Fetal Gestational Age Estimation without Clinical Measurement using Deep Learning

Lok Hin Lee

Hertford College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Hilary 2021

Abstract

Gestational age estimation is a key marker in obstetric care for determination of fetal growth and health. Current clinical estimation of gestational age is based on regression using clinical fetal growth charts and biometry measurements by an experienced sonographer. However, estimation of gestational age using biometry is difficult especially in late pregnancy, and there may be additional salient anatomical information in images to assist gestational age estimation in standard planes which is discarded by clinical biometry measurement. This thesis therefore attempts to use deep neural network to directly regress gestational age from fetal standard planes without pixel size information, thus relying on anatomical appearance alone without fetal biometry.

In this thesis, we first investigate the data augmentation methods commonly used for medical image analysis and implement an data augmentation selection strategy. We find that this optimized augmentation strategy leads to increased performance on the proxy task of standard plane classification compared to conventional hand-crafted data augmentation methods.

We then use this data augmentation strategy selection framework to train a deep convolutional neural network for gestational age estimation from the head, abdomen and femur standard planes, and investigate the performance of different loss functions for deep regression. We demonstrate gestational age estimation across the second and third trimester on the head standard plane with a mean absolute error of 0.6 weeks with the optimized data augmentation policy, and use the trained neural network to visualize anatomically salient areas of the input image.

We then extend the deep learning framework by introducing priors to trainable weights, making a Bayesian neural network. The Bayesian neural network separately estimates aleatoric and epistemic uncertainty during gestational age inference. This allows for calibrated uncertainties during gestational age estimation.

This is the first work to estimate fetal gestational age from ultrasound standard plane images without pixel size or biometry information, and may be potentially of use in low- and medium- income countries where accurate gestational age dating in late pregnancy is important as women present to obstetric care in all stages of pregnancy.

Fetal Gestational Age Estimation without Clinical Measurement using Deep Learning



Lok Hin Lee

Hertford College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2021

Acknowledgements

Personal

I would like to thank my supervisor, Professor Alison Noble, who has continually provided guidance and support through this journey. This thesis would not have been possible without her.

My appreciation also goes out to my colleagues and collaborators, without whom lab research would have been a lot more dull and difficult. In particular to the untiring efforts of Dr. Elizabeth Bradburn, without whom the work would not have been possible. It has been an absolute pleasure working with you.

I am extremely grateful to my family and especially my parents. Without their selfless love and understanding I would not have been able to start and finish this journey.

Institutional

I would like to thank the Croucher Foundation for their assistance in my Oxford studies. I would also like to thank the INTERGROWTH-21st project which has provided the data for this research.

Abstract

Gestational age estimation is a key marker in obstetric care for determination of fetal growth and health. Current clinical estimation of gestational age is based on regression using clinical fetal growth charts and biometry measurements by an experienced sonographer. However, estimation of gestational age using biometry is difficult especially in late pregnancy, and there may be additional salient anatomical information in images to assist gestational age estimation in standard planes which is discarded by clinical biometry measurement. This thesis therefore attempts to use deep neural network to directly regress gestational age from fetal standard planes without pixel size information, thus relying on anatomical appearance alone without fetal biometry.

In this thesis, we first investigate the data augmentation methods commonly used for medical image analysis and implement an data augmentation selection strategy. We find that this optimized augmentation strategy leads to increased performance on the proxy task of standard plane classification compared to conventional hand-crafted data augmentation methods.

We then use this data augmentation strategy selection framework to train a deep convolutional neural network for gestational age estimation from the head, abdomen and femur standard planes, and investigate the performance of different loss functions for deep regression. We demonstrate gestational age estimation across the second and third trimester on the head standard plane with a mean absolute error of 0.6 weeks with the optimized data augmentation policy, and use the trained neural network to visualize anatomically salient areas of the input image.

We then extend the deep learning framework by introducing priors to trainable weights, making a Bayesian neural network. The Bayesian neural network separately estimates aleatoric and epistemic uncertainty during gestational age inference. This allows for calibrated uncertainties during gestational age estimation.

This is the first work to estimate fetal gestational age from ultrasound standard plane images without pixel size or biometry information, and may be potentially of use in low- and medium- income countries where accurate gestational age dating in late pregnancy is important as women present to obstetric care in all stages of pregnancy.

Contents

List of Figures **xi**

List of Tables **xxiii**

List of Abbreviations **xxv**

1 Introduction **1**
1.1 Clinical Motivation 1
1.2 Thesis Contributions 3
1.3 Thesis Structure 4
1.4 Originality 5
1.5 List of Accepted Publications Contributing to This Thesis 6
1.6 List of Accepted Publications Referenced in This Thesis 6

2 Literature Review **7**
2.1 Introduction 8
2.2 Clinical Gestational Age Estimation 8
2.2.1 LMP-Based Gestational Age Estimation 8
2.2.2 Ultrasound Based Clinical Gestational Age Estimation . . . 9
2.2.3 First-Trimester Clinical Fetal Gestational Age Estimation . . 9
2.2.4 Second and Third-Trimester Clinical Fetal Age Estimation . 11
2.2.5 Other Non-Standard Measurements for Gestational Age Estimation 18
2.2.6 Challenges for Clinical Gestational Age Estimation 20
2.3 Classical Automated Gestational Age Estimation 21
2.4 Convolutional Neural Networks 22
2.4.1 Fully-Connected Layers 23
2.4.2 Convolutional Layers 24
2.4.3 Activation Functions 24
2.4.4 Pooling Layers 25
2.4.5 Optimization 26
2.4.6 Deep Residual Networks - ResNet, ResNeXt 28
2.4.7 Data Augmentation 28

2.5 Deep Learning for Regression . . . 31
2.6 Deep Learning based Gestational Age Estimation . . . 33
2.7 Conclusions . . . 34

3 Principled Data Augmentation for Fetal Plane Classification 35
3.1 Introduction . . . 36
3.2 Delving into Data Augmentation Methods used in Biomedical Ultrasound . . . 37
3.2.1 Original Data Manipulation . . . 37
3.2.2 Artifical Data Generation . . . 42
3.2.3 Principled Data Augmentation . . . 45
3.2.4 Quantification of Augmentation Strategies . . . 45
3.3 Dataset . . . 47
3.3.1 Pre-processing and Annotation . . . 47
3.4 Experiments and Results . . . 49
3.4.1 Experiment Implementation Details . . . 49
3.4.2 Quantitative Evaluation on Fetal Plane Classification . . . 53
3.4.3 Qualitative Evaluation on Fetal Plane Classification . . . 58
3.4.4 Evaluation of Augmentation Policies using Affinity and Diversity 61
3.5 Discussion and Conclusion . . . 61

4 Regression of Fetal Gestational Age using CNNs 65
4.1 Introduction . . . 66
4.2 Dataset . . . 67
4.2.1 Pre-processing . . . 67
4.2.2 Pre-processing . . . 68
4.2.3 Annotation . . . 68
4.3 Supervised Regression using Neural Networks . . . 72
4.3.1 The Classification-Based Binning Approach . . . 73
4.3.2 Direct Deep Regression . . . 75
4.3.3 Rank-consistent Ordinal Regression . . . 75
4.4 Experiments and Results . . . 76
4.4.1 SE-ResNeXt-50 . . . 76
4.4.2 Experiments . . . 77
4.4.3 Implementation Details . . . 78
4.4.4 Quantitative Evaluation . . . 79
4.4.5 Qualitative Evaluation . . . 87
4.4.6 Visualization of Anatomical Structures of Interest . . . 89
4.5 Discussion and Conclusion . . . 93

5 Uncertain Gestational Age Estimation of Fetal Head Images with Bayesian Neural Networks 95
5.1 Introduction . . . 96
5.2 Dataset . . . 97
5.2.1 Pre-processing . . . 98
5.2.2 Validation Datasets . . . 99
5.3 Training Bayesian Neural Networks . . . 100
5.3.1 Variational Inference . . . 102
5.3.2 Bayes by Backpropagation . . . 102
5.3.3 Minimizing Gradient Variance using Flipout Monte Carlo Estimator . . . 103
5.3.4 Kullbeck-Leibler Loss Annealing . . . 104
5.4 Aleatoric and Epistemic Uncertainty . . . 105
5.5 Accuracy of Uncertainty Estimates . . . 106
5.5.1 Uncertainty Calibration . . . 108
5.6 Experiments and Results . . . 109
5.6.1 Implementation Details . . . 111
5.6.2 Quantitative Evaluation . . . 111
5.6.3 Visualization of Uncertainty Calibration . . . 114
5.6.4 Evaluation of Aleatoric and Epistemic Uncertainties . . . 115
5.6.5 Out of Distribution Data Detection . . . 120
5.7 Discussion and Conclusion . . . 123

6 Summary and Future Work 125
6.1 Main Thesis Contributions . . . 125
6.2 Future Work . . . 128
6.2.1 Gestational Age Estimation from Concurrent Multi-Anatomy Standard Planes . . . 128
6.2.2 Mixed Example Extended RandAugment on other Ultrasound Tasks . . . 130
6.2.3 End-to-end Solution for Gestational Age Estimation in a Clinical Setting . . . 131

References 133

List of Figures

2.1 **Gestational age estimation in the first trimester via crown rump length (CRL) measurement. Left:** An example of a first trimester measurement plane taken of a fetus at 8+4 weeks of gestational age, and the CRL measurement in red. **Right:** The clinical chart used for gestational age regression based on first trimester crown rump length. The solid blue line represents the mean gestational age for a given crown rump length, with each successive blue area representing the 95% confidence interval and the 99% confidence interval respectively. 10

2.2 **An example of a 22+1 week transthalamic plane of the fetal head. Image is taken from the INTERGROWTH-21st [26] study. Left:** Key biometries used in gestational age regression, including the occipito-frontal diameter (OFD), biparietal diameter (BPD) and head circumference (HC). Markers are placed by an experienced sonographer. **Right:** Some of the required anatomies seen in the transthalamic plane, including the Cavum Septum Pellucidum (CSP), Choroid Plexus (CP), and Thalami. Image is best viewed in colour. 12

2.3 **Fetal Biometry Growth Charts for the Head Circumference Plane derived from the INTERGROWTH-21st study.** Fetal biometry growth charts for the head circumference plane derived from the INTERGROWTH-21st study. Charts are reproduced from [4]. Gestational Ages (GA) from **Top:** Head Circumference (HC); **Middle:** Biparietal Diameter (BPD); **Bottom:** Occipitofrontal Diameter (OCD); All measurements are in millimeters. Different lines on the same chart show the percentile distribution of biometries. The solid blue line represents the mean biometry for a given gestational age, with each successive blue area representing the 95% confidence interval and the 99% confidence interval respectively. 14

2.4 **An example of a 26+4 week fetal abdominal circumference plane. Image is taken from the INTERGROWTH-21st [26] study. Left:** The abdominal circumference (AC) commonly used for gestational age estimation. Markers are placed by an experienced sonographer. **Right:** Some of the required anatomies seen in the abdominal circumference plane, including the fetal stomach, umbilical vein, vertebral body (spine) and ribs. Image is best viewed in colour. 15

2.5 **Fetal Biometry Growth Charts for the Abdominal Circumference Plane derived from the INTERGROWTH-21st study.** Fetal biometry growth charts for gestational age (GA) estimation using the abdominal circumference (AC) derived from the INTERGROWTH-21st study. Charts are reproduced from [4]. The solid blue line represents the mean biometry for a given gestational age, with each successive blue area representing the 95% confidence interval and the 99% confidence interval respectively. 16

2.6 **An example of fetal femur length planes. Images are taken from the INTERGROWTH-21st [26] study. Left:** The femur length plane from a 16+2 week fetus **Right:** The femur length plane from a 26+4 week fetus. 17

2.7 **Fetal Biometry Growth Charts for the Femur Length Plane derived from the INTERGROWTH-21st study.** Fetal biometry growth charts for gestational age (GA) estimation using the femur length (FL) derived from the INTERGROWTH-21st study. Charts are reproduced from [4]. The solid blue line represents the mean biometry for a given gestational age, with each successive blue area representing the 95% confidence interval and the 99% confidence interval respectively. 18

2.8 **Examples of correct and incorrect standard plane selections. Left:** Correctly identified standard planes for the three anatomical planes (1: Femur plane, 2: Abdominal circumference plane, 3: Transthalamic head plane.) **Right:** Examples of misplaced standard planes on the same fetus, leading to misleading measurements. 1 and 3 will lead to shortened measured biometries and hence underestimation of the true gestational age, 2 will lead to an artificially lengthened abdominal circumference measurement and therefore an overestimation of the true gestational age. Figure is reproduced from [60]. 21

2.9 **Architectural overview of AlexNet [77], a convolutional neural network.** AlexNet, typical of most convolutional neural networks, is composed of convolutional layers (depicted in yellow), pooling layers (depicted in orange), and fully-connected layers (depicted in purple). The numbers indicate the size of each feature map and the number of channels. ReLU non-linearities are included after every convolutional layer, but not shown on the figure for brevity. The size of each feature map is not to scale, and the figure is best viewed in colour. 23

2.10 **Base residual block of a ResNet [89] network** The ResNet module consists of a skip connection and multiple convolutional layers, with a ReLU non-linearity in the middle. The figure depicts a ResNet module with two convolutional layers, and the 3×3 within each convolutional block refers to the kernel height and width respectively. 27

2.11 **Examples of common individual augmentation transformations and their effects on an image** The original image in the training set (fetal HCP) is shown on the left. The remaining four training images are generated from the original image using commonly used augmentation transformations. 28

2.12 **Comparison between heatmap regression and deep regression for an illustrative Cavum Septum Pellucidum localization task** The figure shows the comparison between heatmap regression and deep regression for the localization task for the Cavum Septum Pellucidum (CSP) and the differences between the neural network architectures used for this task. **Left:** Heatmap regression neural networks generate heatmaps in the same shape as the input feature map. **Right:** Deep regression neural networks directly regress to the desired continous variables, here the pixel coordinate value locations for the CSP. Convolutional neural networks in both tasks are depicted as example architectures with nomenclature following that of figure 2.9 and are for illustrative purposes only. In practice, the convolutional neural networks used for these tasks will likely be more complex. 31

3.1 **Common image manipulations of training data illustrated on a fetal spine plane image** The figure shows examples of original training data **identity** manipulated with various transformations. **Vertical Flipping, Horizontal Flipping, Rotation, Shear (X-Axis), Shear (Y-Axis), Translation (X-Axis), Translation (Y-Axis)** are examples of affine transformations. **Auto Contrast, Equalize, Invert, Posterize, Solarize, Contrast, Brightness, Sharpness** are examples of image histogram transformations. The original image is of fetal spine in the coronal plane captured in a routine anomaly scan. 38

3.2 **Non-Linear Energy-Normalized Mixed-Example Data Generation** The figure shows the procedure for energy normalized non-linear generation of mixed-example data. Three hyperparameters are used, $\lambda_1, \lambda_2, \lambda_3$ which control the degree of mixing between each original training image x_1 and x_2, generating the output image $\hat{x}$ . . 44

3.3 Example images of fetal planes used in fetal standard plane classification task 48

3.4 **Augmented images through different hyperparameters of the Mix. RA algorithm** The first column shows the original abdominal circumference plane image used for augmentation. The other columns show images augmented with (2, 2), (5, 5) and (7, 7) (m, n) respectively. As can be seen from the figure, increasing the magnitude of distortion for each augmentation and increasing the number of transformations selected drastically changes final augmented image distortion. 52

3.5 **Final training classification accuracies for hyperparameter grid-search using RA for a SE-ResNeXt-50 model trained from random initialization Top Left:** Best performing classification accuracy for various (m, n) tuples. **Top Right:** Variation in accuracy for evaluated models in standard deviation. **Bottom:** Bottom shows the example augmented training set generated with $(m, n) = (5, 3)$. The original ground truth image can be found in figure 3.4. 54

