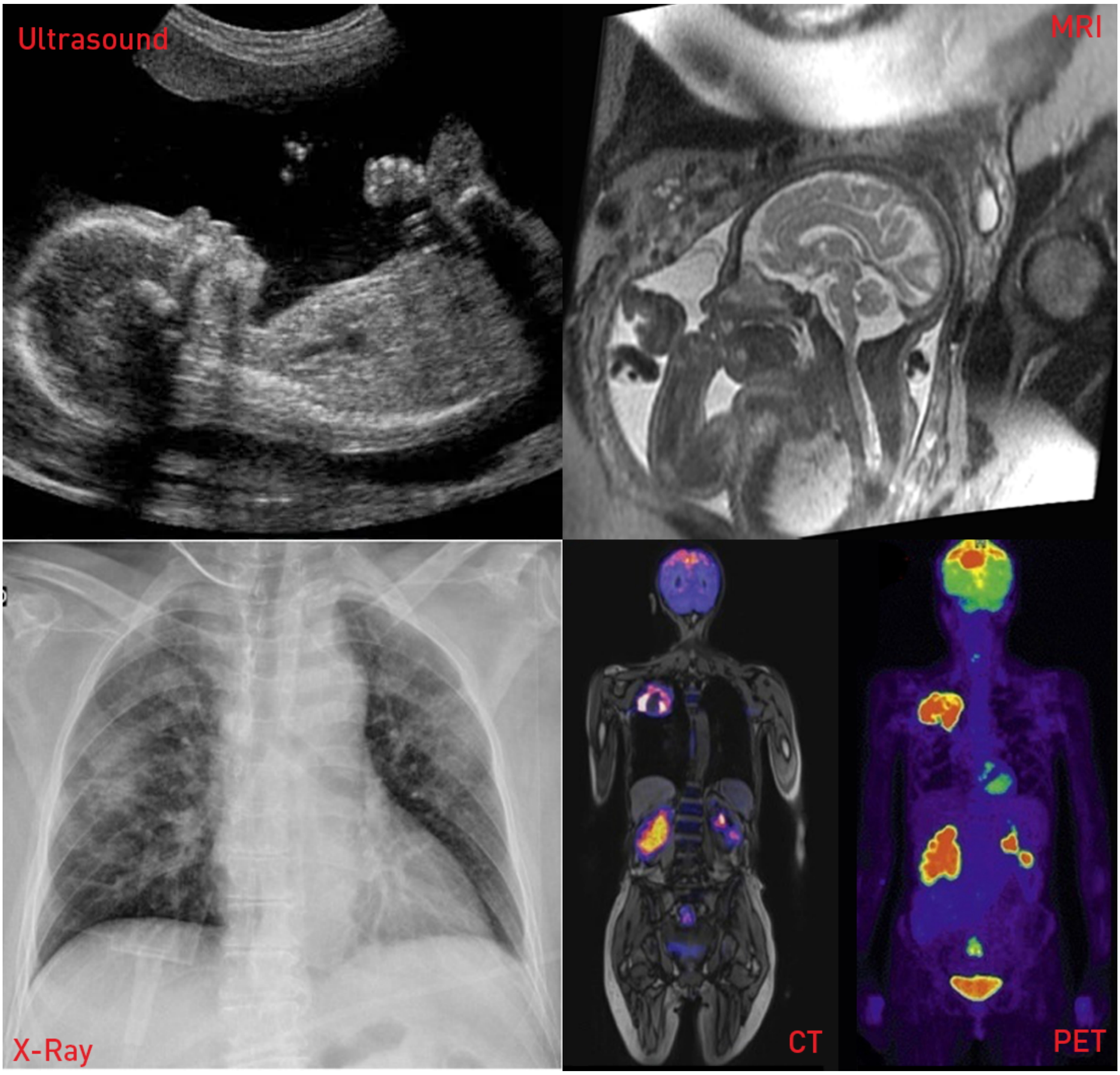


Figure 1.3: Modern Medical Imaging examples. Top Left: Fetal Ultrasound[3]; Top Right: Fetal MRI[4]; Bottom Row Left to Right: X-Ray[5], CT[6] and PET[6].

1.2.1 Computer vision and segmentation

As technology has improved, so has the speed and efficacy with which we can acquire medical images. For many modalities, the bottleneck is no longer in acquiring images, but in analysing them. As such, techniques have developed to automate steps in the analysis of medical images. One such step is segmentation[21]. A major part of many clinical work flows is the delineation of body parts, organs and bones etc within both 2D and 3D images, to enable finer analysis to be performed on those parts alone. Applications are wide ranging from localisation and measurement of tumours, to the design of 3D prosthetics[22]. In the past (and in many settings still today), clinicians and practitioners would perform this segmentation by hand, manually tracing over an image. This is no longer practical, and great efforts have been made to automate segmentation of medical images in a vast array of different scenarios and application settings. Automated segmentation methods have benefited from growing computational power and improving technologies, and classical methods such as Otsu thresholding, Graph-Cut methods and atlas label propagation have provided the groundwork from which all modern segmentation methods derive[23]. More recently, with the proliferation of Deep Learning (DL), a new family of methods has achieved state-of-the-art performance for image segmentation (in both natural images and medical images) that is paving the way for segmentation applications not thought possible just twenty years ago (Figure 1.4).

1.2.2 Deep Learning for computer vision and segmentation

DL is a branch of Machine Learning (ML) that centres around artificial neural networks, inspired by the functionality of neurons within the brain. Through exposure to large amounts of data, DL systems can develop a powerful and flexible understanding of myriad concepts through learning a nested hierarchy of abstract concepts, each learning a representation of more abstract concepts computed in terms of simpler ones[24].

DL has opened new doors for medical image analysis, and in particular segmentation, allowing segmentation tasks that would have previously taken hours for clinicians to complete to be performed in a tiny fraction of that time[25]. Supervised learning methods have enabled new methods to leverage the segmentations performed by experts to develop automated models capable of segmenting regions of interest in almost every imaging modality, and to be done so with a high level of accuracy. While many DL methods have shown great promise in research settings, clinical translation of these methods has proved harder than anticipated[26, 27]. The same performance gains seen in natural image applications have not been realised in medical image domains, for a variety of different reasons. Medical images are naturally more scarce than natural images, protected by privacy policy and not as freely available as natural images[28]. When medical images are available or can be acquired, the amount of data available is usually much less, a major stumbling block for DL methods[28]. Worse still is the availability of high quality image annotations from which supervised methods learn. Medical images are seen as complex and difficult to understand images. Often noisy and full of artefacts, it is assumed that significant expertise is required to reliably annotate most medical images, especially for image segmentation, a challenging and time-consuming task. This requirement that medical experts annotate medical image data has slowed down the creation and dissemination of high quality and high quality datasets. Medical expert time is precious enough as it is, without asking experts to spend significant effort on annotating medical image datasets.

There are many efforts being made to mitigate the negative impacts of smaller annotated dataset sizes in medical imaging[29]. These efforts are being made at every stage of DL enabled clinical workflow, from dataset curation and model building, to prediction interpretation and adjustment.

1.3 Introduction to Symbiotic and Human-in-the-loop Deep Learning

At the core of many DL approaches lies the assumption that humans will play a vital role in both the development, and operation of DL enabled systems. It is these methods that we focus on and leverage in our work, and through which we hope to realise collaborative technologies

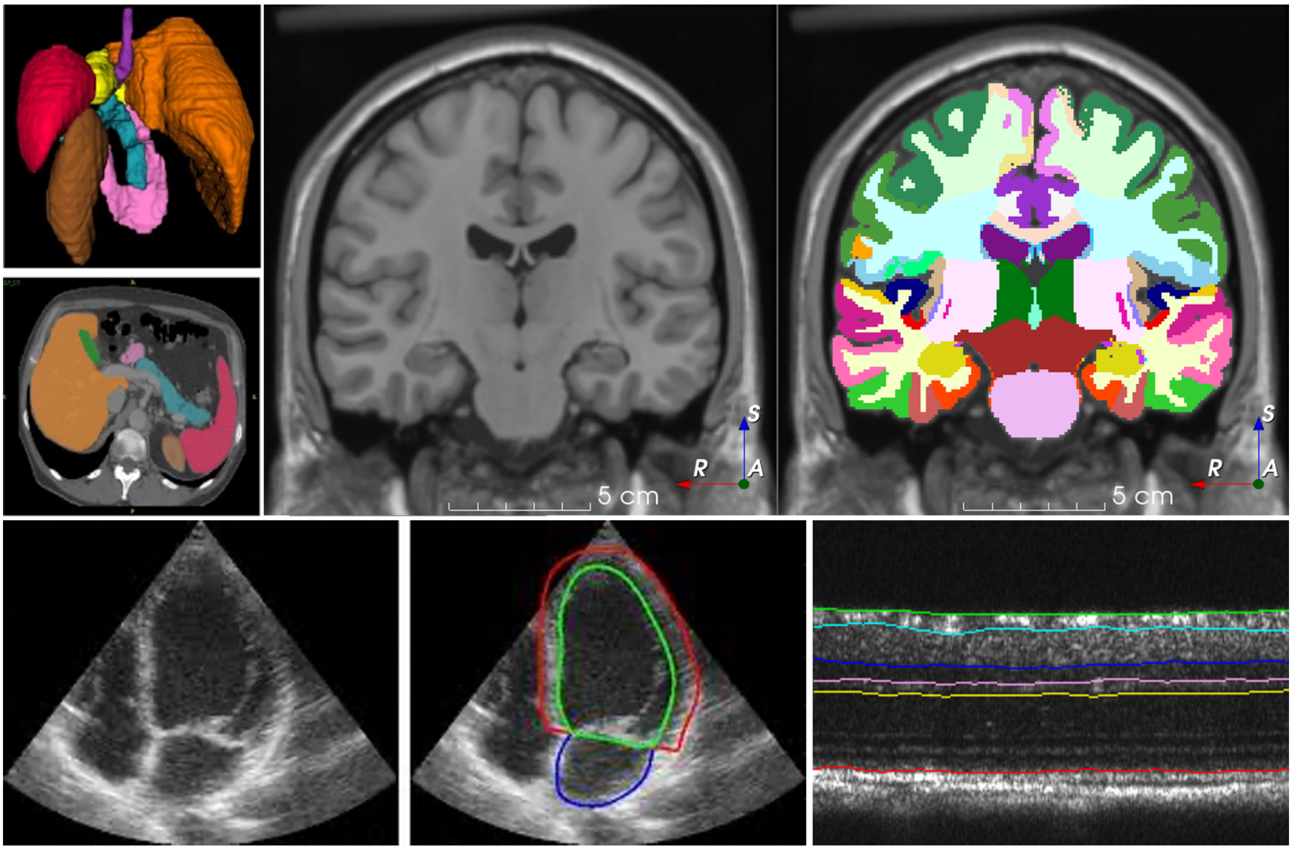


Figure 1.4: Medical Image Segmentation examples. Top Left: CT organ segmentation[7]; Top Right: MRI brain segmentation[8]; Bottom Left: Ultrasound Cardiac segmentation[9] and Bottom Right: OCT retinal layer segmentation[10]

which satisfy a dual symbiosis i.e provides a depth of understanding, and ease of use.

In biology, 'Symbiosis' is defined as the interaction between two different organisms living in close physical association, typically to the advantage of both. In the context of human and machine collaboration we define it as the interaction between a human and a machine, to achieve a task to a higher standard than either could alone. And through this lens it naturally follows that Human-in-the-loop computing systems are integral to achieving symbiosis. Human-in-the-Loop computing can be defined as a technology that can improve a performance objective by engaging its human users, through iterative feedback from the underlying technology to the user, and from the user back to the technology.

Here it is important to consider the life-cycle of a DL enabled system, through development, deployment and operation. We seek a dual symbiosis that enables the successful growth of a model from infancy (development) to adulthood (deployment) to retirement (end of operation). This requires a different focus at every stage, and a different type of interaction between humans and machine.

In model development, we require a depth of understanding from developers to build models that perform well enough for their job. As we have discussed, a large component of achieving this when building supervised DL models is acquiring the necessary annotated data to train such models. This data collection and annotation effort presents many of its own challenges, and many proposed methods seek to reduce the time and costs associated with acquiring enough data to develop an acceptable model and improving the ease with which this can be done (for both developers and those annotating the data)[29].

In model deployment and operation, we require a depth of understanding of model outputs. While state-of-the-art models can achieve astounding results in research settings, this is not guaranteed once deployed in the real-world[26]. For automated models to be used ethically and responsibly in clinical practice, care must be taken to ensure that those operating the systems are made aware of potential short-comings. It is important we develop a trust between human and machine, but this trust must not be blind. Both human and machine have a responsibility to communicate with each other to ensure that outputs of models are acted on with confidence and that the model is performing as expected[30]. This expectation creates a feedback loop between human and machine. Machines must provide not only model predictions, but explanations, and measures of confidence in their predictions, from which humans can accept predictions, or reject predictions and guide the machine towards a more accurate prediction, ensuring the same mistakes are not made in the future. This idea of continual learning through collaboration between human and machine has been shown time and time again with both humans teaching machines, and machines teaching humans, thus it is likely to continue. As our machines grow and learn care must be taken that our machines do not drift from their initial requirements. Extensive monitoring of machine performance is required to maintain trust between human and machine throughout operation[30].

1.3.1 Active Learning

Active Learning (AL) is a branch of ML that seeks to reduce the amount of data required to train a ML model[31]. It is argued that learning from similar data creates redundancy, and that equivalent model performance can be achieved with an optimal subset of available data. Through this assumption, AL methods aim to iteratively select the most informative unannotated data for which to acquire annotations. This reduces the time and costs associated with model development, without impacting model performance[31].

Images are chosen based on their 'Informativeness' to the current model, which in general can be split into two main components: 'Representativeness' and 'Uncertainty'. Here, representativeness measures how similar an image is to other images in our dataset, and images are not annotated if very similar images have already been represented in the training set. Uncertainty measures how confident or uncertain the model was when making a prediction on that image. We want our models to be confident of the predictions they are making (if they are accurate) and thus training on more images for which the model is uncertain should improve the performance of models more than training on images for which the model is already accurate, and certain about it[31].

1.3.2 Interactive Learning

In Interactive Learning (IL) methods, users make iterative corrections to model predictions until the model prediction meets the users required quality. These methods acknowledge that not all model predictions are perfect, and many will require manual intervention to improve their quality[32]. IL methods aim to make this part of the model prediction process, to both enable rapid feedback to be given to the model, and for the model to rapidly update predictions based on that feedback. Many IL methods aim to capture and learn from every user interaction throughout operation to ensure better predictions in the future.

1.3.3 Practical Learning

Practical Learning methods acknowledge that neither data, nor those annotating the data are perfect. These methods seek to mitigate the impact of noisy and imperfect input and annotation data by modelling the quality of the data, through which learning can be guided more heavily by high quality data and annotations[33]. These methods enable a reduction in cost by not expecting only experts to annotate data, and also make considerations for the practical interfaces being used to generate annotations.

1.3.4 Trustable Learning

Trustable Learning methods place their focus on interpretability. These methods acknowledge that almost all DL methods are considered 'black-box' methods for which we have no real knowledge of the decision making processes that have led to a given prediction. Trustable Learning methods seek to inspire the trust of users in their predictions by opening this black box and providing feedback to the users about their internal workings e.g uncertainty or confidence of model predictions[34]. Without such trust being built and maintained we cannot hope to develop models that truly offer the dual symbiosis of depth of understanding and ease of use throughout their life-cycles.

1.3.5 Closing the Loop

In the new data-driven world, establishing a symbiosis between DL systems and humans will require mastery of all the methods and techniques mentioned above. This is a lifetime of work, too vast for any individual, however improvements made in any of these areas stand to make a valuable contribution towards this endeavour. In this thesis we set out to establish the current state-of-the-art in each of these areas, and go on to make contributions towards solving several of the core challenges associated with closing the symbiotic loop between humans and data-driven technology. These challenges are as follows:

- Categorisation Challenge: In order to fully understand where the the state-of-the-art lies in closing the symbiotic loop, we categorise the existing literature and research efforts to better understand the core components needed for a system to achieve symbiosis and be defined as a true human-in-the-loop system. We identify the gaps in knowledge that still need to be filled and highlight areas in which we make a contribution.
- Confidence Challenge: Outputs of DL systems must provide measures of confidence from which their predictions can be valued appropriately. These confidence measures must be communicated with clarity and provide users with a reasoned understanding of the models' behaviour.
- Complexity Challenge: As DL has been shown to perform tasks of growing complexity, so grows the complexity of how we represent our knowledge and data to be learnt from. As the complexity of tasks performed by DL models increases, the efforts to present this information and deciding on how best to use it too becomes more complex.
- Classification Challenge: The core tasks of disease identification and classification remain key goals of many DL models. The mechanisms by which disease can be identified are becoming more complex, but the need to breakdown these predictions into understandable and interpretable components remains vital in ensuring appropriate use of DL predictions, retaining the ability to interrogate complex systems to understand how a prediction has been made.

- Curation Challenge: As supervised learning methods proliferate, the need to large amounts of well annotated data too increases. Acquiring this data in many application settings presents a major challenge due to the limited resources available to us. As such we examine whether we can challenge many assumptions made about curating expert annotated data and whether the costs associated with acquiring this data from experts outweigh the gains in performance from acquiring annotations from non-experts.

To achieve symbiosis in data-driven Human-in-the-Loop systems, all of the above challenges need to be addressed. This motivates the core aims of this thesis. At present, existing research has been unable to consider these five challenges simultaneously, and in this thesis we aim to address these five challenges to understand the interplay of each in the context of real-time medical image analysis and from a more general technical point of view. Through this work we make many contributions to addressing each challenge and ask new questions which need to be answered to achieve symbiosis in data-driven Human-in-the-Loop systems.

1.4 Motivations and applications

The wide array of advancements made in DL for medical image analysis have enabled an ever greater variety of applications to be tackled. A core motivation of DL and generalisable data-driven models is the democratisation of knowledge. We live in a world where connectivity is at the heart of everything, and this notion is becoming more pervasive in healthcare. No longer does every aspect of clinical workflow have to be performed in the same physical location, with more automated analysis becoming more available, clinical decision making is becoming more remote. This democratisation of knowledge in the form of high performing models has the potential to improve healthcare standards across the globe, reducing the level or training and resources required to administer the same level of care[35]. As compute resources proliferate and imaging devices get cheaper, the ability of models to provide a high level of care across the globe too will proliferate, improving equality of care for all.

1.4.1 Fetal and Neonatal Development

Fetal development is fundamentally challenging to monitor. Detection of pregnancy is possible from a very early stage, but tracking development in the womb is a multi-dimensional challenge that is of significant consequence. It is everyone's hope that pregnancy can proceed without complication, but for the unfortunate few for whom this is not the case, the detection of fetal development abnormalities at the earliest stage possible can make a significant impact on the health outcomes of both mother and child[36]. The ongoing tracking of development throughout pregnancy and the early detection of potential adverse health outcomes provide vital information for both parents and doctors that inform the decisions made and care given. The work done to provide this care is paramount to give every child the best possible start in life.

1.4.2 Ultrasound Screening

A major component of fetal screening procedures is regular Ultrasound (US) screening. Throughout pregnancy, at several key growth stages, ultrasound screening is performed to assess fetal development. A series of 'standard views' are acquired, covering key fetal anatomical areas, from which indicative measurements can be taken to track longitudinal fetal growth trajectory, and diagnosis of various diseases and health issues including Congenital Heart Disease (CHD). This screening has proven to be hugely beneficial for the early identification of a wide variety of fetal growth abnormalities, and improving the healthcare of both mother and child in the short and long term[36, 37].

1.4.3 Magnetic Resonance Imaging Screening

Magnetic Resonance Imaging (MRI) is another non-invasive modality being applied to fetal screening scenarios. MRI can provide a much higher resolution than US, and as such is being used to image the fetal brain. MRI of the fetal brain is being used in research to study the development of brain structures in-utero, but also has the potential to provide important longitudinal health predictions and non-invasive diagnosis for several adverse health outcomes at an early stage[38].

1.5 Contributions

Chapter 2: In Chapter 2 we address the categorisation challenge with a thorough literature review of human-in-the-loop DL methods for medical image analysis. We hypothesise that there is significant overlap between related areas of research that can be considered to solve the same problems under the umbrella of human-in-the-loop DL. We introduce the key areas of research concerned with closing the symbiotic loop in data-driven technology and human-in-the-loop DL. We establish a categorisation of methods that will help future researchers to understand the field in a wider context and to understand the gaps of knowledge requiring further research.

Chapter 3: In Chapter 3 we address the confidence challenge. We hypothesise that real-time feedback of DL model predictions can be achieved effectively using probabilistic segmentation methods. We develop a general purpose method for understanding the confidence of DL predictions. We present a method for automated probabilistic image segmentation with real-time feedback on measurement robustness. We evaluate multiple probabilistic DL methods from which an ensemble of segmentations can be produced for each. From each ensemble of predictions, upper and lower bounds of segmentation based measurements can be generated. In addition we derive 'variance scores' through which we can reject measurements and re-acquire images to produce optimal measurements, guiding operators towards optimal image acquisition for more consistent and accurate segmentation based measurements.

Chapter 4: In Chapter 4 we address the complexity challenge. We hypothesise that probablistic segmentation methods can be extended to more complex domains and that they can outperform existing methods with a lower computational cost. We extend probabilistic segmentation methods from 2D to 3D, and introduce multi-task learning to simultaneously learn 3D image segmentation and shape metric learning for two common anatomical shape metrics. We extend our previous efforts to provide confidence scores on more complex DL outputs, and show the robustness of our method for use on highly abnormal structures for which other methods fail, enabling population wide comparisons of outputs to be used in downstream analysis tasks reliably.

Chapter 5: In Chapter 5 we address the classification challenge. We hypothesise that classification methods can achieve improved interpretability by decomposing them into modular components through for which each presents improved interpretability, without the loss of performance over end-to-end DL methods. We develop a classification pipeline that is robust and interpretable for disease classification from single images. We extend an existing method to a multi-task setting in which we are able to jointly segment, register and build a labelled atlas that focuses on relevant features to robustly classify images. We provide a new method for image feature-based classification and show this to be interpretable for downstream analysis. We demonstrate that disease conditioned segmentations show greater correlation to true disease status than naive segmentation methods.

Chapter 6: In Chapter 6 we address the curation challenge. We hypothesise that novice annotators can perform complex medical annotation tasks with minimal training to a high standard. We challenge the assumptions that medical image segmentation data can only be acquired by medical experts and employ a novice workforce to perform this segmentation from a short series of instructions and examples. We assess the upstream impacts of training medical image multi-class segmentation models, and evaluate downstream classification models on these noisy labels from novice annotators compared against gold-standard labels from expert annotators. We demonstrate that in resource constrained settings, greater DL performance may be achieved through the use of novice annotators than with expert annotators alone.

1.6 Statement of Originality

I declare this work original and of original thought where not stated otherwise.

1.7 Publications

Parts of the following chapters have been published as peer reviewed publications. First author publications:

- **Samuel Budd**, Matthew Sinclair, Thomas Day, Athanasios Vlontzos, Jeremy Tan, Tianrui Liu, Jacqueline Matthew, Emily Skelton, John Simpson, Reza Razavi, Ben Glocker, Daniel Rueckert, Emma C. Robinson and Bernhard Kainz, *"Detecting Hypo-plastic Left Heart Syndrome in Fetal Ultrasound via Disease-specific Atlas Maps"*, International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI'21 [39].

- **Samuel Budd**, Thomas Day, John Simpson, Karen Lloyd, Jacqueline Matthew, Emily Skelton, Reza Razavi and Bernhard Kainz, *"Can non-specialists provide high quality gold standard labels in challenging modalities?"*, Affordable Healthcare and AI for Resource Diverse Global Health (FAIR) in Conjunction with MICCAI'21 [40].

- **Samuel Budd**, Emma C. Robinson and Bernhard Kainz, *"A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis"*, Medical Image Analysis, Volume 71, July 2021, 102062 [41].

- **Samuel Budd**, Arno Blaas, Adrienne Hoarfrost, Kia Khezeli, Krittika D'Silva, Frank Soboczenski, Graham Mackintosh, Nicolas Chia and John Kalantari, *"Prototyping CRISP: A causal relation and inference search platform applied to colorectal cancer data"*, Outstanding Paper Award for Oral Presentation, IEEE Global Conference on Life Sciences and Technologies, LifeTech'21 [42].

- **Samuel Budd**, Prachi Patkee, Ana Baburamani, Mary Rutherford, Emma C. Robinson and Bernhard Kainz, *"Surface Agnostic Metrics for Cortical Volume Segmentation and Regression*, Best Paper Honourable Mention, Machine Learning in Clinical Neuroimaging in Conjunction with MICCAI'20 [43].

- **Samuel Budd**, Matthew Sinclair, Bishesh Khanal, Jacqueline Matthew, David Lloyd, Alberto Gomez, Nicolas Toussaint, Emma C. Robinson and Bernhard Kainz, *"Confident Head Circumference Measurement from Ultrasound with Real-Time Feedback for Sonographers"*, International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI'19 [44].

- **Samuel Budd**, Emma C. Robinson and Bernhard Kainz, *"The Cortical Explorer: A Web-based User-interface for the Exploration of the Human Cerebral Cortex"*, Eurographics Workshop on Visual Computing for Biology and Medicine, EG VCBM'17 [45].

Co-authored publications:

- Athanasios Vlontzos, **Samuel Budd**, Benjamin Hou, Daniel Rueckert and Bernhard Kainz, *"3D Probabilistic Segmentation and Volumetry from 2D Projection Images"*, International Workshop on Thoracic Image Analysis in Conjunction with MICCAI'20 [46].

- Jacqueline Matthew, Emily Skelton, Thomas George Day, Veronika A. Zimmer, A Gomez, Gavin Wheeler, Nicolas Toussaint, Tianrui Lio, **Samuel Budd**, K Lloyd, R Wright, Shujie Deng, Nooshin Ghavami, M Sinclair, Qingjie Meng, B Kainz, Julia A. Schnabel, Daniel Rueckert, Reza Razavi, John Simpson and Jo Hajnal, *"Exploring a New Paradigm for the Fetal Anomaly Ultrasound Scan: Artificial Intelligence in Real Time"*, World Congress on Ultrasound in Obstetrics and Gynecology, ISUOG'21 [47].

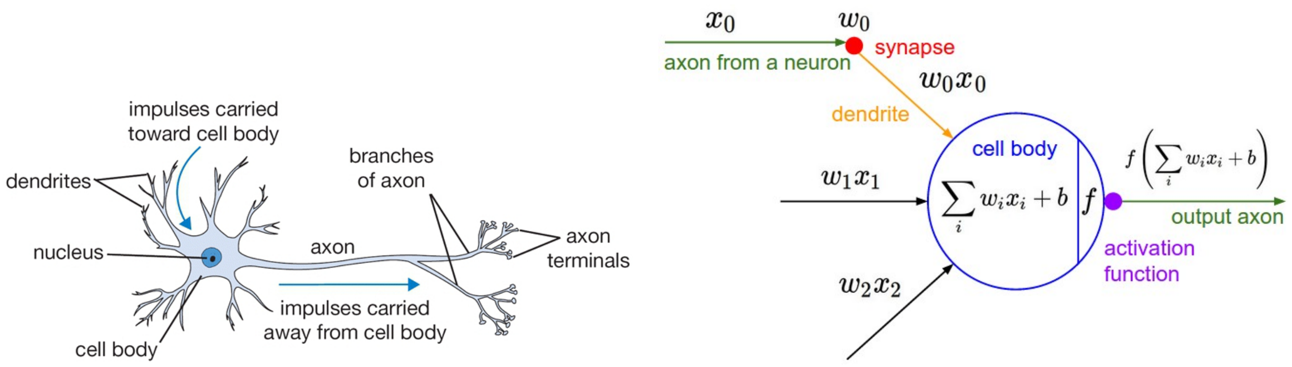
Chapter 2

Background and Related work

In this chapter we summarise the key concepts and previous research from which our work builds. Section 2.1 introduces several ML techniques and DL methods that will be used in Chapters 3, 4, 5 and 6. It gives an overview of DL and how DL is applied to key areas of medical image analysis that are leveraged throughout our work. Section 2.2 provides an in depth literature review of Human-in-the-loop DL, showcasing state-of-the-art uses in medical image analysis and highlighting key areas where additional research is needed, motivating the works completed in Chapters 3, 4, 5 and 6. We focus on the key areas of 'Active Learning', 'Interactive Learning', 'Practical Considerations' and 'Future Prospectives' as these sections together cover a wide range of different areas including but not limited to uncertainty quantification, multi-task learning and explainability implicitly and as such give insights into these areas through the lens of Symbiotic Human-in-the-Loop computing.

2.1 Deep Learning with Neural Networks

DL methods are inspired by the biological inner workings of the human brain, specifically by the networks of highly connected neurons through which we have developed a complex understanding of the world.



(a) (b)

Figure 2.1: Schematic diagram of (a) a biological neuron and (b) the perceptron [11]

2.1.1 Artificial Neural Networks

An Artificial neural network (ANN) is a biologically inspired computational network, composed of collections of artificial neurons which transform a given input to a chosen output. In the fashion in which biological neural networks operate, ANNs consist of simple neurons (or computational units), which are highly interconnected, and the connections betweens neurons that determine the function of the network[48]. Initially proposed by McCulloch and Pitts (1943)[49], in which an early computational model of the neuron was developed (Figure 2.1a), ANNs have been built upon ever since. The concept of a *perceptron* was introduced by Rosenblatt (1958)[50], in which a more developed version of a neuronal computational unit was proposed (Figure 2.1b). In a perceptron, every input x_i is assigned a weight w_i, through which a summation is calculated. After summation, an activation function is applied to determine what amount of information from the summation is allowed to pass through that computational unit to the rest of the network. The output of a perceptron can be written as:

$$y = \varphi(\sum_{i=1}^{m}(x_i * w_i + b))$$

where m is the number of inputs to the neuron, b is the bias term used to alter the decision boundary from the origin and $\varphi(\cdot)$ is an activation function. The activation function plays an important role in ANNs by introducing non-linearity to the output of perceptrons, enabling ANNs to model more complex relationships between input and output. Common activation functions are the sigmoid activation functions, hyperbolic tangent functions (tanh) and rectified

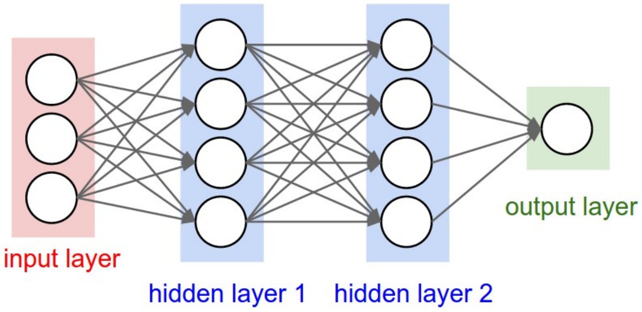


Figure 2.2: Example MLP with two hidden layers, input layer and output layer[11].

linear units (ReLu) among others[51].

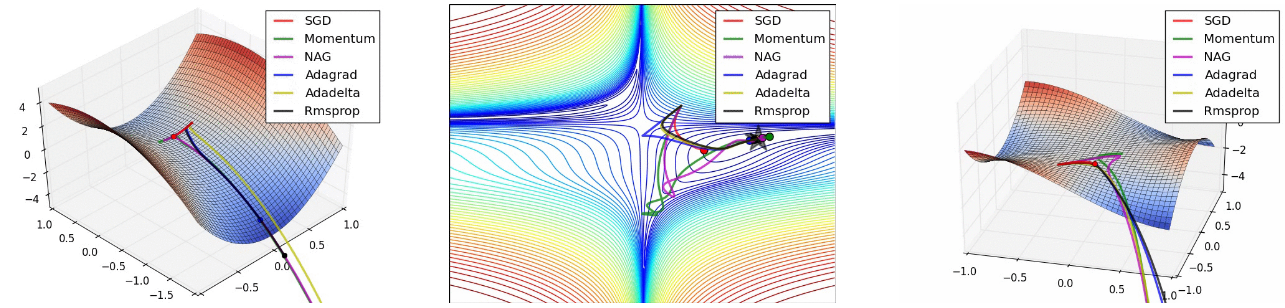
When we combine multiple perceptrons together, we form multi layer perceptrons (MLPs), consisting of at least three layers of nodes: the input layer, a hidden layers and the output layer[48]. Every node in a layer connects to every node in the next layer with an assigned weight. An MLP containing two hidden layers, and input and output layers is shown in Figure 2.2. Neural networks have been shown to be universal general approximators i.e these is guaranteed a neural network which can compute or approximate any function of any combination of inputs to outputs, as demonstrated by the modelling of a sigmoidal function in [52].

Recent computational advances and the use of Graphical Processing units (GPUs) have enabled efficient computation and training of much larger neural networks. As we have introduced more hidden layers to ANNs and these networks have become especially deep, the terms Deep neural networks (DNNs) and DL have emerged as the defacto way to describe these methods in recent literature.

2.1.2 Neural Network Optimisation

Neural Networks are trained by repeatedly updating the weights of the connections between nodes in the network, to minimise a loss function between the current predicted output of the network and the desired output of the network.

The backpropagation (BP) algorithm was proposed by Rumelhart et al. (1986)[53] to compute the gradients of a loss function for training networks with neuronal units. The BP algorithm



(a) Long Valley

(b) Beale's function

(c) Saddle Point

Figure 2.3: Example optimisation trajectories on a variety of loss function surfaces[12]

computes these gradients for use by a gradient descent optimisation algorithm to update the weights of a network to minimise the loss function. The optimal solution of a network is reached the the value of the loss function reaches a global minimum, however, depending on the optimisation algorithm, solutions may fall into local minima, where the gradient of the loss function is zero, causing the network weights to stop updating.

Gradient Descent

The update rule for Gradient descent can be written as:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla J(\theta_t)$$

where θ is the network parameters, η is the learning rate and $\nabla J(\theta_t)$ is the gradient of the loss function $J(\theta_t)$. Each weight is updated through the BP algorithm calculating how much of the loss every node in the network is responsible for, and updating accordingly. The update rule is designed to make each optimisation step more robust e.g faster convergence, avoiding local minima or preventing vanishing or exploding gradients[54]. As such a variety of works provide alternative update rules to gradient descent to address these problems, such as the Adam (adaptive momentum estimation) update rule:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t$$

where $\hat{v}_t$ is the exponentially decaying average of squared gradients and $\hat{m}_t$ is the momentum of descent at update step t[55]. The optimisation trajectory taken by different optimisation algorithms, over various loss surfaces is shown in Figure 2.3.

Loss Functions

In this thesis we focus on supervised ML methods which train models to directly map inputs to outputs through learning from examples of input-output pairs. As such, we build datasets of paired data where for every input we have a corresponding output label. Supervised learning methods can be further broken into two categories: classification and regression, where both map inputs to outputs however in regression, the output is a numerical (real) value, and for classification the output is categorical (discrete). This statement is not true for all tasks, but in the context of this thesis in which we consider classification, regression and segmentation tasks this is a useful categorisation. Tasks such as object detection and image reconstruction that are beyond the scope of this thesis present additional categorisations that we do not consider here.

In regression models the most common types of loss function are the Mean Absolute Error (MAE), the Mean Squared Error (MSE) and the Huber Loss. The MAE and MSE can be written as:

$$\text{MAE} = |y - f(x)|$$

$$\text{MSE} = (y - f(x))^2$$

where y is our output label and $f(x)$ is the output of our neural network. The MAE loss, also known as the L1 loss, calculates the error as the distance between the predicted and desired output values. The MSE loss, also known as the L2 loss, calculates the error as the squared distance between predicted and desired output values, thus penalising large errors more. As a result, the MSE loss function is less robust to outliers than MAE. The gradient of the MAE loss is the same throughout, which allows small errors to carry the same weight as large errors which is also undesirable. In response to this, the Huber loss introduces a tuning parameter δ, in order to switch between MAE and MSE losses depending on the current loss value[56]. Through this we gain the benefits of both MAE and MSE losses at the expense of an additional tunable hyper-parameter. We can write the Huber loss as:

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \leq \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$

In classification models the most common type of loss function is Cross-Entropy. This is typically preceded by an activation function, the choice of which depends on the task the network is being used for e.g Multi-Class or Multi-Label. In Multi-Class problems each input can only belong to one class in a set of classes C, however for Multi-Label problems each input can belong to many classes. We focus only on Multi-class problems in this thesis. A Multi-Class model will output a vector of scores s which represent the probability of the input belonging that class, and out desired output label will be a one hot vector in which only one class is positive and all remaining classes are negative. For Multi-class problems we use the softmax activation function, which squashes our output vector s to lie in the range $[0, 1]$ with the constraint that the sum of all class probabilities sum to 1, ensuring the laws of total probability are satisfied. The softmax function can be written as:

$$f(s_i) = \frac{\exp(s_i)}{\sum_j^C \exp(s_j)}$$

Cross-Entropy is a measure of the difference between two probability distributions for a given random variable:

$$\text{Cross Entropy} = -\sum_i^C t_i \text{log}(s_i)$$

where t_i represents our target desired output. In many medical image segmentation applications, the softmax cross-entropy loss is used[18], as every individual pixel in an image will only belong to one particular class and as such is a suitable loss for function for those applications.

Additionally there are many more loss functions available for both regression and classification, as well as many custom designed loss functions for specific tasks.

2.1.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are among the most popular DL model architectures to be actively researched, they have shown state-of-the-art performance in many fields such as computer vision and medical image analysis.

A typical CNN architecture consists of four main types of layers: convolutional layers, pooling layers, fully-connected layers and normalisation layers. These are usually formulated as a series of 2D/3D convolutional, pooling and normalisation layers, and for classification models, ended with one or more fully connected layers.

Convolutional Layers

Convolutional layers are the key element of CNNs. The main benefit of a CNN comes from being able to learn the spatial and/or temporal dependencies in an image, via the application of convolutional filters. In a convolution layer, all the nodes share the same filters (or convolutional kernel) and biases. These filters are learned during the training process by updating convolution kernel weights and biases, and these act the the feature extractors, which were previously hand crafted such as SIFT[57] or SURF[58]. Each filter has a size which defines its receptive field with respect to the previous layer, this is defined by its height, width, and depth. The number of filters in a convolutional layer describes the number of channels in the output. The number of parameters of a convolutional layer is:

$$N = (h * w * d * n_b) * c$$

where h, w, d are the height, width and depth of a filter, and c is the number of filters. n_b is the number of bias terms for each filter. In the forward pass of the network, each filter of a

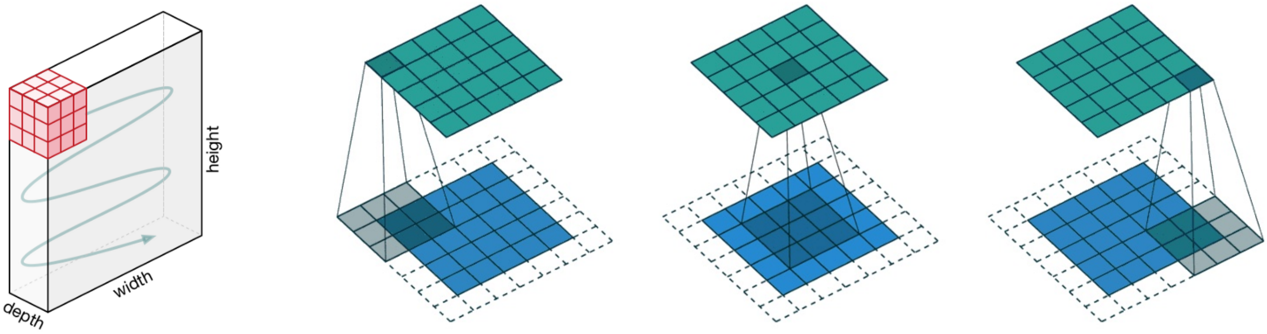


Figure 2.4: Convolutional layers in action[13]

convolutional layer slides across the whole input along the width and height as shown in Figure 2.4.

Pooling Layers

We reduce the computational load required for processing our data via the use of pooling layers, which perform dimensionality reduction while retaining the most relevant and important features, and enabling robustness to shifts in location of important features across an image or feature space. these can be divided into local pooling layers and global pooling layers. Local pooling layers are often used between convolutional layers to reduce the spatial dimension of the input image. Common local pooling layers include max-pooling and average-pooling. Global pooling layers are commonly behind the last convolutional layer to reduce the depth of the inputs, by combining all nodes along the depth of an input feature into a single node. Common global pooling layers include average-pooling.

Fully-connected Layers

Every node of a fully-connected layer connects to every node of the previous layer in the same way as for MLPs. This can sometimes place limitations on the input data size, as such fully-connected layers are often replaced by convolutional layers to allow the whole network to be fully convolutional.

Normalisation Layers

There are four main types of normalisation layers commonly used in CNNs. These are batch normalisation[59], layer normalisation[60], instance normalisation[61] and group normalisation[62]. Their use is actively researched and benefits of each depends largely on the application.

2.1.4 Network Architectures

The described components of general artificial neural networks and convolutional neural networks have enabled many to build powerful architecture designs from which others have drawn inspiration. Here we briefly describe some of the most influential architecture designs to date.

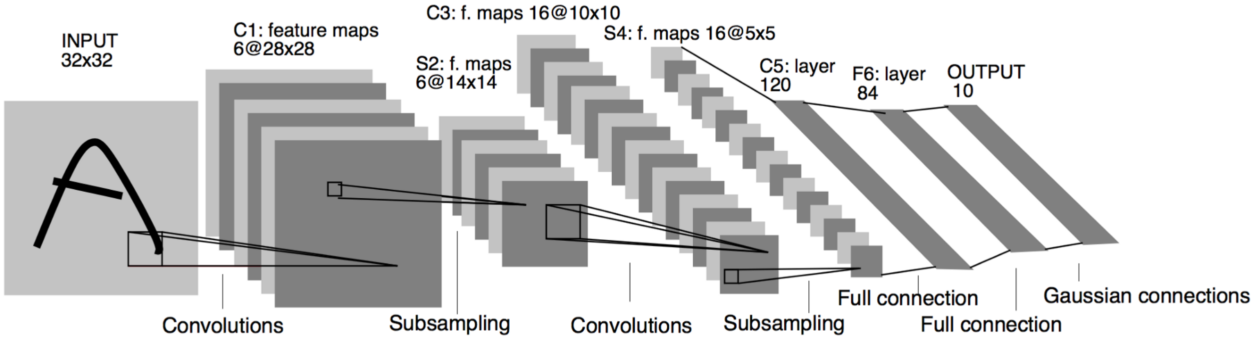


Figure 2.5: The LeNet architecture[14].

Many variants of CNN architectures have been proposed for image classification. Among the first CNN is the **LeNet** model, as proposed by LeCun et al. (1998)[14]. The LeNet architecture is composed of early convolutional and max-pooling layers, followed by fully connected layers, as shown in Figure 2.5. This network was originally designed for handwritten digit classification (MNIST). This is inspired a deeper architecture from Krizhevsky et al. (2012)[63], called **AlexNet**, winner of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), significantly outperforming the runner up.

Following this, **GoogLeNet** (Inception v1) was proposed by Szegedy et al. (2014)[15]. This was a 22-layer network that won the ILSVRC, where the introduction of the inception module, which performs dimensionality reduction and feature extraction with different sized kernels simultaneously, allowed for the detection of features with different sized receptive fields at each depth, greatly improving performance (Figure 2.6). This architecture has since been further improved to Inception v4 (2016).

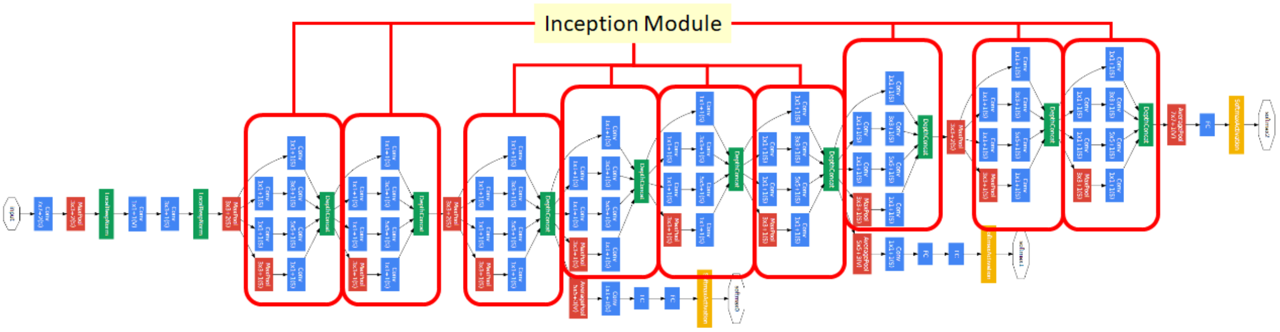


Figure 2.6: The GoogLeNet (Inception v1) architecture[15].

The runner up of ILSVRC 2014 was **VggNet**, proposed by Simonyan et al. (2014)[16], showed that it was possible to train even deeper networks by leveraging pre-training and having shallower networks at the start, and to progressively add more layers during training until it reached the 19-layers as shown in Figure 2.7.