3.6 **Final training accuracies for hyperparameter grid-search using Mix. RA for a SE-ResNeXt-50 model trained from random initialization Top Left:** Best performing accuracy for various (m, n) tuples. **Top Right:** Variation in accuracy for evaluated models in standard deviation. **Bottom:** Bottom shows the example augmented training set generated with $(m, n) = (3, 3)$. The original ground truth image can be found in figure 3.4. 55

3.7 **Class confusion matrix for the best performing augmentation policy found by Mix RA. augmentation policy search Left:** Confusion matrix for accuracies broken down by class. **Right:** The difference in accuracy between a model trained with Mix. RA and a model trained with SN Pol. Both models were trained from random initialization. 56
3.8 **t-SNE visualizations of the extracted high level features from best performing RA and Mix. RA augmentation policies** 59
3.9 **Integrated Gradients based attribution for image classification on RA and Mix. RA** The model trained with the RA augmentation policy mis-classifies these images as background. These images are subsequently correctly classified by the Mix. RA model. The integrated gradients (Int. Grad.) based attribution assigns the attribution for the classification. 60
3.10 **The effects of m and n on augmented image diversity and affinity** Each point represents a (m, n) point in the hyperparameter search space. The best performing m and n hyperparameter for each augmentation policy search method is highlighted in red. Colours indicate final overall test accuracy. 62

4.1 **Ground truth gestational age distribution of the dataset of images from INTERGROWTH-21$^{\text{st}}$.** Ground truth gestational ages are approximately uniformly distributed across the entire gestational age range due to the acquisition protocol defined in the INTERGROWTH-21st study. **Left:** Femur length plane **Middle:** Abdomen circumference plane **Right:** Head circumference plane . . 68
4.2 Examples of the three planes (Abdominal, Femur, Head) used in the gestational age estimation work derived from INTERGROWTH-21$^{\text{st}}$ Images of each plane are taken from the first, second and third trimester in the first, second and third rows respectively. Obvious acoustic shadowing artefacts can be seen in the head and abdomen images for the second and third trimester, indicative of a common features of fetal ultrasound imaging. 69
4.3 **INTERGROWTH Standard Plane Image Cleaning and Pre-Processing.** The standard plane images included text and sonographer markings. These markings were found with template matching and nearest-neighbour averaging. Images were then resized before input into the CNN due to memory limitations. 70

4.4 **The anatomical definition of the femur and abdominal standard planes used in clinical gestational age estimation Top:** The fetal anatomical structures used to define the abdominal plane include a segment of the umbilical vein (1), the stomach bubble (2) and the spine (3). **Bottom:** The femur length is measured between (4), with the greater trochanter (highlighted in red) not included in the femur measurement. 71

4.5 **The anatomical definition of the head circumference standard planes used in clinical gestational age estimation.** The appearance of each standard plane was defined by a central and continuous midline (1), with a symmetrical thalami around the midline (2) and the cavum septi pellucidi along 1/3 of the midline. The fetal head should also take up at least 30% of the image. This example fetal head standard plane has a gestational age of 26.3 weeks. 72

4.6 **The base block of the SE-ResNeXt model [152]** The figure shows the base building block of the SE-ResNeXt model. Each blue node represents a convolutional operation with parameters represented by (input channel number, kernel size, output channel number). The red nodes represent the summation operator, and the green nodes represent a 256-dimensional input feature and output feature respectively. The SE-Block includes a dimensionality reduction ratio r, which is set to 16 by default [152]. 77

4.7 **Predictions against ground truth gestational age for the best performing models trained with regression, classification and ordinal loss Top row:** Results of the best models trained with regression loss. **Middle row:** Results of the best models trained with classification loss. **Bottom row:** Results of the best models trained with ordinal loss. **Left column:** Head circumference plane. **Middle column:** Abdominal circumference plane. **Right column:** Femur length plane. The red line shows the equality line where the predicted gestational age is equal to ground truth gestational age. . 83

4.8 **Bland-Altman plots [155] to compare the distribution of the predicted gestational ages and ground truth gestational ages.** Each chart shows the Bland-Altman plots for the predicted gestational age and the ground truth gestational age **Top:** Models trained with regression loss. **Middle:** Models trained with classification loss. **Bottom:** Models trained with ordinal loss. HCP (left), ACP (middle) and FLP (right) represent the head circumference plane, femur length plane and abdominal circumference plane respectively. 84

4.9 **Absolute predictive error for trained models against ground truth gestational age** Each chart shows the distribution of absolute error against ground truth gestational age. **Top:** Error distribution of models trained with regression loss. **Middle:** Error distribution of models trained with classification loss. **Bottom:** Error distribution of models trained with ordinal loss. HCP (left), ACP (middle) and FLP (right) represent the head circumference plane, femur length plane and abdominal circumference plane respectively. 85

4.10 **Successful examples of gestational age prediction for each fetal standard plane. Top:** The output of neural networks trained to predict gestational age. The bar chart shows the distribution of the final predictions of a model trained with a classification loss. The green line is the prediction of the model trained with the best performing regression loss. The ground truth is indicated by the blue arrow on the x-axis. **Bottom:** Images of the head circumference plane (HCP), abdominal plane (ACP) and femur length plant (FLP) respectively used to generate the regression results. 86

4.11 **Multi-modal results from classification model gestational age estimation Top:** The output of neural networks trained to predict gestational age. The bar chart shows the distribution of the final predictions of a model trained with a classification loss. The green line is the prediction of the model trained with a regressive L_1 loss. The red line is the ground truth. **Bottom:** Images of the head circumference plane (HCP), abdominal plane (ACP) and femur length plant (FLP) respectively. 87

4.12 **Images with high gestational age prediction error due to not adhering to acquisition protocols Left:** The HCP image shows significant shadowing and the fetal head is not fully in view. **Middle:** The abdominal plane does not have a clear umbilical vein section and stomach bubble, and is not fully in view with a portion of the ribs hidden by acoustic shadowing. **Right:** The femur is not clearly visible in the image, and is not situated horizontally. The femur does not constitute 30% of the image. 89

4.13 **Attributions to input images of fetal planes less than 23 weeks of gestational age** Images are from left to right columns the HCP, ACP and FLP planes respectively. The green indicates the degree of attribution to each individual pixel for gestational age estimation. Images are best viewed digitally. 90

4.14 **Attributions to input images of fetal planes between 23 weeks and 31 weeks gestational age.** Images are from left to right columns the HCP, ACP and FLP planes respectively. The green indicates the degree of attribution to each individual pixel for gestational age estimation. Images are best viewed digitally 91
4.15 **Attributions to input images of fetal planes greater than 31 weeks gestational age.** Images are from left to right columns the HCP, ACP and FLP planes respectively. The green indicates the degree of attribution to each individual pixel for gestational age estimation. Images are is best viewed digitally 92

5.1 **Ground Truth Gestational Age Distribution of Fetal Head Images from INTERGROWTH.** Due to the longitudinal study formulation, ground truth gestational ages were well distributed across the entire gestatational age range. Pregnant women included in the study were invited back to acquire additional ultrasound scans every 5 weeks. Instances of fetal head images are therefore distributed in all three trimesters, with the youngest fetal head images at 13 weeks and the oldest at 42 weeks. 98
5.2 **Examples of the INTERGROWTH-21st and PULSE datasets used in Bayesian neural network training and testing. Left:** A picture of a fetal head with GA of 20+2 weeks from the PULSE dataset, used only for validation. **Middle:** A picture of a fetal head drawn from the INTERGROWTH-21st dataset with GA of 21+2 weeks. **Right:** A picture of a fetal femur drawn from the INTERGROWTH-21st dataset with GA 21+2 weeks. 99
5.3 **Differences Between Conventional Convolutional Neural Networks and Bayesian Neural Networks.** Top: In a conventional neural network, each weight in the hidden layer is characterized by a single point value. Bottom: in a Bayesian neural network, probability distributions over weights are learned. 101
5.4 **Confidence Interval Accuracy for a given Neural Network.** The above figures show confidence intervals (in red) for each prediction (in blue) from the same neural network. **Left (Calibrated):** A well-calibrated 80% confidence interval would correctly predict confidence intervals for 80% inferences, which is the case here with five examples. **Right (Mis-Calibrated):** A well-calibrated 40% confidence interval would have an expectation that the ground truth lies within the confidence interval 40% of the time. However, 60% of the predictions lie within the confidence interval, implying that the right figure is mis-calibrated. 106

5.5 **Quantifying Uncertainty Calibration in Deep Learning. Left:** A well-calibrated uncertainty estimate would have the accuracy of prediction similar to the confidence of predictions for every given confidence interval. **Right** Mis-calibrated uncertainty estimates will have accuracies different from estimated confidence intervals. In this case, the model is consistently under-forecasting and estimating uncertainties that are too low. Calibration errors are calculated with Equation 5.6. For illustrative purposes, this figure is shown with $m = 10$ with equally spaced confidence intervals for evaluation. . . . 108

5.6 **Schematic of the VGG-16 neural network architecture used in this chapter.** The neural network architecture used in this chapter for both Bayesian neural networks and convolutional neural networks. Each **conv** block consists of multiple rectified convolutional layers (yellow) followed by a maxpooling layer (orange). In contrast to the VGG-16 network as described in [79], a general average pooling (GAP) layer is used at the top for feature aggregation before the final aleatoric uncertainty and gestational age estimation. Due to the regression objective, the final layer consists of fully connected layer with a linear activation. 110

5.7 **Schematic of the ResNet-18 neural network architecture used in this chapter.** The ResNet-18 neural network is defined in line with [89]. A general average pooling (GAP) layer is used at the top for feature aggregation before the final aleatoric uncertainty and gestational age estimation. Blue arrows represent skip connections, and a convolutional layer with a (1x1) kernel is used in the skip connection when there is a mismatch between feature sizes in the input and the output. The + symbol represents concatenation in the network diagram. 110

5.8 **Training loss progression of the Bayesian Neural Networks architectures** The figure shows the overall training loss of both neural network architectures when trained with Kullbeck Leibler loss cycling. The cycling rate can be easily visible from cyclical nature of the training loss. The VGG-16 network continues to converge with subsequent training batches with this training regime, whilst the ResNet-18 architecture diverges with each subsequent Kullbeck Leibler loss cycle. 113

5.9 **Calibration accuracy and auxiliary isotonic regression model for a VGG-16 based Bayesian neural network. Left:** Uncalibrated overall uncertainties from the VGG-16 based Bayesian neural network trained with Kullbeck Leibler loss cycling on the uncertainty calibration training set. **Middle:** The auxiliary isotonic regression middle learnt from the training set. **Right:** The uncalibrated and calibrated uncertainty estimates on the test set after uncertainty estimates are put through the auxiliary regression model. The uncalibrated blue line represents a calibration error of 0.21, and the calibrated green line has a calibration error of 1.2×10^{-3}. 115

5.10 **Aleatoric and Epistemic Uncertainty Change with Gestational Age. Top:** Aleatoric and Epistemic uncertainty variation with differences in gestational age. **Bottom:** Prediction against ground truth gestational ages for the hold-out INTERGROWTH training set. The red line shows the ideal line where predicted gestational age is equal to the ground truth gestational age. 116

5.11 **Relationship Plot between Gestational Age Error and Uncertainty.** All values in both x- and y- axis are in gestational age weeks. The ground truth ranges are selected based on gestational age quantile ranges, and the validation set is divided into three equal sets for easier visualization. 118

5.12 **Visualization of Fetal Head Images with Low Aleatoric Uncertainty.** Numbers above each image indicate ground truth gestational age. Figure shows fetal head images with the lowest aleatoric uncertainties in each quantile. 119

5.13 **Visualization of Fetal Head Images with High Aleatoric Uncertainty.** Numbers above each image indicate ground truth gestational age. Figure shows fetal head images with the highest aleatoric uncertainties in each quantile. 120

5.14 **Visualization of Fetal Head Images and Epistemic Uncertainties. Left:** Head images in each quantile with the lowest epistemic uncertainty. **Right:** Head images in each quantile with the highest epistemic uncertainty. 121

5.15 **Kernel density estimates of uncertainty estimates on out-of-distribution datasets.** Images in yellow are from the original INTERGROWTH fetal head dataset. Images in green are fetal head images from the PULSE dataset. Images in purple are fetal femur images from the INTERGROWTH-21st dataset. **Left:** Uncertainty distribution for the Bayesian neural network evaluated on the same anatomical plane but with a different acquisition protocol **Right:** Uncertainty distribution for the Bayesian neural network evaluated on a different anatomical plane with the same protocol. 121

6.1 **Examples of neural network architectures used for Multi-Task Learning (MTL) in previous literature.** Convolutional layers are depicted in blue, and aggregation layers are depicted in purple. **Top left:** Two separate neural networks for two different tasks on the same input image. **Top right:** Neural network with early division of convolutional layers for the two tasks. **Bottom left:** Neural network with late division of features for the two tasks. **Bottom right:** Cross-stitch network [174] for self-learnt feature aggregation for tasks. Dashed arrows represent learnt cross-stitch attention modules. 128

6.2 **Examples of common augmentation transformations and the effect of the transformation on a landmark detection task.** The landmark detection task used for illustration here is the identification of the transcerebellar boundary. Affine transformations such as vertical flipping and rotation lead to corresponding changes in the training target landmark, whereas image intensity histogram manipulation and blurring such as Gaussian blur and Gamma adjustments do not. 130

List of Tables

2.1 Equations to estimate the mean and standard deviations of fetal head biometries from gestational age in weeks, reproduced from the INTERGROWTH-21st study. All measurements are in millimeters (mm) and all gestational ages are in weeks [26] 13

2.2 Equations to estimate the mean and standard deviations of abdominal circumference from gestational age in weeks, reproduced from the INTERGROWTH-21st study [26] 16

2.3 Equations to estimate the mean and standard deviations of femur lengths from gestational age in weeks, reproduced from the INTERGROWTH-21st study [26] 17

2.4 Review of non-standard fetal biometries used for gestational age estimation in prior work. Accuracies of biometry regressions to gestational which were not provided in the literature is listed with "NP". The gray rows represent biometry that is performed on aborted fetuses, and may therefore not reflect the accuracy derived from ultrasound-based biometry. 19

3.1 Fetal Anomaly Screening Programme (FASP) [5] derived planes for standard plane classification. 48

3.2 **Macro-precision, recall and F1-scores for evaluated augmentation policies compared to recent work**. Table of final classification performance results. Results in grey include pre-training using external data sources such as gaze and video data, and are included as comparison. Results in black were trained from random initialization. 55

3.3 **Ablation study on the individual components of our *Mix. RA* policy search algorithm on training of a randomly initialized SE-ResNeXt-50 for ultrasound standard plane detection.** All metrics are macro-averages due to the class imbalance. The *Linear Mix. RA* is included as a baseline mixed-example augmentation strategy. 57

4.1 **Table of common approaches for bin selection for regression via classification** 74

4.2 **Results for gestational age regression on individual standard planes** MSE/MAE/Explained Variance metrics are calculated based on maximum likelihood with gestational age estimation for models trained with a classification loss. Models in **bold** indicate the best performing model on all three metrics for each anatomy. The regression model highlighted in grey represents the best performing regression-based loss when the model is trained with for vanilla deep regression. Arrows indicate direction of better performance for each metric. 80

4.3 The p values calculated using Welch's unequal variance t-test [154] to determine the probability that the mean absolute error samples are drawn from different populations. HCP, FLP and ACP represent the head circumference plane, femur length plane and abdominal circumference plane respectively. **Top:** Regression models trained on the HCP. **Middle:** Regression models trained on the FLP. **Bottom:** Regression models trained on the ACP 81

4.4 **Predictive error dependence on gestational age** Results below show the mean absolute error in weeks for different sets of images for the best performing neural network on each anatomical plane. Percentage increase numbers are relative to the gestational age range column on the immediate left. 82

5.1 **Quantitative performance for gestational age estimation for both CNNs and BNNs on the INTERGROWTH dataset.** Best results for each metric are bolded. RMSE and MSE Metrics are calculated with the maximum likelihood for both aleatoric and epistemic uncertainties. The "Calibration Error" column shows uncertainty calibration error pre- and post- calibration respectively. 112

List of Abbreviations

1D, 2D, 3D	One-, two- or three-dimensional
3VT	Three vessel and trachea view
4CH	Four chamber view
AC	Abdominal circumference
ACP	Abdominal circumference plane
ADAM	Adaptive moment estimation
BNN	Bayesian neural network
BPD	Biparietal diameter
BRaTS	Brain tumor segmentation challenge
CI	Celaphic index
CP	Choroid plexus
CRL	Crown-rump length
CSP	Cavum septum pellucidum
CT	Computed tomography
EDD	Estimated delivery date
FASP	Fetal anomaly screening programme
FL	Femur length
FLP	Femur length plane
GA	Gestational age
GAP	General average pooling
GPU	Graphics processing unit
HC	Head circumference
HCP	Head circumference plane
INTERGROWTH-21st	International fetal and newborn growth consortium for the 21st century

LMP . . . . . Last menstrual period

LVOT . . . . . Left ventricular outflow tract

MAE . . . . . Mean absolute error

MRI . . . . . Magnetic resonance imaging

MSE . . . . . Mean squared error

OFD . . . . . Occipitofrontal diameter

PULSE . . . . . Perception ultrasound for learning sonograhpic experience study

RA . . . . . RandAugment

RMSE . . . . . Root mean squared error

RVOT . . . . . Right ventricular outflow tract

ResNet . . . . . Residual network

SE . . . . . Squeeze-and-excitation

SGD . . . . . Stochastic gradient descent

SN . . . . . SonoNet

TC . . . . . Transcerebellar

TT . . . . . Transthalamic

1 Introduction

Contents

1.1 Clinical Motivation 1
1.2 Thesis Contributions 3
1.3 Thesis Structure 4
1.4 Originality 5
1.5 List of Accepted Publications Contributing to This Thesis 6
1.6 List of Accepted Publications Referenced in This Thesis 6

1.1 Clinical Motivation

Gestational age estimation is a key part of obstetric care, as it allows for pregnancy planning, and timing of screenings and tests by the care provider and informs fetal growth predictions [1, 2]. Accurate gestational age estimation allows for screening and management of potential complications in pregnancy such as preterm labour, and allows estimation of fetal viability and scheduling of prenatal tests [3].

Accurate gestational age dating in the clinical setting is based on biometry regression on existing clinical growth charts [4]. In the United Kingdom Fetal Anomaly Screening Programme (FASP) [5], gestational age is estimated in the first trimester using the crown-rump length (CRL), with an optimal time for scanning

between 11+2 weeks and 14+1 weeks of gestational age. First trimester scans clinically result in the most accurate estimate of gestational age [6].

However, first trimester scans are not always available or possible. Women may present to the clinic for pregnancy care too late for accurate dating from the crown-rump length measurement to be accurate [7]. This problem is present particularly for low- and medium-income countries. A recent study in Uganda [8] found that despite 97% of women receiving antenatal care from a skilled provider during pregnancy, only 29% of women had a first trimester scan.

Clinical growth charts are typically generated for second and third trimester based fetal biometry from measurements in the abdomen, femur and head plane [4]. However, gestational age estimation uncertainty increases with fetal gestational age, due to increasing difficulty in measurement of a larger fetus and increased variation in size between fetuses. This leads to gestational age uncertainty of $>$ 14 days in late pregnancy, which may not be as clinically useful [9, 10]. Despite these limitations, ultrasound-based dating is still the preferred dating method for gestational age estimation in late pregnancy, as other clinical dating methods are found to be even less accurate [11].

Because of these challenges, the efficacy of automated fetal gestational age estimation algorithms have been evaluated. However, existing algorithms for gestational age estimation on two-dimensional ultrasound images are based on fetal biometry, such as the femur length and head circumference that is already currently used in the clinic for estimation in late pregnancy. This would then incur the same uncertainty penalty as direct biometry by an experienced sonographer.

This thesis therefore aims to automate feature extraction and gestational age estimation using convolutional neural network based deep learning. Compared to classical computer vision and machine learning approaches, deep learning methods do not rely on hand-crafted features and automatically learn useful representations of the input image for the target task. This means that deep learning based models are not constrained to gestational age estimation using existing clinical growth charts and quantified biometry, but may be able to extract anatomically relevant

information from the ultrasound image itself. The main objective is to generate estimates of gestational age from second and third trimester ultrasound images which are more accurate than current clinical best estimates using deep learning, to contribute to better obstetric care in low-and-medium income countries and to further the understanding of gestational age estimation using deep learning.

1.2 Thesis Contributions

This thesis contributes to the optimizing the data augmentation policy selection during the training of convolutional neural networks in the medical imaging space, and uses the method to establish a series of automated gestational age estimation models on ultrasound fetal standard plane images of the head, abdomen and femur without reliance on clinical growth charts.

The first contribution described in in chapter 3 is a novel automated meta-learning approach for optimizing the data augmentation policy used as input in training convolutional neural networks in the medical image analysis space. An automated augmentation search policy was used to generate an optimized augmentation strategy for the proximal task of ultrasound standard plane classification. A state-of-the-art convolutional neural network model architecture was used and trained to classify ultrasound standard planes from clinical second trimester scans. It is found that the optimized augmentation strategy increased the macro precision, recall and F1-score of the best performing model to 75.1%, 73.4% and 74.0% respectively, compared to a naive augmentation strategy with 70.4%, 64.9% and 67.0% when evaluated. We also evaluate the difference in extracted features in models trained with an optimized augmentation policy, and found that more salient anatomical structures were extracted from input images.

The second contribution described in chapter 4 is the evaluation of deep regression convolutional neural networks on gestational age estimation on a dataset uniformly distributed across the second and third trimesters ranging from 13+0 weeks to 41+0 weeks with an optimized augmentation strategy, building on the first contribution. Multiple loss functions and network architectures for regression were evaluated,

as well as the performance of gestational age estimation for different standard plane anatomies. The best performing deep regression convolutional neural network is able to estimate gestational age to 0.615, 0.831 and 0.806 weeks for the head, femur and abdomen planes respectively across the entire gestational age range with no pixel size information provided. Moreover, we generate saliency attributes for input images, and evaluate the anatomical relevance to gestational age estimation for different input planes.

The third contribution described in chapter 5 is the formulation and training of a Bayesian neural network for regression that generates uncertainty estimates during inference for gestational age estimation of the fetal head. The proposed architecture estimates both aleatoric and epistemic uncertainties, where aleatoric uncertainties is representative of noise inherent in the acquisition, and epistemic uncertainties is unknowns that can be explained with more data [12]. This uncertainty is then calibrated by an auxiliary isotonic regression model. We find that both aleatoric and epistemic uncertainty increases with ground truth gestational age.

1.3 Thesis Structure

Chapter 2 introduces the clinical importance of gestational age estimation, and the current clinical methods for estimating gestational age. It then details the limitations of previous work on classical automated gestational age estimation, and the current state of the art in medical image analysis using deep learning and convolutional neural networks. Finally, it reviews existing deep learning based regression methods and deep learning based gestational age estimation, and the existing research questions that this reveals.

Chapter 3 presents an original automated meta-learning strategy for data augmentation medical imaging, and evaluates the performance of this approach on standard plane classification on second trimester clinical ultrasound videos.

Chapter 4 evaluates the performance of deep vanilla regression neural networks on the task of gestational age estimation. Experiments investigating different loss functions and architectures are described, and saliency maps for different

anatomies are generated and evaluated. Results show improvement of gestational age estimation over existing clinical methods using convolutional neural networks .

Chapter 5 introduces a Bayesian neural network approach for estimating uncertainty during the inference of gestational age estimation for fetal head images. The approach calibrates the inferred uncertainty estimates. The effect of gestational age on aleatoric and epistemic uncertainties are investigated. This pilot study reveals the ability of a Bayesian neural network to learn deep regression.

Chapter 6 concludes the thesis and summarizes the main contributions of the thesis. The strengths and limitations of this work and potential avenues for future work are also discussed.

1.4 Originality

I hereby declare that I am the sole author of this thesis, including the tables and figures herein unless otherwise specified. Critique and constructive criticism was provided by my supervisor, Professor Alison Noble, and collaborator (for chapter 3), Dr. Gao Yuan.

This thesis used two datasets, from the Perception Ultrasound by Learning Sonographic Experience (PULSE) in Chapter 3 and INTERGROWTH-21st in Chapter 4 and 5. For the PULSE dataset, standard plane classification labelling was a collaborative effort jointly in the lab. For the INTERGROWTH-21st dataset, an implementation of ScanNav (ScanNav^{TM}, Intelligent Ultrasound, United Kingdom) was used for initial sorting of standard planes to the three anatomical classifications (head, abdomen and femur). This classification was then validated by my collaborator Dr. Elizabeth Bradburn. I matched ground truth gestational ages from each sorted image to the original master database.

I implemented all image pre-processing operations using open source Python libraries, including the Python Imaging Library (v8.1.2) and OpenCV (v4.5.1.48). Deep learning algorithm implementations were built using the open source Tensorflow (v2.3.1) and PyTorch (v1.8.0) deep learning frameworks.

1.5 List of Accepted Publications Contributing to This Thesis

Chapter 3

Lee L.H., Gao Y., Noble J.A. (2020) Principled Ultrasound Data augmentation for Classification of Standard Planes. International Conference on Information Processing in Medical Imaging (IPMI) 2021 [13].

Lee L.H., Noble J.A. (2020) Generating Controllable Ultrasound Images of the Fetal Head. Proceedings of the International Symposium on Biomedical Imaging (ISBI) 2020 [14].

Chapter 5

Lee L.H., Bradburn E., Papageorghiou A.T., Noble J.A. (2020) Calibrated Bayesian Neural Networks to Estimate Gestational Age and Its Uncertainty on Fetal Brain Ultrasound Images. Proceedings of the International Workshop on Advances in Simplifying Medical Ultrasound (MICCAI ASMUS) 2020 [15].

1.6 List of Accepted Publications Referenced in This Thesis

Chapter 2

Lee L.H.,, Noble J.A. (2020) Automatic Determination of the Fetal Cardiac Cycle in Ultrasound Using Spatio-Temporal Neural Networks. Proceedings of the International Symposium on Biomedical Imaging (ISBI) 2020 [16].

Xu, Y., **Lee L.H.,**, Noble J.A. (2020) Simulating realistic fetal neurosonography images with appearance and growth change using cycle-consistent adversarial networks and an evaluation. Journal of Medical Imaging (JMI) 2020 [17].

2

Literature Review

Contents

2.1 Introduction 8
2.2 Clinical Gestational Age Estimation 8
2.2.1 LMP-Based Gestational Age Estimation 8
2.2.2 Ultrasound Based Clinical Gestational Age Estimation . 9
2.2.3 First-Trimester Clinical Fetal Gestational Age Estimation 9
2.2.4 Second and Third-Trimester Clinical Fetal Age Estimation 11
2.2.5 Other Non-Standard Measurements for Gestational Age Estimation 18
2.2.6 Challenges for Clinical Gestational Age Estimation . . . 20
2.3 Classical Automated Gestational Age Estimation . . . 21
2.4 Convolutional Neural Networks 22
2.4.1 Fully-Connected Layers 23
2.4.2 Convolutional Layers 24
2.4.3 Activation Functions 24
2.4.4 Pooling Layers 25
2.4.5 Optimization 26
2.4.6 Deep Residual Networks - ResNet, ResNeXt 28
2.4.7 Data Augmentation 28
2.5 Deep Learning for Regression 31
2.6 Deep Learning based Gestational Age Estimation . . . 33
2.7 Conclusions 34

2.1 Introduction

This chapter presents the key challenges and previous work on gestational age estimation from ultrasound images. Section 2.2 introduces the current method of clinical gestational age estimation in routine ultrasound scanning, and its strengths and weaknesses. Section 2.3 then discusses the computer-vision based methods that have been used to tackle the problem of automated gestational age estimation. Section 2.4 discuss convolutional neural networks, the current state of the art in medical image analysis. Section 2.5 discusses regression using convolutional neural networks, and Section 2.6 discusses the current state of the art for gestational age estimation using convolutional neural network based approaches. In each section, the strengths and weaknesses of previous work are discussed to place our work in context.

2.2 Clinical Gestational Age Estimation

Gestational age (GA), is clinically defined by the American College of Obstetricians and Gynecologists as 280 days before the best Estimated Due Date (EDD), and is commonly expressed in weeks and days [18]. As discussed in section 1.1, gestational age is a key assessment metric in obstetric care. Accurate determination of gestational age is used for the timing of multiple tasks in pre- and post- natal care, and can inform the care provider of the growth of the fetus. There are two main methods of fetal gestational age estimation: Last Menstrual Period (LMP) based, and prenatal ultrasonography. In this section, we describe each method currently in use, and discuss the key strengths and challenges of each.

2.2.1 LMP-Based Gestational Age Estimation

Last menstrual period based gestational age estimation [19] sets the estimated due date 280 days after the last menstrual period by convention, from which the gestational age can derived. However, this estimation relies on several assumptions. Assumptions include; i) menstrual cycle lengths are regular and consistent; ii) cycle lengths last for 28 days; iii) ovulation occurs on the 14th day after cycle beginning;

and iv) self-reported last menstrual periods are accurate. Some estimates of LMP accuracy suggest that this is only accurate to 15-18% [20, 21].

Therefore previous work [22–24] finds that the LMP is an inaccurate predictor of EDD and GA. Savitz [22] finds that gestational age dating on LMP alone yields accurate (± 1 week) estimated due dates for 48.1% of term births compared to 57.8% for ultrasound-based methods. Wegienka [23] finds that only 56% of women accurately recalled their LMP date. Barr [24] finds that EDD (and therefore gestational age) based on second trimester ultrasound scanning is more accurate than LMP.

2.2.2 Ultrasound Based Clinical Gestational Age Estimation

Due to the limitations outlined in section 2.2.1 for LMP based measurement, biometry measurement on 2D images is currently the routine for gestational age estimation. The introduction of first trimester ultrasound scanning to more precisely determine gestational age reduced post-term labour induction rates [25]. During clinical examination, the sonographer finds fetal planes that are defined based on the appearance and position of anatomical structures in the image [5]. Biometric measurements are then taken from these images of specific well-defined anatomical structures, with knowledge of both pixel size from the ultrasound machine and anatomical boundaries from prior medical knowledge by the sonographer. These measurements can then be regressed to gestational age from population-based fetal growth charts [26, 27]. These fetal growth charts are based on biometric measurements that are previously found to be strongly correlated with gestational age [3].