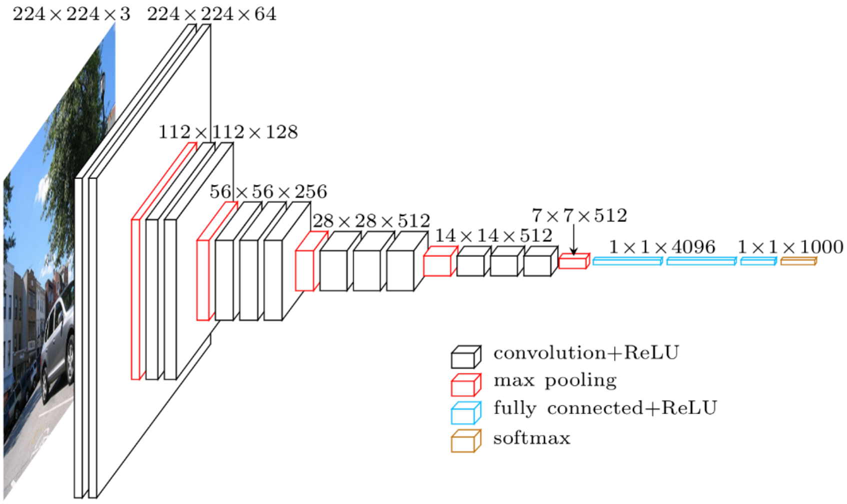


Figure 2.7: The VggNet architecture[16].

The **ResNet** architecture, proposed by He at al. (2016)[17], aimed to combat issues of diminishing gradients by introducing skip connections to the network as residual blocks as shown in Figure 2.8. It was hypothesised that training a network to fit a residual mapping, instead of the desired underlying mapping, would be easier for a network to learn. In a residual block, the input to a convolution, pooling and normalisation layer is combined with the output of those layers to allow gradients to propagate backwards via an identity mapping, improving network training.

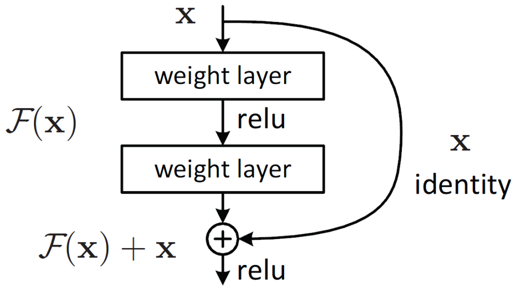


Figure 2.8: Residual block architecture[17].

Finally, **U-Net** as proposed by Ronneberger et al. (2015)[18] is as network specifically designed for medical image segmentation tasks. This network has become among the most popular architecture designs for computer vision. The U-Net belongs to the auto-encoder family of architecture structures, with skip connections introduced at each resolution, as shown in Figure 2.9. The skip-connections enabled features at different resolutions to propagate forward towards the output. The U-Net is trained with classification loss functions for every pixel, resulting in the input and output images to share the same dimensions.

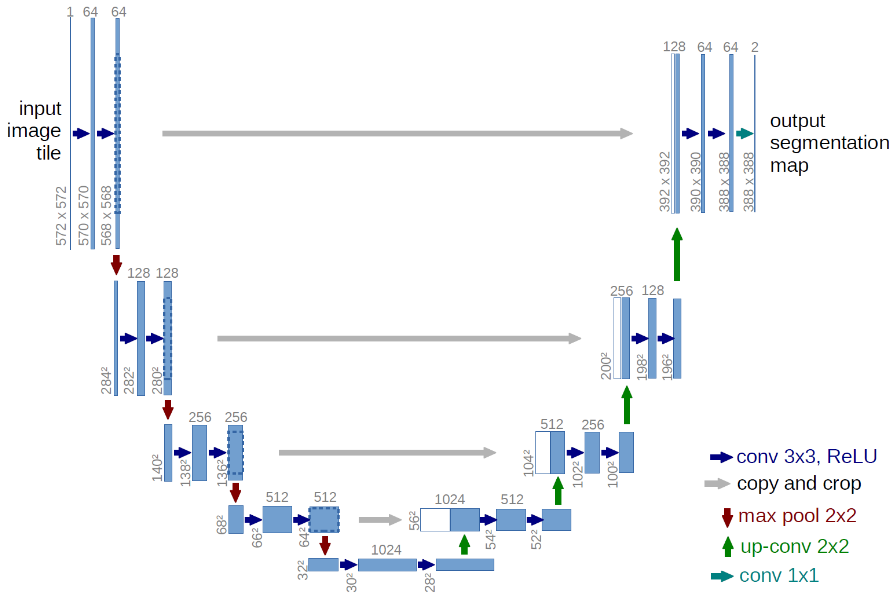


Figure 2.9: The U-Net architecture[18].

2.2 Active Learning and Human-in-the-loop Symbiotic AI

Fully automatic DL has become the state-of-the-art technique for many tasks including image acquisition, analysis and interpretation, and for the extraction of clinically useful information for computer-aided detection, diagnosis, treatment planning, intervention and therapy. However, the unique challenges posed by medical image analysis suggest that retaining a human end-user in any DL enabled system will be beneficial. In this section we investigate the role that humans might play in the development and deployment of DL enabled diagnostic applications and focus on techniques that will retain a significant input from a human end user. Human-in-the-Loop computing is an area that we see as increasingly important in future research due to the safety-critical nature of working in the medical domain. We evaluate four key areas that we consider vital for DL in the clinical practice:

1. *Active Learning* - choosing the best data to annotate for optimal model performance.
2. *Interaction with model outputs* - using iterative feedback to steer models to optima for a given prediction and offering meaningful ways to interpret and respond to predictions.
3. *Practical considerations* - developing full scale applications and the key considerations that need to be made before deployment.
4. *Future Prospective and Unanswered Questions* - knowledge gaps and related research fields that will benefit human-in-the-loop computing as they evolve.

We offer our opinions on the most promising directions of research and how various aspects of each area might be unified towards common goals.

Medical imaging is a major pillar of clinical decision making and is an integral part of many patient journeys. Information extracted from medical images is clinically useful in many areas such as computer-aided detection, diagnosis, treatment planning, intervention and therapy. While medical imaging remains a vital component of a myriad of clinical tasks, an increasing shortage of qualified radiologists to interpret complex medical images suggests a clear need for reliable automated methods to alleviate the growing burden on health-care practitioners [64].

In parallel, medical imaging sciences are benefiting from the development of novel computational techniques for the analysis of structured data like images. Development of algorithms for image acquisition, analysis and interpretation are driving innovation, particularly in the areas of registration, reconstruction, tracking, segmentation and modelling.

Medical images are inherently difficult to interpret, requiring prior expertise to understand. Bio-medical images can be noisy and contain many modality-specific artefacts, acquired under a wide variety of acquisition conditions with different protocols. Thus, once trained, models do not transfer seamlessly from one clinical task or site to another because of an often yawning domain gap [65, 66]. Supervised learning methods require extensive relabelling to regain initial performance in different workflows.

The experience and prior knowledge required to work with such data means that there is often large inter- and intra-observer variability in annotating medical data. This not only raises questions about what constitutes a gold-standard ground truth annotation, but also results in disagreement of what that ground truth truly is. These issues result in a large cost associated with annotating and re-labelling of medical image datasets, as we require numerous expert annotators (oracles) to perform each annotation and to reach a consensus.

In recent years, DL has emerged as the state-of-the-art technique for performing many medical image analysis tasks [29, 67, 68, 69, 70]. Developments in the field of computer vision have shown great promise in transferring to medical image analysis, and several techniques have been shown to perform as accurately as human observers [71, 72]. However, uptake of DL methods within the clinical practice has been limited thus far, largely due to the unique challenges of working with complex medical data, regulatory compliance issues and trust in trained models.

We identify three key challenges when developing DL enabled applications for medical image analysis in a clinical setting:

1. Lack of Training Data: Supervised DL techniques traditionally rely on a large and even

distribution of accurately annotated data points, and while more medical image datasets are becoming available, the time, cost and effort required to annotate such datasets remains significant.

2. The Final Percent: DL techniques have achieved state-of-the-art performance for medical image analysis tasks, but in safety-critical domains even the smallest of errors can cause catastrophic results downstream. Achieving clinically credible output may require interactive interpretation of predictions (from an oracle) to be useful in practice, i.e users must have the capability to correct and override automated predictions for them to meet any acceptance criteria required.

3. Transparency and Interpretability: At present, most DL applications are considered to be a 'black-box' where the user has limited meaningful ways of interpreting, understanding or correcting how a model has made its prediction. Credence is a detrimental feature for medical applications as information from a wide variety of sources must be evaluated in order to make clinical decisions. Further indication of how a model has reached a predicted conclusion is needed in order to foster trust for DL enabled systems and allow users to weigh automated predictions appropriately.

There is concerted effort in the medical image analysis research community to apply DL methods to various medical image analysis tasks, and these are showing great promise [25, 73, 74]. These works primarily focus on the development of predictive models for a specific task and demonstrate state-of-the-art performance for that task.

We give an overview of where humans will remain involved in the development, deployment and practical use of DL systems for medical image analysis. We focus on medical image segmentation techniques to explore the role of human end users in DL enabled systems.

Automating image interpretation tasks like image segmentation suffers from all of the drawbacks incurred by medical image data described above. There are many emerging techniques that seek to alleviate the added complexity of working with medical image data to perform automated segmentation of images. Segmentation seeks to divide an image into semantically meaningful regions (sets of pixels) in order to perform a number of downstream tasks, e.g. biometric measurements. Manually assigning a label to each pixel of an image is a laborious task and as such automated segmentation methods are important in practice. Advances in DL techniques such as Active Learning (AL) and Human-in-the-Loop computing applied to segmentation problems have shown progress in overcoming the key challenges outlined above and these are the studies we focus on. We categorise each study based on the nature of human interaction proposed and broadly divide them between which of the three key challenges they address.

Section 2.2.1 introduces Active Learning, a branch of ML and Human-in-the-Loop Computing that seeks to find the most *informative* samples from an unlabelled distribution to be annotated next. By training on the most informative subset of samples, related work can achieve state-of-the-art performance while reducing the costly annotation burden associated with annotating medical image data.

Section 2.2.2 evaluates techniques used to refine model predictions in response to user feedback, guiding models towards more accurate per-image predictions. We evaluate techniques that seek to improve interpretability of automated predictions and how models provide feedback on their own outputs to guide users towards better decision making.

Section 2.2.3 evaluates the key practical considerations of developing and deploying Human-in-the-Loop DL enabled systems in practice and outlines the work being done in these areas that addresses the three key challenges identified above. These areas are human focused and assess how human end users might interact with these systems.

In Section 2.2.4 we discuss related areas of ML and DL research that are having an impact on AL and Human-in-the-Loop Computing and are beginning to influence the three key challenges outlined. We offer our opinions on the future directions of Human-in-the-Loop DL research and how many of the techniques evaluated might be combined to work towards common goals.

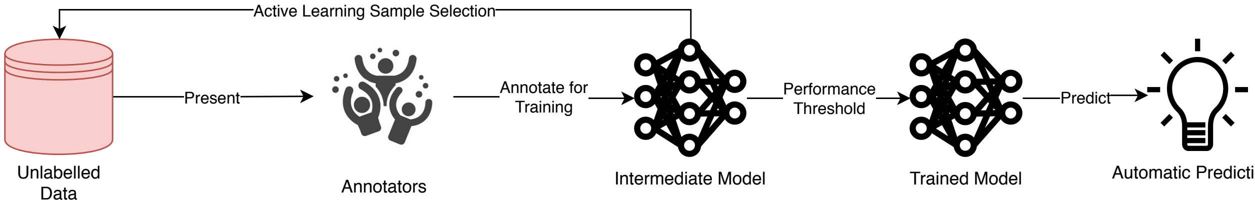


Figure 2.10: Overview of Active Learning frameworks. Active Learning frameworks start with a pool on unlabelled data. Unlabelled data is then presented to 'Annotators' for annotation. Annotated data is then used to train an 'Intermediate Model' using a subset of the Unlabelled Data. This model is then used to select progressively more unlabelled data for annotation using 'Active Learning Sample Selection', after which the model is retrained to improve its performance. This process continues until a certaint 'Performance Threshold' is met, after which the model can be deployed to make 'Automatic Predictions'.

2.2.1 Active Learning

In this section we assume a scenario in which a large pool of un-annotated data U is available to us, and that we have an oracle (or group of oracles) from which we can request annotations for every un-annotated data point x_U to add to an annotated set L. We wish to train some model $f(x|L^*)$ where $L^* \subseteq L$ and consider methods that rely on annotated data to do so. A brute-force solution to this problem would be to ask the oracle(s) to annotate every x_U such that $L^* = L$, but this is rarely a practical or cost-effective solution due to the unique challenges associated with annotating biomedical image data. It is theorised that there is some L^* that achieves equivalent performance to L, i.e. $f(x|L^*) \approx f(x|L)$. A model trained on some optimal subset L^* of a dataset might achieve equivalent performance to a model trained on the entire, annotated dataset. Active Learning (AL) is the branch of ML that seeks to find this optimal subset L^* given a current model $f'(x|L')$, where L' is an intermediate annotated dataset, and an un-annotated dataset U. AL methods aim to iteratively seek the most informative data-points x_i^* for training a model, under the assumption that both the model and the un-annotated dataset will evolve over time, rather than selecting a fixed subset once to be used for training. In a wider context and before the advent of DL, [31] reviewed this field as a state-of-the-art ML methodology.

A typical AL framework, as outlined in Figure 2.10, consists of a method to evaluate the *informativeness* of each un-annotated data point x_U given $f'(x_U|L')$, tied heavily to the choice of *query type*, after which all chosen data-points are required to be annotated. Once new annotations have been acquired, the AL framework must use the new data to improve the model. This is normally done by either *retraining* the entire model using all available annotated data L', or by *fine-tuning* the network using the most recently annotated data-points x_i^*. Using this approach, state-of-the-art performance can be achieved using fewer annotations for several bio-medical image analysis tasks, as shown in the methods discussed in this section, thus widening the annotation bottleneck and reducing the costs associated with developing DL enabled systems from un-annotated data.

Query Types

In every AL framework the first choice to be made is what type of *query* we wish to make using a model and un-annotated dataset. There are currently three main choices available and each lends itself to a particular scenario dependant on what type of un-annotated data we have access to, and what question we wish to ask the oracle(s).

Stream-based Selective Sampling assumes a continuous stream of incoming un-annotated data-points x_U [75, 76]. The current model and an *informativeness* measure $I(x_U)$ are used to decide, for each incoming data-point, whether or not to ask the oracle(s) for an annotation [77]. This query type is usually computationally inexpensive but offers limited performance benefits due to the isolated nature of each decision: the wider context of the underlying distribution is not considered, thus balancing exploration and exploitation of the distribution is less well captured than in other query types. Another disadvantage of this query type is calibrating the threshold to use for the chosen informativeness measure such that we do not request annotations for every incoming data-point, and that we do not reject annotations for too many data-points resulting in valuable information being lost.

Membership Query Synthesis assumes that rather than drawing from a real-world distribution of data-points, we instead generate a data-point x_G^* that needs to be annotated [78]. The generated data-point is what the current model 'believes' will be most informative to itself. This data-point is then annotated by the oracle(s) [79], this can be very efficient in finite domains.

This approach may suffer from the same drawbacks as *Stream-based* methods as a model may have no knowledge of unseen areas of the distribution, and thus be unable to request annotations of those areas. Issues can arise where queries can request annotations for data-points that make no sense to a human oracle [80], and are not representative of the actual distribution that is being modelled, stream based and pool based sampling methods were proposed to overcome these issues [31]. Nevertheless, recent advances of *Generative Adversarial Networks* (GANs) have shown great promise in generating data-points that mimic real-world distributions for many different types of data, including biomedical images, that may go someway to addressing the key issue with using query synthesis for complex distributions, which we discuss in Section 2.2.1. This query type can be advantageous in scenarios where the distribution to generate is fully understood, or domains in which annotations are acquired autonomously instead of from humans [81, 82].

Pool-based Sampling assumes a large un-annotated real-world dataset U to draw samples from and seeks to select a batch of N samples $x_0^*, ..., x_N^*$ from the distribution to request labels for [83]. *Pool-based* methods usually use the current model to make a prediction on each un-annotated data point to obtain a ranked measure of *informativeness* $I(x_U|f'(x_U|L'))$ (where I is our *informativeness measure* for each un-annotated data point x_U given the output of the intermediate model f' where L' is our current labelled dataset) for every data-point in the un-annotated set, and select the top N samples using this metric to be annotated by the oracle(s). Pool based sampling has been applied to several real world tasks, prior to the advent of DL [83, 84, 85, 86, 87]. These methods can be computationally expensive as every iteration requires a metric evaluation for every data-point in the distribution. However, these methods have shown to be the most promising when combined with DL methods, which inherently rely on a batch-based training scheme. Pool based sampling is used in the majority of methods discussed in the rest of this section unless stated otherwise. While pool-based methods hold advantages over other methods in terms of finding the most informative annotations to acquire, scenarios in which stream based or synthesis based queries are advantageous are also common, such as when memory or processing power is limited for example in mobile or embedded devices [31].

Evaluating Informativeness

In developing an AL framework, once a query type has been selected, the next question to ask is how to measure the informativeness $I(x_U)$ of each of the data-points? Many varying approaches have been taken to quantifying the informativeness of a sample given a model and an underlying distribution. Here we sort these metrics by the level of human interpretability they offer.

Traditionally, AL methods employ hand-designed heuristics to quantify what we as humans believe makes something informative. A variety of model specific metrics seek to quantify what the effect of using a sample for training would have on the model, e.g., the biggest change in model parameters. However, these methods are less prevalent than human designed heuristics due to the computational challenge of applying these to DL models with a large number of parameters. Finally some methods are emerging that are completely agnostic to human interpretability of informativeness and instead seek to learn the best selection policy from available data and previous iterations, as discussed in detail in Section 2.2.1.

Uncertainty The main family of informativeness measures falls into calculating uncertainty. It is argued that the more *uncertain* a prediction is, the more information we can gain by including the ground truth for that sample in the training set.

Uncertainty can be broken down into two types of uncertainty: 'Aleatoric' and 'Epistemic' uncertainty, also known as statistical uncertainty and systematic uncertainty respectively. These can be defined as the uncertainty present in the data presented to a model (e.g due to noise), which is unavoidable regardless of how good a model is (aleatoric) and the uncertainty inherent to the model (e.g due to limited capacity) which can be reduced by improving the model but in practice has not been (epistemic). The methods discussed in this thesis address both types, often separately and sometimes together. Evaluation of uncertainty measures is a open research question but methods such as calibration curves are popular to measure the correlation between prediction accuracy and prediction uncertainty.

There are several ways of calculating uncertainty from different ML/DL models. When considering DL for segmentation the most simple measure is the sum of lowest class probability for each pixel in a given image segmentation. It is argued that more certain predictions will have high pixel-wise class probabilities, so the lower the sum of the minimum class probability over each pixel in an image, the more certain a prediction is considered to be:

$$x^*_{LC} = \underset{x}{\operatorname{argmax}}\, 1 - P_\theta(\hat{y}|x)$$

where $\hat{y} = \operatorname{argmax}_y P_\theta(y|x)$. This is a fairly intuitive way of thinking about uncertainty and offers a means to rank uncertainty of samples within a distribution. While the softmax outputs of a model do not directly reflect the actual class probabilities, they are a useful proxy for this information. When used directly as a measure of uncertainty we see varying results but due to the ease of which these are computed, and the often high performance of this approach it has become a useful baseline method. We refer to the method above as *least confident* sampling where the samples with the highest uncertainty are selected for labelling [31]. A drawback of *least confident* sampling is that it only considers information about the most probable label, and discards the information about the remaining label distribution. Two alternative methods have been proposed that alleviate this concern. The first, called *margin sampling* [31], can be used in a multi-class setting and considers the first and second most probable labels under the model and calculates the difference between them:

$$x^*_M = \underset{x}{\operatorname{argmin}}\, P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)$$

where $\hat{y}_1$ and $\hat{y}_2$ are the first and second most probable labels under the current model, respectively. The intuition here is that the larger the margin is between the two most probable labels, the more confident the model is in assigning that label. The second, more popular approach is to use *entropy* [88] as an uncertainty measure:

$$x^*_E = \underset{x}{\operatorname{argmax}} - \sum_i P(y_i|x) logP(y_i|x)$$

where y_i ranges across all possible annotations. Entropy is used to measure the amount of information required to encode a distribution and as such, is often thought of as a measure of uncertainty in ML. For binary classification, all three methods reduce to querying for the data-point with a class posterior closest to 0.5. The ability of *entropy* to generalise easily to probabilistic multi-class annotations, as well as models for more complex structured data-points has made it the most popular choice for uncertainty based query strategies [85].

Using one of the above measures, un-annotated samples are ranked and the most 'uncertain' cases are chosen for the next round of annotation. There have been many recent uses of uncertainty based sampling in AL methods in the DL field and these are discussed next.

[89] propose the Cost-Effective Active Learning (CEAL) method for deep image classification. The CEAL methods is initialised with a set of unlabelled sample U, initially labelled samples L, a choice of pool size K, a high confidence sample selection threshold ω, a threshold decay rate dr, a maximum iteration number T and a fine-tuning interval t. After initialisation, CNN weights W are initialised with L and the model is used to make predictions on each data-point in U. CEAL explores using each of the three uncertainty methods described above to obtain K uncertain data-points to be manually annotated and added to D_L. So far the CEAL method follows very closely the approach outlined in traditional active learning methods as described above, but they introduce an additional training step where the most confident samples (whose *entropy* is less than ω) from U are added to D_H. D_L and D_H are then used to fine-tune W for t iterations. CEAL then updates ω before the *pseudo-labels* from D_H are discarded and each data-point is added back to U, while D_L is added to L. This process repeats for T iterations. The authors describe this approach of simultaneously learning from manual labels of the most uncertain annotations and predicted labels of the least uncertain annotations as *complementary sampling*. The CEAL method showed that state-of-the-art performance can be achieved using less than 60% of available data for two non-medical datasets (CACD and Caltech-256) for face recognition and object categorisation.

[90] propose an active learning method that uses uncertainty sampling to support quality control of nucleus segmentation in pathology images. Their work compares the performance improvements achieved through active learning for three different families of algorithms: Support

Vector Machines (SVM), Random Forest (RF) and Convolutional Neural Networks (CNN). They show that CNNs achieve the greatest accuracy, requiring significantly fewer iterations to achieve equivalent accuracy to the SVMs and RFs.

Another common method of estimating informativeness is to measure the agreement between multiple models performing the same task. It is argued that more disagreement found between predictions on the same data point implies a higher level of uncertainty. These methods are referred to as *Query by consensus* and are generally applied when *Ensembling* is used to improve performance - i.e, training multiple models to perform the same task under slightly different parameters/settings [31]. Ensembling methods have shown to measure informativeness well, but at the cost of computational resources - multiple models need to be trained and maintained, and each of these needs to be updated in the presence of newly selected training samples.

Nevertheless, [91] demonstrate the power of ensembles for active learning and compare to alternatives to ensembling. They specifically compare the performance of acquisition functions and uncertainty estimation methods for active learning with CNNs for image classification tasks and show that ensemble based uncertainties outperform other methods of uncertainty estimation such as 'MC Dropout'. They find that the difference in active learning performance can be explained by a combination of decreased model capacity and lower diversity of MC dropout ensembles. A good performance is demonstrated on a diabetic retinopathy diagnosis task.

[92] propose an active learning approach that exploits geometric smoothness priors in the image space to aid the segmentation process. They use traditional uncertainty measures to estimate which pixels should be annotated next, and introduce novel criteria for uncertainty in multi-class settings. They exploit geometric uncertainty by estimating the entropy of the probability of supervoxels belonging to a class given the predictions of its neighbours and combine these to encourage selection of uncertain regions in areas of non-smooth transition between classes. They demonstrate state-of-the-art performance on mitochondria segmentation from EM images and on an MRI tumour segmentation task for both binary and multi-class segmentations. They suggest that exploiting geometric properties of images is useful to answer the questions of where

to annotate next and by reducing 3D annotations to 2D annotations provide a possible answer to how to annotate the data, and that addressing both jointly can bring additional benefits to the annotation method, however they acknowledge that it would impossible to design bespoke selection strategies this way for every new task at hand.

[93] introduce the use of Bayesian CNNs for Active Learning with 'Bayesian Active Learning by Disagreement' or BALD, and show that the use of Bayesian CNNs outperform deterministic CNNs in the context of Active Learning, and exploit this through the use of a new acquisition function that chooses data-points expected to maximise the information gained about the model parameters i.e maximise the mutual information between predictions and model posterior. This approach uses a Bayesian CNN (induced using Dropout during inference [94]), to produce a single prediction using all parameters of the network for each unlabelled data-point, and a set of stochastic predictions for each unlabelled data-point, generated with dropout enabled. The BALD acquisition function is then calculated as the difference between the entropy of the average prediction and average entropy of stochastic predictions. Intuitively this function selects data-points for which the model is uncertain on average, but there exist model parameters that produce disagreeing predicted annotations with high certainty. They demonstrate their approach for skin cancer diagnosis from skin lesion images to show significant performance improvements over uniform sampling using the BALD method for sample selection. While this method has been shown to be particularly effective for AL, when querying batches of data-points, it often results in many very similar, redundant data-points being acquired when used in a greedy fashion, as such BatchBALD was introduced to alleviate this problem [95]. The BatchBALD approach instead no longer calculates the mutual information between a single sample predictions and model posterior, but instead calculates the mutual information between a batch of samples and the model posterior to jointly score the batch of samples, enabling BatchBALD to more accurately evaluate the joint mutual information and select batches of samples for annotation that result in less redundant data-points being selected together in an acquired batch. This extension is an example of the motivation behind Section 2.2.1 in which we discuss methods that move beyond pure uncertainty based methods and begin to measure diversity among selected samples to reduce redundant annotation.

Representativeness Many AL frameworks extend selection strategies to include some measure of *representativeness* in addition to an uncertainty measure. The intuition behind including a representativeness measure is that methods only concerned with *uncertainty* have the potential to focus only on small regions of the distribution, and that training on samples from the same area of the distribution will introduce redundancy to the selection strategy, or may skew the model towards a particular area of the distribution. The addition of a representativeness measure seeks to encourage selection strategies to sample from different areas of the distribution, and to increase the diversity of samples, thus improving AL performance. A sample with a high representativeness covers the information for many images in the same area of the distribution, so there is less need to include many samples covered by a representative image.

To this end, [96] present Suggestive Annotation, a deep active learning framework for medical image segmentation, which uses an alternative formulation of uncertainty sampling combined with a form of representativeness density weighting. Their method consists of training multiple models that each exclude a portion of the training data, which are used to calculate an ensemble based uncertainty measure. They formulate choosing the most representative example as a generalised version of the maximum set-cover problem (NP Hard) and offer a greedy approach to selecting the most representative images using feature vectors from their models. They demonstrate state-of-the-art performance using 50% of the available data on the MICCAI Gland segmentation challenge and a lymph node segmentation task.

[97] propose *MedAL*, an active learning framework for medical image segmentation. They propose a sampling method that combines uncertainty, and distance between feature descriptors, to extract the most informative samples from an unlabelled data-set. Once an initial model has been trained, the MedAL method selects data-points to be labelled by first filtering out unlabelled data-points with a predictive entropy below a threshold. From this set the CNN being trained is used to generate feature descriptors for each data-point by taking the output of intermediate layers of the CNN, these feature descriptors are then compared amongst each other using a variety of distance functions (e.g 'Euclidian', 'Russellrao', 'Cosine') in order to find the feature descriptors which are most distant from each other. The data-point with the highest average distance to all other unlabelled data-points (above the entropy threshold) is

selected for annotation. In this way, the MedAL acquisition function finds the set of data-points that are both informative to the model, and incur the least redundancy between them by sampling from areas of the input distribution most distant from each other. The MedAL method initialises the model in a novel way by leveraging existing computer vision image descriptors to find the images that are most dissimilar to each other and thus cover a larger area of the image distribution to use as the initial training set after annotation. They show good results on three different medical image analysis tasks, achieving the baseline accuracy with less training data than random or pure uncertainty based methods.

[98] propose a Borda-count based combination of an uncertainty and a representativeness measure to select the next batch of samples. Uncertainty is measured as the voxel-wise variance of N predictions using MC dropout in their model. They introduce new representativeness measures such as 'Content Distance', defined as the mean squared error between layer activation responses of a pre-trained classification network. They extend this contribution by encoding representativeness by maximum entropy to optimise network weights using an novel entropy loss function.

[99] propose a novel method for ensuring diversity among queried samples by calculating the Fisher Information (FI), for the first time in CNNs. Here, efficient computation is enabled by the gradient computations of propagation to allow FI to be calculated on the large parameter space of CNNs. They demonstrate the performance of their approach on two different flavours of task: a) semi-automatic segmentation of a particular subject (from a different group/different pathology not present in the original training data) where iteratively labelling small numbers of voxels queried by AL achieves accurate segmentation for that subject; and b) using AL to build a model generalisable to all images in a given data-set. They show that in both these scenarios the FI-based AL improves performance after labelling a small percentage of voxels, outperforming random sampling and achieved higher accuracy than entropy based querying.

Generative Adversarial Networks for Informativeness

Generative Adversarial Network (GAN) based methods have been applied to several areas of medical imaging such as de-noising, modality transfer, abnormality detection, and for image synthesis, directly applicable to AL scenarios. This offers an alternative (or addition) to the many data augmentation techniques used to expand limited data-sets [100] and a DL approach to *Membership Query Synthesis*.

[101] propose a conditional GAN (cGAN) based method for active learning where they use the discriminator D output as a measure of uncertainty of the proposed segmentations, and use this metric to rank samples from the unlabelled data-set. From this ranking the most uncertain samples are presented to an oracle for segmentation and the least uncertain images are included in the labelled data-set as *pseudo ground truth* labels. They show their method approaches increasing accuracy as the percentage of interactively annotated samples increases - reaching the performance of fully supervised benchmark methods using only 80% of the labels. This work motivates the use of GAN discriminator scores as a measure of prediction uncertainty.

[102] also use a cGAN to generate chest X-Ray images conditioned on a real image, and using a Bayesian neural network to assess the informativeness of each generated sample, decide whether each generated sample should be used as training data. If so, is used to fine-tune the network. They demonstrate that the approach can achieve comparable performance to training on the fully annotated data, using a dataset where only 33% of the pixels in the training set are annotated, offering a huge saving of time, effort and costs for annotators.

[103] present an alternative method of data synthesis to GANs through the use of learned transformations. From a single manually segmented image, they leverage other un-annotated images in a SSL like approach to learn a transformation model from the images, and use the model along with the labelled data to synthesise additional annotated samples. Transformations consist of spatial deformations and intensity changes to enable to synthesis of complex effects such as anatomical and image acquisition variations. They train a model in a supervised way for the segmentation of MRI brain images and show state-of-the-art improvements over other one-shot bio-medical image segmentation methods.

The utility of GAN-based approaches in AL scenarios goes beyond single-modality image synthesis. Many works have demonstrated the capabilities of GANs to perform cross-modality image synthesis, which directly addresses not only problems of limited training data, but also issues of missing modalities which occur in multi-modal analysis scenarios. Methods by which missing modalities can be generated to fill missing data-points enabling the full suite of AL methods to be applied to multi-modal analysis problems.

[104] introduce a GAN based method for super-resolution across different microscopy modalities. This work uses GANs to transform diffraction limited input images into super-resolved ones, improving the resolution of wide-field images acquired using low-numerical-aperture objectives to match the resolution acquired using high-numerical-aperture objectives. This work extends this approach to demonstrate cross-modality super-resolution to transform confocal microscopy images to the resolution acquired with a stimulated emission depletion microscope. This approach enables many types of images acquired at lower resolutions to be super-resolved to match those of higher resolutions, enabling greater performance of multi-modal image analysis methods in both AL and beyond.

[105] introduce a GAN based method for the generation of high-quality PET images which usually require a full dose radioactive tracer to obtain. This work enables a low dose tracer to be used to obtain a low-quality PET images, from which a high quality PET image can be generated using a 3D conditional GAN, conditioned on the low-dose image. Additional to this, a 3D c-GANs based progressive refinement scheme is introduced to further improve the quality of estimated images. Through this work the dose of radioactive tracer required to acquire high-quality PET images is greatly reduced, reducing the hazards to patients and enabling low-dose PET images to be used alongside high-dose images in downstream analysis.

[106] extend existing GAN based methods for improved cross-modality synthesis of MR images acquired under different scanning parameters. Their work introduces edge-aware generative adversarial networks (Ea-GANs), which specifically integrate edge information reflecting the textural structure of image content to depict the boundaries of different objects in images, which goes beyond methods which focus only on minimising pixel or voxel-wise intensity differences. Using two learning strategies they introduce edge information to a generator-induced Ea-GAN (gEa-GAN) and to a discriminator-induced Ea-GAN (dEa-GAN), incorporating edge information via the generator and both generator and discriminator respectively, so that the

edge similarity is also adversarially learned. Their method demonstrates state-of-the-art performance for cross-modal MR synthesis as well as excellent generality to generic image synthesis tasks on facades, maps and cityscapes.

[107] explore the use of GANs to impute missing PET images from corresponding MR images for brain disease identification using a GAN based approach, to avoid discarding data-missing subjects, thus increasing the number of training samples available. A hybrid GAN is used to generate the missing PET images, after which a spatially-constrained Fisher representation network is used to extract statistical descriptors of neuroimages for disease diagnosis. Results on three databases show this method can synthesise reasonable neuroimages and achieve promising results in brain disease identification in comparison to other state-of-the-art methods.

The above works demonstrate the power of using synthetic data conditioned on a very small amount of annotated data to generate new training samples that can be used to train a model to a high accuracy, this is of great value to AL methods where we usually require a initial training set to train a model on before we can employ a data selection policy. These methods also demonstrate the efficient use of labelled data and allow us to generate multiple training samples from a individually annotated image, this may allow the annotated data obtained in AL/Human-in-the-Loop methods to be used more effectively through generating multiple training samples for a single requested annotation, further reducing the annotation effort required to train state-of-the-art models.

Learning Active Learning

The majority of methods discussed so far employ hand designed heuristics of informativeness, but some works have emerged that attempt to learn what the most informative samples are through experience of previous sample selection outcomes. This offers a potential way to select samples more efficiently but at the cost of interpretability of the heuristics employed. Many factors influence the performance and optimality of using hand-crafted heuristics for data selection. [108] propose 'Learning Active Learning', where a regression model learns data selection strategies based on experience from previous AL outcomes. Arguing there is no way to foresee the influence of all factors such as class imbalance, label noise, outliers and distribution shape. Instead, their regression model 'adapts' its selection to the problem without explicitly stating specific rules. [109] take this idea a step further and propose a model that leverages labelled instances from different but related tasks to learn a selection strategy, while simultaneously adapting its representation of the data and its prediction function.

Reinforcement learning (RL) is a branch of ML that enables an 'agent' to learn in an interactive environment, by trial and error, using feedback from its own actions and experiences, working towards achieving the defined goal of the system. Active Learning has recently been suggested as a potential use-case of RL and several works have begun to explore this area.

[110] propose a one-shot learning method that combines with RL to allow the model to decide, during inference, which examples are worth labelling. A stream of images is presented and a decision is made either to predict the label, or pay to receive the the correct label. Through the choice of RL reward function they are able to achieve higher prediction accuracy than a purely supervised task, or trade prediction accuracy for fewer label requests.

[111] re-frame the data selection process as a RL problem, and explicitly learn a data selection policy. This is agnostic to the data selection heuristics common in AL frameworks, providing a more general approach, demonstrating improvements in entity recognition, however this is yet to be applied to medical image data.

RL methods offer a different approach to AL and Human-in-the-Loop problems that is well aligned with aiding real-time feedback between a DL enabled application and its end users, however it requires task specific goals that may not be generalisable across different medical image analysis tasks.

Fine-tuning vs Retraining

The final step of each AL framework is to use newly acquired annotations to improve a model. Two main approaches are used to train a model on new annotations. These are retraining the model using all available data including the newly acquired annotations or to fine-tune the model using only new annotations or the new annotations plus a subset from the existing annotations.

[112] investigate using transfer learning and fine-tuning in several medical image analysis tasks and demonstrate that the use of a pre-trained CNN with fine-tuning outperformed a CNN trained from scratch and that these fine-tuned CNNs were more robust to the size of the training sets. They also showed that neither shallow nor deep tuning was the optimal choice for a particular application and present a layer-wise training scheme that could offer a practical way to reach optimal performance for the chosen task based on the amount of data available. The methods employed in this work perform one-time fine-tuning where a pre-trained model is fine-tuned just once with available training samples, however this does not accommodate an active selection process or continuous fine-tuning.

[113] propose a continuous fine-tuning method that fine-tunes a pre-trained CNN with successively larger datasets and demonstrate that this approach converges faster than repeatedly fine-tuning the pre-trained CNN. They also find that continuously fine-tuning with only newly acquired annotations requires careful meta-parameter adjustments making it less practical across many different tasks.

An alternative approach to retraining from new data that is inspired by the two main approaches described above is to retrain a model using all available data, but using the previous parameters as initialisation, however this approach has not yet been applied to AL in medical image analysis.

Retraining is computationally more expensive than fine-tuning but it provides a consistent means to evaluate AL framework performance. Fine-tuning is used across a number of different ML areas such as one or few shot learning, and transfer learning and the best approach to this is still an open question and as such is less prevalent in AL frameworks, as fine tuning improves we may see a shift towards its use in AL frameworks. It is important to establish baseline fine-tuning and retraining schemes to effectively compare the DL/AL methods in which they are applied in order to isolate the effects of these schemes from the improvements made in other areas.

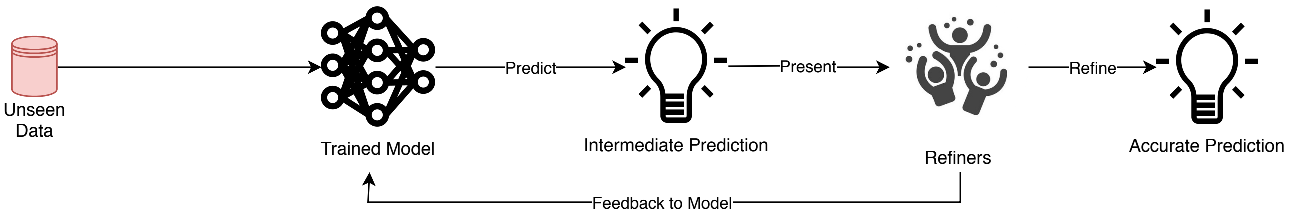


Figure 2.11: Overview of Refinement frameworks.

2.2.2 The Final Percent: Interactive refinement of model outputs

So far we have considered the role of humans in annotating data to be used to train a model, but once a model is trained, we still require a human-in-the-loop to interpret model predictions and potentially to refine them to acquire the most accurate results for unseen data, as outlined in Figure 2.11. In Human-in-the-loop scenarios, a model makes predictions on unseen input, and subject to acceptance criteria, automated predictions may need manual adjustment to meet those acceptance criteria. Communication of information about the prediction is important to allow acceptance criteria to be met with confidence, and form an understanding of the limitations of automated predictions. This communication is two fold i.e. a user must be able to communicate with the model being used to guide predictions to more accurate results or to correct erroneous predictions, and a model must be able to communicate with the user to provide meaningful interpretation of model predictions, enabling users to take the best course of action when interacting with model outputs and to mitigate human uncertainty. This creates the feedback loop as shown in Figure 2.11.

Interactive Refinement

If we can develop accurate, robust and interpretable models for medical image applications we still cannot guarantee automated predictions meet acceptance criteria for every unseen data-point presented to a model. The ability to generalise to unseen input is a cornerstone of DL applications, but in real world distributions, generalisation is rarely perfect. As such, methods to rectify these discrepancies must be built into applications used for medical image analysis. This iterative refinement must save the end user time and mental effort over performing manual annotation or purely manual correction. Many interactive image segmentation systems have been proposed, and more recently these have built on the advances in DL to allow users to refine model outputs and feedback the more accurate results to the model for improvement.

[32] introduced UI-Net, that builds on the popular U-Net architecture for medical image segmentation [18]. The UI-Net is trained with an *active user model*, and allows for users to interact with proposed segmentations by providing *scribbles* over the image to indicate areas that should be included or not, the network is trained using simulated user interactions and as such responds to iterative user scribbles to refine a segmentation towards a more accurate result.

Conditional Random fields have been used in various tasks to encourage segmentation homogeneity. [114] propose CRF-CNN, a recurrent neural network which has the desirable properties of both CNNs and CRFs. [115] propose DeepIGeoS, an interactive geodesic framework for medical image segmentation. This framework uses two CNNs, the first performs an initial automatic segmentation, and the second takes the initial segmentation as well as user interactions with the initial segmentation to provide a refined result. They combine user interactions with CNNs through geodesic distance transforms [116], and these user interactions are integrated as hard constraints into a Conditional Random Field, inspired by [114]. They call their two networks P-Net (initial segmentation) and R-Net (for refinement). They demonstrate superior results for segmentation of the placenta from 2D fetal MRI and brain tumors from 3D FLAIR images when compared to fully automatic CNNs. These segmentation results were also obtained in roughly a third of the time taken to perform the same segmentation with traditional interactive methods such as GeoS or ITK-SNAP.

Graph Cuts have also been used in segmentation to incorporate user interaction - a user provides *seed points* to the algorithm (e.g. mark some pixel as foreground, and another as background) and from this the segmentation is calculated. [117] propose BIFSeg, an interactive segmentation framework inspired by graph cuts. Their work introduces a DL framework for interactive segmentation by combining CNNs with a bounding box and scribble based segmentation pipeline. The user provides a bounding box around the area which they are interested in segmenting, this is then fed into their CNN to produce an initial segmentation prediction, the user can then provide scribbles to mark areas of the image as mis-classified - these user inputs are then weighted heavily in the calculation of the refined segmentation using their graph cut based algorithm.

[118] propose an alternative to BIFSeg in which two networks are trained, one to perform an initial segmentation (they use a CNN but this initial segmentation could be performed with any existing algorithm) and a second network they call interCNN that takes as input the image, some user scribbles and the initial segmentation prediction and outputs a refined segmentation, they show that with several iterations over multiple user inputs the quality of the segmentations improve over the initial segmentation and achieve state-of-the-art performance in comparison to other interactive methods.

The methods discussed above have so far been concerned with producing segmentations for individual images or slices, however many segmentation tasks seek to extract the 3D shape/surface of a particular region of interest (ROI). [119] propose a dual method for producing segmentations in 3D based on a Smart-brush 2D segmentation that the user guides towards a good 2D segmentation, and after a few slices are segmented this is transformed to a 3D surface shape using Hermite radial basis functions, achieving high accuracy. While this method does not use DL it is a strong example of the ways in which interactive segmentation can be used to generate high quality training data for use in DL applications - their approach is general and can produce segmentations for a large number of tasks. There is potential to incorporate DL into their pipeline to improve results and accelerate the interactive annotation process.

[120] propose an interactive segmentation scheme that generalises to any previously trained segmentation model, which accepts user annotations about a target object and the background. User annotations are converted into interaction maps by measuring the distance of each pixel to the annotated landmarks, after which the forward pass outputs an initial segmentation. The user annotated points can be mis-segmented in the initial segmentation so they propose BRS (back-propogating refinement scheme) that corrects the mis-labelled pixels. They demonstrate that their algorithm outperforms conventional approaches on several datasets and that BRS

can generalise to medical image segmentation tasks by transforming existing CNNs into user-interactive versions.

[121] propose modelling the dynamics of iterative interactive refinement as a Markov Decision Process (MDP) and solve this with multi-agent RL. Treating each voxel as an agent with a shared voxel-level behaviour strategy they make voxel-wise prediction tractable in this way. The multi-agent method successfully captures the dependencies among voxels for segmentation tasks, and by passing prediction uncertainty of previous segmentations through the state space can derive more precise and finer segmentations. Using this method they significantly outperform existing state-of-the-art methods with fewer interactions and a faster convergence.

In this section we focus on applications concerned with iteratively refining a segmentation towards a desired quality of output. In the scenarios above this is performed on an un-seen image provided by the end user, but there is no reason the same approach could not be taken to generate iteratively more accurate annotations to be used in training, e.g., using active learning to select which samples to annotate next, and iteratively refining the prediction made by the current model until a sufficiently accurate annotation is curated. This has the potential to accelerate annotation for training without any additional implementation overhead. Much work done in AL ignores the role of the oracle and merely assumes we can acquire an accurate label when we need it, but in practice this presents a more significant challenge. We foresee AL and HITL computing become more tightly coupled as AL research improves it's consideration for the oracle providing the annotations.

It is fairly intuitive how a user might refine segmentations of medical images, but this is not the case for other medical image analysis tasks. Refinements of predictions on clinical tasks involving classification and regression have seen less development than those in segmentation and remains an open area of research. The following works have taken steps towards addressing interactive refinement strategies for classification and regression tasks.