2.2.3 First-Trimester Clinical Fetal Gestational Age Estimation

Due to the limitations of LMP based estimation of gestational age listed in section 2.2.1, measurement of fetal anatomical characteristics in the first trimester (up to 13+6 weeks of gestational age [19]) is the current gold standard in gestational age

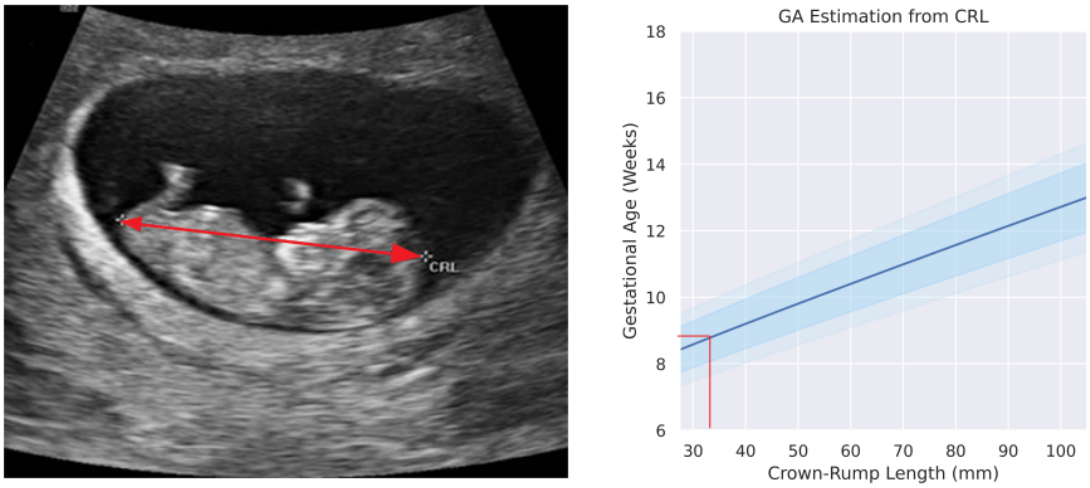


Figure 2.1: Gestational age estimation in the first trimester via crown rump length (CRL) measurement. Left: An example of a first trimester measurement plane taken of a fetus at 8+4 weeks of gestational age, and the CRL measurement in red. **Right:** The clinical chart used for gestational age regression based on first trimester crown rump length. The solid blue line represents the mean gestational age for a given crown rump length, with each successive blue area representing the 95% confidence interval and the 99% confidence interval respectively.

estimation for pregnancies that are not derived from assisted-reproductive technology [28–30]. The parameter measured for gestational age estimation in the first trimester is the Crown Rump Length (CRL), which is acquired based on the length of the measurement between the top of the head (crown) to the bottom of the rump [2, 31]. This is because other anatomical landmarks may not be sufficiently distinct to be used for gestational age determination at this early stage of pregnancy. This measurement can then be regressed back to gestational age based on clinical charts or previous regression formulae based on curve fitting on a previously measured dataset of crown-rump lengths, a review of which can be found by [32].

A recent study by [6] on a multicenter, population-based project produced the following equations for fetal age regression from crown rump length (CRL):

$$GA = 4.1 \times 10^1 + (3.21585 \times CRL^{0.5}) + (0.348956 \times CRL) \tag{2.1}$$

$$\sigma_{GA} = 2.39102 + (0.0193474 \times CRL) \tag{2.2}$$

where GA is the estimated gestational age, and σ_{GA} is the standard deviation of the gestational age. This procedure can be seen in figure 2.1. The measurement of the crown rump length, represented by the red arrow on the left sub-figure, is taken based on anatomical presentation. This is then regressed to estimated gestational age on the right, which is based on equations 2.1 and 2.2.

It should be noted that the uncertainty inherent in gestational age estimation increases with the crown-rump length (and therefore predicted gestational age). A prediction of gestational age of 13 weeks would have a standard deviation of prediction of 3.7 days, compared to 2.6 days for a prediction of gestational age of 8 weeks. Gestational age measurement based on crown rump length becomes less accurate after 13+6 weeks, due to fetal curling, movement and misidentification of the sagittal plane of the fetus [29, 33]. A different approach is therefore used for gestational age estimation for fetuses beyond the first trimester, which is explored in the next section.

2.2.4 Second and Third-Trimester Clinical Fetal Age Estimation

Due to fetal movement and curling post 13+6 weeks of gestational age, crown rump length becomes less accurate for gestational age estimation. Other fetal biometries are then used. Most commonly, measurements from three anatomical planes (also called fetal standard planes) are used for gestational age estimation, which include a fetal head plane, an abdominal plane, and a femur plane.

Measurements from the Fetal Head Plane

A fetal head image is commonly taken transversely at the level of the thalamus, called the transthalamic plane [5], seen in figure 2.2. Visible anatomical structures should include the thalamus in the centre of the skull and the Cavum Septum Pellucidum (CSP) anteriorly visible, as well as the midline of the brain. These structures can be found using standard ultrasonography from the second trimester to the termination of pregnancy. Three measurements can then be measured from the skull to estimate

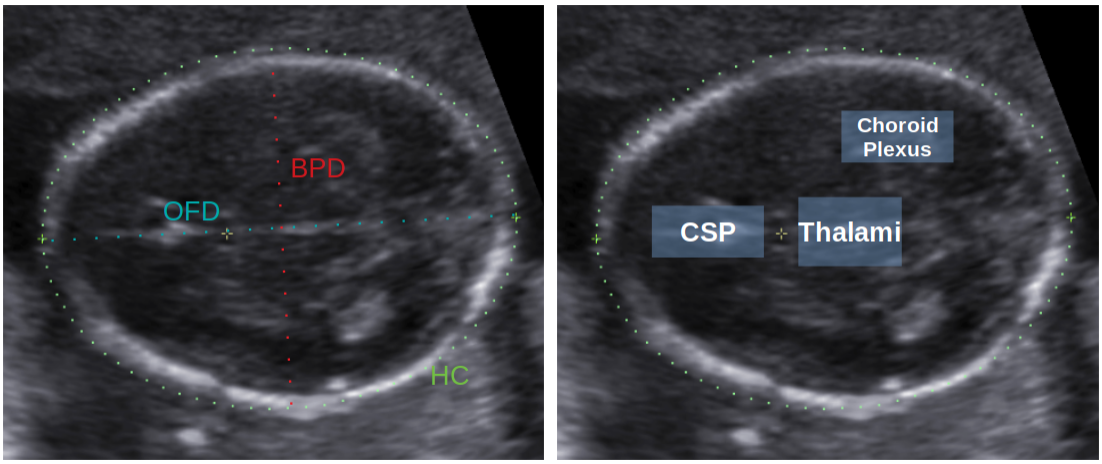


Figure 2.2: An example of a 22+1 week transthalamic plane of the fetal head. Image is taken from the INTERGROWTH-21st [26] study. Left: Key biometries used in gestational age regression, including the occipito-frontal diameter (OFD), biparietal diameter (BPD) and head circumference (HC). Markers are placed by an experienced sonographer. **Right:** Some of the required anatomies seen in the transthalamic plane, including the Cavum Septum Pellucidum (CSP), Choroid Plexus (CP), and Thalami. Image is best viewed in colour.

gestational age, as seen in figure 2.2, which includes the biparietal diameter (BPD), occipito-frontal diameter (OFD), and head circumference (HC) [5, 6].

As fetal clinical growth charts are derived from datasets with different acquisition protocols, the measurement points of BPD and OFD differ between charts but are typically outer leading edge to outer leading edge [34] of the skull. BPD is measured from the widest point of the fetal skull perpendicular to the midline echo. The measurement points of OFD are typically the outermost point of the anterior frontal skull wall to the outer edge of the occipital distal edge, where the measurement can be maximized. The HC, on the other hand, can be approximated with an ellipse-approximating equation (reviewed in [35]) with measurements of the OFD and BPD, or measured directly by ellipse fitting by the sonographer during examination placed typically on the outer edge of the fetal skull.

Variation in these gestational age estimation can come from measurement inaccuracies due to misplacement of sonographer calipers and misidentification of the transthalamus plane. There can also be additional gestational age estimation uncertainty due to variances in fetal growth (from differences in maternal nutrition

Table 2.1: Equations to estimate the mean and standard deviations of fetal head biometries from gestational age in weeks, reproduced from the INTERGROWTH-21st study. All measurements are in millimeters (mm) and all gestational ages are in weeks [26]

HC	$-2.8 \times 10^1 + 1.69 \times GA^2 - 0.40 \times GA^2 \times \log GA$	(2.2.1)
σ_{HC}	$1.99 + 1.4 \times 10^{-2} \times GA^3 - 7.3 \times 10^{-3} \times GA^3 \times \log GA + 9.8 \times 10^{-4} \times GA^3 \times \log(GA)^2$	(2.2.1a)
BPD	$5.6 + 0.16 \times GA^2 - 2.6 \times 10^{-3} \times GA^3$	(2.2.2)
σ_{BPD}	$\exp\{-0.10 + 1.5 \times 10^{-3} \times GA^3 - 7.7 \times 10^{-4} \times GA^3 \times \ln GA + 10.0 \times 10^{-5} \times GA^3 \times \ln GA^2\}$	(2.2.2a)
OCD	$-12.4 + 0.63 \times GA^2 - 0.14 \times GA^2 \times \log GA$	(2.2.3)
σ_{OCD}	$\exp\{-0.88 + 6.3 \times 10^{-2} \times GA^2 - 3.2 \times 10^{-2} \times GA^2 \times \ln GA + 4.1 \times 10^{-3} \times GA^2 \times \ln GA\}$	(2.2.3a)

and population differences) and skull shape, with brachycephalic (circular) skulls shortening apparent OFD for every gestational age, and dolicocephalic (elongated) skulls lengthening apparent OFD for every gestational age. It has been reported [36] that a celaphic index (CI), which is calculated by the ratio of BPD to OFD, has a mean value of 0.783, and a greater than 4.4mm (1 SD) deviation would lead to inaccurate estimations of gestational age from OFD or BPD measurements. An example of the regression equation used to regress gestational age from these measurements can be seen in table 2.1 and figure 2.3, reproduced from the INTERGROWTH-21st study [26].

Uncertainty in gestational age estimates regressed from biometry measurements of the head increase with increasing gestational age, as can be seen from equations

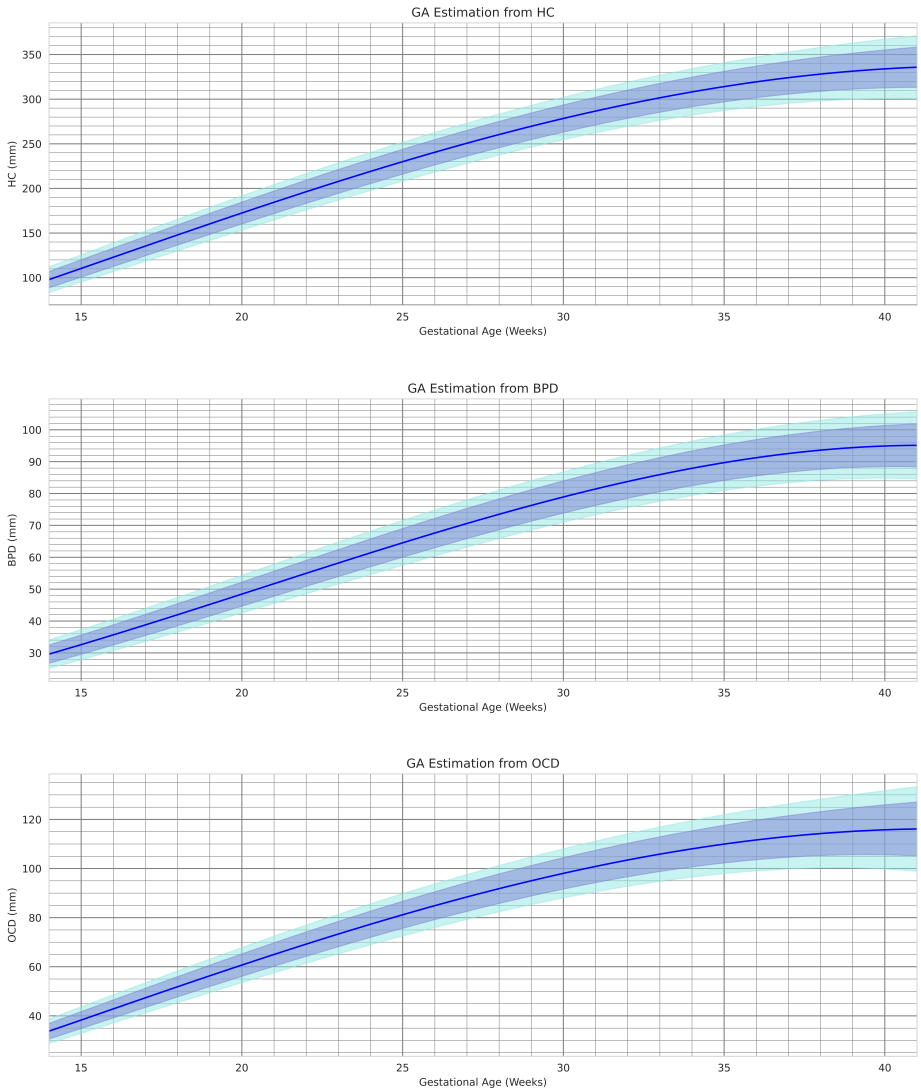


Figure 2.3: Fetal Biometry Growth Charts for the Head Circumference Plane derived from the INTERGROWTH-21st study. Fetal biometry growth charts for the head circumference plane derived from the INTERGROWTH-21st study. Charts are reproduced from [4]. Gestational Ages (GA) from **Top:** Head Circumference (HC); **Middle:** Biparietal Diameter (BPD); **Bottom:** Occipitofrontal Diameter (OCD); All measurements are in millimeters. Different lines on the same chart show the percentile distribution of biometries. The solid blue line represents the mean biometry for a given gestational age, with each successive blue area representing the 95% confidence interval and the 99% confidence interval respectively.

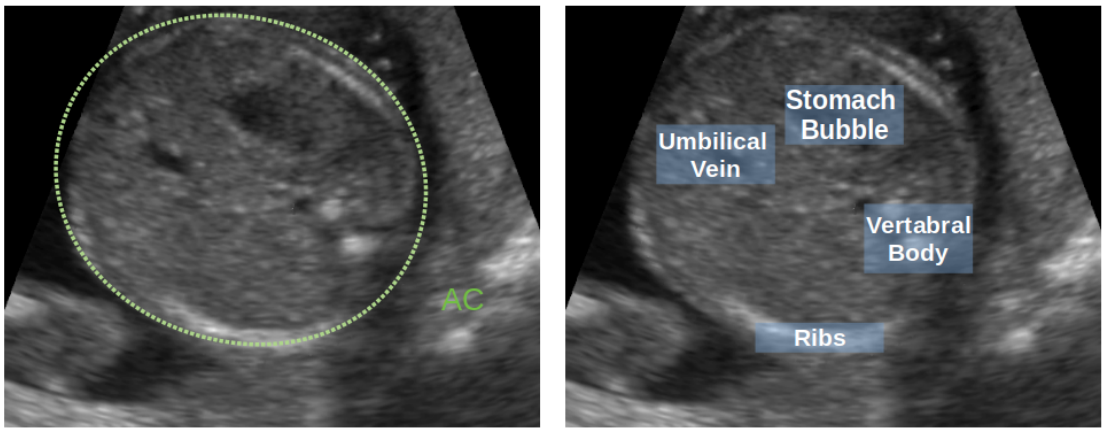


Figure 2.4: An example of a 26+4 week fetal abdominal circumference plane. Image is taken from the INTERGROWTH-21st [26] study. Left: The abdominal circumference (AC) commonly used for gestational age estimation. Markers are placed by an experienced sonographer. **Right:** Some of the required anatomies seen in the abdominal circumference plane, including the fetal stomach, umbilical vein, vertebral body (spine) and ribs. Image is best viewed in colour.

2.2.1a, 2.2.2a and 2.2.3a. This can be due to increased biological variation in fetal presentation with increased gestational age, as well as sonographic measurement difficulty. Late in pregnancy, gestational age estimation can have up to three weeks of gestational age estimation variance compared to ground truth gestational age [9, 37, 38].

Measurement from the Abdominal Circumference Plane

The fetal abdominal circumference can also be used for gestational age estimation in a clinical setting [26, 39]. An example of a fetal abdominal circumference plane including some of the defining anatomical markers can be seen in figure 2.4. In a standard ultrasound examination, the abdominal circumference plane is at the level of the fetal liver and the stomach, with the umbilical vein as a marker. The entirety of a rib should be visible to avoid tilted measurement planes, and the heart and kidneys should not be visible in this view. The abdominal circumference measurement is commonly taken by fitting an ellipse to the outer edges of the abdomen tissues [5], or by using ellipse generating equations from the transverse and anteroposterior measurements of the abdomen.