[122] explore the use of CNN methods for automated diagnosis of Alzheimer's disease and identify that many state-of-the-art methods rely on the pre-determination of informative locations in structural MRI (sMRI). This stage of discriminative localisation is isolated from the latter

stages of feature extraction and classifier construction. Their work proposes a hierarchical fully convolutional CNN (H-FCN) to automatically identify discriminative local patches and regions in whole brain sMRI, from which multi-scale feature representations can be jointly learned and fused to construct classification models. This work enables interactive refinement of patch choice and classifier construction which, if intervened on by human end users could guide the network towards more discriminative regions of interest and thus more effective classifiers.

Similarly, [123] introduce a landmark-based deep multi-instance learning (LDMIL) framework for brain disease diagnosis. Firstly, by adopting a data-driven approach to discover disease related anatomical landmarks in brain MR images, along with nearby image patches. Secondly the framework learns an end-to-end MR image classifier for capturing local structural information in the selected landmark patches, and global structure information derived from all detected landmarks. By splitting the steps of landmark detection and classifier construction, a human-in-the-loop can be introduced to intervene on selected landmarks and to guide the network towards maximally informative image regions. Thus, the resulting classifier can be refined via updating which regions of the image are used as input.

Interactive Interpretation

In the previous section we discussed methods by which the user of a human-in-the-loop system might communicate with a predictive model, in this section we consider methods by which a model might communicate with the user, thus completing the feedback loop in Figure 2.11. 'Interpretation' can mean many different things depending on the context, so here we focus on interpretation of model outputs with the goal of appropriately weighting automated predictions in downstream analysis (e.g uncertainty of predictions) and to enable users to make the most informed corrections or manual adjustments to model predictions (e.g 'Attention Gating'[124]).

While DL methods have become a standard state-of-the-art approach for many medical image analysis tasks, they largely remain black-box methods where the end user has limited meaningful ways of interpreting model predictions. This feature of DL methods is a significant hurdle in the deployment of DL enabled applications to safety-critical domains such as medical image analysis. We want models to be highly accurate and robust, but also explainable and interpretable. This interpretability is vital to mitigate human uncertainty and foster trust in using automated predictions in downstream tasks with real-world consequences.

Recent EU law[1] has led to the 'right for explanation', whereby any subject has the right to have automated decisions that have been made about them explained. This even further highlights the need for transparent algorithms which we can reason about [125, 126, 127].

It is important for users to understand how a certain decision has been made by the model, as even the most accurate and robust models aren't infallible, and false or uncertain predictions must be identified so that trust in the model can be fostered and predictions are appropriately weighted in the clinical decision making process. It is vital the end user, regulators and auditors all have the ability to contextualise automated decisions produced by DL models. Here we outline some different methods for providing interpretable ways of reasoning about DL models and their predictions.

Typically DL methods can provide statistical metrics on the uncertainty of a model output, many of the uncertainty measures discussed in Section 2.2.1 are also used to aid in intepretability. While uncertainty measures are important, these are not sufficient to foster complete trust in DL model, the model should provide human-understandable justifications for its output that allow insights to be drawn elucidating the inner workings of a model. [34] discuss many of the core concerns surrounding model intepretability and highlight various works that have demonstrated sophisticated methods of making a DL model interpretable across the DL field. Here we evaluate some of the works that have been applied to medical image segmentation and refer the reader to [128, 129] for further reading on interpretability in the rest of the medical imaging domain.

[124] and [130] introduce 'Attention Gating' to guide networks towards giving more 'attention' to certain image areas, in a visually interpretable way - potentially aiding in the subsequent refinement of annotations. Attention Gates are introduced into the popular U-Net architecture

[1] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1

[18], where information extracted from coarse scale layers is used in gating to disambiguate irrelevant and noisy responses in skip connections, prior to concatenation, to merge only relevant layer activations. This approach eliminates the need for applying external object localisation models in image segmentation and regression tasks. Coefficients of Attention Gate layers indicate where in an image feature activations will be allowed to propagate through to final predictions, providing users with a visual representation of the areas of an image that a model has weighted highly in making predictions.

[131] propose the application of RL to ultrasound care, guiding a potentially inexperienced user to the correct sonic window and enabling them to obtain clinically relevant images of the anatomy of interest. This human-in-the-loop application is an example of the novel applications possible when combining DL/RL with real-time systems enabling users to respond to model feedback to acquire the most accurate information available.

[132] propose using test-time augmentation to acquire a measure of aleatoric (image-based) uncertainty and compare their method with epistemic (model) uncertainty measures and show that their method provides a better uncertainty estimation than a test-time dropout based model uncertainty alone and reduces overconfident incorrect predictions.

[133] evaluate several different voxel-wise uncertainty estimation methods applied to medical image segmentation with respect to their reliability and limitations and show that current uncertainty estimation methods perform similarly. Their results show that while uncertainty estimates may be well calibrated at the dataset level (capturing epistemic uncertainty), they tend to be mis-calibrated at a subject-level (aleatoric uncertainty). This compromises the reliability of these uncertainty estimates and highlights the need to develop subject-wise uncertainty estimates. They show auxiliary networks to be a valid alternative to common uncertainty methods as they can be applied to any previously trained segmentation model.

Developing transparent systems will enable faster uptake in clinical practice and including humans within the DL clinical pipelines will ease the period of transition between current best practices and the breadth of possible enhancements that DL has to offer.

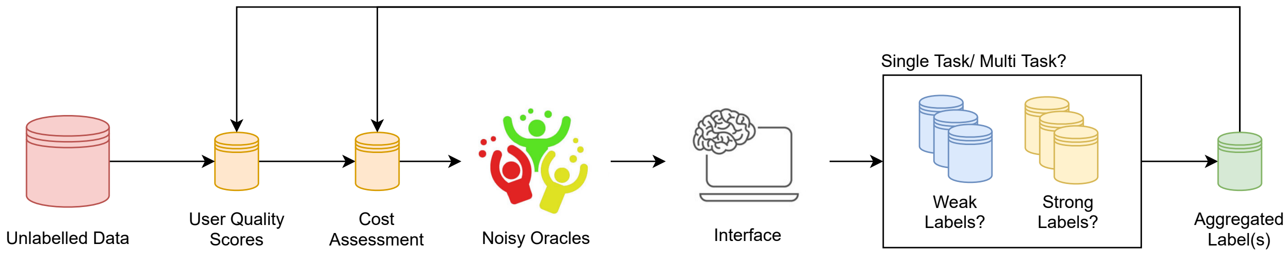


Figure 2.12: Overview of practical considerations

We suggest that ongoing work in improving interpretability of DL models will also have a positive impact on AL, as the majority of methods to improve intepretability are centred on providing uncertainty measures for a models prediction, these same uncertainty measures can be used for AL selection strategies in place of existing uncertainty measures that are currently employed. As intepretability and uncertainty measures improve we expect to see a similar improvement of AL frameworks as they incorporate the most promising uncertainty measures.

The methods discussed in Section 2.2.2 remain open areas of research interest with great implications for the progress of AL development and greater uptake of DL and HITL methods in clinical practice. The study of interaction between users and models is of growing importance and is having a significant impact on the efficacy of Deep Active Learning systems and their deployment to real-world applications, especially in clinical scenarios [134, 135]. The wider study of interpretability and the study of Human Computer Interaction may seem distinct and diverging, however we expect to see these two research fields converge through Active Learning as the feedback loop between human users and machine models becomes of increasing importance.

2.2.3 Practical Considerations

We have so far discussed the core body of work behind AL, model interpretation and prediction refinement, and while the works discussed above go a long way in covering the majority of research being done, there are several practical considerations for developing and deploying DL enabled applications that we must consider. In this section we outline the main practical research areas that are impacting DL enabled application development pipelines (as shown in Figure 2.12) and suggest where we might look next.

Noisy Oracles

Gold-standard annotations for medical image data are acquired by aggregating annotations from multiple expert oracles, but as previously discussed, this is rarely feasible to obtain for large complex datasets due to the expertise required to perform such annotations. Here we ask what effect on performance we might incur if we acquire labels from oracles without domain expertise, and what techniques can we use to mitigate the suspected degradation of annotation quality when using non-expert oracles, to avoid any potential loss in accuracy.

[136] propose active learning method that assume data will be annotated by a crowd of non-expert or 'weak' annotators, and offer approaches to mitigate the introduction of bad labels into the data set. They simultaneously learn about the quality of individual annotators so that the most informative examples can be labelled by the strongest annotators.

[123] propose methods for crowd-sourced learning in two scenarios. Firstly, they aim at inferring instances ground truth given the crowd's annotations by modelling the crowd's expertise and label correlations from two different perspectives: firstly they model expertise based on individual labels, based on the idea that labeller's annotations for similar instances should be similar, and secondly through modelling the crowd's expertise to distinguish the relevance between label pairs. They extend their approach to the active paradigm and offer criteria for instance, label and labeller selected in tandem to minimise annotation cost.

[137] explore using Amazon's MTurk to gather annotations of airways in CT images. Results showed that the novice oracles were able to interpret the images, but that instructions provided were too complex, leading to many unusable annotations. Once the bad annotations were removed, the annotations did show medium to high correlation with expert annotations, especially if annotations were aggregated.

[33] describe an approach to assess the reliability of annotators in a crowd, and a crowd layer used to train deep models from noisy labels from multiple annotators, internally capturing the reliability and biases of different annotators to achieve state-of-the-art results for several crowd-sourced data-set tasks.

We can see that by using a learned model of oracle annotation quality we can mitigate the effects of low quality annotations and present the most challenging cases to most capable oracles. By providing clear instructions we can lower the barriers for non-expert oracles to perform accurate annotation, but this is not generalisable and would be required for every new annotation task we wish to perform.

Weakly Supervised Learning

Most segmentation tasks require pixel-wise annotations, but these are not the only type of annotation we can give an image. Segmentation can be performed with 'weak' annotations, which include image level labels e.g. modality, organs present etc. and annotations such as bounding boxes, ellipses or scribbles. It is argued that using 'weaker' annotation formulations can make the task easier for the human oracle, leading to more accurate annotations. 'Weak' annotations have been shown to perform well in several segmentation tasks, [138] demonstrate obtaining pixel-wise segmentations given a data-set of images with 'weak' bounding box annotations. They propose DeepCut, an architecture that combines a CNN with an iterative dense CRF formulation to achieve good accuracy while greatly reducing annotation effort required. In a later study, [139] examine the impact of expertise required for different 'weak' annotation types on the accuracy of liver segmentations. The results showed a decrease in accuracy with less expertise, as expected, across all annotation types. Despite this, segmentation accuracy was comparable to state-of-the-art performance when using a weakly labelled atlas for outlier correction. The robust performance of their approach suggests 'weak' annotations from non-expert crowds could be used to obtain accurate segmentations on many different tasks, however their use of an atlas makes this approach less generalisable than is desired.

In [140] they examine using super pixels to accelerate the annotation process. This approach uses a pre-processing step to acquire a super-pixel segmentation of each image, non-experts are then used to perform the annotation by selecting which super-pixels are part of the target region. Results showed that the approach largely reduces the annotation load on users. Non-expert annotation of 5000 slices was completed in under an hour by 12 annotators, compared to an expert taking three working days to establish the same with an advanced interface. The non-expert interface is web-based demonstrating the potential of distributed annotation collection/crowd-sourcing. An encouraging aspect of this work is that the results showed high performance on the segmentation task in question compared with expert annotation performance, but may not be suitable for all medical image analysis tasks.

It has been shown that we can develop high performing models using weakly annotated data, and as weak annotations requires less expertise to perform, they can be acquired faster and from a non-expert crowd with a smaller loss in accuracy than gold-standard annotations. This is very promising for future research as datasets of weakly annotated data might be much easier and more cost-effective to curate.

Multi-task learning

Many works aim to train models or acquire training data for several tasks at once, it is argued that this can save on cost as complementary information may result in higher performance over multiple different tasks [141]. [142] propose a dual network for joint segmentation and detection task for lung nodule segmentation and cochlea segmentation from CT images, where only a part of the data is densely annotated and the rest is weakly labelled by bounding boxes, using this they show that their architecture out-performs several baselines. At present this work only handles the case for two different label types but they propose extending the framework for a true multi-task scenario.

This is a promising area but, as of yet, it has not been incorporated into an active learning setting. As such, it may be elucidating to analyse the differences in samples chosen by different AL methods when the model is being training for multiple tasks simultaneously. However, [143] raise concerns over the transferability of actively acquired datasets to future models due to the inherent coupling between active learning selection strategies and the model being trained, and show that training a successor model on the actively acquired dataset can often result in worse performance than from random sampling. They suggest that, as datasets begin to outlive the models trained on them, there is a concern for the efficacy of active learning, since the acquired dataset may be disadvantageous for training subsequent models. An exploration of how actively acquired datasets perform on multiple models may be required to explain the effects of an actively acquired dataset coupled with one model on the performance of related models.

Annotation Interface

So far the majority of Human-in-the-loop methods assume a significant level of interaction from an oracle to annotate data and model predictions, but few consider the nature of the interface with which an oracle might interact with these images. The nature of medical images require special attention when proposing distributed online platforms to perform such annotations. While the majority of techniques discussed so far have used pre-existing data labels in place of newly acquired labels to demonstrate their performance, it is important to consider the effects of accuracy of annotation that the actual interface might incur.

[144] propose a framework for the online classification of Whole-slide images (WSIs) of tissues. Their interface enables users to rapidly build classifiers using an active learning process that minimises labelling efforts and demonstrates the effectiveness of their solution for the quantification of glioma brain tumours.

[145] propose a novel interface for the segmentation of images that tracks the users gaze to initiate seed points for the segmentation of the object of interest as the only means of interaction with the image, achieving high segmentation performance. [146] extend this idea and compare using eye tracking generated training samples to traditional hand annotated training samples for training a DL model. They show that almost equivalent performance was achieved using annotation generated through eye tracking, and suggest that this approach might be applicable to rapidly generate training data. They acknowledge that there is still improvements to be made integrate eye tracking into typical clinical radiology work flow in a faster, more natural and less distracting way.

[147] evaluate the player motivations behind EyeWire, an online game that asks a crowd of

players to help segment neurons in a mouse brain. The gamification of this task has seen over 500,000 players sign up and the segmentations acquired have gone onto be used in several research works [148]. One of the most exciting things about gamification is that when surveyed, users were motivated most by making a scientific contribution rather than any potential monetary reward. However this is very specialised towards this particular task and would be difficult to apply across other types of medical image analysis tasks.

There are many different approaches to developing annotation interfaces and the ones we consider above are just a few that have been applied to medical image analysis. As development increases we expect to see more online tools being used for medical image analysis and the chosen format of the interface will play a large part in the usability and overall success of these applications.

Variable Learning Costs

When acquiring training data from various types of oracle it is worth considering the relative cost associated with querying a particular oracle type for that annotation. We may wish to acquire more accurate labels from an expert oracle, but this is likely more expensive to obtain than from a non-expert oracle. The trade off, of course, being accuracy of the obtained label - less expertise of the oracle will likely result in a lower quality of annotation. Several methods have been proposed to model this and allow developers to trade off between cost and overall accuracy of acquired annotations.

[149] propose a cost-sensitive active learning approach for intracranial haemorrhage detection. Since annotation time may vary significantly across examples, they model the annotation time and optimize the return on investment. They show their approach selects a diverse and meaningful set of samples to be annotated, relative to a uniform cost model, which mostly selects samples with massive bleeds which are time consuming to annotate.

[150] propose a budget based cost minimisation framework in a mixed-supervision setting (strong and weak annotations) via dense segmentation, bounding boxes, and landmarks. Their

framework uses an uncertainty and a representativeness ranking strategy to select samples to be annotated next. They demonstrate state-of-the-art performance at a significantly reduced training budget, highlighting the important role of choice of annotation type on the costs of acquiring training data.

The above works each show an improved consideration for the economic burden that is incurred when curating training data. A valuable research direction would be to assess the effects of oracle expertise level, annotation type and image annotation cost in a unified framework as these three factors are very much linked and may have a profound influence over each other.

2.2.4 Future Prospective and Unanswered Questions

In Sections 2.2.1 & 2.2.2 we discuss methods through which a user might gather training data to build a model, use their model to predict on new data and receive feedback to iteratively refine the model output towards a more accurate result. Each of these techniques assume some human end user will be present to interact with the system at the point of initial annotation, interpretation and refinement. Each of these areas seeks to achieve a shared goal of achieving the highest performing model from as little annotated data as possible - with a means to weigh conclusions of models predictions appropriately.

AL is not the only area of research that aim to learn from limited data. Semi-supervised learning, and Transfer Learning both make significant contributions to extracting the most value from limited labelled data.

In the presence of large data-sets, but the absence of labels, unsupervised and semi-supervised approaches offer a means by which information can be extracted without requiring labels for all the data-points. This could potentially have a massive impact on the medical image analysis field where this is often the case.

In a semi-supervised learning (SSL) scenario we may have some labelled data, but this is often very limited. We do however have a large set of un-annotated instances (much like in active learning) to draw information from, the goal being to improve a model (trained only on the labelled instances) using the un-labelled instances. From this we derive two distinct goals: a) predicting labels for future data (inductive SSL) and b) predicting labels for the available un-annotated data (transductive SSL) [151]. SSL methods provide a powerful way of extracting useful information from un-annotated image data and we believe that progress in this area will be beneficial to AL systems that desire a more accurate model for initialisation to guide data selection strategies.

Transfer Learning (TL) is a branch of DL that aim to use pre-trained networks as a starting point for new applications. Given a pre-trained network trained for a particular task, it has been shown that this network can be 'fine-tuned' towards a target task from limited training data. [152, 153, 151] provide a more general overview of transfer learning in medical imaging, and focus on the use of TL in AL scenarios in the following. [112] demonstrated the applicability of TL for a variety of medical image analysis tasks, and show, despite the large differences between natural images and medical images, CNNs pre-trained on natural images and fine-tuned on medical images can perform better than medical CNNs trained from scratch. This performance boost was greater where fewer target task training examples were available. Many of the methods discussed so far start with a network pre-trained on natural image data.

[154] propose AFT*, a platform that combines AL and TL to reduce annotation efforts, which aims at solving several problems within AL. AFT* starts with a completely empty labelled data-set, requiring no seed samples. A pre-trained CNN is used to seek 'worthy' samples for annotation and to gradually enhance the CNN via continuous fine-tuning. A number of steps are taken to minimise the risk of catastrophic forgetting. Their previous work [113] applies a similar but less featureful approach to several medical image analysis tasks to demonstrate equivalent performance can be reached with a heavily reduced training data-set. They then use these tasks to evaluate several patterns of prediction that the network exhibits and how these relate to the choice of AL selection criteria.

[154] have gone onto to use their AFT framework for annotation of CIMT videos, a clinical technique for characterisation of Cardiovascular disease. Their extension into the video domain presents its own unique challenges and thus they propose a new concept of an Annotation Unit

- reducing annotating a CIMT video to just 6 user mouse clicks, and by combining this with their AFT framework reduce annotation cost by 80% relative to training from scratch and by 50% relative to random selection of new samples to be annotated (and used for fine-tuning).

[155] use TL for supervised domain adaptation for sub-cortical brain structure segmentation with minimal user interaction. They significantly reduce the number of training images from different MRI imaging domains by leveraging a pre-trained network and improve training speed by reducing the number of trainable parameters in the CNN. They show their method achieves similar results to their baseline while using a remarkably small amount of images from the target domain and show that using even one image from the target domain was enough to outperform their baseline.

The above methods and more discussed in this review demonstrate the applicability of TL to reducing the number of annotated sample required to train a model on a new task from limited training data. By using pre-trained networks trained on annotated natural image data (there is an abundance) we can boost model performance and further reduce the annotation effort required to achieve state-of-the-art performance.

A related sub-field of TL worth exploring is domain adaptation (DA). Many DL techniques used in medical image analysis suffer from the domain shift problem caused by different distributions between source data and target data, often due to medical images being acquired on a variety of different scanners, scanning parameters and subject cohorts etc. DA has been proposed as a special type of transfer learning in which the domain feature space and tasks remain the same while marginal distributions of the source and target domains are different. We refer the reader to [156, 157] for an overview of DA methods used for medical image analysis, and hope to see greater application of DA methods in AL scenarios in the future.

In many of scenarios described, models continuously receive new annotations to be used for training, and in theory we could continue to retrain or fine-tune a model indefinitely, but is this practical and cost effective? It is important to quantify the long term effects of training a model with new data to assess how the model changes over time and whether or not performance has improved, or worse, declined. Learning from continuous streams of data has proven more difficult than anticipated, often resulting in ’catastrophic forgetting’ or ’interference’ [158]. We face the *stability-plasticity-dilemma*. Avoiding catastrophic forgetting in neural networks when learning from continuous streams of data can be broadly divided among three conceptual strategies: a) Retraining the the whole network while regularising (to prevent forgetting of previously learned tasks). b) selectively train the network and expand it if needed to represent new tasks, and c) retaining previous experience to use memory replay to learn in the absence of new input [158].

[159] investigate continual learning of two MRI segmentation tasks with neural networks for countering catastrophic forgetting of the first task when a new one is learned. They investigate elastic weight consolidation, a method based on Fisher information to sequentially learn segmentation of normal brain structures and then segmentation of white matter lesions and demonstrate this method reduces catastrophic forgetting, but acknowledge there is a large room for improvement for the challenging setting of continual learning.

It is important to quantify the performance and robustness of a model at every stage of its lifespan. One way to consider stopping could evaluate when the cost of continued training outweighs the cost of errors made by the current model. An existing measure that attempts to quantify the economical value of medical intervention is the Quality-adjusted Life year (QALY), where one QALY equates to one year of healthy life [160]. Could this metric be incorporated into models? At present we cannot quantify the cost of errors made by DL medical imaging applications but doing so could lead to a deeper understanding of how accurate a DL model really ought to be.

As models are trained on more of the end user’s own data, will this cause the network to perform better on data from that user’s system despite performing worse on data the model was initially trained on? Catastrophic forgetting suggests this will be the case, but is this a bad thing? It may be beneficial for models to gradually bias themselves towards high performance for the end user’s own data, even if this results in the model becoming less transferable to other data. [161] explore the role of bias in AL methods. Bias is introduced because the training data no longer follows the population distribution in AL. The authors providing a general method by which unbiased AL estimators may be constructed using novel corrective weights to remove bias. Further to this, an explanation of the empirical successes of existing AL methods which ignore this bias is provided. It is shown that bias introduced by AL methods can be actively helpful when training over-parameterized models like neural networks with relatively little data. This further motivates future work to better understand when the bias introduced by AL could have a positive influence on the performance of AL methods, to the detriment of generalisability to other data sources.

Active learning assumes the presence of a user interface to perform annotations but is only concerned with which data to annotate. Refinement assumes we can generate an annotation through iterative interaction with the current model prediction. Hence, it would be desirable to combine these two in future work. If we can train a model with a tiny amount of training data, and then ask annotators to refine model predictions towards a more accurate label, we can expedite the annotation process by reducing the initial annotation workload and reduce additional interface work for use with unseen data. This would be the same interface used to create the training annotations. By combining the efforts of active learning and iterative refinement into a unified framework we can rapidly produce annotations to train our model, as well as acquiring high quality results from our models from the beginning. This should also have the added side effect of training the model on data from the same distribution that it will be predicting on, reducing domain shift effects in unseen distributions.

By incorporating our end user at each stage of the model life cycle we could also use human feedback on model performance to add a more 'human interpretable' metric of model confidence as each user could rank the performance of the model for each input as it sees it, potentially giving a metric of confidence based on human interpretation of the model output. This of course requires experts to be using the system. One might argue that the models initial predictions may impart some influence over the human user but by crowd-sourcing the initial annotations to a less expert multi-label crowd we could reduce this bias.

Developments in uncertainty quantification will benefit both AL selection heuristics and interpretation of model outputs, but there is no guarantee that the best performing uncertainty metrics for selecting new samples to be annotated will be the same metrics that are the most interpretable to a human user.

There is significant overlap of research goals for many areas of human-in-the-loop computing but there are large gaps that need to be filled in order to understand the relationships between different methods and how these might affect their performance.

As the many areas of DL research converge towards shared goals of working with limited training data to achieve state-of-the-art results, we expect to see more systems emerge that exploit the advances made in the range of sub-fields of ML described here. We have already seen the combination of several methods into individual frameworks but as of yet no works combine all of the approaches discussed into a single framework. As different combinations of approaches begin to appear it is important to consider the measure by which we assess their performance, as isolating individual developments becomes more difficult. Developing baseline human-in-the-loop methods to compare to will be vital to assess the contributions of individual works in each area and to better understand the influences of competing improvements in these areas.

We have explored the large body of emerging medical image analysis work in which a human end user is at the centre. DL has all the ingredients to induce a paradigm shift in our approach to a plethora of clinical tasks. The direct involvement of humans is set to play a core role in this shift. The works presented each offer their own approaches to including humans in the loop and we suggest that there is sufficient overlap in many methods for them to be considered under the same title of Human-in-the-Loop computing. We hope to see new methodologies emerge that combine the strengths of AL and HITL computing into end-to-end systems for the development of DL applications that can be used in clinical practice. While there are some practical limitations as discussed, there are many proposed solutions to such issues and as research in these directions continues it is only a matter of time before DL applications blossom into fully-fledged, accurate and robust systems to be used for daily routine tasks. We are in an exciting era for medical image analysis, with endless opportunity to innovate and improve the current state-of-the-art and to leverage the powers of DL to make a real impact in health care across the board. With diligent research and development we should see more and more applications boosted by DL capabilities finding their way onto the market, allowing users to achieve better results, faster, and with less expertise than before, freeing up expert time to be used on the most challenging cases. The field of Human-in-the-loop computing will play a crucial role to achieve this.

Chapter 3

Confidence Challenge

In this chapter we examine the 'Confidence Challenge' and develop methods by which confidence can be extracted from DL segmentation models' predictions in a generic and widely applicable way. We develop methods for extracting prediction confidence in both qualitative and quantitative forms and show how to communicate both of these types of confidence to guide users towards improved understanding of model predictions, and re-acquisition of images to obtain more accurate results.

We employ our methods in the domain of fetal ultrasound imaging and biometric extraction as this enables us to evaluate both the quality of automated segmentations visually and the quality of automatically extracted biometrics from those segmentations numerically. The real-time nature of ultrasound imaging is an ideal test case for the downstream uses of confidence measures where prediction confidence can be presented in real-time and re-acquisition of input images can be performed immediately.

Manual estimation of fetal Head Circumference, Abdominal Circumference and Femur Length from Ultrasound (US) is a key biometric for monitoring the healthy development of fetuses. Unfortunately, such measurements are subject to large inter-observer variability, resulting in low early-detection rates of fetal abnormalities. To address this issue, we propose a novel probabilistic DL approach for real-time automated estimation of fetal HC. This system feeds back statistics on measurement robustness to inform users how confident a deep neural network is in evaluating suitable views acquired during free-hand ultrasound examination. In real-time scenarios, this approach may be exploited to guide operators to scan planes that are as close as possible to the underlying distribution of training images, for the purpose of improving inter-operator consistency. We demonstrate our framework on three biometric estimation tasks during fetal screening with free-hand ultrasound imaging: Head Circumference (HC), Abdominal Circumference (AC) and Femur Length (FL). We analyse the impact of measurement robustness on fetal weight estimation and show that our framework outperforms state-of-the-art approaches, while providing robust uncertainty estimates to be carried forward into clinical decision making.

3.1 Introduction

Recently, automatic US scanning approaches have been developed using DL [162], which mitigate the problems of manual US measurement through automatic detection of diagnostically relevant anatomical planes. Such systems have allowed development of robust automated methods for estimation of anatomical biometrics [163, 164] in diverse acquisition conditions with various imaging artefacts, outperforming non-DL approaches [165, 166, 167]. Not only is segmentation of anatomical structures a valuable tool for many clinical tasks, the resulting masks are often key to extract physiologically important measurements, such as lung volume [46], fetal weight [168], or tumor extent [169].

Critically, such methods only provide point estimates of HC without confidence or uncertainty measures, and do not provide any means to evaluate the quality of individual measurements during real-time scans. This can lead to many, potentially contradicting, measurements without any means to control the trustworthiness of the predictions during examination or retrospectively.

The most accurate gold-standard for medical image segmentation is to acquire several pixel-wise annotations for each patient from a group of experts and to find a consensus between them. This is often impractical and resource intensive. Thus, usually only a single segmentation is acquired for a given image from a single expert. Acquiring a single segmentation is still time-consuming and requires significant expertise. Segmentation of anatomical structures and deriving biomarkers is inherently ambiguous. Structural boundaries are not always clearly visible and different experts have different styles of delineation. This can result in high inter- and intra-observer variability, which in turn can lead to inconsistent decision making and challenges for longitudinal population-based studies [170].

Automated methods can suffer from performance degradation due to distribution shift, out-of-distribution inputs, adversarial attacks etc. [171]. Key to facilitating consistent biometric measurements across clinical sites is the ability of automated methods to effectively communicate failure modes.

In order for the clinical practice to benefit from the value provided by automated DL segmentation methods, further analysis must be done when evaluating DL methodologies to provide interpretable, quantifiable meta-information for both model-level (epistemic) uncertainty and subject-level (aleatoric) uncertainty [132, 133].

We present a framework for evaluating the ability of automated segmentation methods to A) infer varying plausible solutions for a single input image with stochastic segmentation simulating the inter/intra-observer variability seen in manual observations, B) extract meaningful/well-calibrated uncertainty estimates from a set of plausible solutions, C) improve overall model performance through automated solution acceptance/rejection criteria and D) provide uncertainty estimates such that they can be propagated to downstream analysis tasks.

Anomaly detection and assessment of fetal development from ultrasound scans are required to ensure that the best care is given at the earliest identifiable stage. In many countries a mid-trimester ultrasound scan is carried out between 18-22 weeks of gestation as part of standard prenatal care. ‘Standardized plane’ views are acquired in a primary care setting to measure distinct anatomical features [172] and compare them to known large scale population statistics. Measurements of the head circumference, abdominal circumference and femur length are most commonly used to predict fetal weight. Biometric measurements acquired longitudinally can be used to predict the fetal development trajectory. Unfortunately, rates for early detection of fetal abnormalities can be low due to large intra- and inter-observer variability and regional differences in operator skills [173].

We demonstrate our evaluation framework with current state-of-the-art methods for stochastic segmentation for three fetal screening tasks: Head Circumference (HC), Abdominal Circumference (AC) and Femur Length (FL) estimation. Further, we evaluate a common downstream task: fetal weight estimation that uses HC, AC and FL measurements to predict the weight of a fetus in-utero.

3.2 Related work

3.2.1 Stochastic Segmentation

Several approaches have been proposed for the prediction of varying plausible solutions. These include Monte-Carlo Dropout (MC Dropout). Weights in a deep neural network are 'dropped' randomly during inference with a given probability p which has been shown to approximate Bayesian inference in deep Gaussian processes [94]. Introducing dropout to a network during training has been shown to be an effective regulariser for many DL architectures, reducing over-fitting and boosting performance [94]. Typically dropout is not used during inference to allow the full capacity of the network to make predictions for a given input. However, if dropout is enabled during inference, taking multiple samples from a network for a given input has been shown to simulate an ensemble of different model weights, *i.e.*, as network nodes are randomly dropped each sample is predicted by a different path through the network, all capable of making plausible predictions. Through dropout during inference we can translate almost any deterministic deep neural network architecture into a probabilistic method, with minimal implementation overhead.

It is typical to use a dropout probability of $p = 0.5$ before each layer of a network to induce the largest variety of network configurations during training and inference. Concrete dropout has been shown as an effective method for automatically tuning network dropout probabilities but this has shown to be less effective for convolutional layers and in our experiments each layer converged to a dropout probability $p = 0.5$ indicating a sensible parameter choice for our baseline approach [93].

As an alternative, ensemble approaches produce N prediction samples per input image by training a set of N separate networks for the same task. The results are then combined to produce a final segmentation which seems to offer a good trade-off between robustness and accuracy [65]. However, the computational overhead of several large models is often infeasible for real-time applications.

Several bespoke stochastic segmentation techniques have been developed. The Probabilistic U-Net network architectures represents a generative segmentation model based on a combination of a U-Net-like module with a conditional variational autoencoder. This produces an unlimited number of plausible solutions, reproducing the possible segmentation variants as well as the frequencies with which they occur [174].

The PHiSeg method extends on this approach to model the conditional probability distribution of the segmentations given an input image [175], across several different scales enabling modelling of variation from high-level to low-level structures. The PHiSeg network architecture has been shown to offer a good trade off between segmentation accuracy on binary and non-binary segmentation tasks, and varying plausible predictions for each input[175]. Inspired by the Probabilistic U-Net [174], PHiSeg addresses several problems associated with the Probabilistic U-Net and extends the approach to generate probabilistic outputs at multiple latent and resolution levels, inducing natural variation in predictions of high-level structure and finer low level details. This architecture has demonstrated state-of-the-art performance for prostate segmentation.

Stochastic Segmentation Networks (SSNs) have been developed to model spatially correlated aleatoric uncertainty. In contrast to methods which produce pixel-wise estimates, SSNs model the joint distribution over entire logit maps to generate multiple spatially coherent hypotheses for a single image. A low-rank multivariate normal distribution over the logit space is used to model the probability of a label map given an image to obtain a spatially consistent probability distribution which can be computed efficiently by a neural network without modification to the underlying model architecture [169].

3.3 Contributions

In this chapter, we extend upon a state-of-the-art convolutional DL approach for automatic fetal HC measurement [163] to develop a new approach for automated probabilistic fetal HC with real-time feedback on measurement robustness. Several probabilistic DL methods are evaluated: MC Dropout during inference, Probabilistic U-Net, PhiSeg and SSNs. These are used to return an ensemble of segmentations, from which upper and lower bounds on the measurement are generated. In addition, we propose the derivation of a 'variance score', used to reject acquired images that produce sub-optimal biometric measurements. In this way, the system will guide operators towards acquiring optimal US views, resulting in more consistent and accurate measurements.

Our method is highly modular and extensible, enabling many different methods to be interchanged without compromising the overall approach. We demonstrate different possible components of the framework e.g. segmentation method, metric extraction method, measurement fusion protocol, uncertainty quantification technique and measurement acceptance criteria.

3.4 Method

Our approach is outlined in Figures 3.1 and 3.2, where Figure 3.2 outlines the general stages of our pipeline and Figure 3.1 gives a more detailed example for HC estimation. Our pipeline comprises six generic, interchangeable components. At the highest level these components are

1. A probabilistic/stochastic segmentation method

2. A measurement extraction component

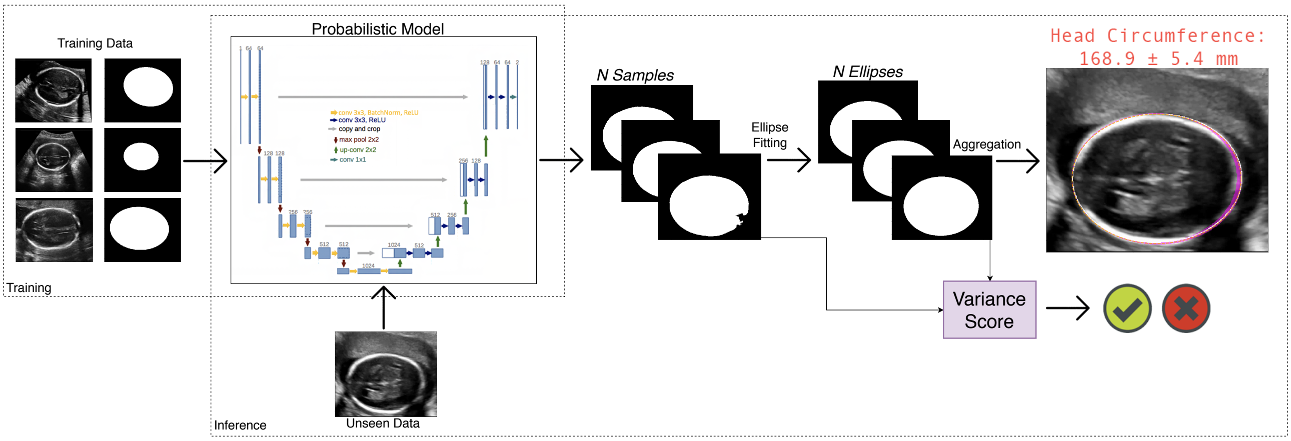


Figure 3.1: Overview of our proposed method. We train a probabilistic model using the available training data. During inference we take N samples from our model, fit ellipses to each sample and aggregate these ellipses to extract a HC value and an upper and lower bound on that HC value. Various outputs of the pipeline are used to calculate different variance scores given a set of N samples. As a proof of concept we extract a threshold such that test cases whose variance score is outside the threshold are rejected, and inside are accepted.

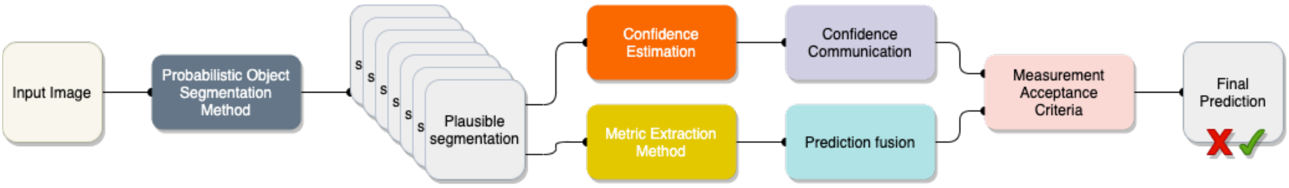


Figure 3.2: Graphical overview of information flow through each component of our proposed framework

3. A segmentation level fusion method and measurement level fusion method

4. A confidence estimation component (image based qualitative representation and measurement based quantitative representation)

5. A confidence communication approach (qualitative and quantitative)

6. A measurement acceptance criteria

3.4.1 Probabilistic segmentation

Our method can incorporate any generic image segmentation model. We place one constraint on choosing a segmentation model and that is the capability to produce stochastic segmentations. Intuitively a probabilistic segmentation method should simulate the inter- and intra-observer variability present when manually segmenting the chosen input object. Here we consider convolutional neural network (CNN) based methods, but any method that meets the above constraint can be used. Given the inherent variability between sonographers' annotations in the training data, we generate a set of N plausible segmentations from a single input using the following methods:

i) ***MC Dropout U-Net***: We train a U-Net architecture using dropout [18]. We randomly drop weights of the network with probability p to predict N segmentation samples. Here, single-sample experiments ($N = 1$) were used to optimise the configuration of the network. This led to implementation of a single dropout layer ($p = 0.6$) before the bottleneck layer of the U-Net during inference.

ii) ***Probabilistic U-Net***: We sample a set of N plausible segmentations using this method [174] where we follow the same training scheme as [174].

iii) ***PhiSeg***: We sample a set of N plausible segmentations using the PhiSeg method which addresses several concerns raised over Probabilistic U-Net, following the training scheme outlined in [175].

iv) ***Stochastic Segmentation Networks***: We sample a single segmentation from each image, from which we obtain the logit map upon which a distribution is fitted. We then sample N varying segmentations from the logit map distribution as in [169].

3.4.2 Biometric estimation

We are concerned with estimating HC, AC and FL from fetal US images and as such present two biometric extraction methods for these tasks: circumference estimation for both HC and AC, and an object length estimation method for FL.

To extract a measurement from an image we first select the largest connected object in the image, and fill in any holes in that object, then pass it on to each of the relevant measurement extraction methods described below.

Object Circumference Estimation: After segmentation of the head or abdomen, an ellipse is fitted to the segmented contours [176] from which the ellipse parameters can be obtained in mm. We extract ellipse centroid co-ordinates (c_x and c_y), major and minor axis radii (a and b) each in pixels, and the angle of rotation (α) and estimate the circumference C using the Ramanujan approximation II [177] as

$$C = \pi(a+b)(1+\frac{3h}{10+\sqrt{4-3h}})s_{xy} \tag{3.1}$$

where

$$h = \frac{(a-b)^2}{(a+b)^2} \tag{3.2}$$

and s_{xy} is the pixel size of the image in mm. The error of this approximation is $O(h^{10})$ which for more circular ellipses is negligible. This ellipse fitting process mimics a sonographer's manual actions when extracting a C measurement during fetal US screening.

Object Length Estimation: After segmentation of the femur, we find the maximum pixel-wise Euclidean distance between any two pixels that have both been classified as part of that object, and multiply this the pixel-wise distance by the pixel spacing to obtain a measurement in millimetres:

$$L = s_{xy}\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}, \tag{3.3}$$

where (x_1, y_1) and (x_2, y_2) are the co-ordinates of the end points of the major axis of segmented object, and s_{xy} is the pixel size of the image in mm.

3.4.3 Segmentation Fusion and 'Extremes' Generation

A typical approach for fusing several segmentations or pixel-wise labels is to use majority voting, where each pixel's final value is determined by which value the majority of segmentations predicts, where each pixel is assigned the objects value if $\geq 50\%$ of predictions assign the pixel that class value, *i.e* by taking the modal value for each pixel for non-binary segmentation tasks - we refer to this as the 'Segmentation Mode'.

To extract reasonable bounds for a single solution, we extract what we call 'Extremes'. We can generate the most extreme bounds by taking the intersection and union of our plausible solutions to form our lower and upper bound segmentations respectively. We argue that the region for which all segmentations agree (intersection) forms the lower bound and the pixels for which at least one segmentation predicts as part our object of interest forms the upper bound segmentation. From these two additional generated segmentations we can extract our upper and lower bound measurement values to be used downstream.

3.4.4 Measurement fusion

We extract measurements from each segmentation to generate a list of possible measurement values.

Given our list of possible measurements we compare different approaches to deciding a final measurement value:

- **Segmentation mode**: measurement extracted from 'segmentation mean' image.
- **Median**: Median of extracted measurements
- **All Mean**: Mean of extracted measurements (including 'extremes' measurements).
- **Sample Mean**: Mean of extracted measurements (not including 'extremes' measurements).
- **Mid-point**: We can take the mid-point or weighted mid-point between a) our generated segmentation upper and lower bound measurements and b) the upper and lower quartile measurements of all measurements.

3.4.5 Variance Estimation and Measurement acceptance criteria

With a probabilistic mapping function $g_P(X) = \hat{X}_i$, in our case a deep probabilistic neural network, we can map a continuous input image to a possible segmentation mask $\hat{X}_i$. We assume a deterministic function $f(\hat{X}_i) = [a, b, \theta, x_c, y_c]^T$, with semi-major axis length a, semi-minor axis length b, angle of orientation θ and center $C(x_c, y_c)$, which provides a least square solution to the ellipse fitting problem to the set of points $\hat{X}$ as proposed by [178]. Based on $f(\hat{X}_i)$ we can evaluate hypotheses for their suitability to act as a metric to measure robustness during inference given N prediction samples from $g_P(X)$. These proposed metrics are

h1) Ellipse parameter variance: $\sum_i^5(\mathrm{Var}(f(\hat{X}_n)_i))$;

h2) 'Extremes' Range width: $M(f(\bigcup_{i=1}^N \hat{X}_i)) - M(f(\bigcap_{i=1}^N \hat{X}_i))$ where $M(x)$ is our measurement from segmentation function;

h3) Range width: $max(M(\hat{X}_i)) - min(M(\hat{X}_i))$;

h4) Total ring area: $\sum(f(\bigcup_{i=1}^N \hat{X}_i) - f(\bigcap_{i=1}^N \hat{X}_i)) \cdot s_{xyz}$, where s_{xyz} scales $\hat{X}_i$ to world space in mm;

h5) Mask classification entropy: $\sum_{x,y}^K \underline{\hat{X}}(x,y)\log(\underline{\hat{X}}(x,y))$, where K is the number of pixels in $\underline{\hat{X}} \in \mathbb{Z}_2$ after $argmax(\hat{X}_i)$ class assignment and $\underline{\hat{X}} = \frac{1}{N} \cdot \sum_i^N \hat{X}_i$; and

h6) Softmax confidence entropy: given $\hat{X}_i \in \mathbb{R}$ before class assignment, after conversion of the network's final layer's logits with $Softmax(x_i) = \frac{\exp(x_i)}{\sum^i \exp(x_i)}$, the resulting $\hat{X}_i^*$ can be interpreted as two-element prediction confidence $[p_f, p_b]_i = \hat{X}_i^*(x, y)$ for foreground p_f and background p_b. Thus we can estimate class-agnostic prediction entropy by $\sum_i^K p_i \log(p_i)$ where $p_i = \sum_i^N \max([p_f, p_b]_i)$.

We calculate each of these heuristics at test time, and normalise them to lie between [0,1]. By thresholding this range we can determine which measurements to accept and which to reject.

3.4.6 Confidence Communication

We estimate the confidence of our network in predicting measurements for each individual image. We communicate this in both quantitative and qualitative ways.

Quantitative uncertainty estimation

Several methods are emerging to quantify prediction uncertainty for deep neural networks. One of the main benefits of probabilistic segmentations and derived measurements is the capability to directly derive confidence estimates from a set of varying predictions.

We align uncertainty prediction with error such that we predict a wide enough bound that we can be confident of the true value being contained within such a bound, while encouraging a small enough bound so that it remains meaningful i.e we can predict a measurement with a large range that will definitely contain the true value, but then the bounds may become meaningless if they do not correlate with prediction error. This is a core concept that our evaluation framework seeks to quantify for a given network and confidence estimation method.

We generate a confidence interval for each measurement using both information derived from fusing many possible segmentations and fusing many possible measurements, we analyse formulating these confidence ranges in the following ways:

- **Upper and lower bounded**: We consider the upper and lower limits of our 'extremes' measurements to be the bounds of our confidence interval.
- **Samples minimum and maximum**: We consider the minimum and maximum measurements extracted from our segmentation samples to be the bounds of our confidence interval.
- **Interquartile range (IQR)**: We consider the 25% and 75% percentile of our extracted measurements to be the bounds of our confidence interval.
- **Two Standard Deviations**: We consider the mean ± twice the standard deviation of the measurements to be the bounds of our confidence interval.

These confidence metrics may then be passed to our acceptance criteria module to decide at test time whether or not to accept the measurement or reject it and re-acquire a more suitable image, or defer to manual annotation or expert referral.

Qualitative uncertainty estimation

By using a probabilistic segmentation we can acquire robust estimations of image based prediction confidence and visualise this simultaneously to the estimated final prediction result.

We generate this visual communication by taking the difference between the union and the intersection of samples generated for a given input, and presenting this region visually to the operator during test time.

We extract a visual representation of confidence by indicating the upper and lower bounds of a particular input segmentation. The user of such a system can therefore see which regions of an image a network is more or less confident about and adjust their input accordingly. This provides in real-time an intuitive indication of prediction confidence.

3.5 Implementation

We conduct two sets of experiments to validate the performance of our methods. Firstly a series of experiments based solely on Head Circumference estimation, and secondly a series of experiments for weight estimation for which estimates of the Head Circumference, Abdominal Circumference and Femur length are required.

3.5.1 Data

Head Circumference Estimation

Our base dataset, named subsequently as Dataset A, consists of 2,724 two-dimensional US examinations from volunteers at 18-22 weeks gestation, acquired and labelled during routine screening by 45 expert sonographers. Several images were taken during each session, including the standard transverse brain view at the posterior horn of the ventricle (TV) plane used for HC measurement.

This data was combined with the HC18 Challenge [179] dataset which consists of 1334 two-dimensional US images of the standard plane that is used to measure HC, each image is 800x540 pixels with a pixel size ranging from 0.052mm to 0.326mm. Each image in the training set has an accompanying manual annotation of the HC (ellipse outline) performed by a single trained sonographer [179]. We resample all images to 320 × 384 pixels, and produce a head mask from the expert ground truth delineation. Training data is randomly flipped both horizontally and vertically, and a random rotation (±5°) is performed.

Weight Estimation

We use two-dimensional ultrasound images from routine examinations on volunteers at 18-22 weeks gestation, acquired and labelled during routine screening by 45 expert sonographers. Sonographers extract each measurement during screening, as well as providing a manual annotation on image stills from which the measurement is derived.

Head Circumference data contains 2,724 two-dimensional images of the standard transverse brain view at the posterior horn of the ventricle (TV) plane used for HC measurement, and an accompanying ellipsoid manual label.

Abdominal Circumference data contains 2,352 two-dimensional images of the standard abdominal circumference plane used for AC measurement, and an accompanying ellipsoid manual label.

Femur Length data contains 2,456 two-dimensional US images of the standard femoral view plane used for FL measurement, and an accompanying manually drawn axis line from which a femoral mask is generated.

Weight Estimation: From the above three datasets we extract 500 cases for which the same subject was used for each measurement as a test set. We must have a measurement for each to estimate fetal weight. We do the same for the validation set used to train the segmentation networks. All remaining cases are used as the training set.

Pre-processing: Training data is randomly flipped both horizontally and vertically, a random rotation (±25°), random translation and random scaling is applied.

3.6 Experiments and Results

In these experiments we evaluate the efficacy of our proposed methods in extracting accurate biometric measurements from ultrasound images. We go on to present experiments to evaluate each of the proposed probabilistic segmentation methods in providing reasonable confidence bounds on their predictions, as well as evaluate the ability of the proposed 'variance metrics' to reject poor performing cases to improve overall network performance metrics.

3.6.1 Head Circumference Estimation

Single-Sampling Experiments

In the first instance, single-sample experiments, generating a single segmentation and HC measurement ($N = 1$) per subject, were used to evaluate the performance of the proposed model against the state-of-the-art [163]. Table 3.1 reports performance measures for all *single-sampling* experiments. These show comparable performance relative to [163] for our U-Net implementation, trained on Dataset A. This result improves further when the same model is trained on Dataset A and HC18 data. MC dropout during training further improves the result. For subsequent analysis, all experiments for MC Dropout (during inference) use the combined data and are trained using MC dropout.

Multi-Sampling Experiments

MC Dropout during inference has been compared against a Probabilistic U-Net. Here, multiple (N) segmentation predictions are made for each US image. From these, the mean and median

Table 3.1: Single sample results of three U-Net's. **Baseline**: Trained on Dataset A data only. **Dataset A + HC18**: Trained on Dataset A data and HC18 Challenge data transformed to same format as Dataset A data. **Dropout**: Trained on Dataset A and HC18 Challenge data with dropout ($p = 0.6$ value found to be best performing in variety of dropout configurations). We compare the Mean absolute difference between the final HC measurement, the DICE overlap of the fitted ellipse with the ground truth ellipse, and the Hausdorff distance between the outline of the fitted ellipse and the outline of the ground truth ellipse. Results calculated on Dataset A test data.

	Mean abs difference ± std (mm)	Mean DICE ± std (%)	Mean Hausdorff distance ± std (mm)
Baseline	2.09 ± 1.97	0.982 ± 0.011	1.289 ± 0.880
Dataset A + HC18	1.90 ± 1.90	0.982 ± 0.010	1.292 ± 0.791
Dropout $p = 0.6$	**1.81 ± 1.65**	**0.982 ± 0.008**	**1.295 ± 0.664**

of the set of fitted ellipse parameters are used to obtain a single HC value for each test case, and the set of N segmentations are used to obtain an upper and lower bound. Table 3.2 shows the performance measures for our *multi-sampling* experiments. Results show that we lose performance through aggregating multiple results using the mean or median, although this is likely due to dropout not being applied during inference for single sample experiments. However, the *multi-sampling* methods do allow us to produce an upper and lower bound on the HC value, with an average difference of $1.82 \pm 1.78mm$ between upper-lower bounds and ground truth HC measurement ($N = 10$ samples), for cases where the ground truth is not within the upper-lower bounds (*MC(inf.)*).

Variance Measure Thresholding

Finally, we experiment with each of the variance scores produced over the test set as a means to accept/reject images at test time. We evaluate their performance by counting the number of accepted/rejected cases for a range of thresholds between zero and one, and how this threshold affects the resulting average performance scores after rejected images are removed from the test set. In this experiment we use only MC dropout during inference ($p = 0.6$) which performs best in our previous experiments.

Table 3.2: Multi-Sampling results for the two methods. We report the performance measures of a single-sampled point-predictor (*Det. (Deterministic)*), mean/median of $N = 10$ samples from the Probabilistic U-Net (*Prob. U-Net (Probabilistic U-Net)*), and our previous best U-Net with Monte-Carlo dropout during inference (*MC(inf.) (Monte Carlo dropout during inference)*, $p = 0.6$). We report the % ground truth HC values that lie in the calculated upper/lower bound range. This percentage varies significantly with N, for *MC(inf.)*: $N = 2$: 14.8%; $N = 1000$: 50.4%. See Supplementary Material Figures 1-3.

	Mean abs difference ± std (mm)	Mean DICE ± std (%)	Mean Hausdorff distance ± std (mm)	$LB \leq HC_{gt} \leq UB(\%)$
Det.				
MC $p = 0.6$	**1.81 ± 1.65**	**0.982 ± 0.008**	**1.295 ± 0.664**	N/A
Prob. UNet				
Mean	2.22 ± 2.15	0.980 ± 0.011	1.413 ± 0.751	20.4
Median	2.21 ± 2.15	0.980 ± 0.011	1.410 ± 0.748	20.4
MC(inf.)				
Mean	2.15 ± 2.09	0.981 ± 0.010	1.313 ± 0.613	27.8
Median	**2.15 ± 2.07**	**0.981 ± 0.010**	**1.307 ± 0.604**	**27.8**

Figure 3.3 shows graphs depicting how each variance measure can be used to reject test cases, and how rejecting high variance cases can lead to improved performance. In each case we normalise the variance score to lie between 0 and 1, and for each threshold between 0 and 1 we ‘reject’ cases whose variance score is above the threshold. Plots show the performance for remaining ‘accepted’ cases, plotted against the number of ‘rejected’ cases. For most variance scores we obtain an initial performance boost from ‘rejecting’ the worst cases, but after an initial improvement, the variance scores do not delineate ‘good’ from ‘bad’ cases very well. Results suggest that higher measurement variance may indicate sub-optimal imaging plane acquisition.

Qualitative evaluation

Figure 3.4 shows examples for successful and less model-compliant images using Dropout during inference to produce the samples, where model-compliance captures the proximity of the image to the training data. Note that the best performing examples produce very narrow upper and lower bounds (in this figure where the upper and lower bounds occupy the same pixels the margin is not visible). The worst performing examples show a wider upper and lower bound range but the ground truth ellipse is often not contained within the predicted range.

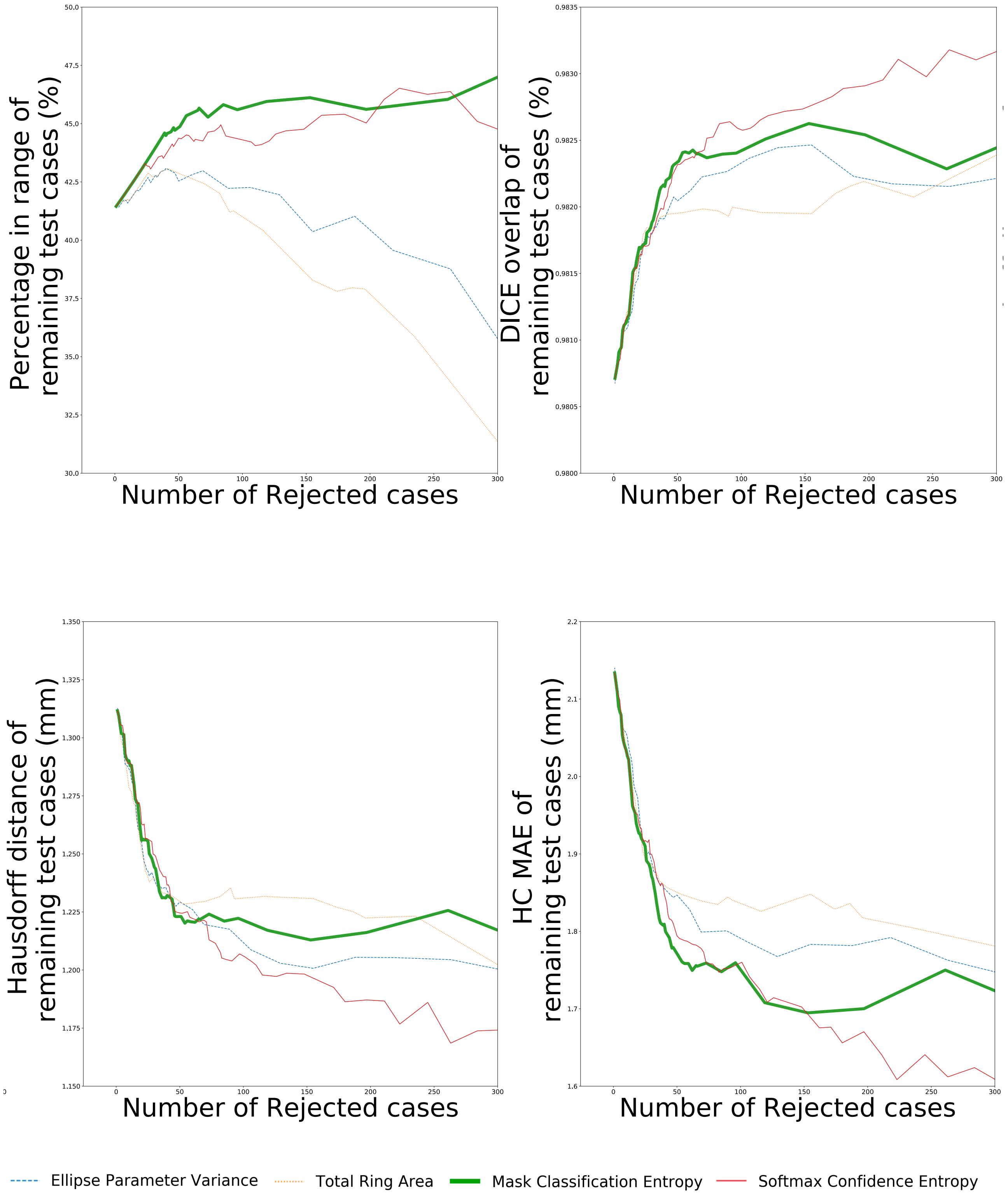


Figure 3.3: Plots showing performance measures against the number of rejected test cases. Each measure shows improvement after removing a few test cases for each score (these thresholds vary for each score). We calculate the performance metric for the entire test set, and then using each variance measure independently we reject progressively more and more data points from the test set by varying the threshold above which test cases should be kept. In this way we simulate using that variance measure to improve overall performance by rejecting the cases for which that variance measure deems worst performing. After removing an initial low performing set, the scores power to discriminate between 'good' and 'bad' images deteriorate. 'Percentage in range' calculated as the percentage of test cases for which the ground truth HC measurement lies within the the predicted upper-lower bounds.

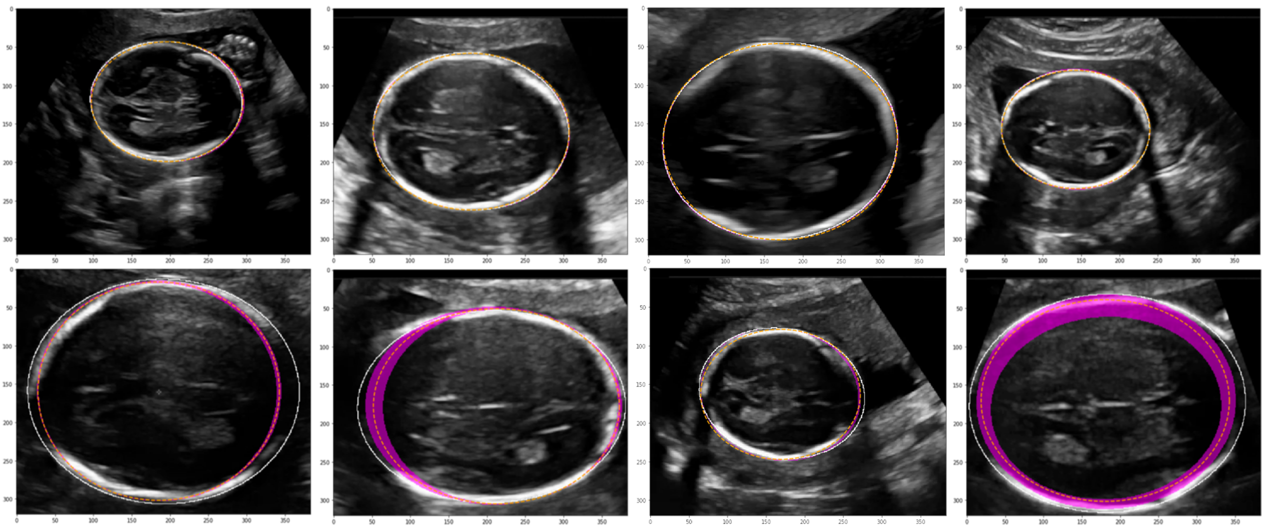


Figure 3.4: Results produced by our model. White line: Ground Truth, Orange dashed line: Mean of sampled ellipse parameters, Pink shaded area: Upper/lower bound range. Top row: High performing images. Bottom row: Low performing images. See Supplementary Material Figures 4 and 5 for more examples and a demo video demonstration.

These images often show a lack of clear white presentation of the skull. However, ambiguous segmentation of the regions with missing signal is often reflected in the confidence margin produced, showing greater variation in those image regions, which can be seen clearly in the second example in the bottom row - a wider upper-lower bound area for image regions with low signal from fetal skull. The example on the bottom far right shows missing signal on both sides, which results in a large uncertainty in the ellipses globally due to the compounded effect of missing signal on both sides of the skull.

3.6.2 Weight Estimation

We apply our framework to three biometric estimation tasks used for downstream weight estimation and demonstrate how our framework can be used to estimate the quality of segmentation confidence estimation approaches. We also show improved performance over state-of-the-art biometric extraction methods for each task using our framework. We then use the automated predictions from our three tasks in a further downstream task to analyse the impacts of good uncertainty estimates on error propagation. We show the efficacy of our framework for the prediction and interpretation of three common fetal biometrics extracted during a typical fetal screening session: HC, AC and FL. For each task, we evaluate each stage of our proposed framework independently. We use a single sample from each segmentation method as a performance baseline for each method, and sample N stochastic segmentations for each input to use in analysis of the remaining stages of the framework. We evaluate performance of automated HC, AC and FL estimates to estimate fetal weight, which is used to place a fetus in a growth percentile for their age which determines whether a fetus may be growing abnormally.

Single Sample Measurement Estimation Baseline

We take a single segmentation sample ($N = 1$) from each model to acquire a single measurement. In this experiment we quantify the performance of the segmentation module of our framework for producing a single prediction for each test input image. Table 3.3 shows three performance measurement for our tasks and we can see that our models achieve state-of-the-art art performance even with a single sample from the each network. The introduction of the stochastic head to the UNet architecture has introduced significant improvement to each task. In this experiment the single sample for a UNet with dropout during inference is obtained by taking a sample without dropout, and obtained for stochastic U-Net by sampling the mean of the fitted distributions. We can see from these results that by taking the mean of the fitted distribution we obtain a more accurate output than by just turning dropout off.

Multi Sample Measurement Estimation and Fusion

We take multiple segmentation samples ($N = 100$) from each model to acquire N different segmentations, and N different measurements for each input image. We generate 'extreme' segmentations and measurements as described in Section 3.4.3. Table 3.4 shows the results of fusing each set of segmentations and measurements into a single measurement using each method described in Section 3.4.3. We see that by taking multiple samples and measurements and averaging between them we achieve superior results to taking a single sample from our network, indicating our networks improve performance by taking multiple samples and fusing them. We can see that the median of probabilistic measurements (not including 'segmenta-

Table 3.3: Metric prediction results from a single sample taken for each test case in 'probabilistic mode'.

Model	DICE ± std	MAE ± std (mm)	Hausdorff distance ± std (mm)
Head			
U-Net	0.980 ± 0.029	3.07 ± 6.58	2.858 ± 3.979
PHiSeg	0.985 ± 0.011	2.17 ± 2.33	1.607 ± 0.652
SSN U-Net	0.984 ± 0.012	2.49 ± 2.64	1.424 ± 0.481
Abdomen			
U-Net	0.964 ± 0.042	6.33 ± 9.54	5.940 ± 6.068
PHiSeg	0.964 ± 0.028	5.83 ± 6.88	6.497 ± 5.980
SSN U-Net	0.972 ± 0.020	4.40 ± 5.44	2.755 ± 2.883
Femur			
U-Net	0.936 ± 0.006	1.64 ± 4.02	1.690 ± 3.536
PHiSeg	0.943 ± 0.058	1.31 ± 2.83	1.669 ± 4.680
SSN U-Net	0.944 ± 0.051	1.14 ± 2.65	1.368 ± 3.860

tion mean') produces the best results, perhaps due to median being more robust to outlier predictions which the network may predict. We also see an interesting effect when looking at the 'Extreme measurements Midpoint' where for the Stochastic U-Net the error is very large, compared to PHiSeg and Dropout U-Net (both are the worst performing for their respective models), indicating that extreme PHiSeg and Dropout U-Net predictions are less biased than those produced by Stochastic U-Net, which produces much larger amounts of variation compared to the other methods, and tend to over-segment the object of interest more often than under-segmenting.

3.6.3 Confidence Range Estimation

We use the outputs of our Multi-sampling experiments to generate confidence intervals on each measurement using each method described in Section 3.4.6. We evaluate how the number of samples impacts the width of each generated interval, and what percentage of generated intervals contain the ground-truth measurement as the width of each interval changes. Figure 3.5 show this relationship clearly. We can interpret these plots as indicating the level to which our network predicts plausible variation in each prediction. A perfect probabilistic model would show 'Percentage in range' rapidly increase to 100% while the corresponding range width

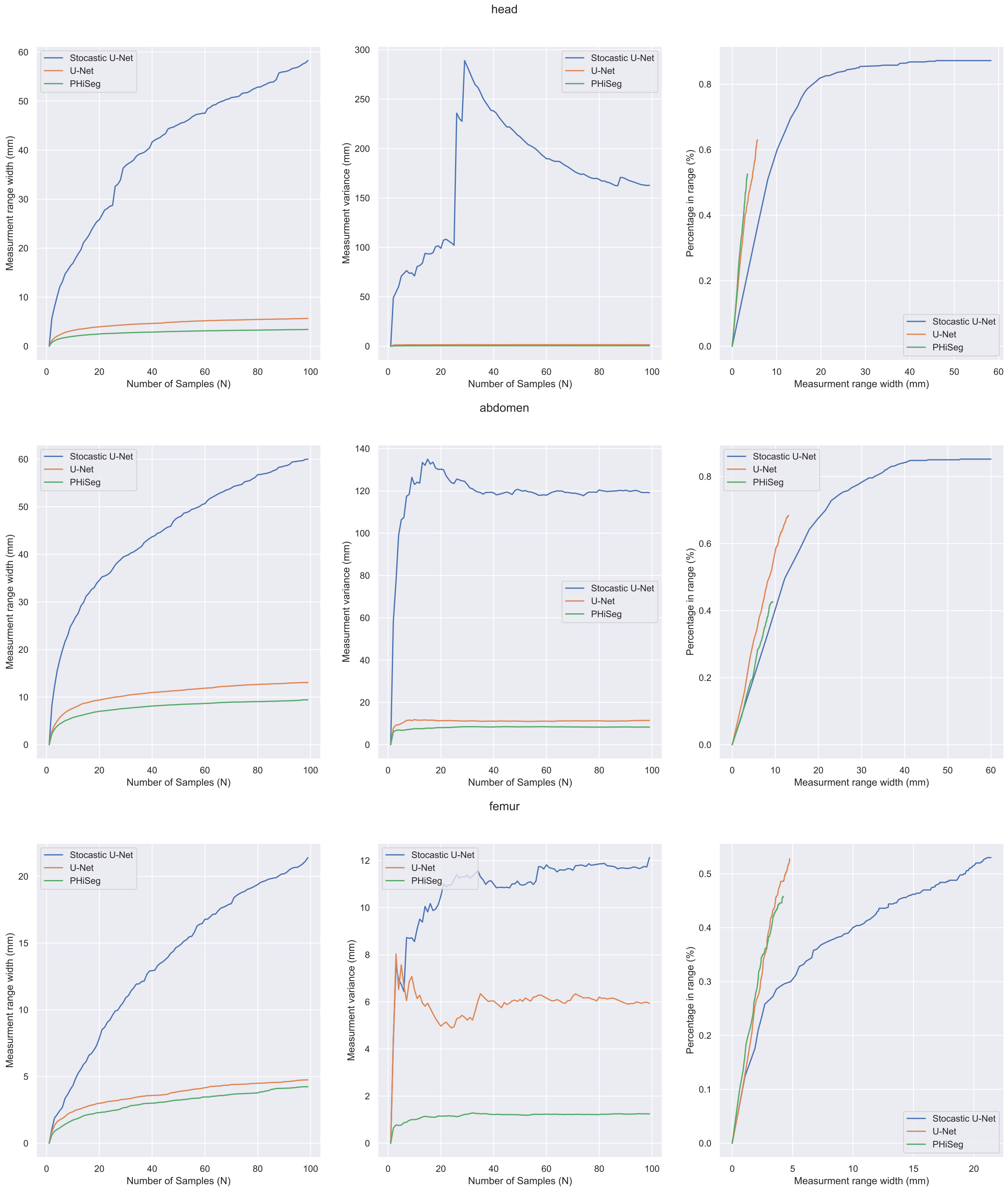


Figure 3.5: Analysis of the number of stochastic samples taken for each image. Left column: Number of Samples vs the measurement range width. Middle column: Number of samples vs measurement variance. Right column: Measurement range width vs Percentage in range.

Fusion Method	Dropout U-Net	PHiSeg	SSN U-Net
Head			
Segmentation Mode	3.065 ± 6.581	2.170 ± 2.335	2.491 ± 2.637
Median	2.143 ± 2.626	2.164 ± 2.325	2.477 ± 2.526
All Measurements Mean	2.155 ± 2.693	2.166 ± 2.316	2.997 ± 3.401
Sample Measurements Mean	2.152 ± 2.689	2.164 ± 2.315	2.859 ± 3.046
Sample Percentile Midpoint	2.144 ± 2.693	2.172 ± 2.320	2.565 ± 2.559
All Measurements Percentile Midpoint	2.145 ± 2.696	2.176 ± 2.320	2.559 ± 2.554
Extreme Measurements Midpoint	2.436 ± 3.162	2.328 ± 2.500	20.819 ± 43.093
Abdomen			
Segmentation Mode	6.325 ± 9.535	5.832 ± 6.883	4.401 ± 5.438
Median	3.885 ± 4.240	5.770 ± 6.666	4.373 ± 5.251
All Measurements Mean	4.001 ± 4.373	5.792 ± 6.657	4.828 ± 5.384
Sample Measurements Mean	3.985 ± 4.354	5.782 ± 6.656	4.734 ± 5.323
Sample Percentile Midpoint	3.926 ± 4.356	5.731 ± 6.680	4.551 ± 5.197
All Measurements Percentile Midpoint	3.926 ± 4.358	5.735 ± 6.679	4.562 ± 5.196
Extreme Measurements Midpoint	5.531 ± 6.373	6.510 ± 7.124	16.898 ± 17.402
Femur			
Segmentation Mode	1.635 ± 4.015	1.310 ± 2.826	1.135 ± 2.652
Median	1.045 ± 2.774	1.261 ± 2.745	1.138 ± 2.648
All Measurements Mean	1.133 ± 2.676	1.330 ± 2.758	1.557 ± 2.593
Sample Measurements Mean	1.117 ± 2.638	1.327 ± 2.764	1.412 ± 2.596
Sample Percentile Midpoint	1.053 ± 2.624	1.332 ± 2.784	1.154 ± 2.659
All Measurements Percentile Midpoint	1.038 ± 2.418	1.332 ± 2.784	1.160 ± 2.659
Extreme Measurements Midpoint	2.344 ± 5.688	1.675 ± 3.098	10.510 ± 7.595

Table 3.4: Fusion experiments: We fuse our set of generated measurements using several methods and report the mean absolute error of each fusion method.

remains below the accepted level of measurement error for each task, indicating that the method successfully captures the variation of manual measurements in a meaningful way. Table 3.5 shows the results of each range generation method as well as a new metric named 'Range AUC' which evaluates the suitability of a measurement range. We calculate this as the area under the curve in Figure 3.5 (c), divided by the maximum measurement range. This metric aims to quantify the trade-off between a high percentage in range and a high range width, methods that produce a high percentage in range with a small range width will result in a higher 'Range AUC'. We can see that Stochastic U-Net introduces significant variation between samples, however this variation quickly exceeds the meaningful bounds we wish to see in a probabilistic segmentation task compared to Dropout U-Net and PHiSeg.

Method	U-Net (Dropout p=0.5)	PHiSeg	SSN U-Net
Head			
Extremes range	91.0 (9.224 ± 4.822)	96.6 (12.118 ± 2.268)	88.4 (73.562 ± 101.310)
Samples range	72.2 (5.609 ± 3.734)	51.8 (3.354 ± 1.079)	87.8 (58.577 ± 91.830)
Samples IQR	47.6 (1.264 ± 0.721)	65.2 (0.860 ± 0.029)	44.8 (4.226 ± 3.192)
± 2 std	29.6 (4.171 ± 2.468)	47.0 (2.750 ± 0.608)	9.8 (31.189 ± 40.361)
Range AUC	0.096	0.132	0.014
Abdomen			
Extremes range	97.0 (27.322 ± 21.750)	93.4 (27.146 ± 11.336)	86.2 (75.312 ± 64.609)
Samples range	70.6 (13.247 ± 11.511)	47.6 (9.516 ± 8.259)	85.8 (60.040 ± 47.067)
Samples IQR	55.4 (2.892 ± 3.047)	48.0 (2.448 ± 3.348)	36.6 (7.350 ± 6.821)
± 2 std	31.6 (10.010 ± 9.269)	37.2 (7.906 ± 8.210)	10.6 (34.998 ± 26.031)
Range AUC	0.045	0.041	0.013
Femur			
Extremes range	84.6 (6.676 ± 15.771)	92.4 (5.363 ± 7.823)	71.2 (22.901 ± 18.854)
Samples range	71.6 (4.774 ± 12.552)	80.6 (3.577 ± 3.766)	70.2 (21.470 ± 17.954)
Samples IQR	53.2 (0.749 ± 1.692)	53.3 (0.882 ± 1.615)	59.8 (0.539 ± 0.537)
± 2 std	31.2 (3.466 ± 8.592)	26.2 (2.841 ± 3.799)	17.0 (10.805 ± 8.745)
Range AUC	0.093	0.157	0.020

Table 3.5: In range results for $N = 100$ samples: Percentage in range (range width), and Range AUC for each method.

3.6.4 Measurement Acceptance Filtering

We evaluate the capability of scoring each probabilistic prediction using the methods described in Section 3.4.5. We take $N = 100$ samples from our segmentation network, and calculate each variance score for each set of predictions. We normalise the set of variance scores calculated to be between zero and one for each, and then by setting progressively higher thresholds on this range, we filter out or 'reject' test-cases for which this score is the highest and recalculate average performance metrics over this new test set. We show that for several methods and tasks we can significantly improve the overall performance by removing just the least-confident 10% of cases, motivating the use of confidence measures as a means of indicating at test time the unreliability of measurements for a given input.

Qualitative Evaluation Figures 3.9-3.11 shows some examples of the best and worst performing test cases for each model. We show the region between 'extreme' segmentations depicted in purple, the ground truth in blue, and fused prediction in green/white. We can see clearly from these images that both PHiSeg and Dropout U-Net produce similar variations in prediction

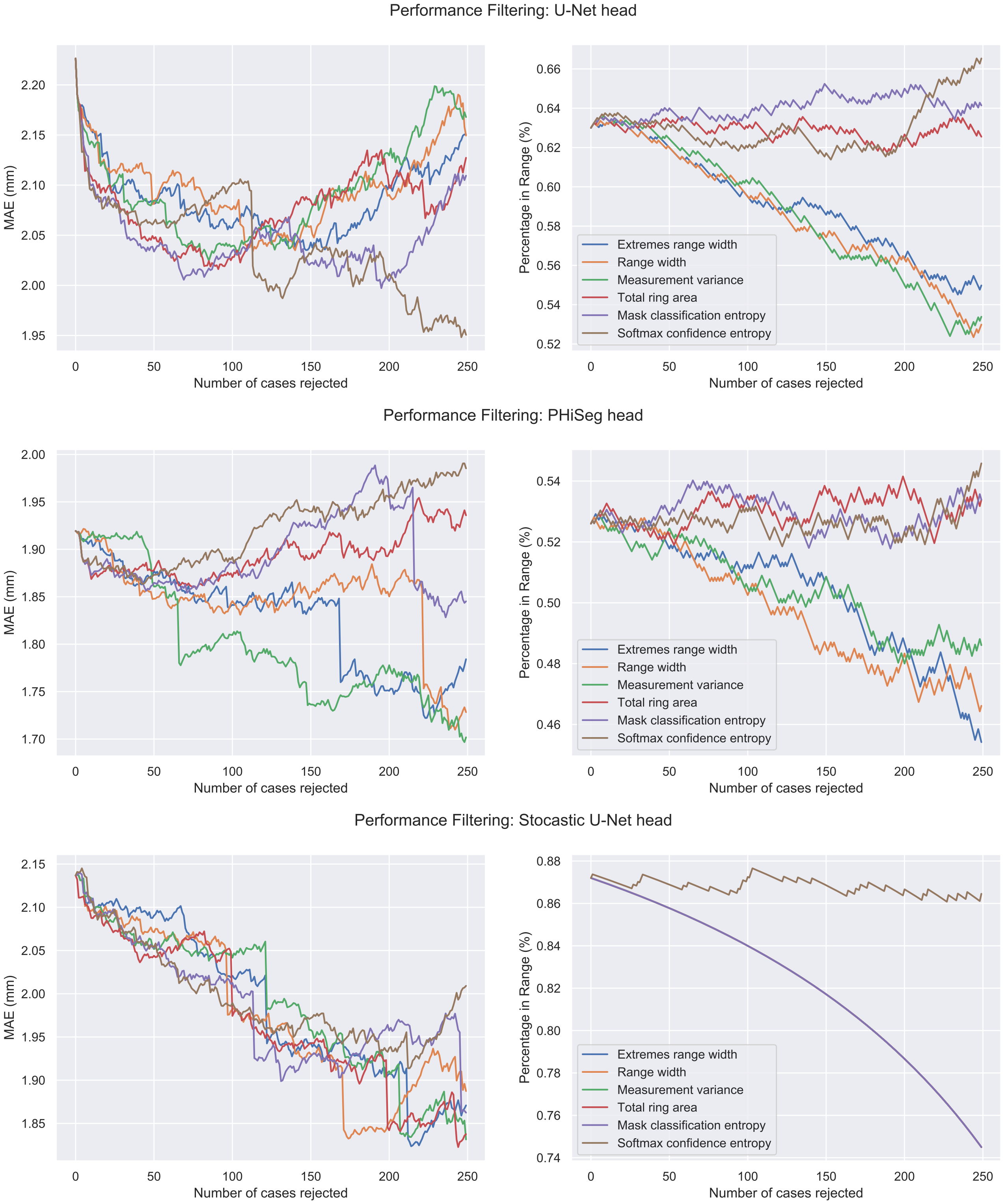
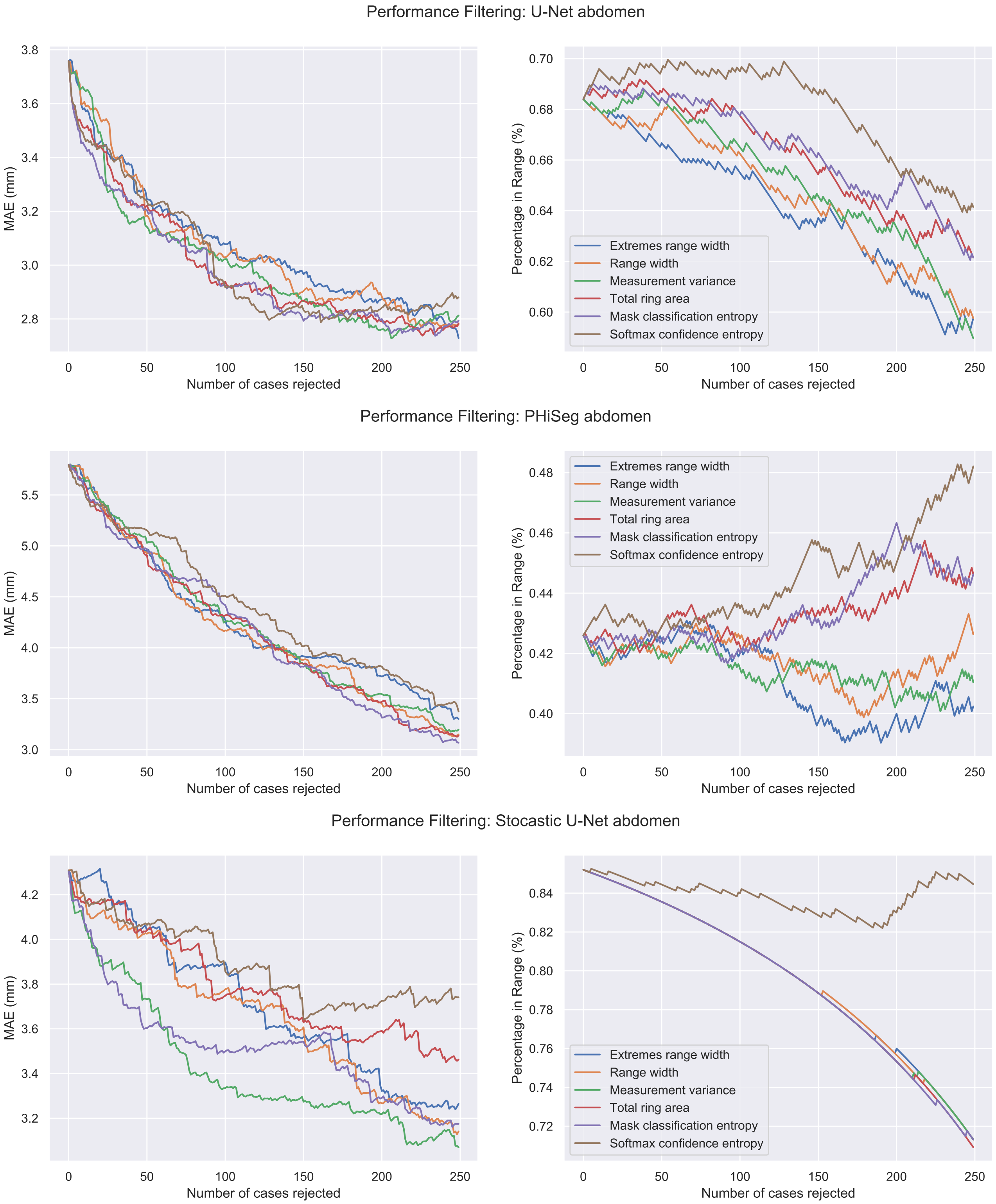


Figure 3.6: Performance filtering (Head): Performance measures as the least confident cases are removed from the test set as per each confidence measure.



Performance Filtering: U-Net abdomen

Performance Filtering: PHiSeg abdomen

Performance Filtering: Stocastic U-Net abdomen

Figure 3.7: Performance filtering (Abdomen): Performance measures as the least confident cases are removed from the test set as per each confidence measure.

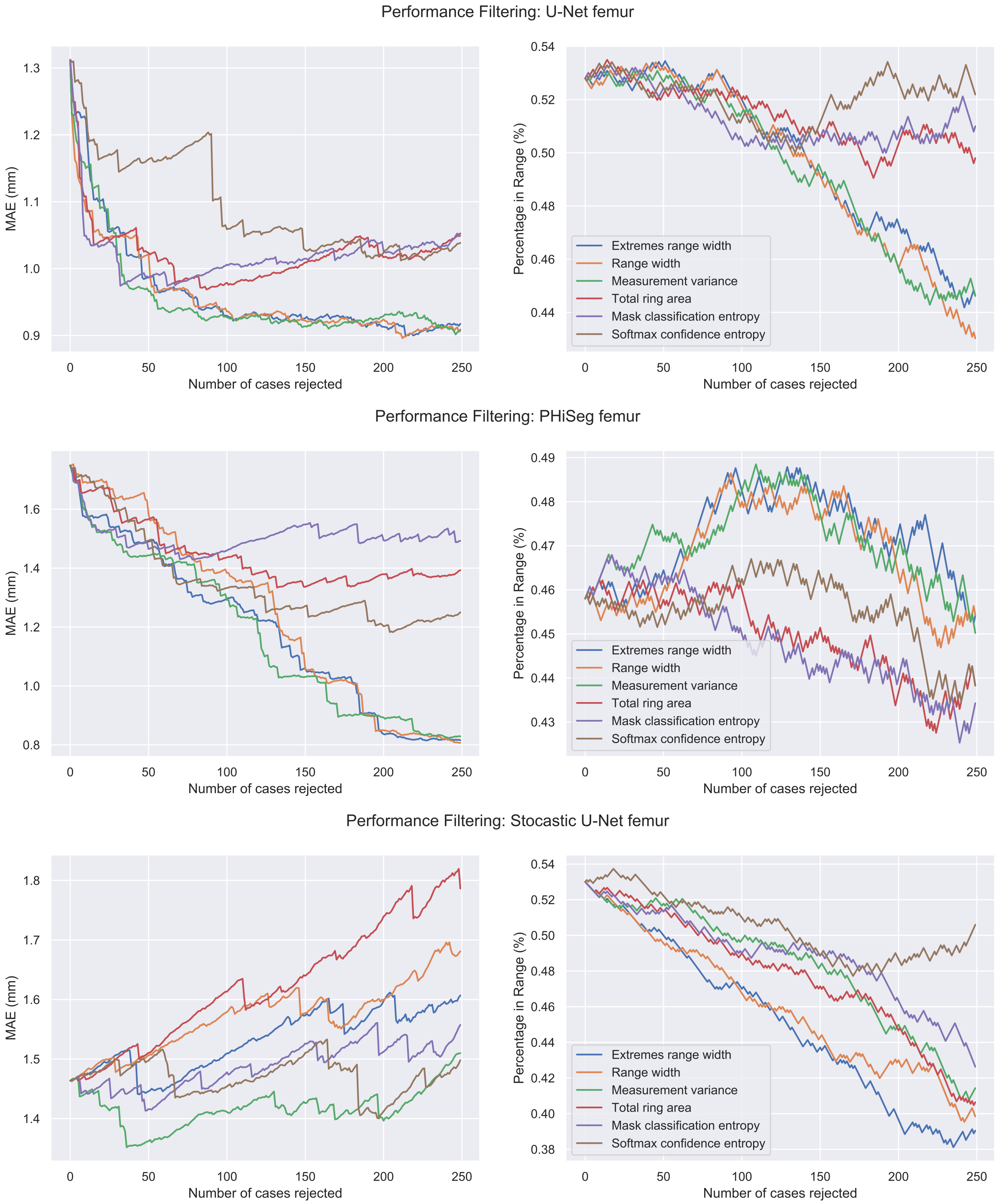


Figure 3.8: Performance filtering (Femur): Performance measures as the least confident cases are removed from the test set as per each confidence measure.