Table 2.2: Equations to estimate the mean and standard deviations of abdominal circumference from gestational age in weeks, reproduced from the INTERGROWTH-21st study [26]

AC	$-8.1 \times 10^{1} + 1.2 \times 10^{1} \times GA - 5.6 \times 10^{-4} \times GA^{3}$	(2.3.1)
σ_{AC}	$-4.4 + 0.12 \times GA^{2} - 1.3 \times 10^{-2} \times GA^{3} + 2.8 \times 10^{-3} \times GA^{3} \times \log GA$	(2.3.1a)

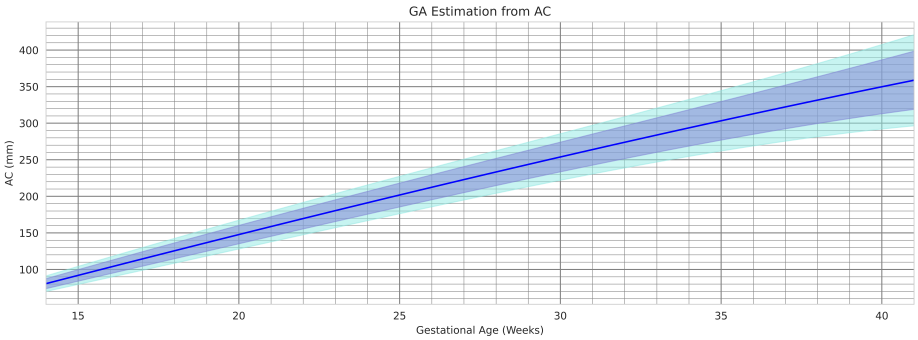


Figure 2.5: Fetal Biometry Growth Charts for the Abdominal Circumference Plane derived from the INTERGROWTH-21st study. Fetal biometry growth charts for gestational age (GA) estimation using the abdominal circumference (AC) derived from the INTERGROWTH-21st study. Charts are reproduced from [4]. The solid blue line represents the mean biometry for a given gestational age, with each successive blue area representing the 95% confidence interval and the 99% confidence interval respectively.

This measurement can then be regressed to gestational age using the equations found in table 2.2 or figure 2.5. It is found that gestational age estimation using abdominal circumference is less accurate than head measurements (Equation 2.3.1a and [40–42]). However, variations in abdominal circumference with other growth parameters has been used for identification of growth abnormalities [40, 43, 44] and identification of gestational diabetes. Gestational age estimation error increases over the pregnancy due to reasons similar to that of the head discussed in Section 2.2.4, including biological differences and measurement difficulties.

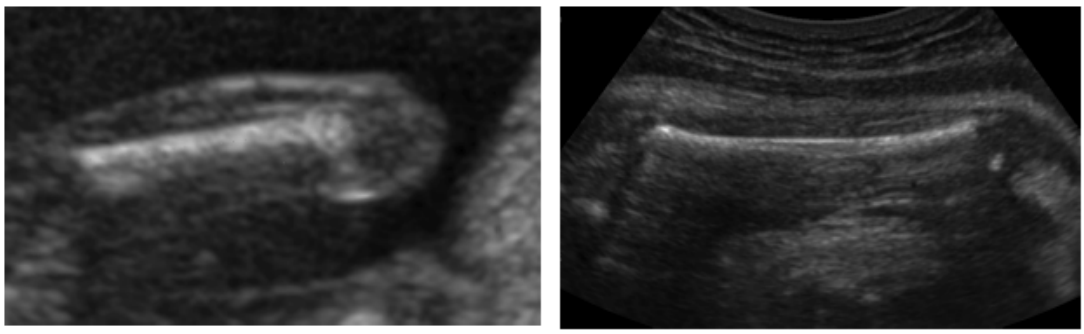


Figure 2.6: An example of fetal femur length planes. Images are taken from the INTERGROWTH-21st [26] study. Left: The femur length plane from a 16+2 week fetus **Right:** The femur length plane from a 26+4 week fetus.

Table 2.3: Equations to estimate the mean and standard deviations of femur lengths from gestational age in weeks, reproduced from the INTERGROWTH-21st study [26]

FL	$-4.0 \times 10^{1} + 4.3 \times GA - 3.8 \times 10^{-2} \times GA^{2}$	(2.4.1)
σ_{FL}	$\exp\{6.1 \times 10^{-1} - 42 \times GA^{-2} + 9.2 \times 10^{-6} \times GA^{3}\}$	(2.4.1a)

Measurement from the Femur Plane

The femur can also be used for gestational age estimation from 14+0 weeks of gestation until delivery, and is measured along the axial length of the entirety of the bone [45–48]. A straight-line measurement is taken between each end of the femur bone, ignoring the curvature of the femur with higher gestational ages as seen on the left of figure 2.6. The gestational age can be estimated from biometry using equation 2.3 or figure 2.7.

Despite the apparent simplicity of the measurement, femur length planes suffer from similar measurement inaccuracies to the abdomen and head planes [45]. These include overestimation of femur length from inclusion of soft-tissue cartilage, inclusion of the femur distal epiphysis, and tangential planes taken of the femur bone.

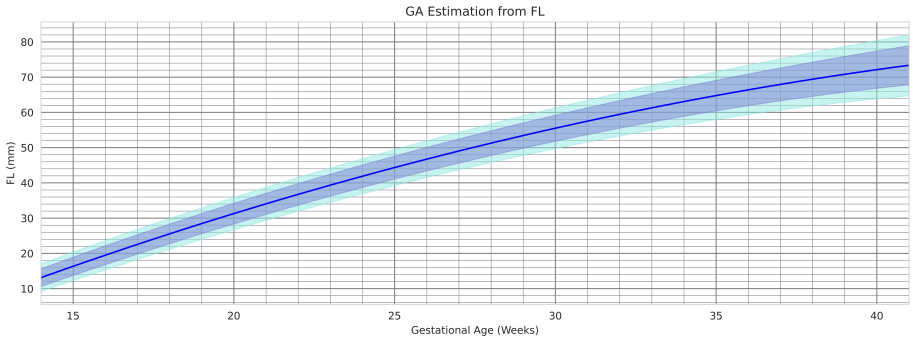


Figure 2.7: Fetal Biometry Growth Charts for the Femur Length Plane derived from the INTERGROWTH-21st study. Fetal biometry growth charts for gestational age (GA) estimation using the femur length (FL) derived from the INTERGROWTH-21st study. Charts are reproduced from [4]. The solid blue line represents the mean biometry for a given gestational age, with each successive blue area representing the 95% confidence interval and the 99% confidence interval respectively.

2.2.5 Other Non-Standard Measurements for Gestational Age Estimation

Due to the uncertainty inherent in the measurements in the standard planes as defined in the prior sections, there has been work on clinical gestational age estimation based on other planes and measurements. Table 2.4 shows some of the other biometry that has been previously used for fetal age estimation and the associated accuracy if given, including the kidney lengths, fetal long bones, intestinal development, heart circumference and foot length. The previous work also suggests that there are other additional sources of information within the anatomy that can be used for gestational age estimation outside of biometry on ultrasound images. Goldstein [49] finds that fetal intestinal peristalsis development and echogenicity have a strong correlation with gestational age greater than that of BPD and abdominal circumference, but these measurements are difficult to take during standard obstetric care.

Table 2.4: Review of non-standard fetal biometries used for gestational age estimation in prior work. Accuracies of biometry regressions to gestational which were not provided in the literature is listed with "NP". The gray rows represent biometry that is performed on aborted fetuses, and may therefore not reflect the accuracy derived from ultrasound-based biometry.

Authors	No. Subjects	Anatomy Measured	Age Range	Ground Truth GA Reference	Accuracy (days)
Konje, J. C. et al [50]	73	Kidney Length Measurement	24 - 38 Weeks	Coincidence of LMP and CRL	±1.5 Weeks
Sagi, J et al [51]	660	Kidney Length Measurement	toshev2014deeppose 15 - 42 Weeks	CRL or BPD	±2.5 Weeks at 20 Weeks GA ±2.2 Weeks at 30 Weeks GA
Jeanty, P. et al [52]	250	Long Bone Lengths (Femur, Humerus, Tibia, Ulna)	12 - 40 Weeks	Coincidence of LMP and CRL	± 1.4 Weeks for Femur, ±1.7 Weeks for Humerus ±1.7 Weeks for Tibia, ±1.8 Weeks for Ulna
Yarkoni, S. et al [53]	85	Clavicle Length	15 - 40 Weeks	Coincidence of BPD and FLP based Gas	± 2.7 Weeks
Sherer D. M. et al [54]	620	Clavicle Length	14 - 42 Weeks	Early Ultrasound	±1.3 Weeks at 20 weeks GA ±2.1 Weeks at 30 weeks GA
Dilmen, G. et al [55]	343	Scapula Length	16 - 41 Weeks	LMP	±1.05 Weeks between 16 to 28 weeks GA ±1.78 Weeks between 29 to 41 weeks GA
Goldstein, I. et al [49]	289	Intestinal Development (Transverse Colon Diameter, Echogenicity)	16 - 40 Weeks	NR	NR
Hill, L. M. et al [56]	262	Heart Circumference	15 - 42 Weeks	Coincidence of LMP and CRL	± 1.92 Weeks
Mercer, B. M. et al [57]	223	Foot Length	11 - 43 Weeks	Fetal birth date	±1.6 Weeks at 20 weeks GA ±2.4 Weeks at 30 weeks GA
Stevens, K. et al [58]	610	Foot Length	14 - 21 Weeks	Self-reported LMP or CRL, whichever available	NR
Drey, E. A. et al [59]	1,099	Foot Length	19 - 23 Weeks	BPD Regression	±1.2 weeks

2.2.6 Challenges for Clinical Gestational Age Estimation

Clinical gestational age estimation is challenging for obstetric sonographers. Ultrasound imaging as a modality has greater variations in image quality than the other commonly used medical imaging modality in fetal ultrasound, magnetic resonance imaging (MRI). These variations include differences in image uniformity, echoes, lower spatial and constrast resolution, as well as speckle noise [61–63]. Despite the above, ultrasound is still the default imaging modality in obstetric care due to its safety. Ultrasound is also lower-cost and more portable than MRI scans.

Sonographers therefore need to contend with finding accurate standard planes (as defined in Figures 2.2, 2.4 and 2.6). However, this is non-trivial, and there can be significant inter-sonographer measurement differences during scanning [64–67]. In particular, Sarris [67] quantified interobserver variability in biometry for the same fetus, and found that it increases with gestational age, and are ±4.9%, ±8.8% and ±11.1 for the HC, AC and FL measurements respectively in millimetres for the third trimester. This may have been due to differences in final standard plane determination or differences in measurement.

Inaccurate scan plane detection can lead to inaccurate fetal biometry, as illustrated in figure 2.8. Planes that differ from the correct standard plane location and orientation can lead to both overestimations and underestimations of the true gestational age. Anatomical structures such as the umbilical vein, stomach bubble, and kidney can be confused with each other during scanning. This problem is exacerbated in earlier gestational ages, where small changes in tilt and location can lead to greater changes and measurement, and the correct anatomical plane is harder to find due to the smaller size of the fetus. However, in late pregnancy, there is a greater degree of ultrasound image degradation from the ossification of bones leading to image echo and shadowing.

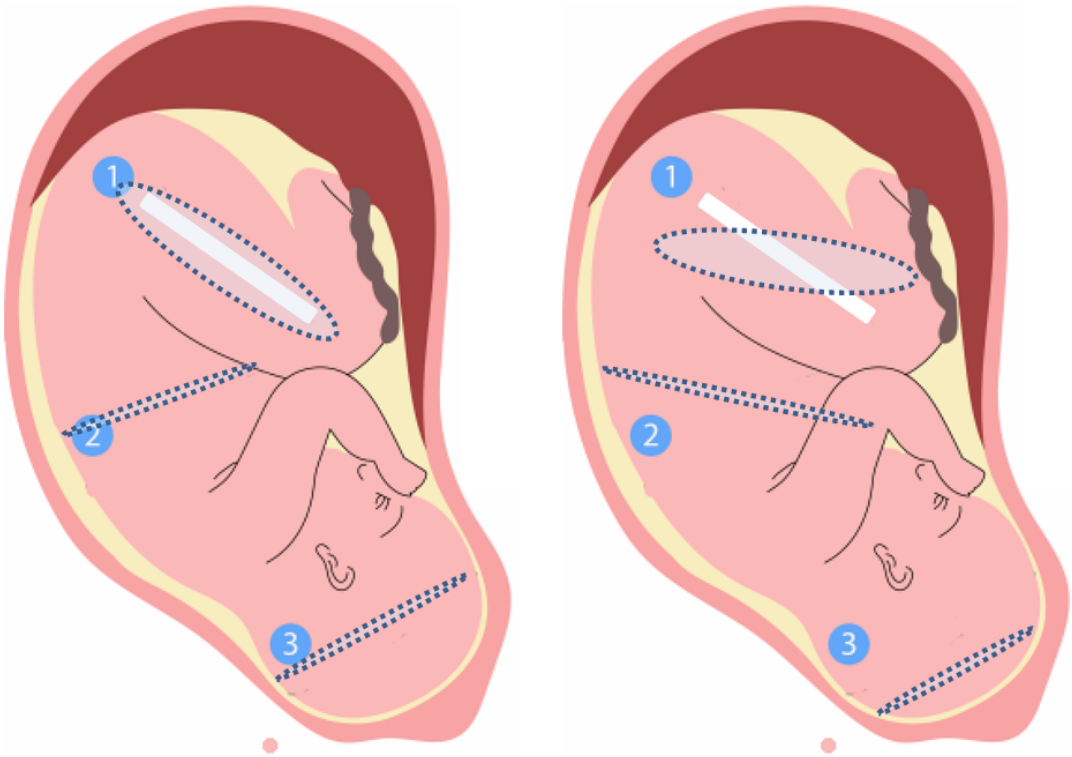


Figure 2.8: Examples of correct and incorrect standard plane selections. Left: Correctly identified standard planes for the three anatomical planes (1: Femur plane, 2: Abdominal circumference plane, 3: Transthalamic head plane.) **Right:** Examples of misplaced standard planes on the same fetus, leading to misleading measurements. 1 and 3 will lead to shortened measured biometries and hence underestimation of the true gestational age, 2 will lead to an artificially lengthened abdominal circumference measurement and therefore an overestimation of the true gestational age. Figure is reproduced from [60].

2.3 Classical Automated Gestational Age Estimation

There have been attempts at using classical computer vision methods for gestational age estimation in previous literature.

Previous methods largely relied on automated extraction of relevant clinical biometry from 2D standard planes for gestational age estimation. This includes segmentation of the femur length using the Hough Transform [68], automated ellipse fitting to the HC [69], constrained probabilistic boosting trees to segment fetal features [70], K-means and an iterative Hough transform for fetal head measurement [71], and segmentation of the fetal head by formulating the segmentation as a

continuous min-cut problem [72]. Automated extraction of fetal biometry has been set as a segmentation challenge previously in the IEEE International Symposium on Biomedical Imaging, and methods are reviewed in Sylvia et al. [73].

However, these methods above all rely on existing clinical fetal growth charts for conversion of biometry to gestational age, and are therefore subject to the same growth-based uncertainties. Furthermore, features are usually hand-crafted and may not identify regions which are otherwise useful for gestational age estimation.

Furthermore, all the methods above rely on clinical fetal growth charts for conversion of fetal measurements to GA, and therefore are subject to the same uncertainties.

Namburette [74] use regression forests with bespoke features to regress gestational age from 3D ultrasound scans of the fetal head without using clinical fetal growth charts. But this method relies on costly 3D ultrasound volumes and probes.

We therefore look towards deep learning for medical image analysis, which has been shown to be a powerful method of self-learnt feature extraction in prior literature [75, 76].