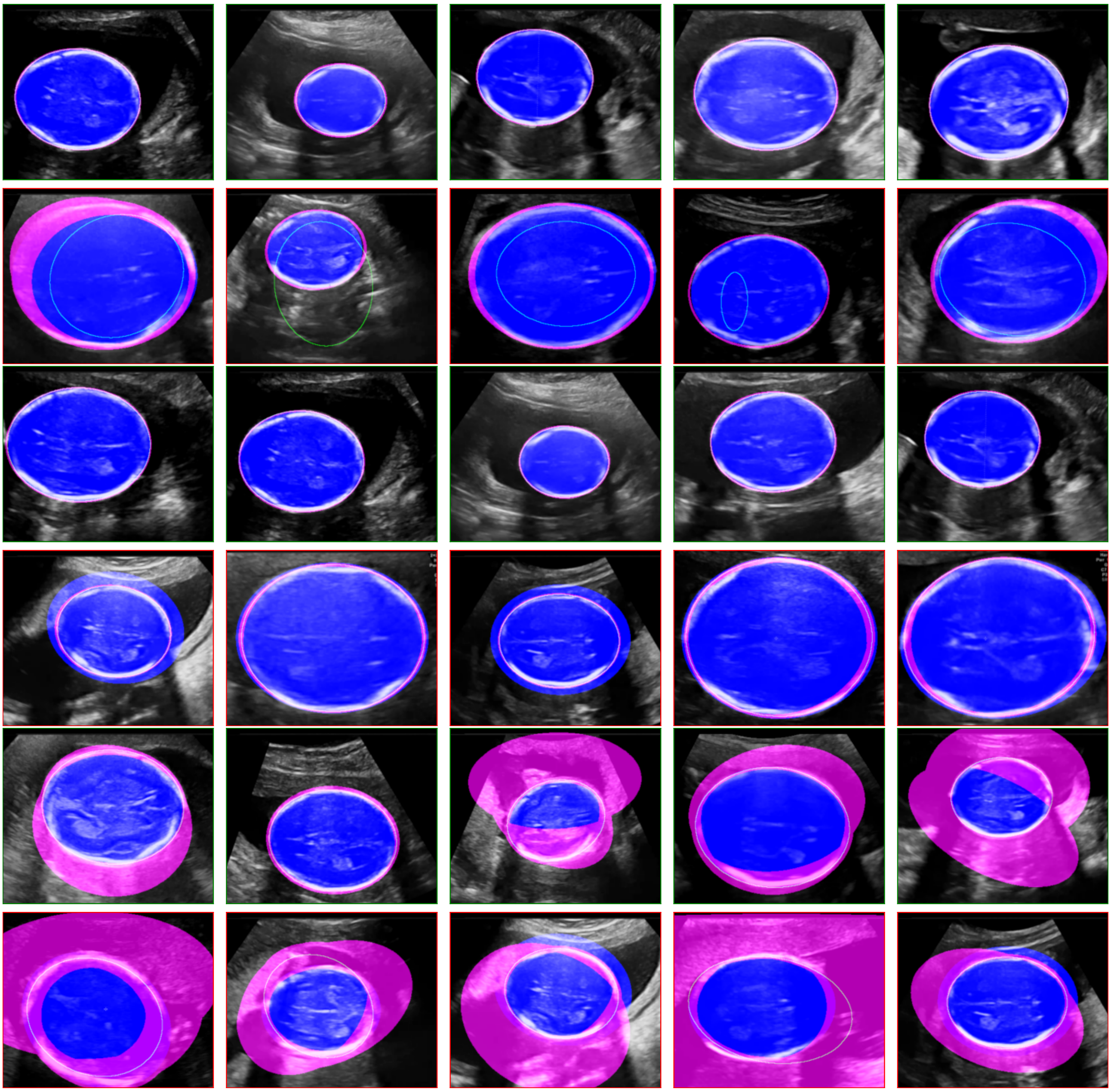


Figure 3.9: Head: Top and Bottom 5 performing (DICE on 'mean segmentation') test cases for U-Net (Top) and PHiSeg (Middle) and Stochastic U-Net (Bottom). Green border around best cases, red around worst cases. Blue Ellipse ground truth, white/green outline fused prediction, purple region between upper and lower bounds

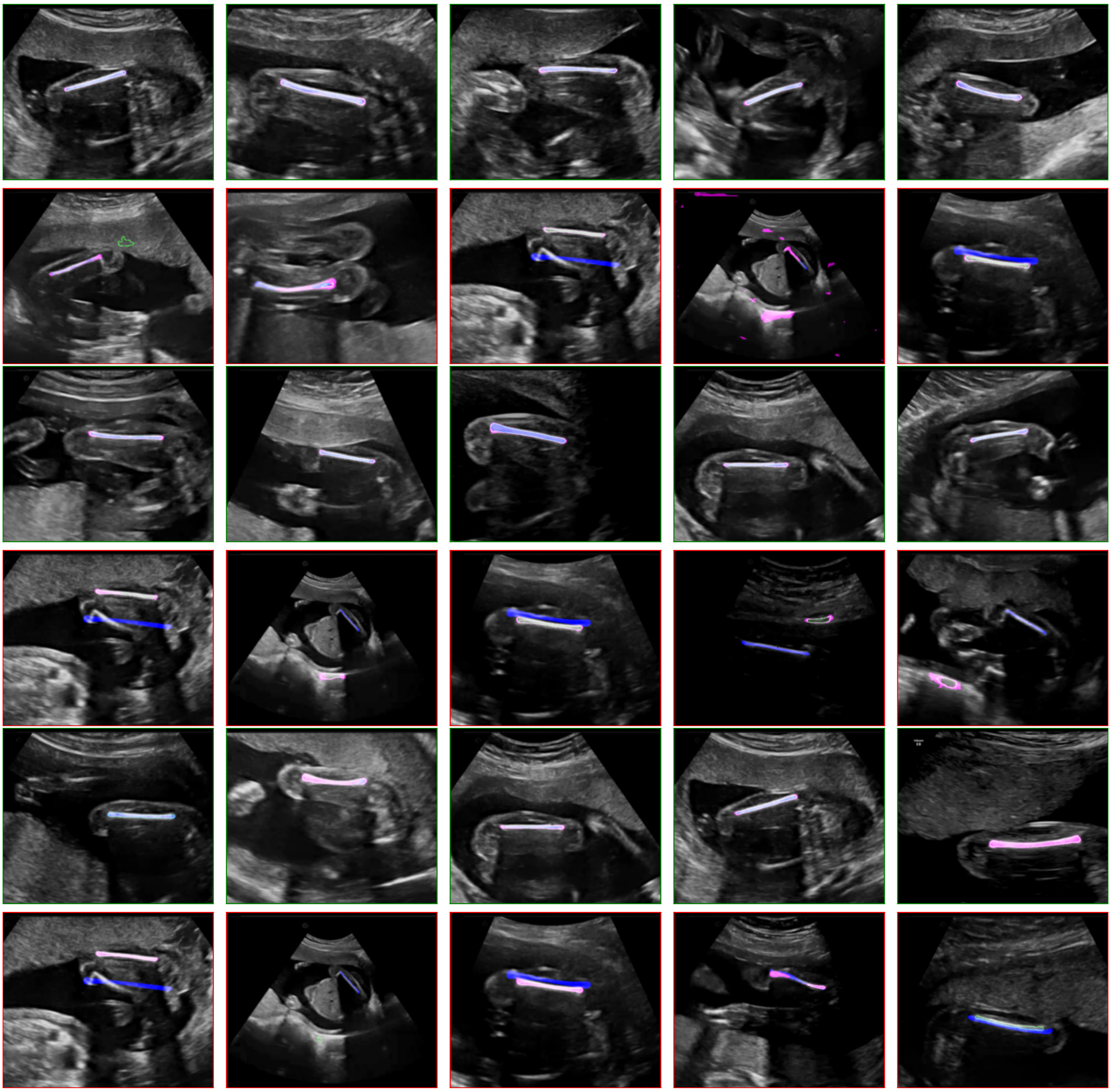


Figure 3.10: Femur: Top and Bottom 5 performing (DICE on 'mean segmentation') test cases for U-Net (Top) and PHiSeg (Middle) and Stochastic U-Net (Bottom). Green border around best cases, red around worst cases. Blue Ellipse ground truth, white/green outline fused prediction, purple region between upper and lower bounds

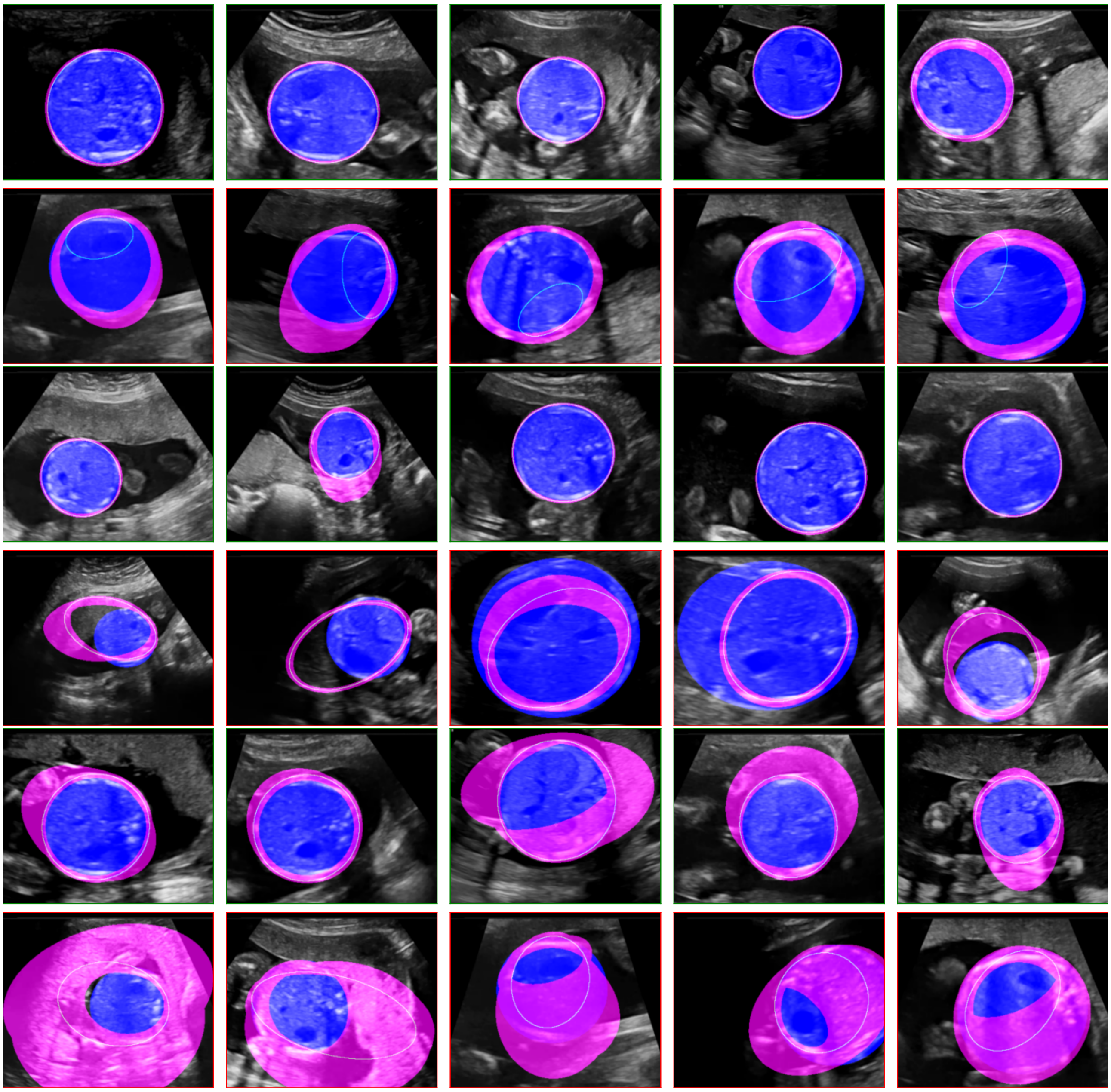


Figure 3.11: Abdomen: Top and Bottom 5 performing (DICE on 'mean segmentation') test cases for U-Net (Top) and PHiSeg (Middle) and Stochastic U-Net (Bottom). Green border around best cases, red around worst cases. Blue Ellipse ground truth, white/green outline fused prediction, purple region between upper and lower bounds

Method	MAE (mm)	% in Range
U-Net (Dropout)	18.39 ± 34.24	80.4 (201.749 ± 2861.13)
PHiSeg	22.35 ± 25.97	70.0 (59.303 ± 38.484
Stochastic U-Net	20.06 ± 24.44	95.2 (478.732 ± 848.56)

Table 3.6: Hadlock's weight estimation: Results for each method. We report MAE (mm) and standard deviation, and Percentage in range (Range width and standard deviation)

where worse performing cases produce more variation than the most accurate, while Stochastic U-Net produces as much larger level of variation for each input, whether it performs well or not.

3.6.5 Hadlocks Weight estimation

We continue our evaluation by using the three measurement estimations to predict fetal weight. As per the World Health Organisation (WHO) this is performed using Hadlock's formula C [168]. We make a prediction of fetal weight for each subject using each methods trio of measurements, as well as a min and max weight calculated using the min and max of each measurement from each method. Table 3.6 shows the outcomes predicting fetal weight from head, abdomen and femur measurement estimations from each model. We can see that each model has a similar amount of accuracy, however only PHiSeg produces weight ranges of a sensible width. We perform a similar experiment to that of each task and estimate the performance after replacing the least confident measurements with the ground truth to simulate a manual intervention on that measurement, shown in Figure 3.12.

3.7 Discussion

While we cannot claim our proposed 'variance scores' represent model uncertainty directly, they show some capability to 'reject' particularly low performing test cases. In this way, the 'variance scores' can be described as a measurement for the proximity to the variance of the training data of an unseen test sample, which is also desirable, showing the confidence of the network with respect to its capacity and seen training examples. Scenarios in which an operator is

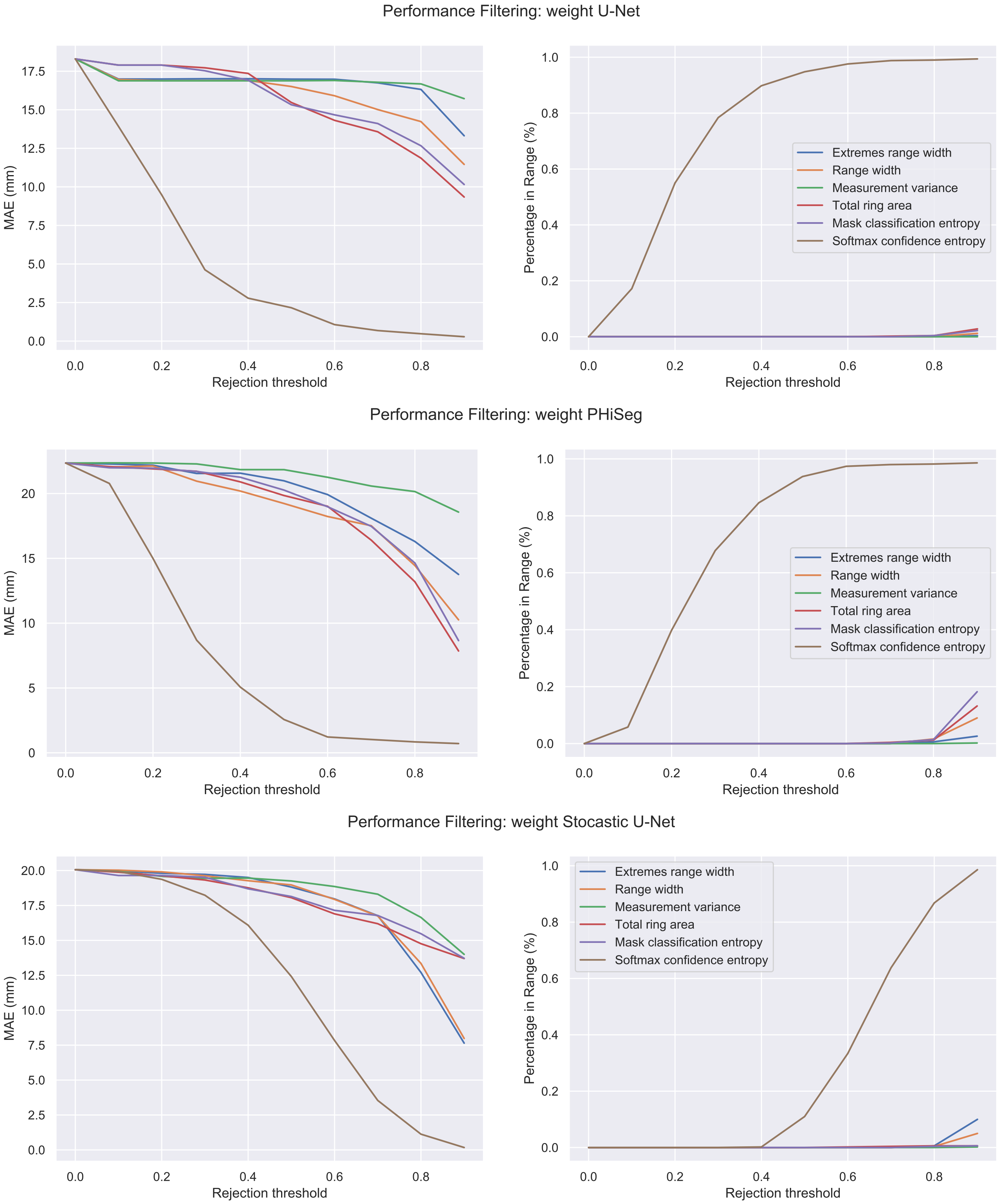


Figure 3.12: Performance filtering (weight): At each threshold, measurements of the head, abdomen or femur are rejected and replaced with the ground truth measurement before recalculating the estimated fetal weight to simulate a manual intervention on the measurement.

present stand to benefit practically using methods introduced in this work, prompting operators to reject sub-optimal measurements by providing real-time feedback during acquisition, thus improving inter-operator consistency. This work lays the foundations for methods by which this can be achieved.

We demonstrate performance on a number of application tasks and show superior performance and interpretability to state-of-the-art methods. We provide an evaluation pipeline for analysing the impacts of probabilistic model design on the distributions of predictions it produces, enabling future researches to better understand their models and whether or not the variation produced by probabilistic methods in aligned with human expectations.

We provide a way to accept and reject poor performing test cases during acquisition as the first indicator of a sub-optimal measurement, and provide visual feedback guidance to promote inputs that are optimal for the current model. We go further than this and demonstrate how once a model successfully captures plausible variation in predictions, this can be leveraged as a measure of proximity to training data and an indicator of potential pathology where models have been trained on healthy cohorts.

Our work presents several limitations. The degree to which our 'variance metrics' estimate model accuracy is only shown in this work through testing through their use to reject poor performing cases, however this could be evaluated further through the use of calibration curves. While our methods produce accurate results, further work is needed to evaluate whether this performance meets clinical expectations of measurement accuracy.

3.8 Summary

We demonstrate the effectiveness of probabilistic CNNs to automatically generate HC measurements from US scans, and produce upper-lower bound confidence intervals in real-time. Using multi-sampling probabilistic networks we derive ‘variance scores’, which indicate how confident our network is in generating measurements for a given image. This approach could be used to derive a system which rejects images collected from sub-optimal views, forcing sonographers to take measurements from a view for which the network performs optimally. This could lead to techniques for automated fetal HC measurement, which outperform manual approaches in terms of accuracy and consistency.

Future directions of this work include exploring alternative methods for multi-sampling networks, alternative segmentation fusion strategies and alternative 'variance scores'. Analysis of new datasets to investigate network bias towards particular datasets is valuable, as well as analysis of cases with anomalous anatomy to evaluate performance in the presence of pathologies, clinically the most important cases to identify.

Chapter 4

Complexity Challenge

In this chapter we examine the 'Complexity Challenge' and develop methods by which we can use DL models to predict more complex information from medical images. In this chapter we extend previous segmentation approaches from 2D to 3D and introduce automated simultaneous prediction of complex shape properties from medical image volumes. As such we choose to use fetal brain MRI as our input to the developed algorithms, from which 3D segmentation and extraction of 3D shape properties is a key part of many clinical pathways. The extension to 3D and the introduction of shape properties bring added complexity in automated prediction, and also an additional challenge of how best to present the data and extracted predictions for ease of understanding.

The cerebral cortex performs higher-order brain functions and is thus implicated in a range of cognitive disorders. Current analysis of cortical variation is typically performed by fitting surface mesh models to inner and outer cortical boundaries and investigating metrics such as surface area and cortical curvature or thickness. These, however, take a long time to run, and are sensitive to motion and image and surface resolution, which can prohibit their use in clinical settings. In this chapter, we instead propose a DL solution, training a novel architecture to predict cortical thickness and curvature metrics from T2 MRI images, while additionally returning metrics of prediction uncertainty. Our proposed model is tested on a clinical cohort (Down Syndrome) for which surface-based modelling often fails. Results suggest that deep convolutional neural networks are a viable option to predict cortical metrics across a range of brain development stages and pathologies.

4.1 Introduction

Irregularities in cortical folding and micro-structure development have been implicated in a range of neurological and psychiatric disorders including: Autism, where disruptions to folding cortical and thinning of the cortex has been found in regions associated with social perception, language, self-reference and action observation [180]; Down Syndrome, where smoother cortical surfaces and abnormal patterns of cortical thickness are linked to impaired cognition [181]; Epilepsy, where malformations in cortical development are associated with seizure onset [182] and psychosis, which is associated with abnormal functional behaviour of the pre-frontal cortex [183].

There is a strong need to model cortical development in at-risk neonatal populations. However, due to the heterogeneous and highly convoluted shape of the cortex, it has proved highly challenging to compare across populations. Recent consensus has been that cortical features are best studied using surface-mesh models of the brain [184], as these better represent the true geodesic distances between features on the cortex. However, these require running of costly multi-process pipelines which perform intensity-based tissue segmentation, followed by mesh tessellation and refinement.

For developmental cohorts, the fitting of surface mesh models is even more challenging due to the relatively low resolution and likely motion corruption of these datasets. This leads to artifacts and partial volume effects or blurring across tissue boundaries. Methods to tackle these problems individually exist [185] but they are highly tuned to high-resolution, low motion research data sets and do not always transfer well to clinical populations as outlined in Figure 4.1.

As an alternative, several groups have proposed techniques for extracting cortical surface metrics from volume data directly [186, 187, 188]. Specifically, Tustison et al [187, 188] show that an ANTs-based extension for the estimation of cortical thickness from volumetric segmentation generates thickness measures which outperform FreeSurfer (surface) metrics when applied to predictive tasks associating thickness with well-studied phenotype relations such as age and gender. Nevertheless, volumetric fitting approaches such as [187] have not yet been evaluated on developmental data and their slow run times limit their utility for clinical applications. In this chapter we therefore seek to develop a novel algorithm for cortical metric prediction from clinical, developmental data, through DL.

4.2 Related work

The cerebral cortex is a thin layer of grey-matter tissue at the outer layer of the brain. Studying it is important for improved understanding of cognitive and neurological disorders but doing so is challenging due to it's complex shape and patterns of micro-structural organisation.

Currently, most studies of the cortex use surface mesh models [185, 189, 190], which fit mesh models to inner and outer cortical boundaries following pipelines which perform tissue segmentation, followed by surface tessellation with intensity based refinement (Figure 4.1). Summary measures of cortical thickness and curvature may then be estimated from the Euclidean distance between mesh surfaces, and principal curvatures of the white-matter surface respectively.

By contrast, ANTs (Advanced Normalisation Tools) and CAT (Computational Anatomy Toolbox) propose volume-based pipelines for cortical thickness estimation. Specifically, the ANTs pipeline estimates cortical thickness in five steps: 1) Initial N4 bias correction of input MRI; 2) Segmentation/Registration based brain extraction; 3) Alternating prior-based segmentation and weighted bias correction using Atropos and N4; 4) DiReCT-based cortical thickness estimation; 5) Optional normalisation to template and multi-atlas cortical parcellation [187, 188]. CAT uses segmentation to estimate white matter (WM) distance, and then projects the local maxima (equal to cortical thickness) to other grey matter voxels by using a neighbour relationship defined by the WM distance [186].

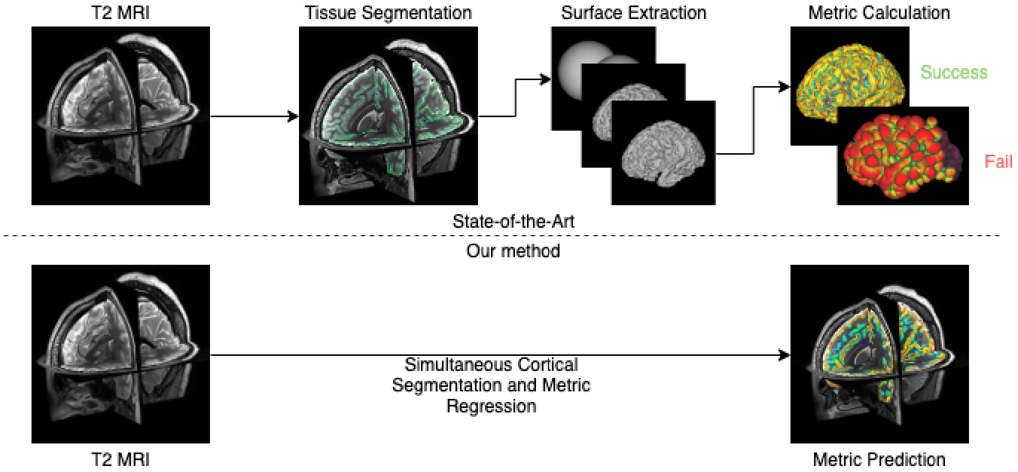


Figure 4.1: Overview figure of current state-of-the-art approach to extracting cortical metrics from MRI volumes showing a success and failure case vs. our method. The segmentation and surface extraction components of the pipeline are prone to fail, thus to produce artifacts as displayed in the fail case above.

4.3 Contributions

The key contributions of this chapter are: 1) We propose the first probabilistic DL tool for cortical segmentation and metric learning. 2) We evaluate the method against cortical thickness and curvature prediction for data from the Developing Human Connectome Project (dHCP); 3) The tool is used to predict cortical metrics for a Down Syndrome cohort in which surface-based analysis often fails; 4) Our probabilistic approach returns confidence maps which can inform clinical researchers of areas of the brain where measurements are less reliable.

4.4 Method

Segmentation and Regression Network: Our proposed network architecture augments the popular U-Net architecture into a 3D Multi-Task prediction network [18]. We introduce a branch of fully connected layers prior to the final convolution of the U-Net as shown in Figure 4.2. The network predicts a cortical segmentation through the standard U-Net architecture while simultaneously regressing a cortical metric value for every voxel in the image [191]. These two tasks are strongly coupled and as such we design the regression branch of our network to see a large amount of information from the segmentation path. We use a cross-entropy loss function for the segmentation task and consider two different loss functions for the regression task: Mean squared error (MSE) and Huber/L1 Loss, the latter is considered to encourage smoothness of regression predictions across neighbouring pixels by being robust to outliers [56], a property that should hold for cortical metric predictions.

We propose a probabilistic extension of our network in which we introduce Dropblock to our network during training and test time, this approach ensures variation in our network predictions during inference, this forms the baseline for our probabilistic experiments [192]. We propose an alternative probabilistic segmentation and regression architecture based on the PHiSeg network [175], which extends from the probabilistic U-Net [174] to model prediction variation across multiple scales and generate multiple plausible predictions for each input. We extend the PHiSeg architecture with fully connected layers in the same way we extended the initial 3D U-Net architecture to regress cortical metrics predictions for each voxel.

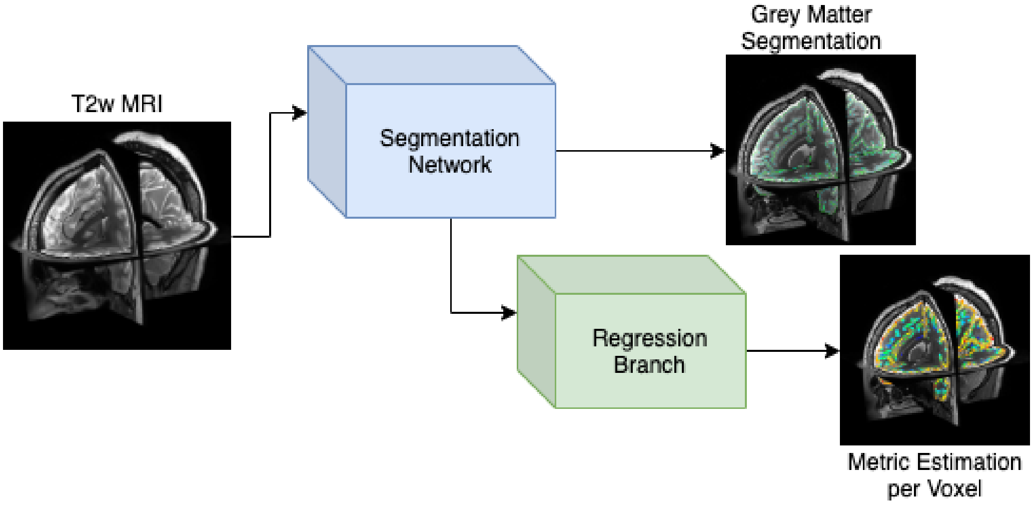


Figure 4.2: Architecture Diagram for the proposed simultaneous segmentation and regression network: The network predicts a cortical segmentation through the chosen segmentation architecture while simultaneously regressing a cortical metric value for every voxel in the image via the fully connected regression branch

We generate confidence maps for each prediction by sampling multiple times from each probabilistic network. This results in a range of segmentations and a range of metric predictions for each voxel. We argue that the larger the range of values predicted for a given voxel, the less confidence our network has in predicting a value for that voxel. We quantify the confidence of each prediction as the variance of each prediction made during inference. We seek the confidence map with the greatest correlation to prediction accuracy. From each range we can also define a metric 'Percentage in range' where we measure the percentage of voxels for which the ground truth value lies within the predicted range of values, where we seek the smallest ranges for which the ground truth is contained as introduced in Chapter 3.

4.5 Experimental Methods and Results

In this section we evaluate our proposed methods for automatic segmentation of the cortex and simultaneous estimation of the cortical metrics: thickness and curvature. We then evaluate the performance of our methods when used on a test set for which non-DL methods often fail and compare this with population wide distributions to evaluate our methods consistency when used on challenging populations in comparison to a widely used non-DL method.

Data: Data for this study comes from the Developing Human Connectome Project (dHCP) acquired in two stacks of 2D slices (in sagittal and axial planes) using a Turbo Spin Echo (TSE) sequence [193]. The used parameters were: TR=12s, TE=156ms, SENSE factor 2.11 (axial) and 2.58 (sagittal) with overlapping slices. The study contains 505 subjects aged between 26-45 weeks. Tissue segmentations and surface metrics for training are derived form the dHCP surface extraction pipeline [185][1].

A second clinical Down Syndrome cohort of 26 subjects was collected with a variety of scanning parameters, aged between 32-45 weeks, most subjects were acquired in the sagittal and transverse planes using a multi-slice TSE sequence. Two stacks of 2D slices were acquired using the scanning parameters: TR=12s, TE=156ms, slice thickness = 1.6 mm with a slice overlap = 0.8 mm; flip angle = 90° and an in-plane resolution: 0.8x0.8 mm.

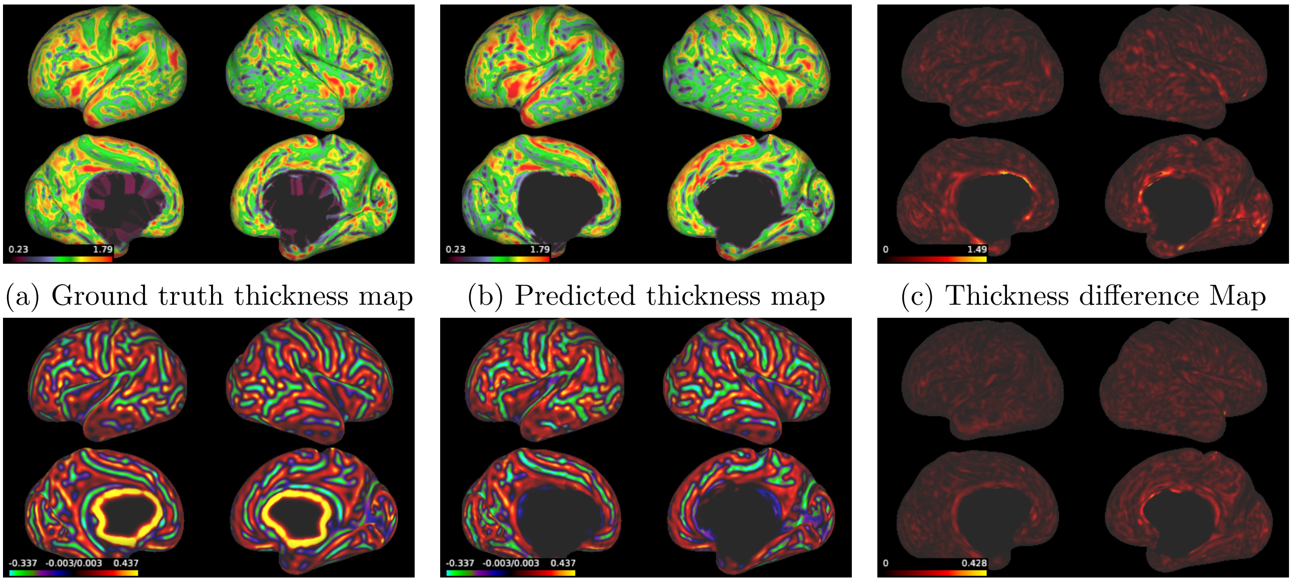
Preprocessing: We project surface-based representations of cortical biometrics into a volumetric representation using a ribbon constrained method[2] provided by the HCP project [194]. This operation is performed for both hemispheres of the brain and then combined into a single

[1] https://github.com/BioMedIA/dhcp-structural-pipeline

[2] https://www.humanconnectome.org/software/workbench-command/-metric-to-volume-mapping

volume, where overlapping metric values are averaged. These volumes (together with the tissue segmentations) represent the training targets for the learning algorithm. T2w input volume and corresponding metric volumes are then resampled to a isotropic voxel spacing of 0.5 to ensure a prediction of physically meaningful values for each subject, *i.e*, each voxel of our input image represents the same physical size in millimetres. Each T2 volume is intensity normalised to the range [0,1].

Training: 400 subjects are used for training; 50 for validation; 55 for test. During training we sample class-balanced 64x64x64 patches (N=12) from each subject's volume pair. We test on the entire volume of the image using a 3D sliding window approach. We conduct our experiments to predict two different cortical metrics: Thickness and Curvature.



(a) Ground truth thickness map (b) Predicted thickness map (c) Thickness difference Map

(d) Ground truth curvature map (e) Predicted curvature map (f) Curvature difference Map

Figure 4.3: Qualitative results on a dHCP subject: Here we have used ground truth surfaces to re-project our predicted metric volumes and difference map back into a surface representation for ease of comparison.

Deterministic experiments: We establish a baseline for simultaneous estimation of cortical segmentation and metric regression. Table 4.1 reports performance measures for all deterministic experiments. We find minimal difference in performance using the Huber loss function instead of MSE loss on the regression task module. Figure 4.3 show example outputs produced by our best performing model in comparison to the ground truth. Our network successfully extracts accurate cortical metric predictions directly from the input MRI, maintaining the

Experiment	Mean DICE ± std (%)	Mean error ± std	Median error ± std
Thickness (mm)			
UNetMSE	**0.946 ± 0.010**	**0.179 ± 0.025**	**0.125 ± 0.021**
UNetHuber	0.939 ± 0.012	0.197 ± 0.027	0.139 ± 0.022
Curvature			
UNetMSE	**0.945 ± 0.010**	**0.042 ± 0.005**	**0.030 ± 0.003**
UNetHuber	0.935 ± 0.011	0.045 ± 0.006	0.036 ± 0.004

Table 4.1: Results for metric prediction: Metric error calculated as average voxel-wise difference between prediction and ground truth only for voxels within the cortex. DICE score is reported over the segmentation.

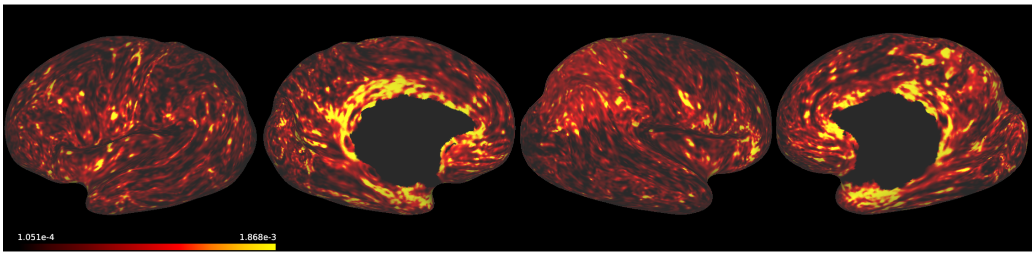


Figure 4.4: Confidence map generated with PHiSeg for a dHCP subject. Brighter areas indicate decreased confidence.

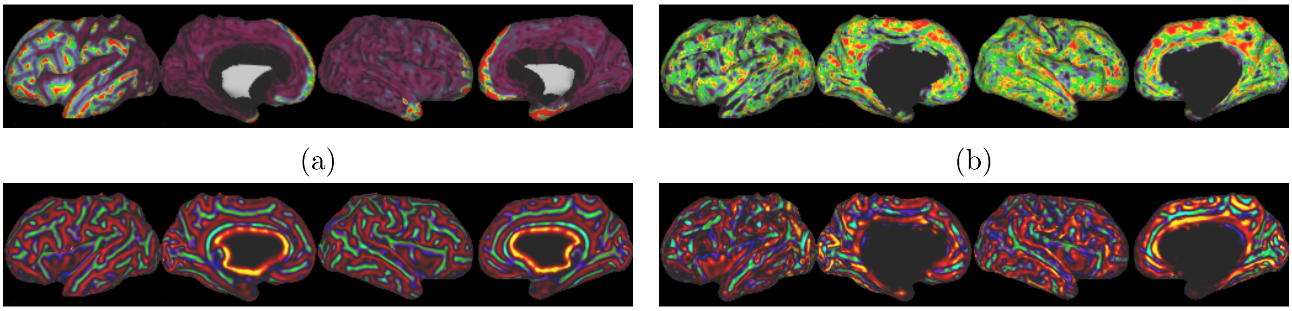
structural variation we expect across the cortex. We notice that some extreme values have not been accurately predicted, such as in cortical regions at the bridge between the two brain hemispheres (for curvature prediction). However the extreme values that are present in the ground truth data are an artifact of the surface based metric prediction method, hence it is less important to replicate this precisely. In Figure 4.6 we report test-set wide metrics comparing predicted global metric distributions in comparison to ground truth distributions, our method predicts a similar distribution of results to the ground truth.

Probabilistic experiments: We consider probabilistic extensions of our previous best performing method. Table 4.2 reports performance measures for all probabilistic experiments. We find that introducing DropBlock layers into our network has improved segmentation accuracy, but metric estimation accuracy has declined. PHiSeg has not improved either segmentation or metric estimation performance. In these experiments, PHiSeg training was often unstable, and took much longer to converge than other methods. While our error has increased, the ability to generate confidence maps for these predictions increases their value. Figure 4.4 shows a generated confidence map using PHiSeg.

Experiment	Mean DICE ± std (%)	Mean Mean error ± std	Mean Median error ± std	% in Range
Thickness (mm)				
UNetDropBlock	**0.945 ± 0.009**	**0.268 ± 0.017**	**0.373 ± 0.017**	**59.83%**
PHiSeg	0.774 ± 0.075	0.567 ± 0.096	0.550 ± 0.052	19.60%
Curvature				
UNetDropBlock	**0.946 ± 0.009**	**0.054 ± 0.004**	**0.071 ± 0.003**	**56.19%**
PHiSeg	0.796 ± 0.022	0.102 ± 0.006	0.102 ± 0.006	24.98%

Table 4.2: Probabilistic Prediction results: We show dice scores and the mean metric error when taking mean and the median of multiple (N=5) predictions as the final output.

Down Syndrome experiments: We evaluate the performance of our method on a challenging Down Syndrome MRI dataset for which the dHCP pipeline fails to extract metrics correctly for many subjects. Since the 'ground truth' for this dataset is error prone, we demonstrate the performance of our method qualitatively and through population comparisons to the healthy dHCP test set used in previous experiments. Figure 4.5 shows that our method produces more reasonable estimates of cortical thickness than the dHCP pipeline, thus showing evidence for our method's robustness to challenging datasets. Population statistics indicating that our method produces metric values in a sensible range for cortical thickness are shown in Figure 4.6. However our predicted distribution for cortical curvature is not consistent with healthy patients, indicating curvature and other more complex metrics remain challenging.



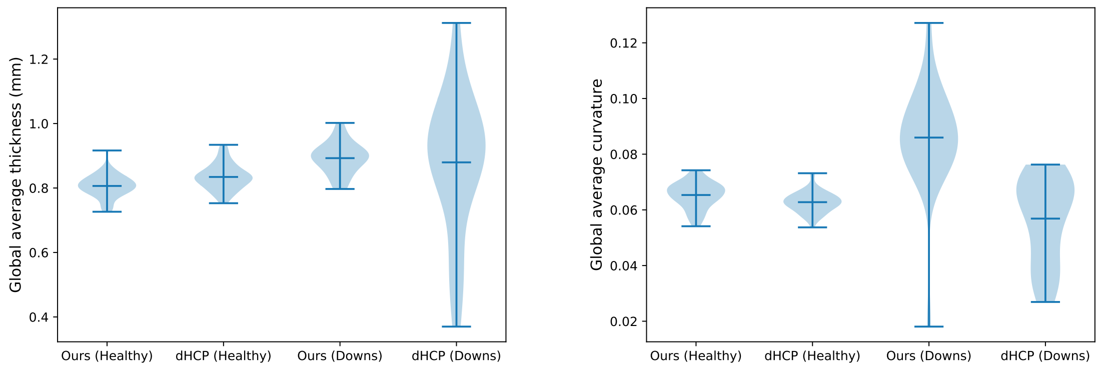
(a)

(b)

(c)

(d)

Figure 4.5: Qualitative results on Down Syndrome dataset: a) dHCP predicted thickness; b) Our method predicted thickness; c) dHCP predicted curvature; d) Our method predicted curvature.



(a) Thickness distributions

(b) Curvature distributions

Figure 4.6: Global average metric distributions a) Thickness and b) Curvature. Our method produces a sensible distribution of values for the Down Syndrome data set compared to the dHCP pipeline for thickness prediction. However curvature prediction remains challenging. Pathological cases (right two in each plot) cannot be quantitatively evaluated because of missing ground truth.

4.6 Discussion

We propose an automatic, pathology-robust method to predict cortical metric values from any T2w image stack. Our method is fast, robust and precise. We experiment with multi-task variants of the well known U-Net architecture and demonstrate how readily applicable DL is to predict cortical metric values in a reproducible way. Many architecture extensions are possible, which can for example be explored with neural architecture search methods [195]. In order to fully utilise the predictions of our network an imminent extension of our method is to automatically extract surface meshes from unseen data to enable proper visualisation of our predictions from images without ground truth surfaces.

Probably the biggest advantage of our method is that we can produce cortical measurements for pathological cases. However, the curvature example in Fig.4.6b shows that more complex metrics remain challenging and that we will need to further evaluate the robustness of our method in a clinical setting. Fine tuning may allow to generate disease-discriminative biomarkers directly from the network's latent space. At present our pipeline optimises cortical curvature and thickness prediction which naturally extends to sulcal depth and myelination prediction. There is potential to combine all metric predictions into a single model as it can be argued that prediction of different cortical properties are strongly correlated and would benefit each other as natural regularisers.

4.7 Summary

We offer an open-source framework for the extraction of cortical biometrics directly from T2 MRI images. This is the first of its kind that shows potential to be independent of age, image quality or presence of pathologies, and likely extendable to any cortical properties. We have tested our approach on a challenging pathological dataset for which we have not been able to reliably extract metrics with conventional methods like the dHCP processing pipeline. We expect that our future work will open new avenues for the analysis of cortical properties related to human brain development and disease in heterogenous populations.

Chapter 5

Classification Challenge

In this chapter we examine the ’Classification Challenge’ and develop methods by which complex classifications of disease can be made using DL models, but also provide means with which to interrogate those classifications and develop a reasoned understanding of how that classification has been made. We do this by breaking down the classification pipeline into interpretable parts that can be easily understood by users, without sacrificing accuracy of prediction.

We employ our methods in the specialised domain of fetal cardiac screening for Congenital Heart Disease detection, specifically Hypo-plastic Left Heart Syndrome. We choose this application as diagnosis is based upon the presentation of cardiac structure from a variety of ultrasound views during fetal screening, and as such, segmentation of cardiac structures can be leveraged as a means to acquire discriminative features known to sonographers from ultrasound views of the heart for diagnostic purposes.

Fetal ultrasound screening during pregnancy plays a vital role in the early detection of fetal malformations which have potential long-term health impacts. The level of skill required to diagnose such malformations from live ultrasound during examination is high and resources for screening are often limited. We present an interpretable, atlas-learning segmentation method for automatic diagnosis of Hypo-plastic Left Heart Syndrome (HLHS) from a single ‘4 Chamber Heart’ view image. We propose to extend the recently introduced Image-and-Spatial Transformer Networks (Atlas-ISTN) into a framework that enables sensitising atlas generation to disease. In this framework we can jointly learn image segmentation, registration, atlas construction and disease prediction while providing a maximum level of clinical interpretability compared to direct image classification methods. As a result our segmentation allows diagnoses competitive with expert-derived manual diagnosis and yields an AUC-ROC of 0.978 (1043 cases for training, 260 for validation and 325 for testing).

5.1 Introduction

Fetal Ultrasound (US) screening is a key part of ensuring the ongoing health of fetuses during pregnancy. Assessment of fetal development and accurate anomaly detection from US scans are integral in diagnosing potential fetal development issues at the earliest time possible to ensure the best care may be given. For these reasons a mid-trimester US scan is carried out between 18-22 weeks gestation in many countries as part of standard prenatal care procedures. During screening 'standard plane' views are used to acquire images in which key anatomical features may be examined, biometrics extracted and diagnosis of developmental issues may be made [172]. Several of these standard views and surrounding frames are used to make the diagnosis of Hypo-plastic Left Heart Syndrome (HLHS). Antenatal diagnosis of congenital heart disease such as HLHS has been shown to result in reduced mortality and morbidity of affected infants [196, 197]. Unfortunately, antenatal detection of HLHS is not universal, due to the high level of skill required to make the diagnosis accurately from often noisy and inconsistent US views, which vary with gestational age, among other factors.

Recently, automatic ultrasound US scanning methods have been developed using DL, mitigating the difficulties of manual US screening through automatic detection of diagnostically relevant anatomical planes [162]. These systems have enabled the development of robust automated methods for estimation of anatomical biometrics and diagnosis of fetal structural malformations such as HLHS, under diverse acquisition scenarios with various imaging artefacts. Critically, these methods still provide limited interpretability of predictions, and as such reasoning about diagnosis and appropriate interventions remains a challenge even in the presence of accurate predictions of anatomical features and diagnosis of development issues [198, 199, 200, 201].

5.2 Related work

Automated segmentation of anatomical structures in US images has been the topic of significant research, with CNN based methods [163, 164] often outperforming non-DL approaches [165, 166, 167]. Many of these methods perform well despite having no prior knowledge of the anatomical structure under consideration. However, in cases where performance drops, the resulting segmentations often bear no resemblance to the expected anatomical structure, resulting in segmentations that are not suitable for downstream analysis. As such, recent work to mitigate this fact has been introduced.

Methods such as Stochastic Segmentation Networks (SSNs) [169] aim to enforce continuity between anatomical structure segmentations (an assumption that holds in our case) to force a prediction to segment structures such that they remain connected and allow for sampling multiple plausible solutions to any given image segmentation. Similarly, [202, 203] introduces topological priors to enforce continuity between segmented regions. Another recent approach aims to automatically learn an atlas of the anatomical structure under consideration during training of a segmentation model. Predicting both an image segmentation and a transformation between the automatically constructed atlas and the predicted segmentation, forces the resulting segmentation to retain the expected anatomical structure. In the presence of imaging artefacts or other image features the above behavior may result in a worse segmentation performance [19]. The aforementioned methods provide accurate segmentations of anatomical structures familiar to sonographers, however at present, these are not used to perform diagnosis of CHD or provide any means for disease-specific conditioning.

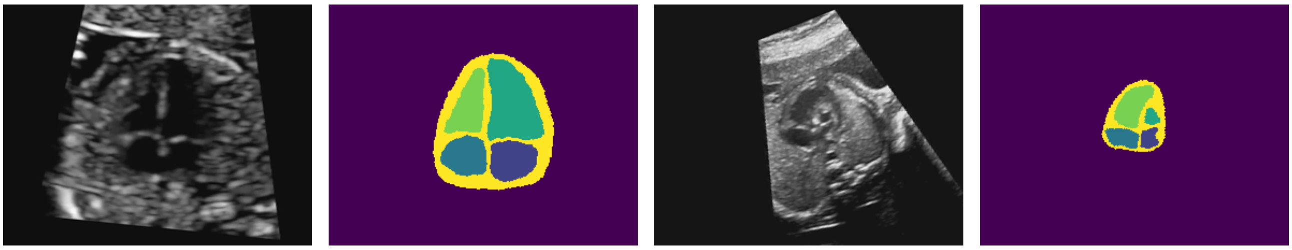
Deep Ultrasound Classification is currently the only option that has been explored in literature to perform automated diagnosis of CHD directly from US images. Deep classification methods achieve high accuracy [204, 205, 198], but rely on large curated datasets [205] or additional views of the heart to support multitask learning [198]. Unfortunately, conventional image classification is difficult to apply to fetal ultrasound because only very specific “standard planes” contain sufficient diagnostic information [172].

Classifying non-diagnostic frames (that should not be considered healthy or diseased) could lead to erroneous diagnosis or obscure the signal from the true diagnostic frames. As such, direct classification models have very little utility if clinicians cannot clearly interpret and assess the validity of the classification or find a view in pathological cases that would correspond to the defined anatomical standard [172].

5.3 Contributions

In this chapter we introduce a novel method for the diagnosis of HLHS from US images using pathology-robust segmentation. To the best of our knowledge, we present for the first time a segmentation network that is able to jointly segment, register and build a labeled atlas that focuses on relevant features to robustly diagnose HLHS for fetal 4-chamber views in ultrasound imaging. By extending the recently proposed Atlas-ISTN framework [19] with an additional classification module and corresponding component in the loss function, our method provides an interpretable and accurate option for HLHS diagnosis compared to direct image classification approaches through segmentation of anatomical structures known to sonographers.

We evaluate the quality of our segmentations in the downstream task of inferring HLHS status from ventricular areas. From ground truth expert annotations we evaluate the possible correlation of this approximation to true disease status, which is confidently known from post-natal outcome records. We compare this correlation to using naive segmentation for area parameter extraction and our proposed method.



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

Figure 5.1: Example ultrasound images and manual segmentations of anatomical areas. (a) Healthy patient 4CH ultrasound view; (b) manual segmentation of anatomical areas of healthy heart in (a); (c) HLHS patient's approximation of a 4CH ultrasound view; (d) manual segmentation of anatomical areas in (c).

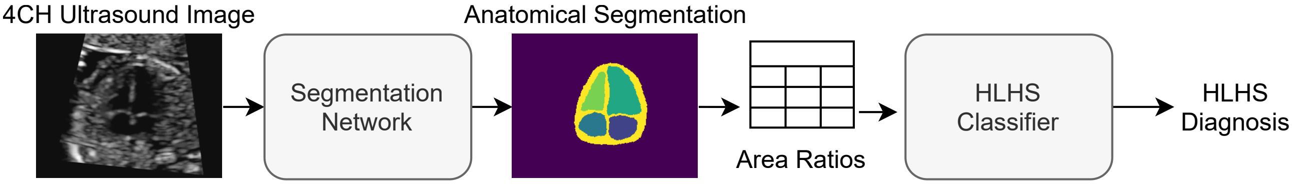


Figure 5.2: Classification via deep segmentation and quantitative area ratio feature extraction

5.4 Method

We propose a new method for the automatic diagnosis of HLHS from a single US image of the '4-Chamber Heart View' (4CH view). Our system is inspired by current clinical practice and can be broken down into three major modules. First, for a given 4CH image, we seek a model that can provide accurate segmentations for 5 anatomical areas: *'Whole Heart'*, *'Left Ventricle'*, *'Right Ventricle'*, *'Left Atrium'* and *'Right Atrium'*. Figure 5.1 shows an example for the differences in these areas between a healthy fetus and a baby with HLHS.

Secondly, the resulting segmentation is used to extract simple image features informative of HLHS diagnosis. For each class, we calculate the ratio of that region's area to the area of every other segmented region to obtain a set of scalar features representative of that image. Finally, we use the quantitative features to construct a classifier for HLHS using: 1) class-weighted Logistic Regression Classifier as baseline; 2) Gaussian Process classifier with a radial basis function kernel. Figure 5.2 shows an overview over the diagnostic approach.

For the task of **Robust segmentation**, we adopt a recently proposed method, Atlas-ISTN [19], that generates a segmentation as well as learns a label atlas that both ensures robustness and

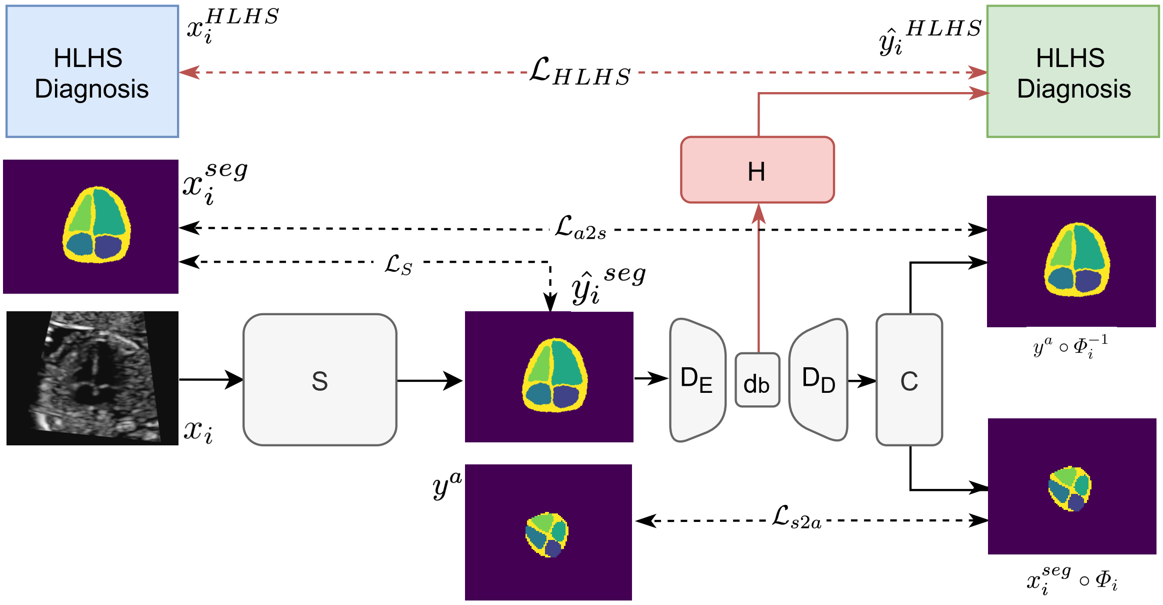


Figure 5.3: Disease conditioned Atlas-ISTN network architecture. **S** the segmentation network; **D** indicates the atlas to image mapping module; **C** is the transformation computation Module; and **H** is a disease prediction branch, highlighted in red, which is the key difference to [19]. In this figure, L_{HLHS} indicates the loss computed on the disease prediction branch H between x_i^{HLHS} (the ground truth label) and $\hat{y}_i^{HLHS}$ (the predicted label); L_s indicates the segmentation loss computed on the segmentation network S between x_i^{seg} (the ground truth segmentation) and $\hat{y}_i 6 seg$ (the predicted segmentation); L_{a2s} (atlas to segmentation loss) indicates the loss computed between x_i^{seg} (the ground truth segmentation) and $y^a \circ \phi_i^{-1}$ (the current atlas deformed to image space); and L_{s2a} (segmentation to atlas loss) indicates the loss computed between y^a (the current segmentation atlas) and $x_i^{seg} \circ \phi_i$ (the ground truth segmentation deformed to atlas space).

can be used to inspect the inner beliefs of the network. Conditioning is used to sensitise the atlas generation to regions that are most relevant for the downstream task of disease classification.

Our detailed model is outlined in Figure 5.3. As input we use uncropped 4-chamber ultrasound images, ground truth segmentation maps and a binary disease label in

$$X = \{x_i, x_i^{seg}, x_i^{HLHS}\}.$$

The model aims to learn:

$$\{\hat{y_i}^{seg}, \hat{y_i}^{HLHS}, y^a\} = \mathbf{M}(x_i),$$

where $\mathbf{M}$ is the entire model, $\hat{y_i}^{seg}$ are the logits of a predicted segmentation describing five cardiac labels with one background channel as defined in $x_i{}^{seg}$, $\hat{y_i}^{HLHS}$ are probabilities for discrete disease categories in ${x_i}^{HLHS} \in [0, 1]$, and y^a is an automatically optimised atlas label map. **M** consists of four modules. Image to segmentation mapping is obtained through

$$\hat{y_i}^{seg} = \mathbf{S}_{\theta_s}(x_i),$$

which we define as a 2D UNet [18] with a SSN module [169]. The concatenation of $\hat{y_i}^{seg}$ and the atlas label map y^a is used to establish the atlas to image transformation:

$$\mathbf{d}_{b_i} = \mathbf{D}_{E,\theta_{enc}}(\hat{y_i}^{seg}, y^a),$$
$$\{v_i, T_i\} = \mathbf{D}_{D,\theta_{dec}}(\mathbf{d}_{b_i}).$$

v_i is a stationary velocity field and T_i an affine transformation matrix that is processed to a deformation field with a Transformation Computation Module **C** according to [19]. Thus, **C** yields forward and inverse transformations, Φ_i and Φ_i^{-1}. To steer the atlas generation process and emphasise disease-relevant labels, we predict:

$$\hat{y_i}^{HLHS} = \mathbf{H}_{\theta_\mathbf{h}}(\mathbf{d}_{b_i})$$

where $\mathbf{H}_{\theta_h}$ are three fully connected layers with ReLU activations. Additionally to the image transformer loss:

$$\mathcal{L}_S = \frac{1}{N}\left(\sum_{i=1}^{N} ||{x_i}^{seg} - \hat{y_i}^{seg}||^2\right),$$

the atlas-to-segmentation loss:

$$\mathcal{L}_{a2s} = \frac{1}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{c} ||{x_{i,j}}^{seg} - {y_j}^a \circ \Phi_i^{-1}||^2\right),$$

and the segmentation-to-atlas loss:

$$\mathcal{L}_{s2a} = \frac{1}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{c} ||{x_{i,j}}^{seg} \circ \Phi_i - {y_j}^a||^2\right),$$

where j indicates the individual labels, we introduce a cross entropy loss term:

$$\mathcal{L}_{HLHS} = -(x_i^{HLHS} \log(\hat{y_i}^{HLHS}) + (1 - x_i^{HLHS}) \log(1 - \hat{y}_i^{HLHS}))$$

to enforce disease-sensitive atlas generation. Thus, our final loss function is the Atlas-ISTN loss [19] with its regularisation loss term, $\mathcal{L}_{reg} = \sum_i^N ||\nabla\phi||^2$, that encourages smoothness of the non-rigid deformation fields, paired with $\mathcal{L}_{HLHS}$:

$$\mathcal{L} = \mathcal{L}_S + \omega(\mathcal{L}_{a2s} + \mathcal{L}_{s2a} + \lambda\mathcal{L}_{reg}) + \gamma\mathcal{L}_{HLHS},$$

where λ adjusts smoothness of Φ, ω influences the contribution of the deformation terms similar to [19], and γ steers how much the atlas should be specific to the targeted disease category.

HLHS classification from fetal cardiac 4-chamber view segmentations: We extract numerical features from $\hat{y_i}^{seg}$ in order to classify HLHS vs. Healthy patients from interpretable features $f = \{f_0, f_1, ..., f_N\}$ where $f_i = r_{ab} = A_a/A_b$ if $a \neq b$ and r_{ba} is not in f already. We represent the ratio between two quantities as r_{ab} and consider r_{ab} and r_{ba} to contain equivalent information and as such exclude the latter from f. Here A_a is the count of pixels belonging to class a in $\hat{y_i}^{seg}$ which acts as an estimate to the area.

We apply two common classification algorithms to classify the extracted segmentation area ratio features as healthy vs HLHS. We first use an L2 regularised, class weight balanced Logistic regression classifier implementation. Secondly we use a Gaussian Process classifier based on Laplace approximation [206].

The Atlas-ISTN method is very sensitive to the alignment of input images and segmentations in the training set. Initial experiments showed the constructing an atlas with images and segmentations with the heart located arbitrarily across the image, with arbitrary rotation and scaling resulted in a meaningless atlas image bearing no resemblance to any single image in the dataset. As such, we performed a Procrustes alignment of the images to bring each heart to the centre of the image, with a similar rotation and scaling. Using manual line annotations of the

’Apex Base’ and ’Spine-Sternum’ we were able to construct suitable shape information sufficient for Procrustes alignment. We use classical Procrustes alignment where we manually selected a reference shape to align each image to, as opposed to generalised Procrustes alignment which infers a mean shape to align from all available shapes [207].

5.5 Experiments and Results

In this section we evaluate the performance of our proposed method in both segmentation and classification tasks. We evaluate our methods segmentation performance against an expert derived ground-truth segmentation. We evaluate our methods classification performance against the classification performance of our method when applied to the expert derived ground-truth segmentations.

Data and Pre-processing: We use a private, de-identified dataset of 1628 4CH US images (1560 healthy controls, 68 HLHS), with 1043 for training, 260 for validation, 325 for testing with equivalent class imbalance within each set (42, 10 and 16 HLHS cases respectively), acquired on Toshiba Aplio i700, i800 and Philips EPIQ V7 G devices. Class imbalance reflects the prevalence of HLHS observed in our tertiary care referral clinic ($\sim 3-4\%$), which is a specialised centre, thus the incident rate is relatively high. HLHS is rare, $\sim$ 3 in 10000 live births [208], thus this condition can be challenging to identify for primary care sonographers. Our images are taken from volunteers at 18-24 weeks gestation, acquired in a fetal cardiology clinic, where patients are given advanced screening due to their family history. Each image has been hand-picked from ultrasound videos by an expert sonographer, representing a best possible 4CH view. A fetal cardiologist and three expert sonographers delineated the images using Labelbox [209]. The images have been resampled to 288×224 pixels, centred on the heart and aligned along the cardiac main axis.

Robust segmentation We compare several methods for automated segmentation by average DICE score achieved for each anatomical class and summarise the results in Table 5.1. We show that each of the compared methods is effective for the segmentation of anatomical structures from ultrasound image views. The question remains, which method produces the most informative segmentation for downstream disease diagnostics?

Expert derived single image classification: To establish human performance on the segmentation task, heart segmentation features (area ratios of each anatomical class) are extracted from the manual ground truth segmentations. These are used to train a linear classifier and a Gaussian process classifier to predict HLHS diagnosis. We report the confusion matrices for each method using the ground truth segmentations shown in Figure 5.5. Table 5.1 reports F1-score for positive and negative HLHS classification as well as ROC-AUC for manual as well as automated segmentation.

F1 and AUC scores in Table 5.1 show that our 'Area Ratios' classification method achieves state-of-the-art performance for HLHS classification over previous classification methods. Classification performance of 'area ratios' extracted from automated segmentations is on par with those extracted from expert manual segmentations. The addition of a disease-conditioned branch to the Atlas-ISTN improves the downstream 'area ratios' classification task performance over both expert segmentations and previous segmentation methods.

Figure 5.5 shows the performance of the 'area ratios' classification using segmentations produced by experts and by each tested segmentation method. Subfigures (5.5e-5.5f) and (5.5k-5.5i) highlight the improved sensitivity (fewer false negatives) of Atlas-ISTNs with a disease conditioning branch over expert segmentations (5.5a,5.5g) and other segmentation methods (5.5b-5.5d, 5.5h-5.5j). Our application is for fetal screening and as such sensitivity is the desired metric to improve, and due to the low prevalence of HLHS, F1 scores for HLHS across all methods may seem low.

Table 5.1 shows our diagnostic branch (**H**) is competitive with previous image classification

	DICE Score							F1 Score		ROC
Method	**BG**	**LA**	**RA**	**LV**	**RV**	**WH**		**NC**	**HLHS**	**AUC**
Expert	1.000	1.000	1.000	1.000	1.000	1.000	**LR:**	0.970	0.550	0.944
(Std)	–	–	–	–	–	–	**GP**	0.989	0.741	0.954
UNet [18]	0.993	0.768	0.804	0.793	0.794	0.635	**LR:**	0.972	0.585	0.922
(Std)	(0.007)	(0.192)	(0.185)	(0.184)	(0.153)	(0.105)	**GP:**	0.974	0.579	0.928
SSN [169]	0.993	0.761	0.800	0.793	0.794	0.632	**LR:**	0.955	0.471	0.883
(Std)	(0.007)	(0.196)	(0.192)	(0.194)	(0.154	(0.108)	**GP:**	0.974	0.579	0.923
$Atlas^{\gamma=0}_{\lambda=1}$	0.991	0.767	0.789	0.801	0.783	0.626	**LR:**	0.942	0.451	0.895
(Std)	(0.007)	(0.187)	(0.192)	(0.191)	(0.172)	(0.106)	**GP:**	0.981	0.625	0.970
$Atlas^{\gamma=1}_{\lambda=10^3}$	0.993	0.764	0.789	0.791	0.790	0.648	**LR:**	0.958	0.528	0.929
(Std)	(0.007)	(0.185)	(0.184)	(0.196)	(0.146)	(0.110)	**GP:**	0.974	0.619	0.973
	–	–	–	–	–	–	**H** :	0.967	0.565	0.883
$Atlas^{\gamma=1}_{\lambda=1}$	0.993	0.760	0.783	0.784	0.788	0.637	**LR:**	0.950	0.500	0.974
(Std)	(0.007)	(0.197)	(0.200)	(0.208)	(0.164)	(0.110)	**GP:**	0.974	0.636	0.978
	–	–	–	–	–	–	**H** :	0.982	0.667	0.905

Table 5.1: DICE scores and standard deviation (Std) for all segmentation methods (left) and performance of downstream disease predictors (right). (**BG** = background; **LA** = left atrium; **RA** = right atrium; **LV** = left ventricle; **RV** = right ventricle; **WH** = whole heart; **LR** = Logistic regression; **GP** = Gaussian process; **H** = disease prediction branch; **NC** = Normal control; **HLHS** = hypo-plastic left heart syndrome.)

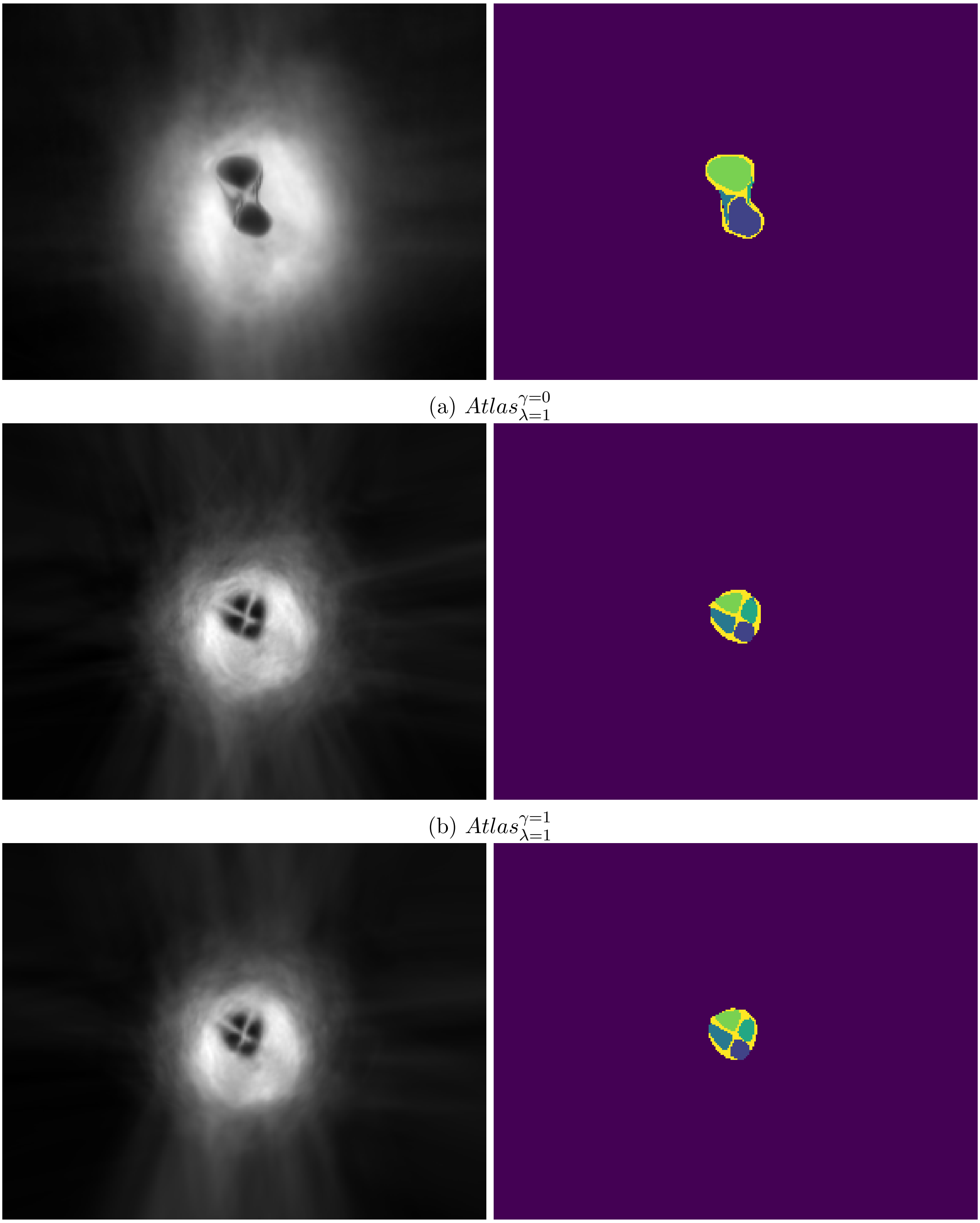
approaches, further to this our method uses only a single 4CH image as opposed to previous methods that use multiple heart view US images or video sequences. Our method provides greater interpretability by producing a segmentation (from which 'area ratios' classification is performed) and a disease specific atlas for free.

Examples for constructed atlases with different configuration are shown in Figure 5.4.

Implementation: PyTorch 1.7.1+cu110 with two Nvidia Titan RTX GPUs used to train segmentation and atlas models ($\sim 10^6$ parameters) in 24-48 hours; scikit-learn [210] for the LR and GP models.

5.6 Discussion

Assuming a linear downstream model for clinical decision making, our results show that automated segmentation methods are en-par with human-generated annotations for the accurate

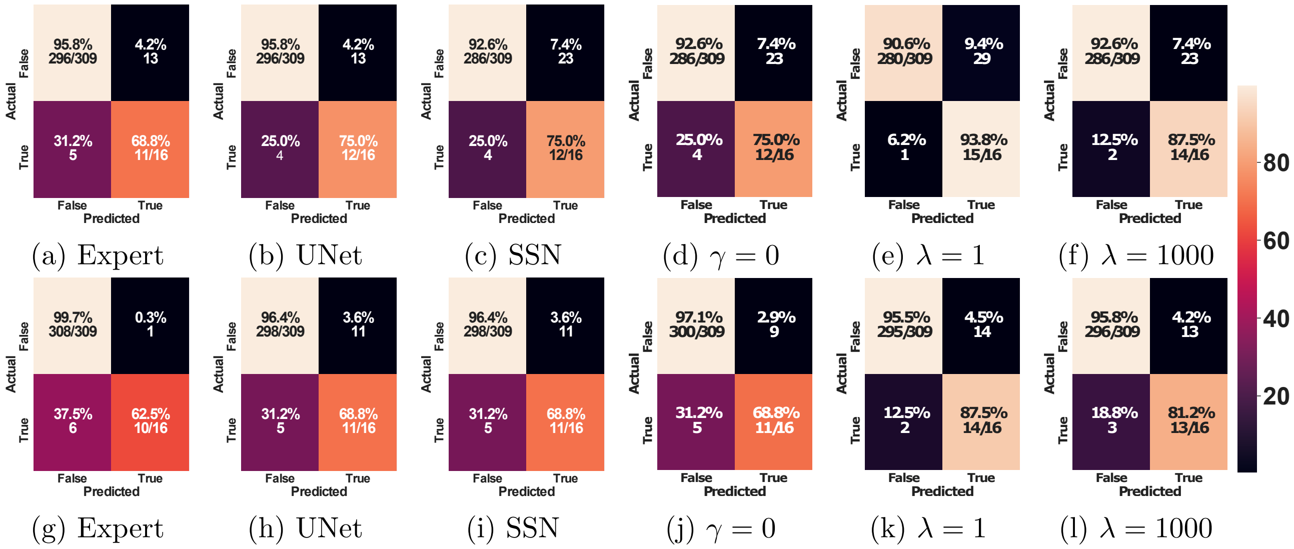


(a) $Atlas_{\lambda=1}^{\gamma=0}$

(b) $Atlas_{\lambda=1}^{\gamma=1}$

(c) $Atlas_{\lambda=10^3}^{\gamma=1}$

Figure 5.4: Example automatically constructed atlas images and atlas label maps in the tested configurations.



(a) Expert (b) UNet (c) SSN (d) $\gamma = 0$ (e) $\lambda = 1$ (f) $\lambda = 1000$

(g) Expert (h) UNet (i) SSN (j) $\gamma = 0$ (k) $\lambda = 1$ (l) $\lambda = 1000$

Figure 5.5: Confusion matrices for expert derived classification: Top Row shows logistic regression, bottom row shows Gaussian Process. $\gamma = 1$ for (e-f)(k-l).

identification of HLHS patients. We believe the reason for higher than expert performance is due to more consistent automated segmentation results that aid the following linear model in contrast to predicting from ground truth segmentations, which have been generated by different observers. An interesting observation is that a reasonable DICE score is sufficient to achieve excellent performance in diagnostic follow-up tasks.

A limitation of our study is that we require input images that resemble a 4CH acquisition orientation in a healthy subject. This can be challenging for severely affected patients. However, for cases with severely abnormal hearts, manual detection of CHD would likely be trivial at the point of care, also without segmentation analysis. Good views for borderline cases, which are in focus here, can be identified either manually or with automated view classification [162].

For this work we rigidly aligned all the data to a canonical orientation relative to the heart. This can be achieved in the clinical practice through automated localisation/segmentation/spatial transformer approaches. We observed that this data curation step has a significant impact on all models' performances compared to unaligned images, in which fetuses may present in arbitrary orientation. Accounting for flipped probe orientations paired with hyper-parameter tuning for ω, λ, γ would likely lead to further improvements.

Another limitation is that we do not consider inherent spatio-temporal information of ultrasound imaging. Experienced fetal cardiologists can derive valuable secondary information from how the heart moves. This knowledge can inform future work on the topic. In the clinical practice, still images, as used in our work are common practice to report and document cases, thus a direct application to retrospective quality control and diagnosis support for primary care is in reach.

5.7 Summary

We have discussed how segmentation models can be used as clinically interpretable alternative to direct image classification methods for the diagnosis of hypo-plastic left heart syndrome during routine ultrasound examinations. We test a new approach that facilitates disease-status information to bias an automatically constructed atlas label map for robust segmentation and apply Atlas-ISTNs to the problem of fetal cardiac segmentation from ultrasound images for the first time. Our analysis shows that our interpretable approach is en-par with direct image classification, for which ROC-AUC of up to 0.93 is reported [198]. Future work will investigate the true effectiveness of such methods in a prospective clinical trial, which is currently implemented in our clinic.

Chapter 6

Curation Challenge

In this chapter we examine the 'Curation Challenge' and test key assumptions that are commonplace within medical image analysis regarding how we should curate annotated medical image datasets for supervised DL problems.

Supervised DL dominates performance scores for many computer vision tasks and defines the state-of-the-art. However, medical image analysis lags behind natural image applications. One of the many reasons is the lack of well annotated medical image data available to researchers. One of the first things researchers are told is that we require significant expertise to reliably and accurately interpret and label such data. We see significant inter- and intra-observer variability between expert annotations of medical images. Still, it is a widely held assumption that novice annotators are unable to provide useful annotations for use by clinical DL models. In this work we challenge this assumption and examine the implications of using a minimally trained novice labelling workforce to acquire annotations for a complex medical image dataset. We study the time and cost implications of using novice annotators, the raw performance of novice annotators compared to gold-standard expert annotators, and the downstream effects on a trained DL segmentation model's performance for detecting a specific congenital heart disease (hypoplastic left heart syndrome) in fetal ultrasound imaging.

6.1 Introduction

It is commonly believed that domain experts are the only reliable source for annotating medical image data. This assumption has resulted in a dearth of annotated medical image datasets due to the time and high costs associated with expert labelling time. In this chapter we challenge this assumption and employ novice annotators to perform a complex multi-class fetal cardiac ultrasound (US) segmentation task.

A core goal of medical image analysis is to free up experts' time for more challenging tasks and time with patients. Our current view is that expert annotation efforts that aid in the development of models, will save expert time in the long term. However we hypothesise that in many cases, this annotation effort can be performed by novice annotators at a lower cost, saving both resources and experts' time, with minimal impact on the performance of automated downstream models.

Segmentation is widely regarded as among the most labour intensive medical image analysis tasks, requiring pixel-level labels to enable supervised learning methods to learn complex segmentation tasks. In this study we use a multi-class fetal cardiac US segmentation task as our initial test case, as this task is challenging in both anatomy and modality (noisy, heterogeneous and often contains artefacts). This makes the task of annotating fetal US images challenging for both experts and novices, and an ideal test case for comparing the efficacy of novice annotations. Segmentation of the fetal heart from '4-Chamber view' images provides quantitative biomarkers that can be used for the diagnosis of Hypoplastic Left Heart Syndrome (HLHS). As such we include in our dataset several HLHS cases. The presence of pathology within our dataset makes this annotation task even more challenging, and enables us to compare the performance of novice and expert annotations on a segmentation-informed diagnostic classification task.

We provide evidence that the reliability of novice annotators is greater than expected and that this approach might be a viable option for annotation of medical image datasets in the future.

6.2 Related work

Significant work has been done to mitigate for a lack of well-annotated medical imaging data. Learning from fewer labels, unsupervised learning and AL are all valuable contributions in this and their benefits go beyond the scope of this chapter. Advances in these fields can only benefit from the increasing sizes of annotated medical image datasets, and as such we do not challenge these approaches. More tightly related to our work are methods for learning from crowd-sourced noisy labels, where annotations of varying quality are acquired [211, 212, 29, 147, 33, 137]. While there are many works related to crowdsourcing medical image annotations, very few exist that directly compare the quality and downstream implications of crowd-sourced novice vs expert annotations, and it is this gap in literature that we seek to address.

In [213] it is shown that novice annotators are comparable to expert annotators for a series of natural language annotation tasks, and that only a small number of novice annotations are necessary to equal the performance of expert annotators. In [214] it is shown that novice annotators are able to effectively prune non-informative text from training data for sentiment classifiers to improve classification performance of trained models.

In [215] it is shown that crowd-sourcing many noisy labels for heavily class imbalanced text classification datasets is expensive and the usual benefits of redundant labelling seen in crowd-sourcing scenarios is lesser in imbalanced settings. [215] provide techniques for discarding redundant instances such that annotations can be acquired in a cost-effective way over a five-way majority vote aggregation.

In [216] the authors evaluate the effects of aggregating progressively more labels per instance on model performance for mitotic figure detection from histologic images. They show that high accuracy can be achieved with a single annotation per image, and improved by aggregating three annotations per image, while aggregating beyond three annotations per image results in only minor very minor performance increases.

In [217] criteria are proposed by which the suitability of a text sentiment classification task for crowdsourcing can be evaluated (1. Noise level, 2. Inherent Ambiguity and 3. Informativeness

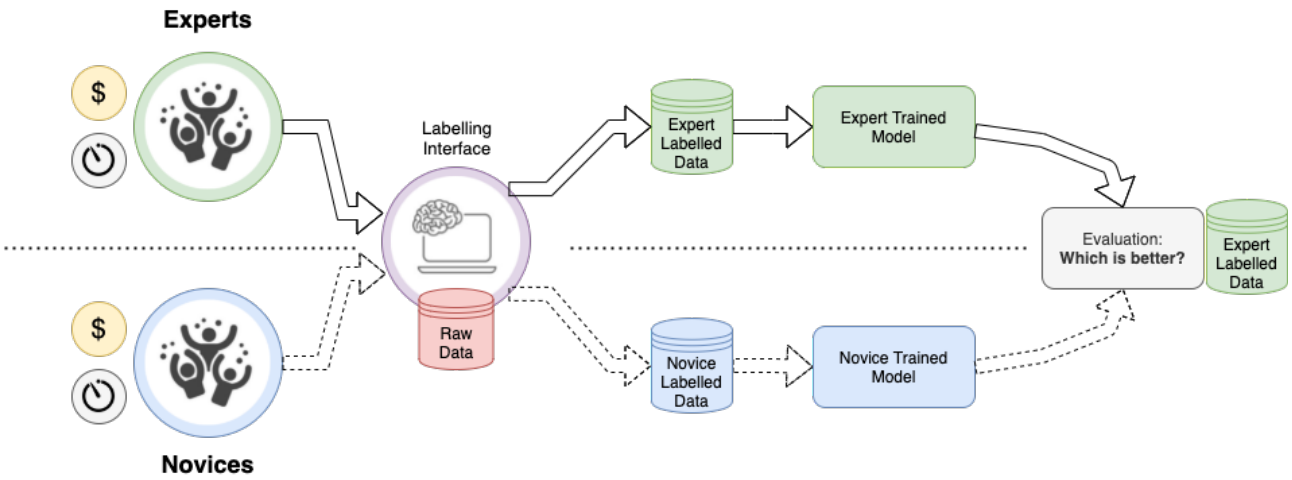


Figure 6.1: Graphical overview of our process: We evaluate the upstream and downstream impacts of using novice annotations in place of expert annotations on a challenging medical image segmentation task. Each set of annotations is acquired, pre-processed and used to train models in the same way. We evaluate both expert and novice models against an expert annotated test set.

to the model). Models trained on expert and novice annotations are compared. By considering the three proposed criteria, it is shown that comparable model performance can be achieved using expert or novice annotations.

For a 3D segmentation correction task, there is evidence for little to no difference between novices and expert performance (engineers with domain knowledge, medical students, and radiologists) in the ability to detect and correct errors made by a segmentation algorithm [218], although novice annotators need significantly more time per annotation.

6.3 Contributions

We evaluate the upstream and downstream impacts of training medical image multi-class segmentation models, and downstream classification models on noisy labels from novice annotators compared against gold-standard labels from expert annotators.

We show that novice annotators are capable of performing complex medical image annotation tasks to a high standard, and that variability between novices and experts is comparable to that amongst experts themselves. We show that models trained on novice labels are comparable to those trained on expert labels for multi-class segmentation and downstream classification.

We analyse the time and costs associated with using expert vs. novice labels to show that using novice annotations is more resource efficient, and that the major parameter governing model performance is dataset size, rather than label quality in this setting. This will enable clinical and translational researchers to develop a greater understanding of the trade-offs associated with acquiring medical image annotations with respect to cost, time and supervised learning method performance.

6.4 Method

Annotation Labels collection: The current paradigm for collecting annotations for medical image data is to present experts with un-annotated data in an annotation interface that allows them to delineate structures of interest in every image (Figure 6.1). Once complete, the annotations and input can be exported for use. In this work we employ novice annotators to perform the same task using the same annotation tools on the same data to provide us with novice annotated for later use, as shown in the bottom half on Figure 6.1. We use the Labelbox web-based interface as our annotation tool [209].

Segmentation model: From a single US image of the '4-Chamber Heart View' (4CH view) acquired during fetal screening, we train a model to delineate 5 anatomical areas: *'Whole Heart'* (WH), *'Left Ventricle'* (LV), *'Right Ventricle'* (RV), *'Left Atrium'* (LA) and *'Right Atrium'* (RA) (Figure 6.2).

We use the UNet architecture as our segmentation network [18], known to perform well for US segmentation. We train using dropout [94], for a fixed number of epochs, then select the best performing model on the validation set. Random horizontal and vertical flipping, cropping,

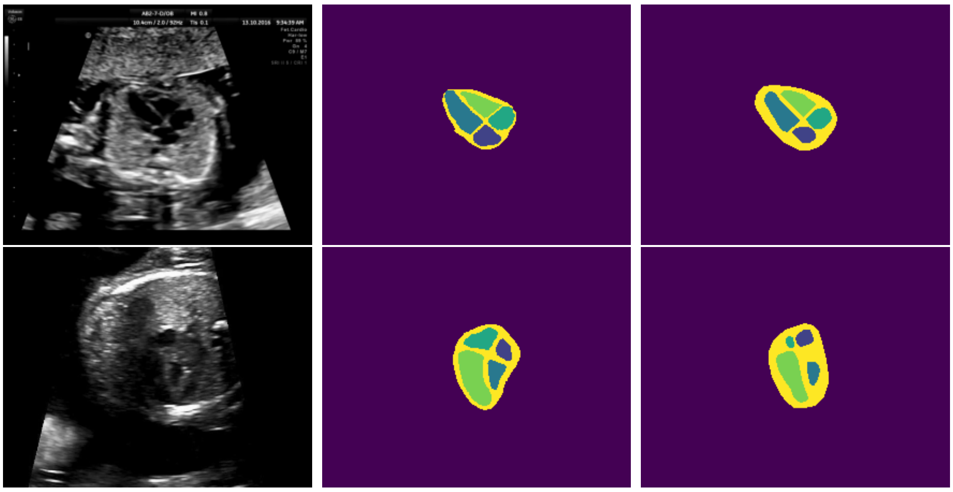


Figure 6.2: Example US images and manual segmentations of anatomical areas. Top row: Healthy image, expert manual label and novice manual label (left to right). Bottom row: HLHS image, expert manual label and novice manual label (left to right)

translation, rotation and scaling is applied during training.

Classification model: We extract numerical features from $\hat{y}_i^{seg}$ (manual or automated segmentation) in order to classify HLHS vs. healthy patients from interpretable features $f = \{f_0, f_1, ..., f_N\}$ where $f_i = r_{ab} = A_a/A_b$ if $a \neq b$ and r_{ba} is not in f already. Here r_{ab} is the ratio between two quantities and consider r_{ab} and r_{ba} to contain equivalent information and exclude the latter from f. A_a is the count of pixels belonging to class a in $\hat{y}_i^{seg}$ which acts as an estimate to the area as in Chapter 5.

We apply an L2 regularised, class weight balanced logistic regression classifier implementation to classify the extracted segmentation area ratio features as healthy vs. HLHS.

Statistical analysis: Here we pose the questions answered in this paper and outline our approach to answering them. For tests of statistically significant difference between distributions we use a two-tailed Z-test for the null hypothesis of identical means:

$$Z = \frac{\hat{X} - \mu_0}{s}$$

where Z is our test statistic, μ is our population mean and s is our population standard deviation.

Q1: Are novice annotations as similar to experts as expert annotations are to other experts? We answer this question by computing the average DICE similarity coefficient between novice and expert annotations, and between pairs of expert annotations. We calculate the Dice score for each class separately, and test for statistically significant difference between the two sets.

Q2: How different are automated segmentations trained on experts annotations to automated segmentations trained on novice annotations? We show evidence by computing the average DICE similarity coefficient between novice and expert trained model predictions and an expert annotated test set. We calculate the DICE score for each class separately, and test for statistically significant difference between the two sets.

Q3: How different are classification predictions trained on either manual, or model based segmentations from novices compared to experts? We train a classifier using training data from manual expert and manual novice annotations, as well as expert model and novice model predictions. We test each classifier on our expert test set and compare key performance metrics to evaluate the discrimitive powers of novice vs. expert based segmentations.

Q4: In resource limited scenarios are expert or novice annotations more cost effective to attain the same model performance? We observe the time/cost/quality trade-off by measuring the DICE scores obtained by models trained using novice and expert data on progressively more labels (50 to 1000 labels) using a UNet with 200 epochs. We use DICE scores on the test set as a measure of prediction quality, and use time taken and estimated financial cost to acquire each annotation, to plot the time vs. cost vs. performance of our models for both experts and novice annotations.

6.5 Experiments and Results

Data and Pre-processing:

- Raw images: We use a private and ethics/IP-restricted, de-identified dataset of 2380 4CH US images, with 1000 for training, 380 for validation, 1000 for testing acquired on Toshiba Aplio i700, i800 and Philips EPIQ V7 G devices.
- Expert segmentations: A fetal cardiologist and three expert sonographers delineated the images using Labelbox [209]. Multiple expert annotations for 319 images were acquired.
- Non-expert segmentations: A novice workforce with no experience annotating medical US data was employed to delineate the images using Labelbox [209]. Three novice annotations for every image in the training set were acquired. An unknown number of novice annotators were used to acquire the image annotations, however it was enforced that no annotator annotate the same image twice. Each annotator was provided with an instruction sheet as shown in Appendix A.
- Time: Experts annotated images in an average time of 127s per image, and Novices annotated images in an average time of 253s per image.
- Cost: Experts costs were set at $60 per labelling hour, and Novices cost $6 per labelling hour.

During analysis 10 cases were found to have two hearts visible (split screen view), resulting in zero DICE agreement amongst experts and experts and novices, having annotating different sides of the image. These cases have been removed. A significant proportion of the worst performing remaining cases are a result of mislabelling of left/right atriums and ventricles resulting in very low DICE scores for those cases.

Q1: In Table 6.1 the average DICE scores for expert-expert and novice-expert segmentations show that no statistical difference is found between the variability of annotators on three out

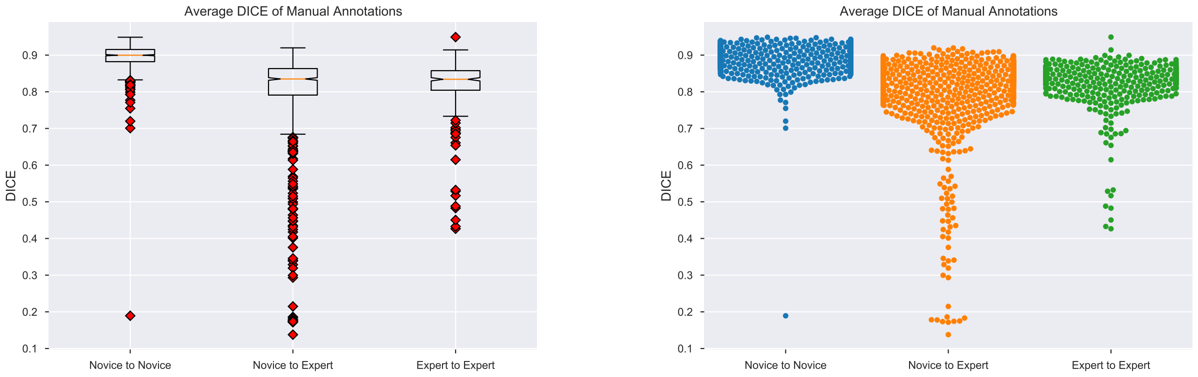


Figure 6.3: Distributions of DICE scores for Novice to Novice labels, Novice to Expert labels and Expert to Expert Labels.

DICE	LV	RV	LA	RA	WH
Expert to Expert	0.807	0.787	0.764	0.808	0.887
Novice to Expert	0.778	0.761	0.757	0.806	0.894
p-value	**0.009**	**0.005**	0.551	0.866	0.359

Table 6.1: Mean DICE scores of manual annotations performed by Experts compared with DICE scores of manual annotation performed by Novices. Statistically significant (95%) results shown in bold.

of five annotated classes. This shows that novice annotators are better at annotating complex medical data than is assumed and the variability between experts and novices is similar to that amongst experts for these three classes. Figure 6.3 highlights the similarity in DICE distributions between novice-novice and expert-novice annotations, indicating it may not be possible to avoid variation in annotations even when using experts annotations alone.