2.4 Convolutional Neural Networks

In this section we discuss the common building-blocks of a convolutional neural network, and their applicability to medical image analysis. Two of the major contributions of this thesis involves the usage of convolutional neural networks for gestational age regression.

Convolutional Neural Networks (CNNs) are deep learning based models that learn image features during the training process, with no need for hand-crafted image features. CNNs are composed of layers of linear and non-linear transformations, the weights of which are learnt with training data and ground truths using an optimization algorithm. An example of a convolutional neural (AlexNet [77]) network can be seen in figure 2.9, which won in the ImageNet Large Scale Visual Recognition Challenge in 2012, achieving a top 5 accuracy of 84.6%.

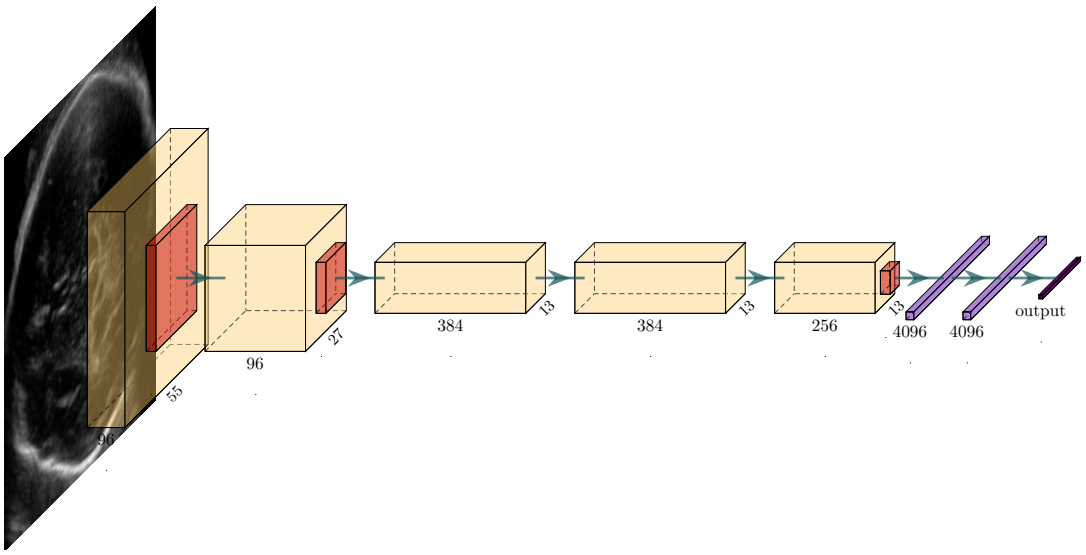


Figure 2.9: Architectural overview of AlexNet [77], a convolutional neural network. AlexNet, typical of most convolutional neural networks, is composed of convolutional layers (depicted in yellow), pooling layers (depicted in orange), and fully-connected layers (depicted in purple). The numbers indicate the size of each feature map and the number of channels. ReLU non-linearities are included after every convolutional layer, but not shown on the figure for brevity. The size of each feature map is not to scale, and the figure is best viewed in colour.

2.4.1 Fully-Connected Layers

As suggested in the name, fully-connected layers connect each neuron in the previous layer to the current layer. The governing equation is,

$$G = \sum_j w_j f_j + b \tag{2.3}$$

where G is the output of the fully-connected layer, f is the input feature to the fully-connected layer, and w, b are the learned weights and biases for the fully-connected layer. and j is the index of each neuron. Fully-connected layers can therefore be seen as a learnt affine transformation between input features and output features. Fully-connected layers are commonly used [77–79] after convolutional layers that act as feature extractors, which can be seen depicted in purple in figure 2.9.

2.4.2 Convolutional Layers

Convolutional neural networks generally contain convolutional layers, which perform convolution operations on input features with learnt kernel weights from the optimization process, which is depicted in yellow in figure 2.9. A convolutional layer applies the following operation to an input array;

$$G_{(m,n)} = (f \star h)_{(m,n)} = \sum^{j} \sum^{k} h_{j,k} f_{m-j,n-k} \tag{2.4}$$

which represents the discrete convolutional operator $\star$ between the learnt kernel h and the input feature map f generating the output feature map G with m, n being the row and column index. In practice, convolutional layers contain multiple convolutional kernels h per layer, thus generating multiple output features G. In figure 2.9, the left-most first convolutional layer generates an output feature of size 55×55, but with 96 11×11 convolutional kernels learnt during the training process. Convolutional kernels learn spatially local features for each input feature as part of the training process, as the kernel performs convolution on local pixel groups. Convolutional layers are also significantly more sparse than fully-connected layers, which reduces computational requirements and memory requirements.

2.4.3 Activation Functions

Both convolutional layers and fully-connected layers are linear operators. Multiple layers therefore decompose into linear combinations of the input matrix, and such a neural network would only be able to learn linear combinations of input features. Non-linearities are therefore introduced in neural networks in order to allow convolutional neural networks to model the class of functions with non-linear decision boundaries with a small number of layers, which is the case for most tasks.

Rectified linear units (ReLU, [80]) are commonly used in convolutional neural networks in order to increase the expressivity of these networks, and consist of the following activation function;

$$g(z) = \max(0, z) \tag{2.5}$$

where z is the input and $g(z)$ is the output of the activation function.

ReLU units are used due to the efficiency of computation during forward and back propagation. ReLU units also reduce the propensity for the vanishing gradient problem compared to other non-linear activation functions such as the *tanh* function or the sigmoid function.

2.4.4 Pooling Layers

Convolutional layers are dependent upon the exact local spatial location of each input feature. Small changes in feature location and distribution in the original image may therefore lead to undesired changes in the output feature map. Pooling layers are therefore used in order to build a strong prior for invariance to small translation and rotation operations. The work in this thesis uses two types of pooling layers commonly used in convolutional neural networks, average pooling [81] and max pooling [82].

Max pooling is defined by the following equation:

$$G_{(m,n)} = \max \begin{bmatrix} f_{m,n} & f_{m,n+1} & \cdots \\ \vdots & \ddots & \\ f_{m+k,n} & & f_{m+k,n+k} \end{bmatrix} \tag{2.6}$$

and average pooling is defined by the following equation:

$$G_{(m,n)} = \text{mean} \begin{bmatrix} f_{m,n} & f_{m,n+1} & \cdots \\ \vdots & \ddots & \\ f_{m+k,n} & & f_{m+k,n+k} \end{bmatrix} \tag{2.7}$$

where G is the pooled output feature map, f is the input feature map, k is the size of the pooling operator, and m, n represent the row and column indicies respectively.

The pooling operation is invariant to small changes in rotation and translation, and therefore leads to greater invariance in the overall convolutional neural network. Furthermore, pooling operations increase the receptive field (defined as the region of the original input image that the convolutaional kernel operates on) of convolutional kernels deeper in the neural network which allows the neural network to learn higher level abstract features. As an example, AlexNet [77] includes two max-pooling operations and an average pooling operation.

2.4.5 Optimization

Model weights in the convolutional and fully-connected layers can be learnt using optimization algorithms, of which Stochastic Gradient Descent (SGD) [83], and Adaptive Moment Estimation (Adam) [84], are commonly used.

Mini-batch Stochastic Gradient Descent iteratively optimizes the trainable weights of a model with respect to training data by iterating piecemeal over sampled mini-batches. The stochasticity comes from the mini-batch optimization during the training loop compared to gradient decent, which optimizes over the entire training dataset. This may not always be possible due to memory limitations. The stochastic gradient descent algorithm can be seen in Algorithm 1.

Algorithm 1: Mini-batch Stochastic Gradient Descent

Input: Learning Rate: α
Input: Neural network trainable parameters: W
while *training* **do**
  for *number of batches* η **do**
    Generate $D_b = \{x_b, y_b\}$ by sampling a mini-batch from the full training dataset $D = \{x, y\}$;
    Generate update gradients with respect to data and parameters $\hat{g}(x_b, y_b, W) :\leftarrow \frac{1}{\eta}\nabla_W \sum_i \mathcal{L}(f(x^i; W), y^i)$;
    Update trainable parameters: $W \leftarrow W - \alpha\hat{g}$;
  end
end

$\mathcal{L}(f(x^i; W), y^i)$ represents the loss calculated from the current weight parameters and the training data. However, the choice of learning rate α is critical to the convergence using stochastic gradient descent [85]. A learning rate too large would lead to weight oscillation during training and may lead to divergence, whereas a learning rate that is too small may lead to the loss being stuck in local minima and slow convergence. Due to the complex paramterization of neural networks, loss surfaces are often of high dimension and convexity is not guaranteed. Learning rate determination is therefore difficult, and there has been prior work on learning rate scheduling [86–88] and learning rate optimization [85].

Algorithm 2: Mini-Batch Adaptive Moment Estimation [84]

Input: Learning Rate: α
Input: Moment estimate decay rates: $\beta_1, \beta_2 \in [0, 1]$
Input: Neural network trainable parameters: W
Initialize moment vectors and timestep $m_0, v_0, t \leftarrow 0$
while *training* **do**
  for *number of batches* η **do**
    $t \leftarrow t + 1$ Generate $D_b = \{x_b, y_b\}$ by sampling a mini-batch from the full training dataset $D = \{x, y\}$;
    Generate update gradients with respect to data and parameters $\hat{g}(x_b, y_b, W) :\leftarrow \frac{1}{\eta}\nabla_W \sum_i \mathcal{L}(f(x^i; W), y^i)$;
    Generate biased first moment estimate: $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1)\hat{g}$;
    Generate biased second moment estimate: $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2)\hat{g}^2$;
    Correct first moment estimate bias: $\hat{m}_t \leftarrow \frac{m_t}{1-\beta_1^t}$
    Correct second moment estimate bias: $\hat{v}_t \leftarrow \frac{v_t}{1-\beta_2^t}$
    Update trainable parameters: $W_t \leftarrow W_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}}$;
  end
end

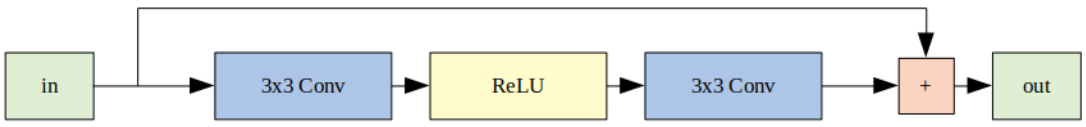


Figure 2.10: Base residual block of a ResNet [89] network The ResNet module consists of a skip connection and multiple convolutional layers, with a ReLU non-linearity in the middle. The figure depicts a ResNet module with two convolutional layers, and the 3×3 within each convolutional block refers to the kernel height and width respectively.

Adam [84] (Algorithm 2) is therefore also popular as an optimizer for convolutional neural network training. Adam optimization calculates adaptive learning rates for every parameter, which allows for different relative learning rates for different parameters. An exponentially decaying average of past gradients m_t is also used to scale the final weight parameter update. This allows for acceleration of the learning rate on loss surfaces, and lower sensitivity of optimization to initial learning rate selection.

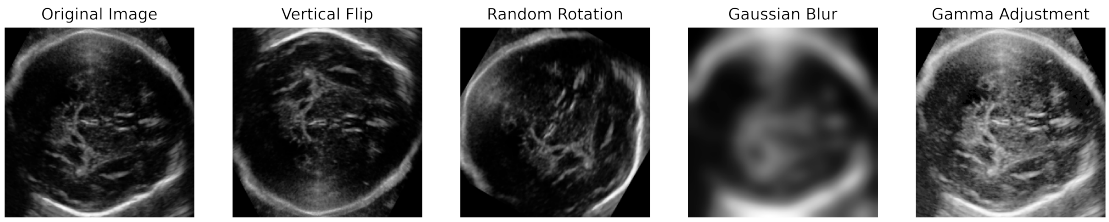


Figure 2.11: Examples of common individual augmentation transformations and their effects on an image The original image in the training set (fetal HCP) is shown on the left. The remaining four training images are generated from the original image using commonly used augmentation transformations.

2.4.6 Deep Residual Networks - ResNet, ResNeXt

Convolutional layers higher in the hierarchy therefore aggregate spatial features at higher representation orders, which lead to learning of more abstract features with higher layer counts. This suggests that increasing convolutional layer count would lead to greater outperformance compared to a similar but shallower network. However, naïve stacking of additional convolutional layers does not lead to continued performance increase, due to gradient vanishing or explosion [90, 91]. Gradient vanishing and explosions are a result of gradients accumulating through layers during backpropagation typically increasing or decreasing exponentially, the problem of which is exacerbated with additional layers.

ResNet [89] variants therefore include a short-cut connection with each convolutional layer, which aids in gradient backpropagation, seen in figure 2.10. This short-cut (residual module) allows deeper layers to have direct connections with both input and output features, which aids in minimizing gradient vanishing or explosion problems. This leads to the successful outperformance of deeper network models compared to shallower networks, which can be seen in the performance of ResNet-50, composed of 48 convolutional layers on the ImageNet Large Scale Visual Recognition Challenge, achieving a top 5 accuracy of 96.5%.

2.4.7 Data Augmentation

The optimization of weights within a convolutional neural network is dependent upon the training dataset, and it is found that the performance of neural networks

increase with increases in the number of examples in the training dataset for any given convolutional neural network model if the distribution of the training dataset and the testing dataset remain the same.

However, data acquisition for training of a neural network may be costly in time and money. Data acquisition and cleaning is often a manual task requiring some form of human input during labelling, and data acquisition can also be costly. This is particularly a problem in medical image analysis, where datasets are more limited in size compared to datasets used in computer vision and labelling requires expensive domain experts. Collection of medical image datasets may also be bound by legal and ethical concerns.

Data augmentation techniques generate additional artificial examples of training data using existing training data as examples, as seen in figure 2.11. This is partially inspired by the fact that the performance of neural networks can improve on training with more data, even if the data is has noisy labels or is otherwise not clean [92, 93]. Data augmentation has therefore been used for model training in most computer vision tasks. Lecun [78] uses image-warping based data augmentation for handwritten digit recognition. Both AlexNet [77] and ResNet [89] use random horizontal flipping, RGB intensity variation and random cropping in order to achieve the stated accuracy in the ImageNet Large Scale Visual Recognition Challenge. AlexNet attributed over 1% of error rate improvement to image augmentation.

Data Augmentation for Medical Image Analysis

Data augmentation is extensively used in deep-learning based medical image analysis works. Malepa [94] explores the data augmentation strategies used for image analysis of a challenge dataset (Brain Tumor Segmentation Challenge, BraTS), and found that most works used affine augmentation transformation as a baseline augmentation, with some works generating additional examples of training data using generative adversarial networks and other synthetic generative methods. Hussain [95] finds that the effectiveness of augmentation strategies for medical image analysis differ with different augmentation transforms; and flipping and

Gaussian filters increase validation performance on tumour classification whilst noise addition reduced validation performance. Augmentation strategies therefore differ in effectiveness, and different imaging modalities and tasks may have different optimum augmentation strategies. Data augmentation in the medical image analysis space is therefore based on domain knowledge.

Meta-Learning for Data Augmentation

Data augmentation policies described in section 2.4.7 are based on hand-crafted augmentation policies for generating artificial examples of data. This, along with the findings from Hussain [95] suggest that automated selection of augmentation strategy may be an area of potential model performance improvement, and suggests that there is a space for optimization of of the most effective data augmentation strategy outside of neural network optimization (meta-learning).

Cubuk [96] uses a reinforcement learning based algorithm to optimize the augmentation strategy consisting of up to ten augmentation transformations, which include affine image transformations and intensity histogram manipulations. It is found that the optimal augmentation strategy is dependent on the dataset, model, and task. On the ResNet-50 network evaluated on the ImageNet Large Scale Visual Recognition Challenge, the optimized augmentation strategy had a top 5 error rate of 93.8%, when compared to the original augmentation strategy of 93.1%. Cubuk [97] further introduces a stochastic augmentation strategy method named RandAugment (RA) that is parameterized by two hyperparameters that can be learned with grid search instead of reinforcement learning, which leads to similar levels of performance improvement without the computational costs imposed by reinforcement learning based optimization.