Q2: Average DICE scores and segmented class sizes for expert trained vs. novice trained models show there is no statistical difference in the performance of the models on three out of

Size (px)	LV	RV	LA	RA	WH
Expert Model	806	630	468	612	4732
Novice Model	737	536	428	546	5130
p-value	**0.0051**	**2.45e-07**	**0.0083**	**0.0009**	**0.0018**

Table 6.2: Class average sizes in pixels of model predictions, comparing models trained using expert annotations and models trained using novice annotations. Statistically significant (95%) results shown in bold.

five classes (Tables 6.3, and 6.2). Expert models average higher DICE scores on all but one class, and one reason for this better performance is that both models are tested against an expert annotated test set. Models trained on novice annotations can perform almost equally well as those trained on expert annotations for multi-class US segmentations problems. We see a significant difference between average class sizes predicted by the two models, most noticeably in the right ventricle class (RV), however their overall similarity is highlighted in Figures 6.4 and Figures 6.5-6.6 where both DICE and sizes appear very similar across all classes.

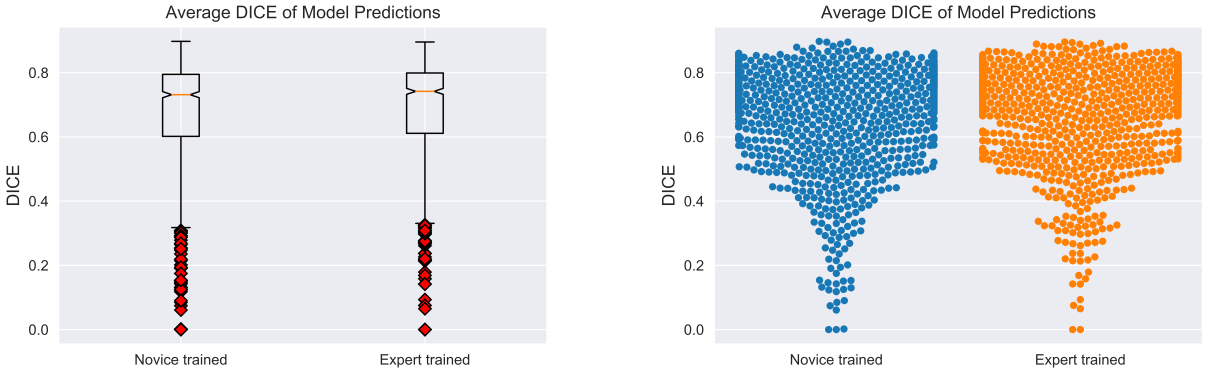


Figure 6.4: Distributions of segmentation predictions DICE scores against the expert test set

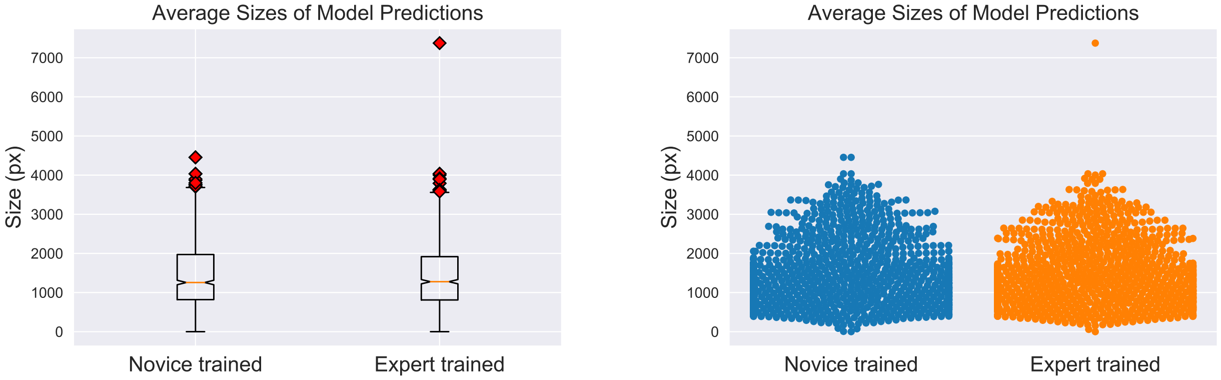


Figure 6.5: Distributions of segmentation predictions average pixel sizes.

Q3: The results of HLHS classification methods trained on manual and automated expert and novice segmentations show that novice trained models attain very similar results to those trained by experts, in both manual and model cases (Table 6.4). We see a slight improvement from expert annotations in Precision and F1 scores but the overall performance is remarkably similar. This result again shows the viability of acquiring a significant proportion of medical

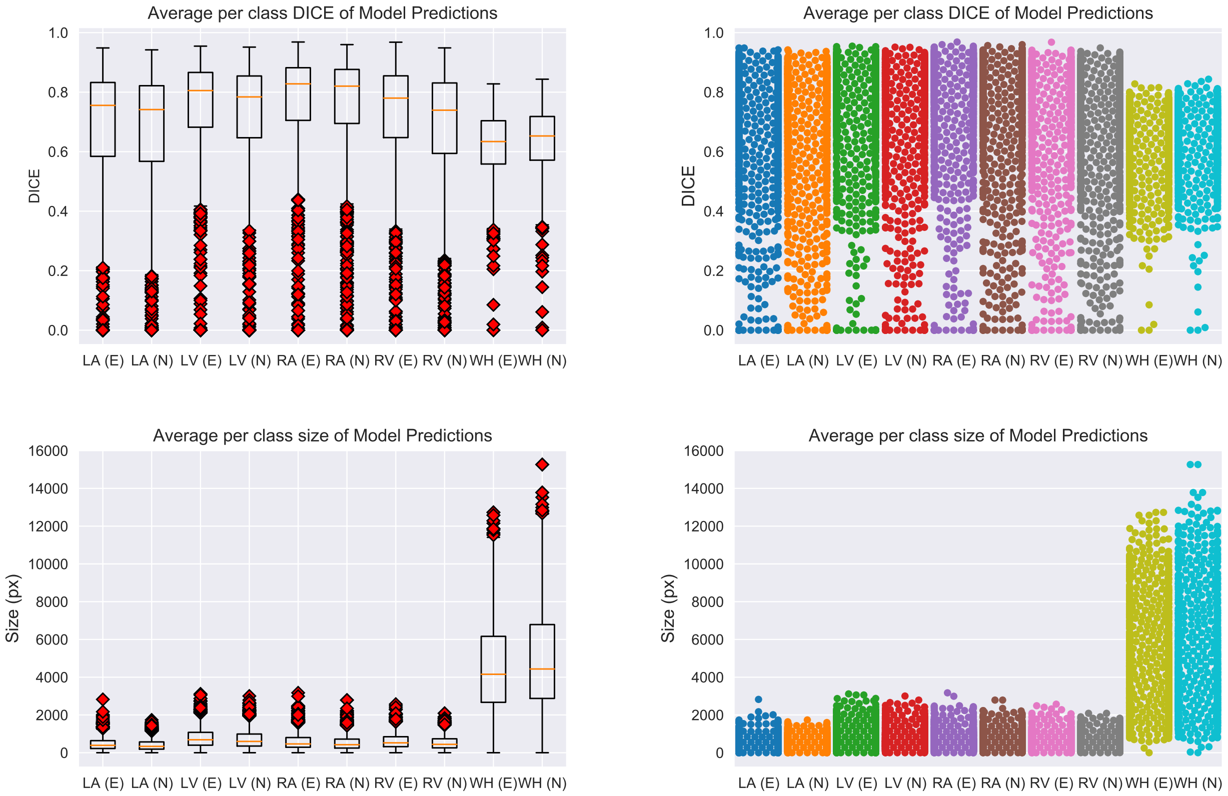


Figure 6.6: Top row: Distributions of per class segmentation DICE scores. Bottom row: Distributions of per class segmentation size.

image annotations from non-experts during annotation efforts. Figure 6.7 highlights that in some scenarios novice manual annotations may out-perform expert annotations on some metrics. Both ROC curves and Precision-Recall Curves for experts and novices follow very similar trajectories demonstrating the similarity in their performance for classification.

Q4: Figure 6.8a shows the consistent increase of both expert and novice trained models as the size of the dataset increases, demonstrating that collecting initial annotations from novices

DICE	LV	RV	LA	RA	WH
Expert Model	0.721	0.707	0.663	0.749	0.617
Novice Model	0.708	0.679	0.652	0.731	0.634
p-value	0.174	**0.003**	0.321	0.071	**0.001**

Table 6.3: Class average DICE scores of model predictions and class average sizes in pixels of model predictions, comparing models trained using expert annotations and models trained using novice annotations. Statistically significant (95%) results shown in bold.

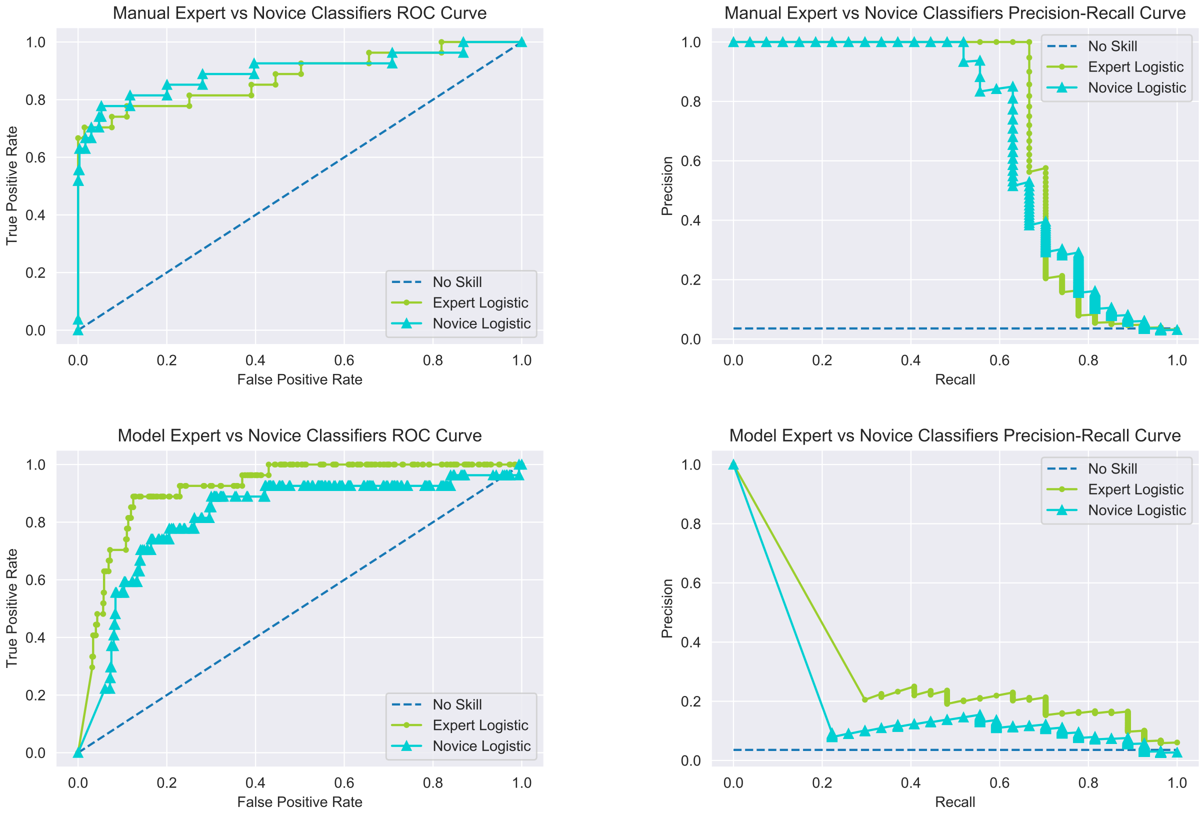
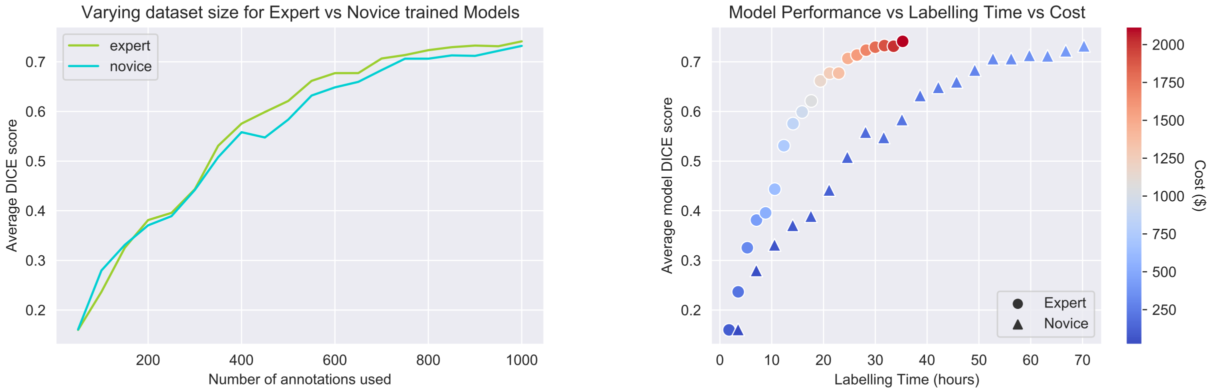


Figure 6.7: Top row: Classification performance for manual annotations predicted classifications. Bottom row: Classification performance for model predicted classifications. Left to right: ROC Curves and Precision-Recall curves.

may well suffice to achieve a good accuracy in many tasks. We calculate the cost per image for both novices and experts using the average cost of an hour of labelling work and the average time each annotation took to create. Figure 6.8b shows how when time is the priority, then expert annotators achieve higher quality models in a shorter time-span, however this comes at a much greater financial cost. If cost is the priority then novice annotators achieve higher quality models at a much smaller financial cost, however the same number of annotations take longer to acquire from novices than from experts. We can see from this that the dominant driving force of improving model quality is dataset size, regardless of whether annotations come from experts or novices, indicating that to train high performing models in a resource efficient way that novice annotations are a useful mechanism by which this can be achieved.

	Expert Manual	Novice Manual	Expert Model	Novice Model
TP	19	20	24	22
FP	47	64	126	224
TN	926	909	847	749
FN	8	7	3	4
Precision	0.288	0.242	0.16	0.091
Recall	0.704	0.753	0.889	0.827
F1	0.409	0.367	0.271	0.165
AUC-ROC	0.879	0.900	0.915	0.829

Table 6.4: Classification results: Precision, Recall and F1 scores are reported for the positive prediction class (HLHS)



(a) DICE scores as we increase the training dataset size from 50 to 1000 images

(b) Time, Cost and DICE scores as we increase the training dataset size from 50 to 1000 images

Figure 6.8: Analysis of the Time/Cost/Model performance trade-off.

6.6 Discussion

We have evaluated the upstream and downstream effects of acquiring complex medical image segmentation annotations from novices compared to experts. We have found that raw novice annotations are of remarkable quality, and that novice trained models show only a minor performance decrease compared expert trained models. Our results highlight that annotations performed by novices are of great utility for complex tasks such as segmentation and classification. A time and cost analysis for using limited resources more efficiently is provided, guiding practitioners in acquiring annotations to give the best performing models under their constraints. Through future studies on other complex tasks, we aim to develop protocols through which confidence can be given that novice annotations are sufficient in many use cases.

Additional combination of crowd-sourcing from novice labels with models incorporating measures of annotator skill and merging of multiple annotations show great promise in enabling highly accurate models to be developed on a wide variety of tasks for which expert annotated data has been infeasible to acquire at a large enough scale.

We note that we are unsure of how representative our Labelbox workforce is of the wider novice annotator community. Through our engagement with Labelbox they were made aware of our intentions with the annotated data and it is our hope that no special measures were taken to improve the quality of annotations beyond that the wider novice annotator community. Similarly, when comparing costs of annotating large datasets, we must consider the ethical implications of employing low-cost workers to perform these tasks - while the low cost makes using workforce services appealing, care must be taken to ensure that workers are paid fairly and under suitable working conditions. Limited information given regarding the locations and working conditions of annotation workforces creates difficultly in making this judgement.

6.7 Summary

We have demonstrated that novice annotators are capable of performing complex medical image segmentation tasks to a high standard, with a comparable variability to experts as experts show to themselves. We have shown that training models with novice annotations is both resource efficient and can give comparable models in terms of prediction performance against expert annotations for both segmentation and downstream classification tasks. While conclusions that generalise across many different tasks and application domains are hard to draw from this work, we conclude that the assumption that novice annotators cannot perform complex annotations for medical imaging tasks is false, and that the level of expertise required to perform different annotation tasks should be evaluated on a task by task basis. We foresee that in combination with existing methods that better handle noisy annotations, and AL methods selectively choosing the most informative annotations to acquire next, that novice annotations will play a vital role in developing high-performing models at a fraction of the cost of using expert annotations.

Chapter 7

Conclusion and Future Work

In this chapter we summarise the research addressed in this thesis. Following this we highlight the methodological contributions made in each chapter and finally discuss the limitations of our methods and propose possible directions for future work, before providing concluding remarks.

7.1 Summary

In this thesis we have examined five key challenges facing data-driven human-in-the-loop computing, and made contributions to progress the field in each of these.

In Chapter 2 Section 2.2 we look at the ’Categorisation Challenge’ and how the fields of Symbiotic DL and Human-in-the-Loop DL are emerging. We have categorised the existing literature and identified the key areas in which progress still needs to be made for DL systems to reach a state of symbiosis, and highlighted several areas in which progress in the wider DL field may assist in this effort.

In Chapter 3 we look at the ’Confidence Challenge’ and how we have shifted from formula driven computing to data-driven computing. Reasoning about the decisions made by machines has become a significant challenge. This poses several problems for the main stream use of data-driven methods in safety critical domains as the reasoning provided for the predictions made are increasingly abstract and more difficult to understand. In this chapter we aimed to propose novel methods for extracting measures of confidence from data-driven methods and propose methods by which these can be used to both better communicate data-driven predictions, and a means by which these can be used to improve the downstream performance of those predictions.

In Chapter 4 we look at the 'Complexity Challenge' and how the increasing computational power of data-driven methods has enabled more complex analysis of challenging data types. This further increases the need for robust and consistent methods by which populations of data can be compared faithfully. In this chapter we aimed to propose novel methods for extracting meaningful and robust multi-task predictions of structures from high-dimensional data, which perform consistently across multiple populations and enable cross-population comparison for downstream analysis.

In Chapter 5 we look at the 'Classification Challenge' and how data-driven methods are starting to provide means to robustly classify complex phenomenon that is beyond the capabilities of humans. This poses a significant challenge as we wish to understand what makes these phenomenon classifiable for data-driven methods, but not for traditional methods. In this chapter we aimed to propose novel methods for robust classification that also provide interpretable features that lead to a deeper understanding of the classification task, and to sensitise state-of-the-art methods related to that task for multi-task classification.

In Chapter 6 we look at the 'Curation Challenge' and how one of the main bottlenecks in data-driven model development may rely on false assumptions about who is capable of annotation challenging data. The reliance on experts to annotate many types of data has slowed the uptake of data-driven methods in many fields due to the difficulty in finding the experts to perform such annotation. In this chapter we aimed to evaluate if it may be feasible to use novice annotators in place of experts for a challenging annotation task and evaluate the downstream effects of using novice annotated data to train our models.

7.2 Achievements

As described above this thesis focuses on improving data-driven methods to provide better communication and downstream utility of model predictions for complex multi-task problems. The main methodological achievements are summarised here.

7.2.1 Categorisation Challenge

The following achievements were made in addressing the 'Categorisation Challenge':

Identification of research gap and overlap

We explored a wide range of literature across many areas of medical image analysis, specifically in segmentation tasks focused on introducing Human-in-the-Loop elements to DL model development. We identified the the gaps in literature that most urgently need addressing to achieve symbiosis of Human-in-the-Loop DL methods. We identified key areas of overlap where solutions to specific problems will generalise across multiple fields and application areas. This will enable future researchers to better focus their research efforts and provide greater impact of their work across multiple related domains.

7.2.2 Confidence Challenge

The following achievements were made in addressing the 'Confidence Challenge':

Probabilistic Segmentation

Segmentation of medical images is used in a variety of downstream tasks such as biometric estimation, however deterministic segmentation methods only provide the capability to produce point estimates of these biometrics. In practice, experts produce variable estimates of these biometrics due to inter and intra observer variability. In Chapter 3 we introduce dropout during inference to produce various plausible automated segmentations in order to acquire a distribution of biometric estimates that better reflect what is produced manually by experts.

Interpretable Prediction Bounds visualisation

The interpretation of metric error previously relied on traditional +/- reporting (standard deviation or other etc), however in real-time scenarios we can provide more information to better guide users. In Chapter 3 we go on to leverage the distribution of segmentation predictions to produce visualisable and interpretable biometric prediction bounds that indicate to end users which parts of an image are less confident than others, guiding them towards acquiring new more suitable images.

Percentage in Range and Range AUC performance metrics

We wish to evaluate the suitability of probabilistic prediction methods for a given task, and in order to do this we introduce a new metric in Chapter 3 that allows us to capture the ability of a probabilistic prediction to contain the true value. The 'Percentage in range' metric measures what percentage of true values are contained within the predicted upper and lower bounds. We go on to derive the Range AUC metric that accounts for the fact that an infinitely large difference between upper and lower bounds would always contain the true value, but this is clearly not desirable, so instead the Range AUC metric measures the ability to have a high proportion of true values contained within upper and lower bounds of an acceptable width.

Variance scores quantification and filtering

When using automated metric estimates, it is vital that we are made aware of failure modes of data-driven methods and as such we derive a series of 'Variance scores' used to quantify the confidence of probabilistic segmentation and use these in Chapter 3 to automatically filter out the worst performing cases in prediction tasks. This enables scenarios in which it is possible to defer the worst performing cases for manual intervention or re-acquisition of input to our models.

7.2.3 Complexity Challenge

The following achievements were made in addressing the 'Complexity Challenge':

3D Simultaneous Multi-task Segmentation Regression

The computational power of data-driven methods enables previously computationally infeasible biometrics to be estimated in a fraction of the time with a fraction of the compute power. This has enabled more complex tasks to be addressed, and introduced the additional tasks of how to present the information derived from these tasks in an interpretable way. In Chapter 4 we extend existing state-of-the-art segmentation methods to 3D multi-task variants in which we simultaneously estimate a segmentation, and regression of complex shape properties of complex brain structures that previously relied on computationally expensive and error-prone non DL pipelines.

Robust comparison of challenging populations

The robustness of our automated volume-based method for estimation of complex shape properties to abnormal shape structures enables population wide comparison in downstream analysis not previously possible. This will enable methods for non-invasive classification for a wide variety of abnormal shape properties to be derived.

7.2.4 Classification Challenge

The following achievements were made in addressing the 'Classification Challenge':

Classification sensitive Atlas-based Segmentation

Atlas-based segmentation provides a valuable guarantee of anatomical consistency in segmentation, however the flexibility of atlas-based methods sometimes limits their suitability for modelling abnormal anatomical shapes and structures. In Chapter 5 we extend state-of-the-art Atlas Image-Spatial-Transformer networks to the additional task of classification, and use this classification branch to guide the atlas construction towards disease sensitised outputs. This led to atlas based segmentations that performed better than previous segmentations for the task of HLHS classification from interpretable features extracted from those segmentations.

Interpretable feature-based classification of HLHS

In Chapter 5 we introduce a linear classification method for HLHS disease prediction. We extract 'area ratios' from segmentations of the '4-Chamber Heart' view in order to robustly classify HLHS with high accuracy. This provides a new method for quantifying likelihood of HLHS in prenatal scans from a single image, which previously relied on expert inspection of several views and video stream of ultrasound scans.

7.2.5 Curation Challenge

The following achievements were made in addressing the 'Curation Challenge':

Time vs. cost vs. quality when acquiring annotations for challenging tasks

The acquisition of complex and challenging annotations remains as a bottleneck in developing high performing data-driven models. It is assumed that only experts can provide these annotations which results in a very high cost to acquire these annotations. In Chapter 6 we challenge this assumption and show that the variation between experts and other experts may be similar to the variation between novices and experts, and the downstream impacts of training models using novice annotations are less severe than assumed. We show that in resources limited scenarios that using novice annotations may be much more cost effective than expert annotations while still resulting in models of equivalent performance.

7.3 Limitations, Future Work and Applications

In this section we elaborate on the main limitations of our presented works and propose suggestions of future work that could address these limitations.

In Chapter 3 we make contributions towards quantifying confidence of data-driven predictions and have developed several heuristics for filtering the worst performing cases at test time. While these contributions are valuable, they are limited in their utility at present, as the developed heuristics only show a mild correlation with prediction accuracy. As such, further work is needed to develop well calibrated uncertainty/confidence metrics that show a stronger correlation to prediction accuracy.

A further limitation is that our high predictive accuracies are made on images from a limited set of scanners and scanning protocols, with the majority of these scans being from healthy subjects. Further work is needed to evaluate the performance of our models on images from different domains and to evaluate performance on images that exhibit abnormal shape/structure as these are the most important to identify during screening.

In Chapter 4 we make contributions towards learning complex shape properties in a robust way using data-driven methods. However, we've seen that some shape properties are easier to learn than others and more work is needed to understand why this is, and the develop more robust methods capable of learning many different types of properties with equal accuracy. A limitation of our work is in the presentation of volume-based predictions, as the state-of-the-art approach for visualising cortical properties is using surface models. Methods to extract those surfaces automatically from medical image volumes are needed in order to develop the final steps of state-of-the-art analysis pipelines in a data-driven manner.

Valuable future works are enabled by robust methods for biomarker prediction in challenging populations. It will be possible in the future to generate disease-discriminative biomarkers from these populations through cross-population comparison, and we hope this will lead to a deeper understanding of the differences between healthy and pathological populations.

In Chapter 5 we developed a robust and powerful method for the interpretable classification of HLHS from a single ultrasound image. A limitation of our work is that it relied on each image to be pre-aligned before input to our model, a future work will include learning this alignment in a data-driven way so automate this step. There will be value in improving the robustness of this method to highly-misaligned images in atlas construction to enable different and more challenging tasks to be tackled with this framework, especially due to the usual abnormal appearance of pathological cases, for which it is most important to have high performance. In ultrasound screening, live video of multiple views is used to make diagnosis, and the developement of video atlas' for common views during common stages of motion such as the heartbeat may provide additional information from which more robust or additional diagnosis can be made.

In Chapter 6 we challenged the assumption that only experts can provide useful annotations for a challenging medical image segmentation task and found that in many scenarios this is a feasible option to take and can result in equivalent performance from data-driven models. This study is limited in the fact it only addressed one task, as the same findings may not hold in other tasks. Future work should include developing protocols by which this decision can be made responsibly on a task by task basis as it is likely that as more medical image datasets become available, the need for quality annotators also increases. Our study was also limited in the number of annotations we were able to acquire, we found that using all the annotations acquired resulted in similar performance when using novice and expert annotations, but further study is needed to evaluate whether model performance diverges as the number of annotations grows larger.

There is still a long way to go in closing the loop between humans and machines in data-driven methods. Throughout this thesis we have explored methods by which we can develop algorithms, and present their outputs in a way that facilitate greater understanding of their functionality, and promote more responsible decision making when relying on ML outputs. There still remains a significant amount of work in order to distil this improved human reasoning about ML outputs into a form that can be faithfully passed back into our algorithms in order to close the loop and create a symbiosis between humans and data-driven methods.

In this thesis we have centred our algorithm development around developing applications for fetal screening - we have shown how our contributions make progress in overcoming some of the barriers currently faced in developing data-driven solutions to fetal screening problems and hope that these may lead to improved applications in this area.

There are many other applications where this work will be relevant. Beyond medical image computing, other domains such as disaster prevention or other safety critical domains where image interpretation remains a bottleneck in decision making, either through the required expertise for interpreting data, or the sheer scale on which imagery is produced where it becomes infeasible for direct manual inference over this data e.g earth observation imagery and space imagery, each stand to benefit from the methods developed in this thesis.

7.4 Conclusion

In summary, in this thesis we have contributed several methods by which data-driven algorithms can communicate richer and more interpretable information to human end users. The primary innovation of this thesis is in providing methods by which each stage of closing the symbiotic loop can be achieved. This thesis shows the capabilities of data-driven technology to complete the first half of the requirements for a Human-in-the-loop computing system to achieve symbiosis, i.e developing models which can communicate rich information to a human user, through which they are provided the information needed to provide reasoned feedback to the model, in order to complete the loop and establish a symbiosis of man and machine. We have examined the state-of-the-art literature to categorise and stratify human-in-the-loop-based work to better understand where more needs to be done to close the loop between human and machine in data-driven methods. We have provided means by which data-driven methods can express confidence in their predictions, and presented this confidence in multiple intuitive ways. We have presented ways in which we can use the confidence to defer to alternative approaches to ensure the highest quality of prediction. We have shown how to extend these methods into complex prediction scenarios where traditional methods perform well but often fail on edge cases scenarios where data-driven methods do not. We have shown methods by which data-driven methods can be used to extract interpretable features that provide a clearer reasoning over themselves than black-box data-driven methods alone. And finally we challenged the main assumptions that governs the development of data-driven methods for complex and challenging tasks and discussed that new work is needed to answer where our finite resources can be best utilised to achieve new and greater human decision making support in the clinical practice.

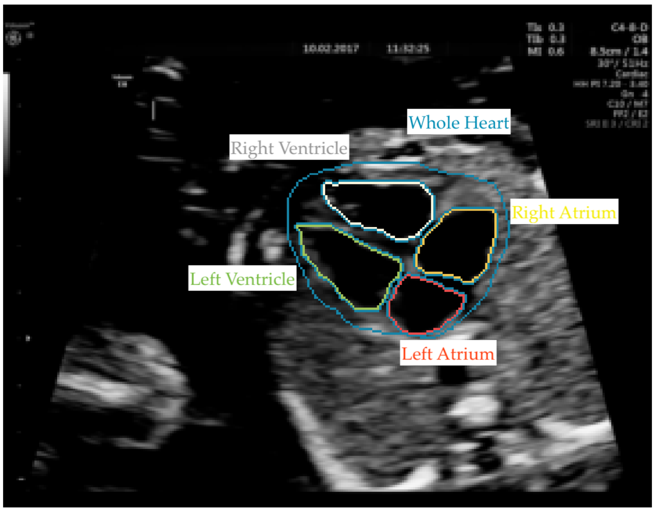
Appendix A

Instructions for Labelbox workforce

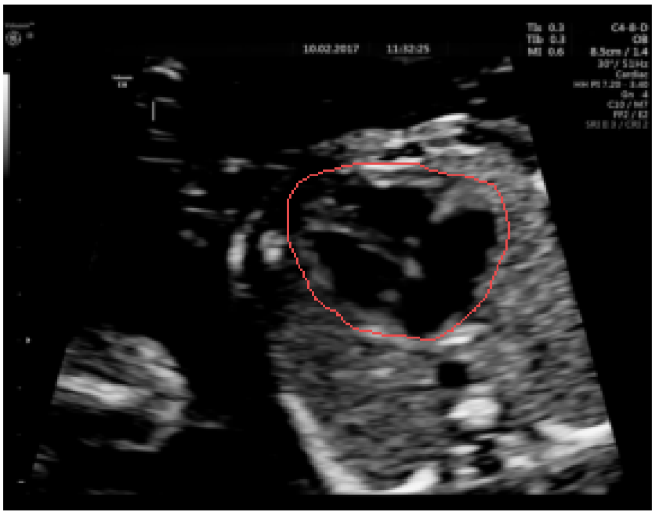
On the following pages, the original PDF instructions provided for the Labelbox novice workforce are provided for reference.

Instructions for Labelbox Workforce

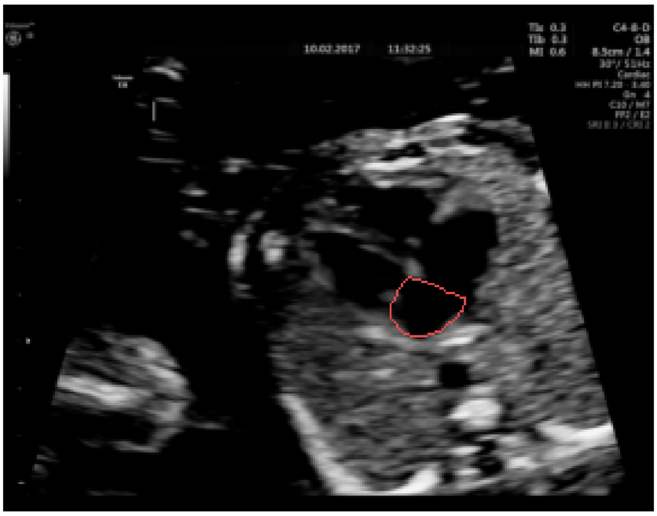
The task at hand is to segment 5 anatomical classes within a fetal ultrasound view (the 4 chamber view), The 5 classes to segment are the 'Whole heart', the 'left ventricle', 'right ventricle', 'left atrium' and 'right atrium'. This is not an easy task, but when done accurately can be of great help in diagnosing congenital heart disease, which is the motivation behind this task.



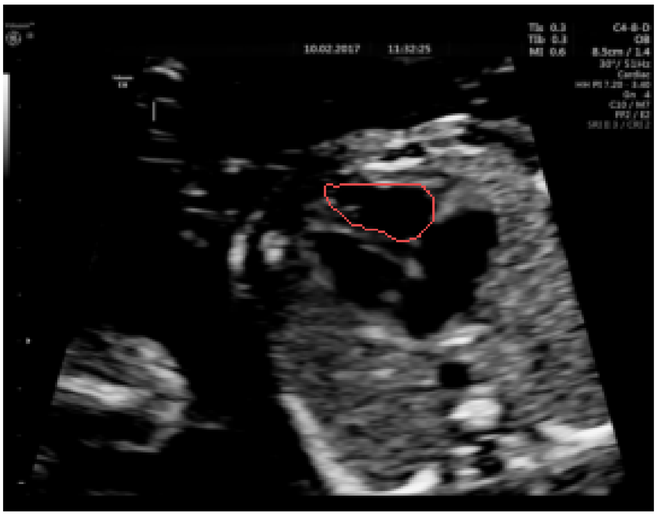
The first class is the 'Whole heart' class that encompasses the four chambers of the heart:



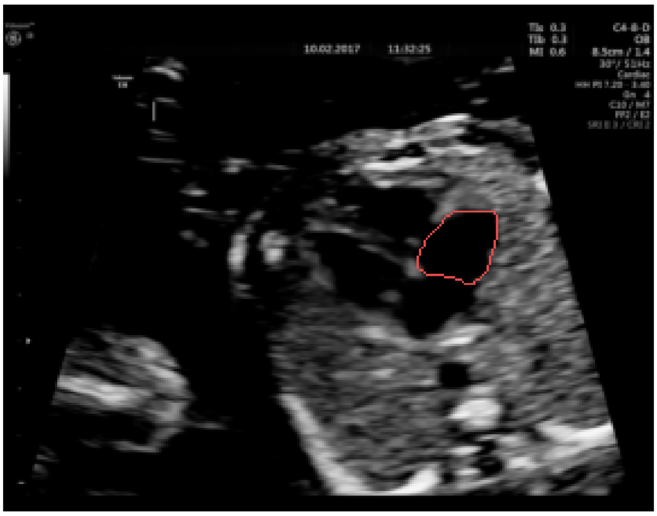
Next is the ‘Left Atrium’ class:



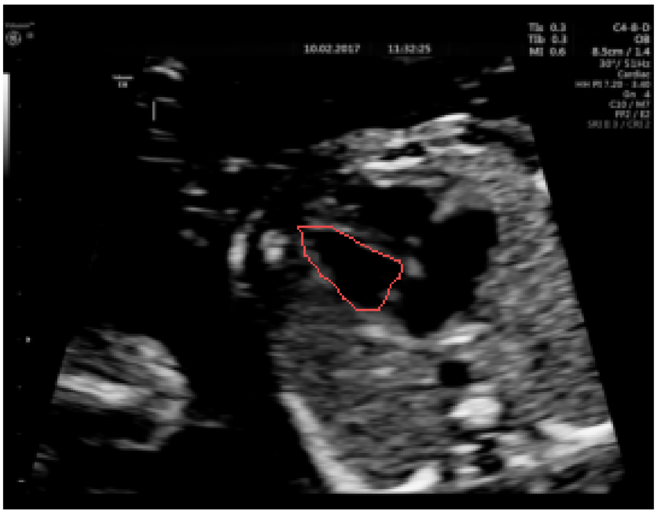
The ‘Right Ventricle’ class:



The ‘Right Atrium’ class:

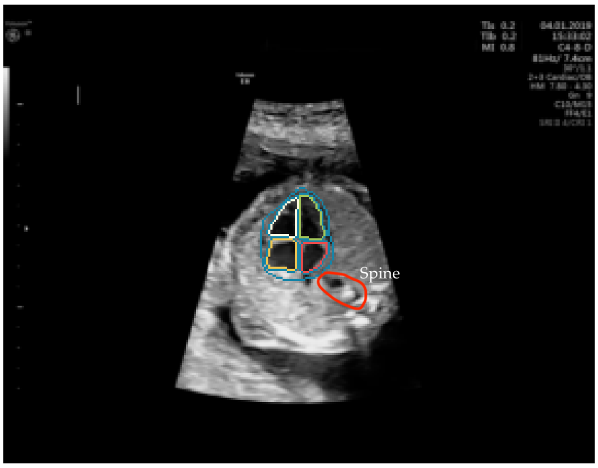


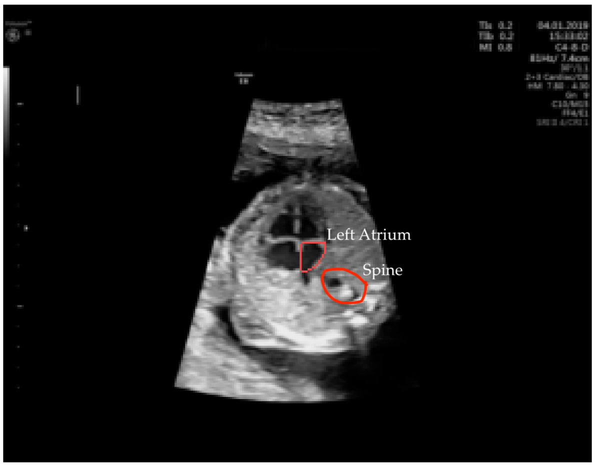
Finally the ‘Left Ventricle’ class:



Challenge 1: Orientation

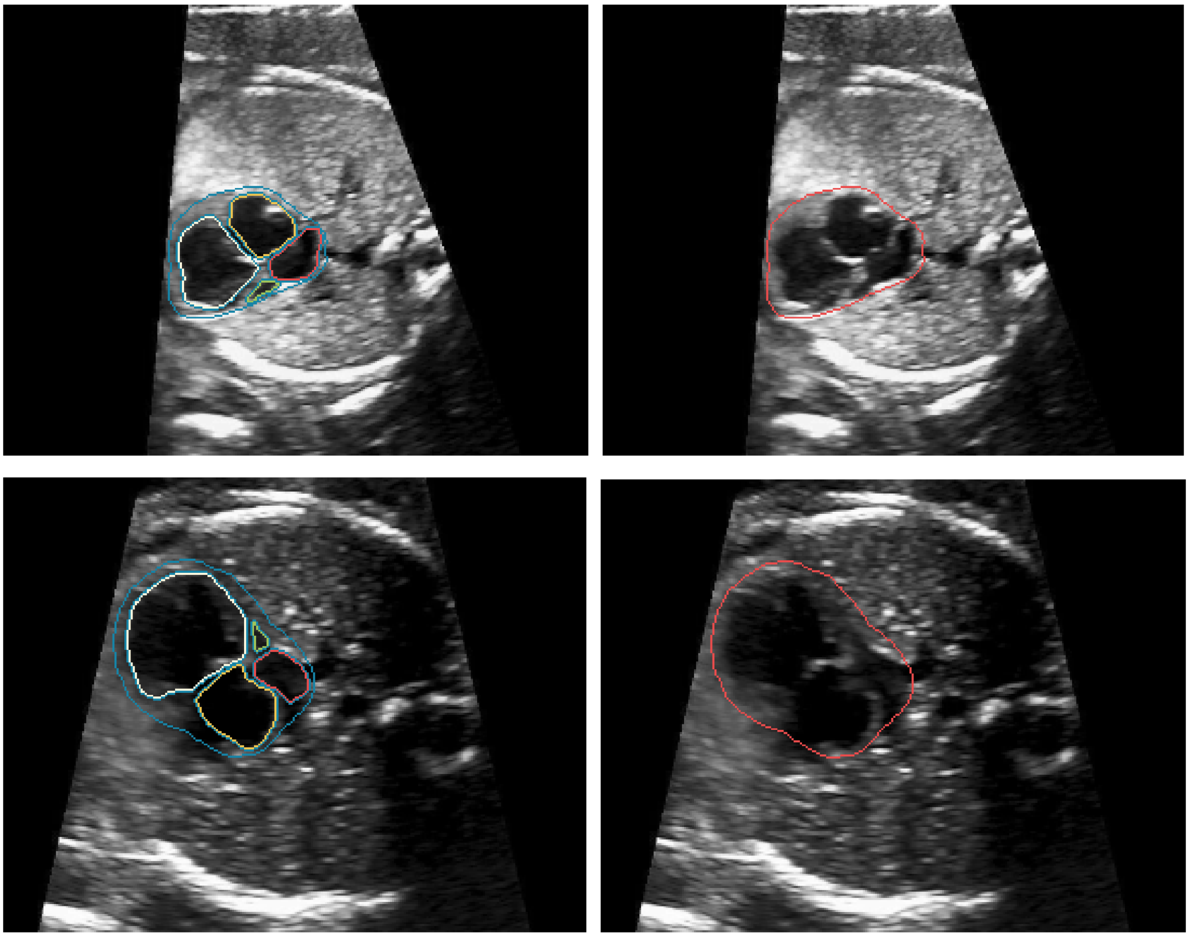
When annotating the left/right atrium and ventricles it is important that we correctly identify the orientation of the heart so that the left and right atrium and ventricles can be correctly annotated. This can be done by locating the spine, from which the heart chamber closest to the spine is the ‘Left Atrium’, and the chamber furthest from the spine is the ‘Right Ventricle’, and the other two chambers can be annotated with respect to these two chambers. In general for healthy hearts, the ventricles are more elongated than atriums, this rule of thumb can be used to distinguish between atriums and ventricles.