However, to the best of the author's knowledge, there is no work on meta-learning augmentation strategies for medical image analysis. Data augmentation for medical imaging is therefore further explored as an original contribution of this thesis in section 3.2, and effective and principled meta-learning data augmentation strategies for ultrasound standard plane classification forms a key contribution of this thesis.

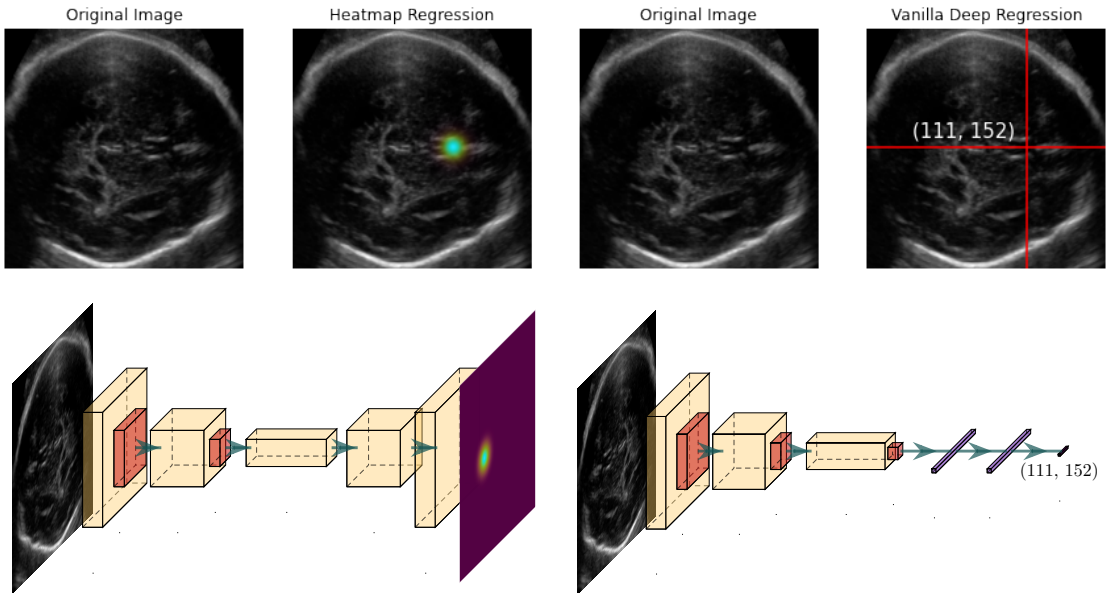


Figure 2.12: Comparison between heatmap regression and deep regression for an illustrative Cavum Septum Pellucidum localization task The figure shows the comparison between heatmap regression and deep regression for the localization task for the Cavum Septum Pellucidum (CSP) and the differences between the neural network architectures used for this task. **Left:** Heatmap regression neural networks generate heatmaps in the same shape as the input feature map. **Right:** Deep regression neural networks directly regress to the desired continous variables, here the pixel coordinate value locations for the CSP. Convolutional neural networks in both tasks are depicted as example architectures with nomenclature following that of figure 2.9 and are for illustrative purposes only. In practice, the convolutional neural networks used for these tasks will likely be more complex.

2.5 Deep Learning for Regression

Convolutional neural networks are commonly used for classification [98–100] and segmentation [101–103] in both computer vision and medical image analysis. Recently, deep learning convolutional neural network based approaches have also been shown to be performant on regression tasks, including facial point detection and landmarking [104], human pose estimation and age estimation [105, 106], and head pose estimation [107].

However, there is a distinction that can be between regression tasks. Regression tasks can be split into heatmap regression and lower-dimensional vector direct deep regression, which can be seen in figure 2.12, with different target representations of the same final desired result [108].

Heatmap regression maintains a spatial relationship between the input feature and the desired output feature map. In this case, the target representation is generally a heatmap generated with a Gaussian-filter like function [105, 109], and training is performed with a mean-squared, mean-absolute or Kullbeck-Leibler divergence loss function. The convolutional neural network then learns to generate heat maps which localize desired points. Heatmap regression is commonly used where there is strong spatial correlation between the input image and the desired output points and is generally found to be more effective than lower-dimension vector deep regression [106, 111]. Heatmap regresion also has the benefit of the prediction by the convolutional neural network is inherently bounded in the space of possible predictions, uncertainty can be inherently modelled by the spread and distribution of the predicted heatmap, and multiple prediction instances can be generated by a heatmap prediction. Examples of medical heatmap regression tasks include mass detection in mammography [110], muscle fiber orientation from ultrasound images [113], retinal boundary regression [114], and counting of cells in microscopy [115].

However, this may not always be possible for all tasks, as some computer vision and medical tasks do not have a such a spatial correlation between the desired regressed variable and the input image. Deep regression based tasks in medical imaging are therefore tasks for which heatmap regression is not possible, including cross-modality image registration based on regressed transformation parameters [116–118], pediatric bone age estimation from X-ray images [119, 120] and regression for cardiac phase estimation in fetal heart ultrasound [121], fetal heart phase estimation [16] and gestational age estimation [112]. Lower-dimensional vector direct deep regression is then used to directly generate the desired continuous variable, which is used when the spatial structure of the input feature is not maintained, seen on the right of figure 2.12.

2.6 Deep Learning based Gestational Age Estimation

Previous work on fetal gestational age estimation bypasses the difficulty of training deep regression neural networks by formulating gestational age estimation as a heatmap regression or segmentation problem. Sobhaninia [122] generate ellipse shape-constrained segmentations of scans of the fetal head to estimate fetal head circumference. Ravishankar [123] use a CNN to propose regions for hybrid segmentation of the abdomen to estimate the abdominal circumference. Maraci [124] segment the transcerebellum from candidate frames in a free-hand video sweep and use the transcerebellum diameter as fetal biometry. Van den Heuvel [125] use a head detection and segmentation algorithm optimized for low-cost ultrasound probes, which can then be used to estimate gestational age.

However, these approaches are based on regressing gestational age from clinical fetal biometry growth charts. These charts have associated uncertainties due to both inter-sonographer measurement error and growth variation, the first of which may not be applicable to convolutional neural network based measurement. Furthermore, as discussed in Table 2.4, other fetal anatomical structures can be potential candidates for gestational age estimation. This is collaborated with the work presented by Namburete [74], that finds the appearance of the Sylvian fissure, cingulate and callosal sulci strongly age discriminating. However, due to the hand-selected biometry used for gestational age regression in the previous work, the information inherent in the anatomical presentation of structures not used for biometry measurement is discarded and not used for gestational age estimation.

The question then becomes whether or not it is possible to use deep regression to directly regress gestational age from standard plane images, instead of the heatmap-regression like approach in prior literature. This would use the self-learnt feature extractor capabilities of a convolutional neural network to extract salient anatomical features from a standard plane image that can be used for gestational age estimation. Most closely related, Namburette [126] finds that a 3-dimensional convolutional neural network is able to directly regress gestational age using vanilla

deep regression, but this work requires the use of expensive three dimensional ultrasound probes. We investigate in chapter 4 different loss functions and the efficacy of anatomies of different standard planes that can be used for gestational age estimation, and in chapter 5 the use of Bayesian neural networks to add an uncertainty estimate to the gestational age estimate.

2.7 Conclusions

In section 2.2 of this literature review, established clinical methods for gestational age estimation were discussed, including the challenges in gestational age estimation in the second and third trimesters using ultrasonography. Section 2.3 discussed previous classical automated gestational age estimation methods. Section 2.4 discussed the basic building blocks of convolutional neural networks, and data augmentation methods which are used to increase the performance of convolutional neural networks especially on smaller medical datasets. The use of convolutional neural networks for regression is discussed, including the difference between heatmap based regression and vanilla deep regression, and prior approaches for gestational age estimation using deep learning. It is found that previous approaches for gestational age estimation using deep learning relied clinical fetal biometry growth charts. This thesis therefore addresses the use of vanilla deep regression networks for gestational age estimation, and their application for extracting useful anatomical features during gestational age estimation by evaluation of image regions that the convolutional neural network focusses on.

This is an ancient kabbalistic technique called notarikon in which words are generated from the initials of longer phrases

— Scott Alexander's *Unsong*

3

Principled Data Augmentation for Fetal Plane Classification

Contents

3.1 Introduction 36
3.2 Delving into Data Augmentation Methods used in Biomedical Ultrasound 37
3.2.1 Original Data Manipulation 37
3.2.2 Artifical Data Generation 42
3.2.3 Principled Data Augmentation 45
3.2.4 Quantification of Augmentation Strategies 45
3.3 Dataset 47
3.3.1 Pre-processing and Annotation 47
3.4 Experiments and Results 49
3.4.1 Experiment Implementation Details 49
3.4.2 Quantitative Evaluation on Fetal Plane Classification . 53
3.4.3 Qualitative Evaluation on Fetal Plane Classification . . 58
3.4.4 Evaluation of Augmentation Policies using Affinity and Diversity 61
3.5 Discussion and Conclusion 61

In this chapter, we address the task of principled image data augmentation from obstetric ultrasound images using a learned stochastic process. This current state-of-the-art learned stochastic process is evaluated on a key fetal ultrasound tasks - fetal standard plane classification and the effect on increasing generalization and performance over conventional data augmentation policies. Morever, a recent

data augmentation quantification method is used to evaluate our learned stochastic process for data augmentation. We show that our method leads to better augment image diversity without decreasing the affinity of the augmented image dataset distribution from the original image dataset. We also show that this stochastic method is able to out-perform some methods which require additional data for self-supervised pre-training. An early version of the work presented in this chapter has been reported in:

Lee L.H., Gao Y., Noble J.A. (2020) Principled Ultrasound Data augmentation for Classification of Standard Planes. International Conference on Information Processing in Medical Imaging (IPMI) 2021 [13].

3.1 Introduction

Data augmentation is a method by which artificial images are generated from the original training dataset in order to expand the availability of training examples. Recent work has demonstrated the power of data augmentation in biomedical image analysis [94] by reducing the tendency of over-parameterized neural networks to overfit on data. The process for data augmentation policy selection is often hand-crafted, and is often based upon prior knowledge of what a valid medical image would look like. However, recent research has found that the most effective data augmentation methods in the natural image space do not follow these constraints, and the greatest performance increase is seen with data augmentation methods that lead to un-realistic looking artificial images, with multiple affine transformations and histogram manipulations[96, 97]. We therefore explore the question of optimal augmentation policies for ultrasound image classification and specifically whether or not unrealistic image augmentation can lead to better fetal ultrasound standard plane classification.

3.2 Delving into Data Augmentation Methods used in Biomedical Ultrasound

Neural networks have been found to be effective in both the natural image space and for biomedical image analysis. However, it is found that neural network performance greatly increases with increased amounts of training data and overfit less with more data. However, acquisition of high-quality ground truth images can be expensive, especially in the medical field, where the usage of a medical expert for medical data annotation can be time-consuming and expensive. Biomedical data may also not be available for acquisition at the scale needed for CNNs due to ethical or logistical concerns.

Data augmentation is therefore used as to increase the amount of available training data and reduce the propensity for neural networks to overfit. Data augmentation methods aims to increase the amount of training data by generating synthetic examples of data that are derived from the original training data.

As explored in section 2.4.7, data augmentation methods can be broadly split into two method categories, explored in the subsections below, which also explore how each augmentation works and the output images of each transformation. The first class of data augmentation methods, explored in section 3.2.1, refers to methods where image transformations are applied to the original data in order to obtain more examples of artificially generated training data. The second class of data augmentation methods, explored in section 3.2.2, refers to methods where an auxiliary generative model is trained to model the distribution of potential training data examples, and artificial examples of training data are then generated from this auxiliary generative model during training of the original model.

3.2.1 Original Data Manipulation

Transformations in this section aim to encode invariances to the neural network during training to various image transformations. Examples of the transformations below can be seen in figure 3.1, and have all been used in ultrasound imaging to augment training data. The figure also shows the increase in training examples that

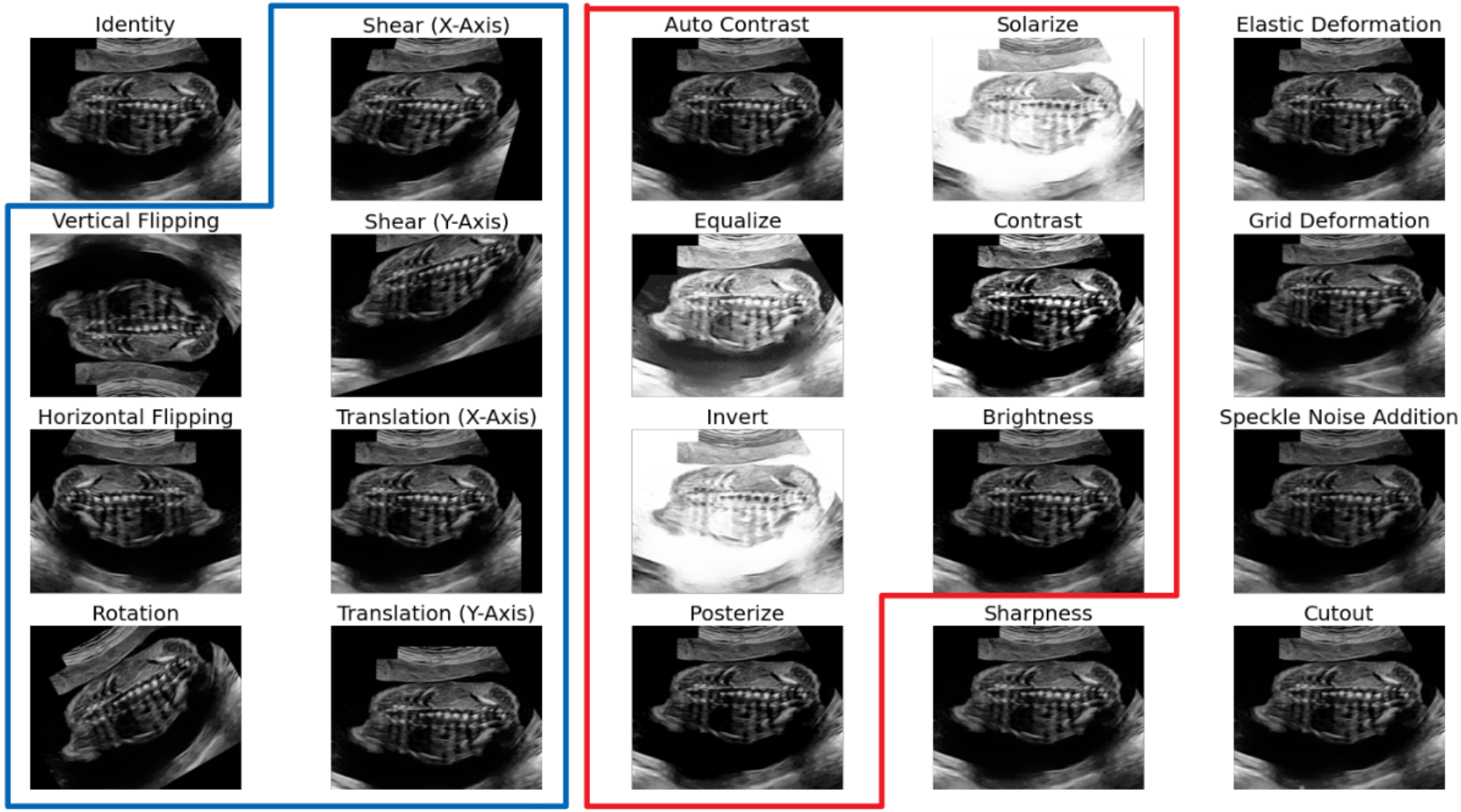


Figure 3.1: Common image manipulations of training data illustrated on a fetal spine plane image The figure shows examples of original training data **identity** manipulated with various transformations. **Vertical Flipping, Horizontal Flipping, Rotation, Shear (X-Axis), Shear (Y-Axis), Translation (X-Axis), Translation (Y-Axis)** are examples of affine transformations. **Auto Contrast, Equalize, Invert, Posterize, Solarize, Contrast, Brightness, Sharpness** are examples of image histogram transformations. The original image is of fetal spine in the coronal plane captured in a routine anomaly scan.

can be generated via transformations on the original image. Detailed descriptions of each transformation follows below. For all the transformation matrix equations and algorithms in the following sections, x and y refer to the coordinates of the original image, $\hat{x}, \hat{y}$ refer to the coordinates of the transformed image.