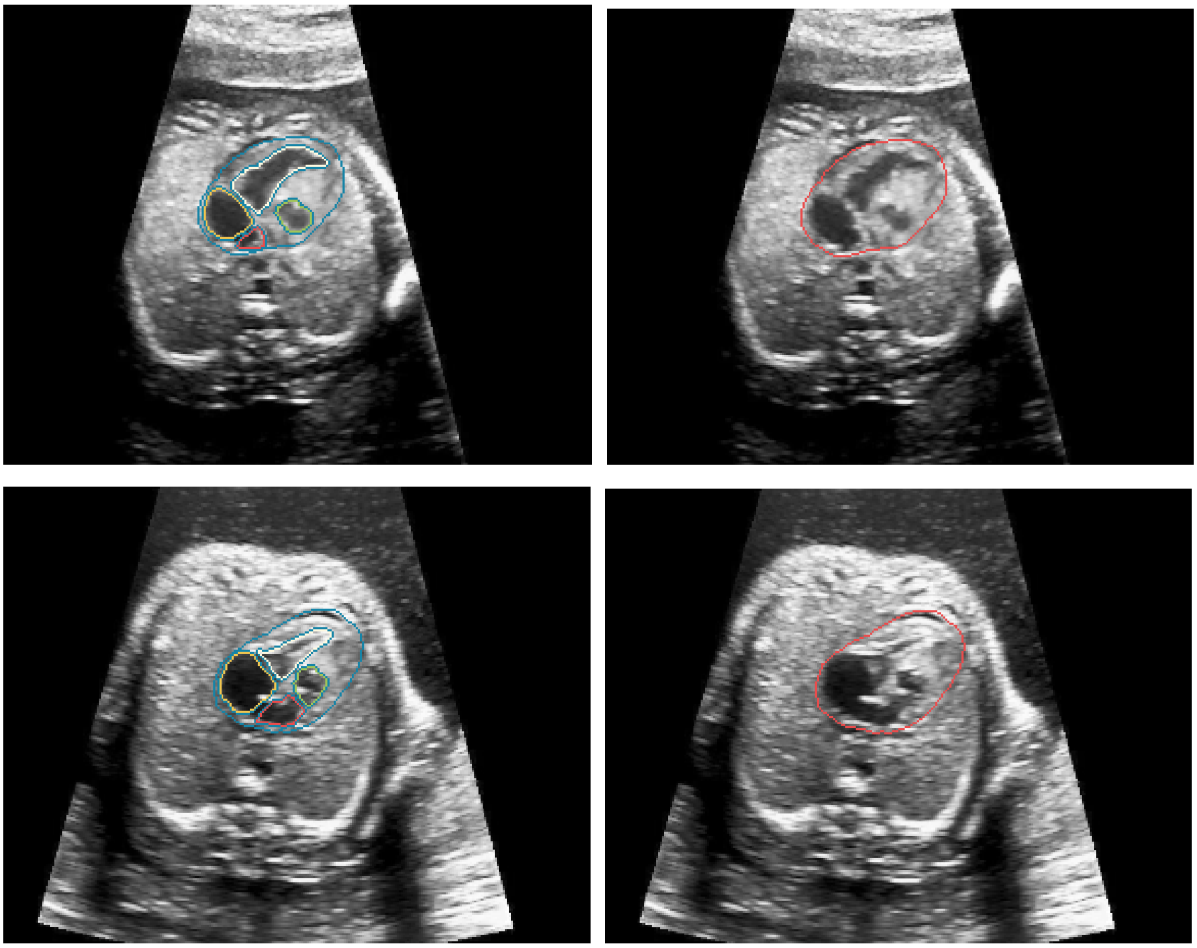




Challenge 2: Abnormal heart structure

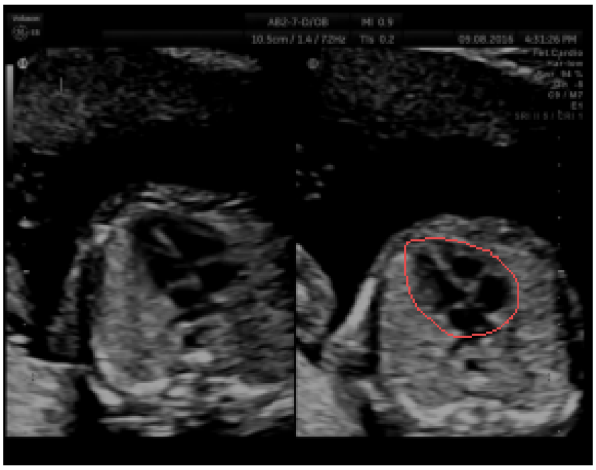
Not all the hearts in this dataset are healthy, and as such several cases presented to you may exhibit shapes unlike the majority of hearts you will see, however it is important that the true shape of the heart is annotated as it is this difference in shape that can be so important in diagnosis congenital heart disease. We provide some examples of abnormal heart structures below:

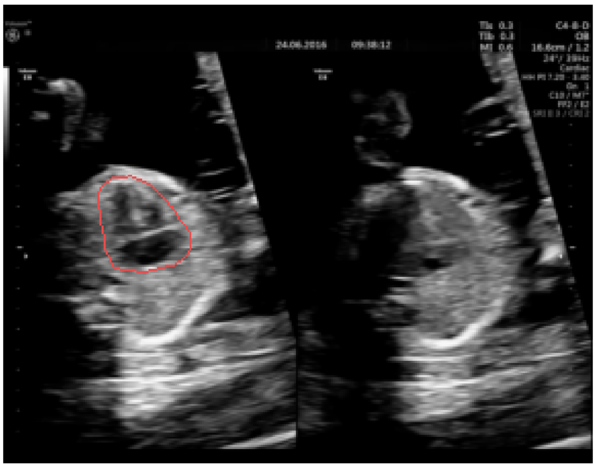


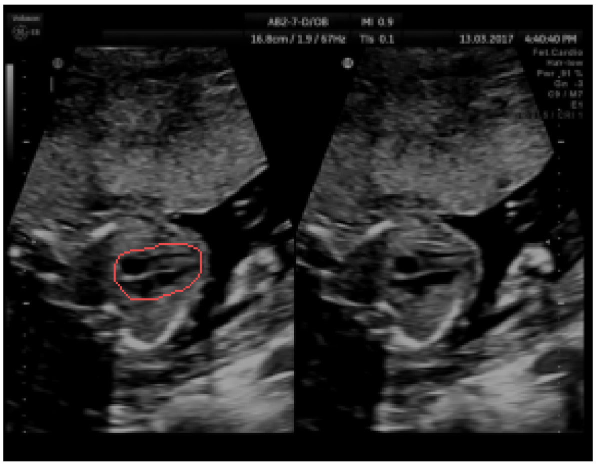


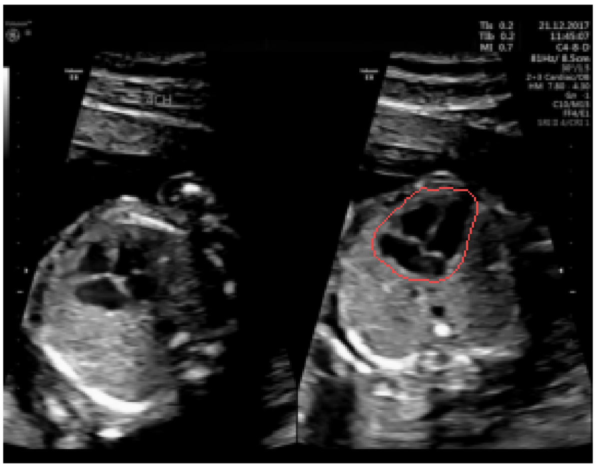
Challenge 3: Split pane images

Several images in this dataset show two slightly different views of the heart side by side, in this case, only one of the hearts should be annotated. Usually one of the sides will show a clearer image of the heart, and this image should be annotated, i.e the heart in which it is easier to delineate between the chambers should be annotated (All five classes should be annotated).









Thank you for your annotation efforts, and we want you to know that these annotations will be directly contributing to improving clinical care in fetal ultrasound screening.

Bibliography

[1] H. Gray and H. V. Carter, “Figure 159.- muscles of the left hand,” *Anatomy: Descriptive and Surgical*, 1859.

[2] O. Glasser, “The hand of mrs. wilhelm roentgen: the first x-ray image, 1895,” *Wilhelm Conrad Röntgen and the early history of the Roentgen rays*, 1933.

[3] M. C. Staff, “Fetal ultrasound - mayo clinic,” *Mayo Clinic Patient Care and Health Information*, 2020.

[4] B. B. R. Group, “Fetal mri study,” *UCSF Department of Radiology and Biomedical Imaging Fetal MRI Study*, 2021.

[5] D. Ippolito, A. Pecorelli, C. Maino, C. Capodaglio, I. Mariani, T. Giandola, D. Gandola, I. Bianco, M. Ragusi, C. T. Franzesi, R. Corso, and S. Sironi, “Diagnostic impact of bedside chest x-ray features of 2019 novel coronavirus in the routine admission at the emergency department: case series from lombardy region,” *European Journal of Radiology*, vol. 129, 8 2020.

[6] P. S. Mumbai, “All you need to know about mri and pet/ct scan,” *PET Scan Mumbai Blog*, 2020.

[7] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren, “Niftynet: a deep-learning platform for medical imaging,” *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 113–122, 5 2018.

[8] F. Pérez-García, R. Sparks, and S. Ourselin, “Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning,” *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106236, 9 2021.

[9] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, J. D’hooge, L. Lovstakken, and O. Bernard, “Deep learning for segmentation using an open large-scale dataset in 2d echocardiography,” *IEEE transactions on medical imaging*, vol. 38, pp. 2198–2210, 8 2019.

[10] W. Janpongsri, J. Huang, R. Ng, D. J. Wahl, M. V. Sarunic, and Y. Jian, “Pseudo-real-time retinal layer segmentation for high-resolution adaptive optics optical coherence tomography,” *Journal of Biophotonics*, vol. 13, p. e202000042, 8 2020.

[11] F.-F. Li, R. Krishna, and D. Xu, “Cs231n convolutional neural networks for visual recognition,” *Stanford Spring 2021 Online Course*, 2021.

[12] A. Radford, “Visualizing optimization algos gifs,” *Visualizing Optimization Algos*, 2014.

[13] S. Saha, “A comprehensive guide to convolutional neural networks,” *Towards Data Science: A Comprehensive Guide to Convolutional Neural Networks*, 2018.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, pp. 2278–2323, 1998.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 1–9, 10 2015.

[16] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, 12 2016.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, 2015.

[19] M. Sinclair, A. Schuh, K. Hahn, K. Petersen, Y. Bai, J. Batten, M. Schaap, and B. Glocker, "Atlas-istn: Joint segmentation, registration and atlas construction with image-and-spatial transformer networks," *arXiv*, 12 2020.

[20] J. H. Scatliff and P. J. Morris, "From röntgen to magnetic resonance imaging," *North Carolina Medical Journal*, vol. 75, pp. 111–113, 3 2014.

[21] N. Sharma and L. M. Aggarwal, "Automated medical image segmentation techniques," *Journal of Medical Physics / Association of Medical Physicists of India*, vol. 35, p. 3, 1 2010.

[22] K.-P. Wong, "Medical image segmentation: Methods and applications in functional imaging," *Handbook of Biomedical Image Analysis*, pp. 111–182, 4 2005.

[23] F. Y. Shih, "Image segmentation," *Encyclopedia of Database Systems*, pp. 1389–1395, 2009.

[24] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, and M. Dehmer, "An introductory review of deep learning for prediction models with big data," *Frontiers in Artificial Intelligence*, vol. 0, p. 4, 2 2020.

[25] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *Journal of Digital Imaging*, vol. 32, pp. 582–596, 8 2019.

[26] O. Oren, B. J. Gersh, and D. L. Bhatt, "Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints," *The Lancet Digital Health*, vol. 2, pp. e486–e488, 9 2020.

[27] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Scientific Reports 2020 10:1*, vol. 10, pp. 1–16, 8 2020.

[28] M. D. Kohli, R. M. Summers, and J. R. Geis, "Medical image data and datasets in the era of machine learning—whitepaper from the 2016 c-mimi meeting dataset session," *Journal of Digital Imaging 2017 30:4*, vol. 30, pp. 392–399, 5 2017.

[29] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 7 2020.

[30] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *npj Digital Medicine 2021 4:1*, vol. 4, pp. 1–9, 1 2021.

[31] B. Settles, "Active learning literature survey," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2009.

[32] M. Amrehn, S. Gaube, M. Unberath, F. Schebesch, T. Horz, M. Strumia, S. Steidl, M. Kowarschik, and A. Maier, "Ui-net: Interactive artificial neural networks for iterative image segmentation based on a user model," *Eurographics Workshop on Visual Computing for Biology and Medicine*, 2017.

[33] F. Rodrigues and F. C. Pereira, "Deep learning from crowds," pp. 1611–1618, 2018.

[34] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models: A survey of results," pp. 1–6, IEEE, 8 2017.

[35] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine 2019 25:1*, vol. 25, pp. 24–29, 1 2019.

[36] L. M. Carlson and N. L. Vora, "Prenatal diagnosis: Screening and diagnostic tools," *Obstetrics and gynecology clinics of North America*, vol. 44, p. 245, 6 2017.

[37] S. G, "Fetal cardiac screening and variation in prenatal detection rates of congenital heart disease: why bother with screening at all?," *Future cardiology*, vol. 8, pp. 189–202, 3 2012.

[38] Y.-S. Sohn, M.-J. Kim, J.-Y. Kwon, Y.-H. Kim, and Y.-W. Park, "The usefulness of fetal mri for prenatal diagnosis," *Yonsei Medical Journal*, vol. 48, p. 671, 8 2007.

[39] S. Budd, M. Sinclair, T. Day, A. Vlontzos, J. Tan, T. Liu, J. Matthew, E. Skelton, J. Simpson, R. Razavi, B. Glocker, D. Rueckert, E. C. Robinson, and B. Kainz, "Detecting hypo-plastic left heart syndrome in fetal ultrasound via disease-specific atlas maps," 7 2021.

[40] S. Budd, T. Day, J. Simpson, K. Lloyd, J. Matthew, E. Skelton, R. Razavi, and B. Kainz, "Can non-specialists provide high quality gold standard labels in challenging modalities?," 7 2021.

[41] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, 7 2021.

[42] S. Budd, A. Blaas, A. Hoarfrost, K. Khezeli, K. D'Silva, F. Soboczenski, G. Mackintosh, N. Chia, and J. Kalantari, "Prototyping crisp: A causal relation and inference search platform applied to colorectal cancer data," pp. 517–521, Institute of Electrical and Electronics Engineers Inc., 3 2021.

[43] S. Budd, P. Patkee, A. Baburamani, M. Rutherford, E. C. Robinson, and B. Kainz, "Surface agnostic metrics for cortical volume segmentation and regression," vol. 12449 LNCS, pp. 3–12, Springer Science and Business Media Deutschland GmbH, 10 2020.

[44] S. Budd, M. Sinclair, B. Khanal, J. Matthew, D. Lloyd, A. Gomez, N. Toussaint, E. C. Robinson, and B. Kainz, “Confident head circumference measurement from ultrasound with real-time feedback for sonographers,” vol. 11767 LNCS, pp. 683–691, Springer, 10 2019.

[45] S. Budd, E. Robinson, and B. Kainz, “The cortical explorer : A web-based user-interface for the exploration of the human cerebral cortex,” *Eurographics Workshop on Visual Computing for Biology and Medicine*, 2017.

[46] A. Vlontzos, S. Budd, B. Hou, D. Rueckert, and B. Kainz, “3d probabilistic segmentation and volumetry from 2d projection images,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12502 LNCS, pp. 48–57, 10 2020.

[47] J. Matthew, E. Skelton, T. G. Day, V. A. Zimmer, A. Gomez, G. Wheeler, N. Toussaint, T. Liu, S. Budd, K. Lloyd, R. Wright, S. Deng, N. Ghavami, M. Sinclair, Q. Meng, B. Kainz, J. A. Schnabel, D. Rueckert, R. Razavi, J. Simpson, and J. Hajnal, “Exploring a new paradigm for the fetal anomaly ultrasound scan: Artificial intelligence in real time,” *SSRN Electronic Journal*, 3 2021.

[48] Y. S. Park and S. Lek, “Artificial neural networks: Multilayer perceptron for ecological modeling,” *Developments in Environmental Modelling*, vol. 28, pp. 123–140, 1 2016.

[49] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics 1943 5:4*, vol. 5, pp. 115–133, 12 1943.

[50] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” *Psychological Review*, vol. 65, pp. 386–408, 11 1958.

[51] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” *Artificial Intelligence Review 2020 53:8*, vol. 53, pp. 5455–5516, 4 2020.

[52] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Mathematics of Control, Signals and Systems 1989 2:4*, vol. 2, pp. 303–314, 12 1989.

[53] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature 1986 323:6088*, vol. 323, pp. 533–536, 1986.

[54] S. Ruder, "An overview of gradient descent optimization algorithms," 9 2016.

[55] D. P. Kingma, "Adam: A method for stochastic optimization.," 2015.

[56] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, pp. 73–101, 1964.

[57] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[58] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3951 LNCS, pp. 404–417, 2006.

[59] S. Ioffe, "Batch renormalization: Towards reducing minibatch dependence in batch-normalized models," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[60] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 7 2016.

[61] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 7 2016.

[62] Y. Wu and K. He, "Group normalization," *International Journal of Computer Vision 2019 128:3*, vol. 128, pp. 742–755, 7 2019.

[63] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[64] T. R. C. of Radiologists, "Clinical radiology uk workforce census 2017 report," 2017.

[65] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "Unsupervised domain adaptation

in brain lesion segmentation with adversarial networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10265 LNCS, pp. 597–609, 2017.

[66] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning 2009 79:1*, vol. 79, pp. 151–175, 10 2009.

[67] H. R. Tizhoosh and L. Pantanowitz, "Artificial intelligence and digital pathology: Challenges and opportunities.," *Journal of pathology informatics*, vol. 9, p. 38, 2018.

[68] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis.," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[69] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 12 2017.

[70] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological Physics and Technology*, vol. 10, pp. 257–273, 9 2017.

[71] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, L. Uhlmann, C. Alt, M. Arenbergerova, R. Bakos, A. Baltzer, I. Bertlich, A. Blum, T. Bokor-Billmann, J. Bowling, N. Braghiroli, R. Braun, K. Buder-Bakhaya, T. Buhl, H. Cabo, L. Cabrijan, N. Cevic, A. Classen, D. Deltgen, C. Fink, I. Georgieva, L.-E. Hakim-Meibodi, S. Hanner, F. Hartmann, J. Hartmann, G. Haus, E. Hoxha, R. Karls, H. Koga, J. Kreusch, A. Lallas, P. Majenka, A. Marghoob, C. Massone, L. Mekokishvili, D. Mestel, V. Meyer, A. Neuberger, K. Nielsen, M. Oliviero, R. Pampena, J. Paoli, E. Pawlik, B. Rao, A. Rendon, T. Russo, A. Sadek, K. Samhaber, R. Schneiderbauer, A. Schweizer, F. Toberer, L. Trennheuser, L. Vlahova, A. Wald, J. Winkler, P. Wölbing, and I. Zalaudek, "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, pp. 1836–1842, 8 2018.

[72] V. J. Mar and H. P. Soyer, "Artificial intelligence for melanoma diagnosis: how can we deliver on the promise?," *Annals of Oncology*, vol. 29, pp. 1625–1628, 8 2018.

[73] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, pp. 102–127, 5 2019.

[74] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, pp. 611–629, 8 2018.

[75] L. E. Atlas, D. A. Cohn, and R. E. Ladner, "Training connectionist networks with queries and selective sampling," pp. 566–573, Morgan-Kaufmann, 1990.

[76] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, pp. 201–221, 5 1994.

[77] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," *Machine Learning Proceedings 1995*, pp. 150–157, 1 1995.

[78] D. Angluin, "Queries and concept learning," *Machine Learning*, vol. 2, pp. 319–342, 1988.

[79] D. Angluin, "Queries revisited," *Lecture Notes in Computer Science*, vol. 2225, pp. 12–31, 11 2001.

[80] K. Lang and E. Baum, "Query learning can work poorly when a human oracle is used," *International Joint Conference on Neural Networks*, 1992.

[81] R. D. King, K. E. Whelan, F. M. Jones, P. G. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver, "Functional genomic hypothesis generation and experimentation by a robot scientist," *Nature*, vol. 427, pp. 247–252, 1 2004.

[82] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan, and A. Clare, "The automation of science," *Science*, vol. 324, pp. 85–89, 4 2009.

[83] D. D. Lewis and J. Catlett, “Heterogeneous uncertainty sampling for supervised learning,” *Machine Learning Proceedings 1994*, pp. 148–156, 1 1994.

[84] A. McCallum and K. Nigam, “Employing em and pool-based active learning for text classification,” pp. 350–358, Morgan Kaufmann Publishers Inc., 1998.

[85] B. Settles and M. Craven, “An analysis of active learning strategies for sequence labeling tasks,” *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1070–1079, 2008.

[86] C. Zhang and T. Chen, “An active learning framework for content-based information retrieval,” *IEEE Transactions on Multimedia*, vol. 4, pp. 260–268, 6 2002.

[87] A. G. Hauptmann, W. H. Lin, R. Yan, J. Yang, and M. Y. Chen, “Extreme video retrieval: Joint maximization of human and computer performance,” *Proceedings of the 14th Annual ACM International Conference on Multimedia, MM 2006*, pp. 385–394, 2006.

[88] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, pp. 623–656, 1948.

[89] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, “Cost-effective active learning for deep image classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, pp. 2591–2600, 12 2017.

[90] S. Wen, T. M. Kurc, L. Hou, J. H. Saltz, R. R. Gupta, R. Batiste, T. Zhao, V. Nguyen, D. Samaras, and W. Zhu, “Comparison of different classifiers with active learning to support quality control in nucleus segmentation in pathology images.,” *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2017, pp. 227–236, 2018.

[91] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, “The power of ensembles for active learning in image classification,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9368–9377, 12 2018.

[92] K. Konyushkova, R. Sznitman, and P. Fua, “Geometry in active learning for binary and multi-class image segmentation,” *Computer Vision and Image Understanding*, vol. 182, pp. 1–16, 5 2019.

[93] Y. Gal, R. Islam, and Z. Ghahramani, “Deep bayesian active learning with image data,” pp. 1183–1192, 3 2017.

[94] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” pp. 1050–1059, PMLR, 6 2016.

[95] A. Kirsch, J. van Amersfoort, and Y. Gal, “Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning,” *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 6 2019.

[96] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, “Suggestive annotation: A deep active learning framework for biomedical image segmentation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10435 LNCS, pp. 399–407, 9 2017.

[97] A. Smailagic, H. Y. Noh, P. Costa, D. Walawalkar, K. Khandelwal, M. Mirshekari, J. Fagert, A. Galdran, and S. Xu, “Medal: Deep active learning sampling method for medical image analysis,” *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018.

[98] F. Ozdemir, Z. Peng, C. Tanner, P. Fuernstahl, and O. Goksel, “Active learning for segmentation by optimizing content information for maximal entropy,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11045 LNCS, pp. 183–191, 9 2018.

[99] J. Sourati, A. Gholipour, J. G. Dy, S. Kurugol, and S. K. Warfield, “Active deep learning with fisher information for patch-wise semantic segmentation,” *Deep learning in medical image analysis and multimodal learning for clinical decision support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, S...*, vol. 11045, pp. 83–91, 9 2018.

[100] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, 12 2019.

[101] M. Ravanbakhsh, V. Tschernezki, F. Last, T. Klein, K. Batmanghelich, V. Tresp, and M. Nabi, "Human-machine collaboration for medical image segmentation," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2020-May, pp. 1040–1044, 5 2020.

[102] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and M. Reyes, "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11071 LNCS, pp. 580–588, 9 2018.

[103] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 8535–8545, 6 2019.

[104] H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydın, L. A. Bentolila, C. Kural, and A. Ozcan, "Deep learning enables cross-modality super-resolution in fluorescence microscopy," *Nature Methods*, vol. 16, pp. 103–110, 1 2019.

[105] Y. Wang, B. Yu, L. Wang, C. Zu, D. S. Lalush, W. Lin, X. Wu, J. Zhou, D. Shen, and L. Zhou, "3d conditional generative adversarial networks for high-quality pet image estimation at low dose," *NeuroImage*, vol. 174, pp. 550–562, 7 2018.

[106] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, "Ea-gans: Edge-aware generative adversarial networks for cross-modality mr image synthesis," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1750–1762, 7 2019.

[107] Y. Pan, M. Liu, C. Lian, Y. Xia, and D. Shen, "Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 2965–2975, 9 2020.

[108] K. Konyushkova, R. Sznitman, and P. Fua, "Learning active learning from data," *Advances in Neural Image Processing*, pp. 4225–4235, 2017.

[109] P. Bachman, A. Sordoni, and A. Trischler, "Learning algorithms for active learning," pp. 301–310, PMLR, 7 2017.

[110] M. Woodward, C. Finn, and B. A. Research, "Active one-shot learning," *arXiv preprint arXiv:1702.06559*, 2017.

[111] M. Fang, Y. Li, and T. Cohn, "Learning how to active learn: A deep reinforcement learning approach," *arXiv preprint arXiv:1708.02383*, 2017.

[112] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1299–1312, 5 2016.

[113] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally," *Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2017, p. 4761, 7 2017.

[114] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," 2015.

[115] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren, "Deepigeos: A deep interactive geodesic framework for medical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 1559–1572, 7 2019.

[116] A. Criminisi, T. Sharp, and A. Blake, "Geos: Geodesic image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5302 LNCS, pp. 99–112, 10 2008.

[117] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren, "Interactive medical image segmentation using deep learning with image-specific fine-tuning," 2017.

[118] G. Bredell, C. Tanner, and E. Konukoglu, "Iterative interaction training for segmentation editing networks," vol. 11046 LNCS, pp. 363–370, Springer Verlag, 9 2018.

[119] T. Kurzendorfer, P. Fischer, N. Mirshahzadeh, T. Pohl, A. Brost, S. Steidl, and A. Maier, "Rapid interactive and intuitive segmentation of 3d medical images using radial basis function interpolation †," *Annual Conference on Medical Image Understanding and Analysis*, pp. 11–13, 2017.

[120] W.-D. Jang and C.-S. Kim, "Interactive image segmentation via backpropagating refinement scheme," pp. 5297–5306, CVPR, 2019.

[121] X. Liao, W. Li, Q. Xu, X. Wang, B. Jin, X. Zhang, Y. Wang, and Y. Zhang, "Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning," pp. 9391–9399, CVPR, 2020.

[122] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 880–893, 4 2020.

[123] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Landmark-based deep multi-instance learning for brain disease diagnosis," *Medical Image Analysis*, vol. 43, pp. 157–168, 1 2018.

[124] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. Mcdonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[125] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a right to explanation," 6 2016.

[126] L. Edwards and M. Veale, "Slave to the algorithm? why a right to explanation is probably not the remedy you are looking for," *SSRN Electronic Journal*, 5 2017.

[127] L. Edwards and M. Veale, "Enslaving the algorithm: From a right to an explanation to a right to better decisions?," *SSRN Electronic Journal*, 2017.

[128] D. Stoyanov, Z. Taylor, S. M. Kia, I. Oguz, M. Reyes, A. Martel, L. Maier-Hein, A. F. Marquand, E. Duchesnay, T. Löfstedt, B. Landman, M. J. Cardoso, C. A. Silva, S. Pereira, and R. Meier, eds., *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, vol. 11038. Springer International Publishing, 2018.

[129] A. Holzinger, B. Malle, P. Kieseberg, P. M. Roth, H. Müller, R. Reihs, and K. Zatloukal, "Towards the augmented pathologist: Challenges of explainable-ai in digital pathology," 12 2017.

[130] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 4 2019.

[131] F. Milletari, V. Birodkar, and M. Sofka, "Straight to the point: Reinforcement learning for user guidance in ultrasound," vol. 11798 LNCS, pp. 3–10, Springer, 10 2019.

[132] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, 4 2019.

[133] A. Jungo, F. Balsiger, and M. Reyes, "Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation," *Frontiers in Neuroscience*, vol. 14, p. 282, 4 2020.

[134] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis, "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," pp. 1–12, Association for Computing Machinery (ACM), 4 2020.

[135] M. Amrehn, S. Steidl, R. Kortekaas, M. Strumia, M. Weingarten, M. Kowarschik, and A. Maier, "A semi-automated usability evaluation framework for interactive image segmentation systems," *International Journal of Biomedical Imaging*, vol. 2019, 2019.

[136] C. Zhang and K. Chaudhuri, "Active learning from weak and strong labelers," pp. 703–711, 10 2015.

[137] V. Cheplygina, A. Perez-Rovira, W. Kuo, H. A. Tiddens, and M. de Bruijne, "Early experiences with crowdsourcing airway annotations in chest ct," vol. 10008 LNCS, pp. 209–218, Springer Verlag, 10 2016.

[138] M. Rajchl, M. C. H. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, and D. Rueckert, "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," 2016.

[139] M. Rajchl, M. C. H. Lee, F. Schrans, A. Davidson, J. Passerat-Palmbach, G. Tarroni, A. Alansary, O. Oktay, B. Kainz, and D. Rueckert, "Learning under distributed weak supervision," 2016.

[140] M. Rajchl, L. M. Koch, C. Ledig, J. Passerat-Palmbach, K. Misawa, K. Mori, and D. Rueckert, "Employing weak annotations for medical image analysis problems," 2017.

[141] P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, "Deep learning for multi-task medical image segmentation in multiple modalities," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9901 LNCS, pp. 478–486, 10 2016.

[142] D. Wang, M. Li, N. Ben-Shlomo, C. E. Corrales, Y. Cheng, T. Zhang, and J. Jayender, "Mixed-supervised dual-network for medical image segmentation," vol. 11765 LNCS, pp. 192–200, Springer, 10 2019.

[143] D. Lowell, Z. C. Lipton, and B. C. Wallace, "Practical obstacles to deploying active learning," 2019.

[144] M. Nalisnik, D. A. Gutman, J. Kong, and L. A. Cooper, "An interactive learning framework for scalable classification of pathology images.," *Proceedings : ... IEEE International*

Conference on Big Data. IEEE International Conference on Big Data, vol. 2015, pp. 928–935, 2015.

[145] N. Khosravan, H. Celik, B. Turkbey, R. Cheng, E. McCreedy, M. McAuliffe, S. Bednarova, E. Jones, X. Chen, P. Choyke, B. Wood, and U. Bagci, "Gaze2segment: A pilot study for integrating eye-tracking technology into medical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10081 LNCS, pp. 94–104, 10 2016.

[146] J. N. Stember, H. Celik, E. Krupinski, P. D. Chang, S. Mutasa, B. J. Wood, A. Lignelli, G. Moonis, L. H. Schwartz, S. Jambawalikar, and U. Bagci, "Eye tracking for deep learning segmentation using convolutional neural networks," *Journal of Digital Imaging*, vol. 32, pp. 597–604, 8 2019.

[147] R. Tinati, M. Luczak-Roesch, E. Simperl, and W. Hall, "An investigation of player motivations in eyewire, a gamified citizen science project," *Computers in Human Behavior*, vol. 73, pp. 527–540, 2017.

[148] J. S. Kim, M. J. Greene, A. Zlateski, K. Lee, M. Richardson, S. C. Turaga, M. Purcaro, M. Balkam, A. Robinson, B. F. Behabadi, M. Campos, W. Denk, H. S. Seung, and the EyeWirers, "Space–time wiring specificity supports direction selectivity in the retina," *Nature*, vol. 509, pp. 331–336, 5 2014.

[149] W. Kuo, C. Häne, E. Yuh, P. Mukherjee, and J. Malik, "Cost-sensitive active learning for intracranial hemorrhage detection," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11072 LNCS, pp. 715–723, 9 2018.

[150] M. P. Shah, Y. S. Bhalgat, and S. P. Awate, "Annotation-cost minimization for medical image segmentation using suggestive mixed supervision fully convolutional networks," 2018.

[151] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 5 2019.

[152] M. A. Morid, A. Borjali, and G. D. Fiol, "A scoping review of transfer learning research on medical image analysis using imagenet," *Computers in Biology and Medicine*, vol. 128, p. 104115, 1 2021.

[153] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[154] Z. Zhou, J. Shin, R. Feng, R. T. Hurst, C. B. Kendall, and J. Liang, "Integrating active learning and transfer learning for carotid intima-media thickness video interpretation," *Journal of Digital Imaging*, vol. 32, pp. 290–299, 2019.

[155] K. Kushibar, S. Valverde, S. González-Villà, J. Bernal, M. Cabezas, A. Oliver, and X. Lladó, "Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction," *Scientific Reports*, vol. 9, p. 6742, 12 2019.

[156] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," 2 2021.

[157] A. Choudhary, L. Tong, Y. Zhu, and M. D. Wang, "Advancing medical imaging informatics by deep learning-based domain adaptation," *Yearbook of medical informatics*, vol. 29, pp. 129–138, 8 2020.

[158] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 5 2019.

[159] C. Baweja, B. Glocker, and K. Kamnitsas, "Towards continual learning in medical imaging," 2018.

[160] N. I. for Health and C. Excellence, "Judging whether public health interventions offer value for money — guidance and guidelines — nice," *2013*.

[161] S. Farquhar, Y. Gal, and T. Rainforth, “On statistical bias in active learning: How and when to fix it,” 2021.

[162] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert, “Sononet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound,” *IEEE Transactions on Medical Imaging*, vol. 36, pp. 2204–2215, 11 2017.

[163] M. Sinclair, C. F. Baumgartner, J. Matthew, W. Bai, J. C. Martinez, Y. Li, S. Smith, C. L. Knight, B. Kainz, J. Hajnal, A. P. King, and D. Rueckert, “Human-level performance on automatic head biometrics in fetal ultrasound using fully convolutional neural networks,” pp. 714–717, IEEE, 7 2018.

[164] L. Wu, Y. Xin, S. Li, T. Wang, P.-A. Heng, and D. Ni, “Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation,” pp. 663–666, IEEE, 4 2017.

[165] G. Carneiro, B. Georgescu, S. Good, and D. Comaniciu, “Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree,” *IEEE Transactions on Medical Imaging*, vol. 27, pp. 1342–1355, 9 2008.

[166] J. Li, Y. Wang, B. Lei, J.-Z. Cheng, J. Qin, T. Wang, S. Li, and D. Ni, “Automatic fetal head circumference measurement in ultrasound using random forest and fast ellipse fitting,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, pp. 215–223, 1 2018.

[167] S. Rueda, S. Fathima, C. L. Knight, M. Yaqub, A. T. Papageorghiou, B. Rahmatullah, A. Foi, M. Maggioni, A. Pepe, J. Tohka, R. V. Stebbing, J. E. McManigle, A. Ciurte, X. Bresson, M. B. Cuadra, C. Sun, G. V. Ponomarev, M. S. Gelfand, M. D. Kazanov, C.-W. Wang, H.-C. Chen, C.-W. Peng, C.-M. Hung, and J. A. Noble, “Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: A grand challenge,” *IEEE Transactions on Medical Imaging*, vol. 33, pp. 797–813, 4 2014.

[168] F. P. Hadlock, R. B. Harrist, R. S. Sharman, R. L. Deter, and S. K. Park, “Estimation of fetal weight with the use of head, body, and femur measurements-a prospective study,” *American Journal of Obstetrics and Gynecology*, vol. 151, pp. 333–337, 2 1985.

[169] M. Monteiro, L. L. Folgoc, D. C. de Castro, N. Pawlowski, B. Marques, K. Kamnitsas, M. van der Wilk, and B. Glocker, “Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty,” 6 2020.

[170] M. A. Bochicchio, A. Longo, L. Vaira, and S. Ramazzina, “Online data analysis of fetal growth curves,” vol. 8286 LNCS, pp. 149–156, Springer, Cham, 12 2013.

[171] Q. Meng, J. Matthew, V. A. Zimmer, A. Gomez, D. F. Lloyd, D. Rueckert, and B. Kainz, “Mutual information-based disentangled neural networks for classifying unseen categories in different domains: Application to fetal ultrasound imaging,” *IEEE Transactions on Medical Imaging*, vol. PP, pp. 1–1, 11 2020.

[172] N. Fasp, “Nhs fetal anomaly screening programme handbook valid from august 2018,” p. 133, 2018.

[173] I. Sarris, C. Ioannou, P. Chamberlain, E. Ohuma, F. Roseman, L. Hoch, D. G. Altman, A. T. Papageorghiou, I. Fetal, and N. G. C. for the 21st Century (INTERGROWTH-21st), “Intra- and interobserver variability in fetal ultrasound measurements,” *Ultrasound in Obstetrics and Gynecology*, vol. 39, pp. 266–273, 3 2012.

[174] S. Kohl, B. Romera-Paredes, C. Meyer, J. D. Fauw, J. R. Ledsam, K. Maier-Hein, S. M. A. Eslami, D. J. Rezende, and O. Ronneberger, “A probabilistic u-net for segmentation of ambiguous images,” pp. 6965–6975, 2018.

[175] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu, “Phiseg: Capturing uncertainty in medical image segmentation,” vol. 11765 LNCS, pp. 119–127, Springer, 10 2019.

[176] A. Fitzgibbon, M. Pilu, and R. Fisher, “Direct least squares fitting of ellipses,” pp. 253–257 vol.1, IEEE, 1996.

[177] R. W. Barnard, K. Pearce, and L. Schovanec, "Inequalities for the perimeter of an ellipse," *Journal of Mathematical Analysis and Applications*, vol. 260, pp. 295–306, 8 2001.

[178] D. K. Prasad, M. K. Leung, and C. Quek, "Ellifit: An unconstrained, non-iterative, least squares based geometric ellipse fitting method," *Pattern Recognition*, vol. 46, pp. 1449–1465, 5 2013.

[179] T. L. A. van den Heuvel, D. de Bruijn, C. L. de Korte, and B. van Ginneken, "Automated measurement of fetal head circumference using 2d ultrasound images," *PLOS ONE*, vol. 13, p. e0200412, 8 2018.

[180] D. Y. Yang, D. Beam, K. A. Pelphrey, S. Abdullahi, and R. J. Jou, "Cortical morphological markers in children with autism: A structural magnetic resonance imaging study of thickness, area, volume, and gyrification," *Molecular Autism*, vol. 7, p. 11, 1 2016.

[181] N. R. Lee, E. I. Adeyemi, A. Lin, L. S. Clasen, F. M. Lalonde, E. Condon, D. I. Driver, P. Shaw, N. Gogtay, A. Raznahan, and J. N. Giedd, "Dissociations in cortical morphometry in youth with down syndrome: Evidence for reduced surface area but increased thickness.," *Cerebral cortex (New York, N.Y. : 1991)*, vol. 26, pp. 2982–90, 2016.

[182] R. J. Leventer, R. Guerrini, and W. B. Dobyns, "Malformations of cortical development and epilepsy," *Dialogues in Clinical Neuroscience*, vol. 10, pp. 47–62, 2008.

[183] P. Mukherjee, A. Sabharwal, R. Kotov, A. Szekely, R. Parsey, D. M. Barch, and A. Mohanty, "Disconnection between amygdala and medial prefrontal cortex in psychotic disorders.," *Schizophrenia bulletin*, vol. 42, pp. 1056–67, 2016.

[184] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. V. Essen, "A multi-modal parcellation of human cerebral cortex," *Nature Publishing Group*, vol. 536, 2016.

[185] A. Makropoulos, E. C. Robinson, A. Schuh, R. Wright, S. Fitzgibbon, J. Bozek, S. J. Counsell, J. Steinweg, K. Vecchiato, J. Passerat-Palmbach, G. Lenz, F. Mortari, T. Tenev,

E. P. Duff, M. Bastiani, L. Cordero-Grande, E. Hughes, N. Tusor, J. D. Tournier, J. Hutter, A. N. Price, R. P. A. Teixeira, M. Murgasova, S. Victor, C. Kelly, M. A. Rutherford, S. M. Smith, A. D. Edwards, J. V. Hajnal, M. Jenkinson, and D. Rueckert, "The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction," *NeuroImage*, vol. 173, pp. 88–112, 6 2018.

[186] R. Dahnke, R. A. Yotter, and C. Gaser, "Cortical thickness and central surface estimation," *NeuroImage*, vol. 65, pp. 336–348, 1 2013.

[187] N. J. Tustison, P. A. Cook, A. Klein, G. Song, S. R. Das, J. T. Duda, B. M. Kandel, N. van Strien, J. R. Stone, J. C. Gee, and B. B. Avants, "Large-scale evaluation of ants and freesurfer cortical thickness measurements," *NeuroImage*, vol. 99, pp. 166–179, 10 2014.

[188] S. R. Das, B. B. Avants, M. Grossman, and J. C. Gee, "Registration based cortical thickness measurement," *NeuroImage*, vol. 45, pp. 867–879, 4 2009.

[189] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. V. Essen, and M. Jenkinson, "The minimal preprocessing pipelines for the human connectome project," *NeuroImage*, vol. 80, pp. 105–124, 2013.

[190] B. Fischl and A. M. Dale, "Measuring the thickness of the human cerebral cortex from magnetic resonance images," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, pp. 11050–11055, 9 2000.

[191] Özgün Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," vol. 9901 LNCS, pp. 424–432, Springer Verlag, 6 2016.

[192] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Dropblock: A regularization method for convolutional networks," *Advances in Neural Information Processing Systems*, vol. 2018-December, pp. 10727–10737, 10 2018.

[193] E. J. Hughes, T. Winchman, F. Padormo, R. Teixeira, J. Wurie, M. Sharma, M. Fox, J. Hutter, L. Cordero-Grande, A. N. Price, J. Allsop, J. Bueno-Conde, N. Tusor, T. Arichi, A. D. Edwards, M. A. Rutherford, S. J. Counsell, and J. V. Hajnal, "A dedicated neonatal brain imaging system," *Magnetic Resonance in Medicine*, vol. 78, pp. 794–804, 8 2017.

[194] D. S. Marcus, J. Harwell, T. Olsen, M. Hodge, M. F. Glasser, F. Prior, M. Jenkinson, T. Laumann, S. W. Curtiss, and D. C. V. Essen, "Informatics and data mining tools and strategies for the human connectome project," *Frontiers in Neuroinformatics*, vol. 5, p. 4, 6 2011.

[195] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *Journal of Machine Learning Research*, vol. 20, 8 2018.

[196] H. BJ, M. JA, and W. CR, "Prenatal diagnosis of critical congenital heart disease reduces risk of death from cardiovascular compromise prior to planned neonatal cardiac surgery: a meta-analysis," *Ultrasound in obstetrics and gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*, vol. 45, pp. 631–638, 6 2015.

[197] C. J, A. N, M. S, P. MH, J. I, and B. D, "Impact of prenatal diagnosis on neurocognitive outcomes in children with transposition of the great arteries," *The Journal of pediatrics*, vol. 161, 2012.

[198] J. Tan, A. Au, Q. Meng, S. FinesilverSmith, J. Simpson, D. Rueckert, R. Razavi, T. Day, D. Lloyd, and B. Kainz, "Automated detection of congenital heart disease in fetal ultrasound screening," vol. 12437 LNCS, pp. 243–252, Springer Science and Business Media Deutschland GmbH, 10 2020.

[199] F. Miceli, "A review of the diagnostic accuracy of fetal cardiac anomalies," *Australasian Journal of Ultrasound in Medicine*, vol. 18, pp. 3–9, 2 2015.

[200] R. Arnaout, L. Curran, E. Chinn, Y. Zhao, and A. Moon-Grady, "Deep-learning models improve on community-level diagnosis for common congenital heart disease lesions," *arXiv*, 9 2018.

[201] L. Yeo and R. Romero, “Fetal intelligent navigation echocardiography (fine): a novel method for rapid, simple, and automatic examination of the fetal heart,” *Ultrasound in Obstetrics and Gynecology*, vol. 42, pp. 268–284, 9 2013.

[202] J. R. Clough, I. Oksuz, N. Byrne, J. A. Schnabel, and A. P. King, “Explicit topological priors for deep-learning based image segmentation using persistent homology,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11492 LNCS, pp. 16–28, 2019.

[203] X. Hu, F. Li, D. Samaras, and C. Chen, “Topology-preserving deep image segmentation,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[204] T. V. Sushma, N. Sriraam, P. M. Arakeri, and S. Suresh, “Classification of fetal heart ultrasound images for the detection of chd,” *Lecture Notes on Data Engineering and Communications Technologies*, vol. 59, pp. 489–505, 2021.

[205] R. Arnaout, L. Curran, Y. Zhao, J. Levine, E. Chinn, and A. Moon-Grady, “Expert-level prenatal detection of complex congenital heart disease from screening ultrasound using deep learning,” *medRxiv*, p. 2020.06.22.20137786, 6 2020.

[206] C. E. Rasmussen and H. Nickisch, “Gaussian processes for machine learning (gpml) toolbox,” *Journal of Machine Learning Research*, vol. 11, pp. 3011–3015, 2010.

[207] N. Duta, “Procrustes shape distance,” *Encyclopedia of Biometrics*, pp. 1278–1279, 2015.

[208] E. Commision, “Prevalence charts and tables,” *EU RD Platform*, 2021.

[209] Labelbox, “Labelbox: The leading training data platform for data labeling,” 2021.

[210] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[211] J. C. Chang, S. Amershi, and E. Kamar, “Revolt: Collaborative crowdsourcing for labeling machine learning datasets,” *Conference on Human Factors in Computing Systems - Proceedings*, vol. 2017-May, pp. 2334–2346, 5 2017.

[212] S. Yu, M. Chen, E. Zhang, J. Wu, H. Yu, Z. Yang, L. Ma, X. Gu, and W. Lu, “Robustness study of noisy annotation in deep learning based medical image segmentation,” *Physics in Medicine and Biology*, vol. 65, p. 175007, 9 2020.

[213] R. Snow, B. O’Connor, D. Jurafsky, and A. Y. Ng, “Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks,” pp. 254–263, 2008.

[214] J. Fang, B. Price, and L. Price, “Pruning non-informative text through non-expert annotations to improve aspect-level sentiment classification,” pp. 37–45, 2010.

[215] E. Jamison and I. Gurevych, “Needle in a haystack: Reducing the costs of annotating rare-class instances in imbalanced datasets,” pp. 244–253, 2014.

[216] F. Wilm, C. A. Bertram, C. Marzahl, A. Bartel, T. A. Donovan, C.-A. Assenmacher, K. Becker, M. Bennett, S. Corner, B. Cossic, D. Denk, M. Dettwiler, B. G. Gonzalez, C. Gurtner, A. Lehmbecker, S. Merz, S. Plog, A. Schmidt, R. C. Smedley, M. Tecilla, T. Thaiwong, K. Breininger, M. Kiupel, A. Maier, R. Klopfleisch, and M. Aubreville, “How many annotators do we need? – a study on the influence of inter-observer variability on the reliability of automatic mitotic figure assessment,” *arXiv*, 12 2020.

[217] P.-Y. Hsueh, P. Melville, and V. Sindhwani, “Data quality from crowdsourcing: A study of annotation selection criteria,” *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pp. 27–35, 2009.

[218] E. Heim, T. Roß, A. Seitel, K. März, B. Stieltjes, M. Eisenmann, J. Lebert, J. Metzger, and G. Sommer, “Large-scale medical image annotation with crowd-powered algorithms,” *Journal of Medical Imaging*, vol. 5, p. 1, 9 2018.