Image Augmentation by Flipping

Vertical or horizontal flipping mirrors the original image across the y- and x- axis respectively, and are calculated with the following transformation matrices:

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{1} \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \qquad (3.1) \qquad \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \qquad (3.2)$$

for a flip around the x-axis and y-axis respectively. It is obvious that horizontally flipping an ultrasound image generates an anatomically possible additional candidate for neural network training, as the probe can be rotated freely during image acquisition to have acquired a horizontally flipped image. However, due to the curvilinear nature of common ultrasound probes used in obstetrics, scan line density decreases with image depth and vertically flipping an ultrasound image generates "unrealistic" looking ultrasound images.

Image Augmentation by Rotation

Rotation of an image involves applying the following transformation:

$$\begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \qquad (3.3)$$

where α refers to the degree of rotation in radians. Similar to vertical flipping, due to the curvilinear nature of ultrasound probes, there exists rotation of the original image can generate unrealistic looking ultrasound images during training. Missing pixels are often zero-padded, as illustrated in figure 3.1, and resultant rotated pixel values interpolated from original pixel values.

Image Augmentation by Translations

Images may also be transformed with image translations using the equation below, where T_x and T_y refer to the translation in the x- and y- axis respectively.

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & T_x \\ 0 & 1 & T_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.4}$$

Image translations maintain the shape and size of fetal anatomical shapes that are imaged, which improves spatial invariance for the neural network, but may remove salient information in edge of the image.

Image Augmentation by Shearing

Shearing deforms the image by each pixel proportional to its distance from the centre, and is governed by the equation below where S_x and S_y refer to the magnitude of shearing in the x- and y- axis respectively.

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{1} \end{bmatrix} = \begin{bmatrix} 1 & S_x & 0 \\ S_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.5}$$

Image shearing maintains the final area of every anatomy in an ultrasound image, but distorts the shape. As shearing is an unrealistic deformation of the original training image, it is rarely used to augment medical image data.

Image Augmentation by Intensity Histogram Manipulation

Unlike the methods above, transformations on the pixel intensity histogram of an image do not affect the final geometric appearance of the image. These include changing the contrast and brightness of the original image. These transformations are often used in ultrasound image data augmentation, as changes in brightness and contrast can be often seen during routine clinical ultrasound image acquisition.

Other image histogram manipulations include **Posterize**, which quantizes the image intensity histogram to a limited number of pre-defined levels, **Equalize** which forces the distribution of image intensity values in the image histogram to be uniformly distributed, **Invert** which inverts every pixel intensity in the image

and **Solarize** which inverts pixel intensities above a certain threshold, examples of which can all be found in figure 3.1.

These image transformations are unlikely to be found in the original training data distribution, but force the neural network to be invariant to grey-level and pixel distortions, which could be useful in the ultrasound image context as this incentivizes networks learning anatomically relevant geometric shapes.

Image Augmentation by Image Distortion

Non-rigid distortions can also be used to transform an original image. Figure 3.1 shows two examples: elastic deformations [127] and grid distortions [128]. The algorithm for elastic transformations is as follows:

Algorithm 3: Algorithm for applying an elastic distortion onto a training image.

Input: Original image, elasticity coefficient σ, transformation scale α
Output: Transformed image $\hat{x}, \hat{y}$

1 Generate displacement fields $\delta x \sim \mathcal{U}(-1, 1)$, $\delta y \sim U(-1, 1)$
2 Generate Gaussian kernel $G = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$
3 Compute $\Delta x = \alpha \odot \delta x * G$
4 Sample transformed image $\hat{x}, \hat{y}$ based on final displacement field Δx

The algorithm for grid distortions is as follows:

Algorithm 4: Algorithm for applying a grid distortion onto a training image.

Input: Original image x, y, grid steps n, maximum distortion p
Output: Transformed image $\hat{x}, \hat{y}$

1 Partition original image into $n \times n$ regular grid governed by axis coordinates O_x, O_y
2 Sample distortions $P_n \sim \mathcal{U}(-p, p)$
3 $\hat{O}_x = P_n \times O_x$
4 $\hat{O}_y = P_n \times O_y$
5 Sample original image x, y based on $\hat{x_n}, \hat{y_n}$ distorted grid to get transformed image $\hat{x}, \hat{y}$

These transformations non-linearly distort the image leading to small topological

differences in the final augmented image. This may be realistic image behaviour in soft tissue which may deform under probe pressure and movement. However, such distortions are unrealistic in hard tissues, and as an example, elastic deformations of the fetal skull in routine clinical ultrasound image acquisition is unlikely.

Image Augmentation by Noise Addition

Noise can be added to pixel intensities can also be used as a form of data augmentation (e.g. Gaussian noise or salt and pepper noise). In this work, we explore the introduction of speckle noise which is a unique form of distortion to ultrasound within modalities specific to biomedical image analysis. Artificial speckle noise [129] can be inserted into augmented images by the following equation:

$$I_{sp}(x,y) = \left|\mathcal{F}^{-1}\left\{\mathcal{F}\left\{I(x,y)\cdot e^{j\,\varphi(x,y)}\right\}\cdot \mathrm{rect}_d(x,y)\right\}\right| \tag{3.6}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ refer to the Fourier and inverse Fourier transformations, rect_d is the rectangular window function with edge length d and I, I_{sp} are the source and transformed image respectively. This simulates scattering of the ultrasound waves at the observed anatomy with independent random phases due to scatterers embedded in the tissue. This scattering interferes with the in-coming and reflected wave and results in the characteristic speckle pattern found in ultrasound images.

3.2.2 Artifical Data Generation

However, there are problems related to generating transformed images from original data manipulation. As outlined above, some methods generate unrealistic looking images when used in biomedical ultrasound. Furthermore, Shin [130] suggests that such transformations on the original image fundamentally produce very correlated image features in the convolutional neural network. There is therefore research into training an external auxiliary model to generate artificial examples training data. In general, such methods train an auxiliary model to learn the distribution of the original training data, either via generative adversarial model-based methods

([14]) or variational autoencoder based methods. However, such methods are computationally expensive, and drastically increase model training time.

Mixed-Example Data Generation

There therefore has been work on mixed-example data generation, which aims to generate new examples of training data by interpolating between multiple examples of original training data [131, 132]. Zhang [132] uses an linear interpolation method with a linear interpolation of the final image, whilst Summers [131] uses a non-linear mixup method that is shown in the figure 3.2 which was found to be more effective than the linear interpolation method when used as part of an augmentation strategy. Furthermore, the general convergence and stability of generative adversarial networks are a field of active research [133–135] as generative adversarial model based methods are fragile during implementation.

As illustrated in the figure, instead of pixel-wise linear averaging as suggested by Zhang [132], the bottom λ_1 fraction of image x_1 is concatenated with the top $1-\lambda_1$ fraction of image x_2, the right λ_2 fraction of image x_1 is concatenated with the left $1-\lambda$ fraction of image x_2, and the two concatenated images are mixed with proportion λ_3. Formally, given initial images $x_{1,2}$ with image intensity means and standard deviations of $\mu_{1,2}$ and $\sigma_{1,2}$, the generated mixed-example image $\hat{x}_{1,2}$ would be:

$$\hat{x}_{1,2} = \begin{cases} x_1(i,j) - \mu_1 & \text{if } i \leq \lambda_1 H \text{ and } j \leq \lambda_2 W \\ \frac{c}{\phi}[x_1(i,j) - \mu_1] + \frac{1-c}{\phi}[x_2(i,j) - \mu_2] & \text{if } i \leq \lambda_1 H \text{ and } j > \lambda_2 W \\ \frac{1-c}{\phi}[x_1(i,j) - \mu_1] + \frac{c}{\phi}[x_2(i,j) - \mu_2] & \text{if } i > \lambda_1 H \text{ and } j \leq \lambda_2 W \\ x_2(i,j) - \mu_2 & \text{if } i > \lambda_1 H \text{ and } j > \lambda_2 W \end{cases} \tag{3.7}$$

where c is the mixing coefficient $(1 + \frac{\sigma_1}{\sigma_2} \cdot \frac{1-\lambda_3}{\lambda_3})^{-1}$ and ϕ is the normalization term defined as $\sqrt{c^2 + (1-c)^2}$. The row index and column index is represented by i, j and the height and width of the images are represented by H, W. This method is able to drastically expand the quantity of training data, and is found to improve performance in the natural image space.

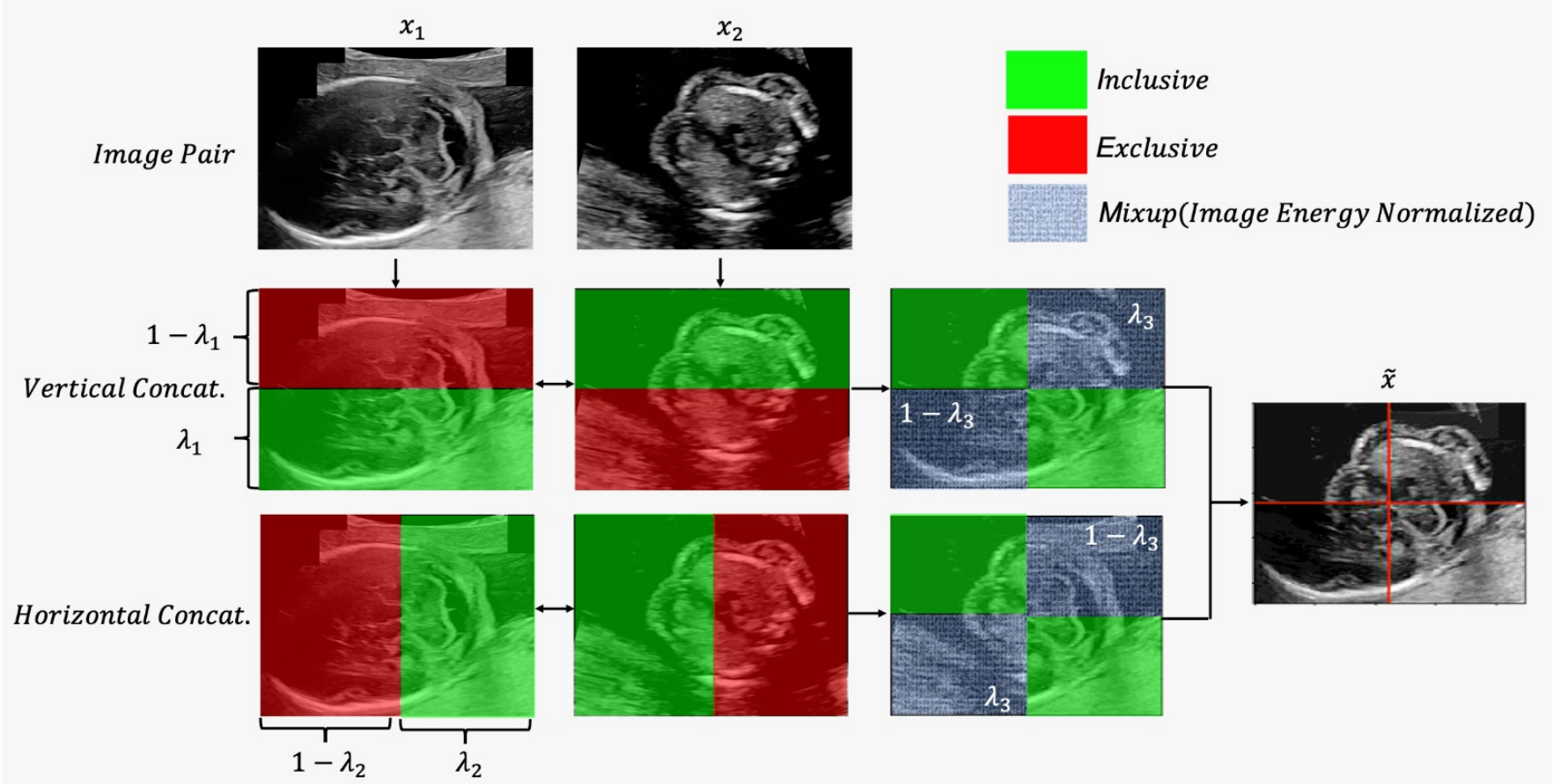


Figure 3.2: Non-Linear Energy-Normalized Mixed-Example Data Generation The figure shows the procedure for energy normalized non-linear generation of mixed-example data. Three hyperparameters are used, $\lambda_1, \lambda_2, \lambda_3$ which control the degree of mixing between each original training image x_1 and x_2, generating the output image $\hat{x}$

3.2.3 Principled Data Augmentation

We can define an augmentation strategy as the series of transformations that are used on the training set in order to generate an augmented training set, and principled data augmentation as optimizing for an augmentation strategy that maximizes final test performance. Augmentation strategy selection in biomedical imaging is currently done in an ad-hoc basis. Augmentation strategies with original data manipulation in ultrasound image analysis is typically justified by stating [99] visually realistic ultrasound images will be simulated, despite the usage of physically impossible transformations such as image rotations and shears on images acquired from a curvilinear probe. However, recent research in the natural image space [96, 97] have found that "unnatural" image transformations that generate data that are visually outside of the training data distribution can improve performance beyond that of augmentation strategies that generate "realistic" images. Furthermore, augmentation strategy optimization methods outperform hand-crafted augmentation strategies [96, 97].

The AutoAugment augmentation strategy search method [96] uses reinforcement learning to find an optimum augmentation strategy for a given model and dataset to maximize test accuracy, but requires significant computational power due to the size of the search space. On the other hand, the RandAugment strategy [97] ties the augmentation strategy to two hyperparameters; the number of augmentation transforms and the intensity of each transform. It is found that for the natural image space and MNIST, optimization of these two hyperparameters is sufficient to outperform the AutoAugment strategy and hand-optimized augmentation strategies despite the drastically reduces search space.

3.2.4 Quantification of Augmentation Strategies

We also employ two quantifiable measures of augmentation strategies and evaluate the impact of each on the final performance of the model, diversity and affinity [136]. These measures can be used to quantifiably evaluate the differences between hand-crafted augmentation strategies and automated augmentation strategy selection.

Affinity refers to the data distribution shift between the augmented training set and the original training set for any given augmentation strategy with respect to a model trained on the original baseline training set. Given a model m trained on the original clean training set, the affinity can be calculated by:

$$A = L(m, D_{val}^S) - L(m, D_{val}) \tag{3.8}$$

where A is the final calculated affinity, D_{val} is the validation dataset, $L(m, D)$ is the loss when evaluated with the model m and the dataset D, and D_{val}^S is the augmented validation dataset when the dataset is augmented with the augmentation strategy S.

Intuitively, the more negative affinity A is, the greater the distribution shift between the augmented dataset D_{val}^S and the validation dataset D_{val}, which suggests that the augmented data is less representative of the original training dataset. Furthermore, affinity is calculated both with respect to the training data distribution and the model architecture, which makes it a more specific distance measure for our task.

Diversity is then a measure of the variation given by the chosen augmentation strategy process. The more varied the training data, the harder it would be for a model to overfit on the dataset. Similar to affinity, this can be calculated with respect to the model architecture chosen for the task.

$$Div = \mathbb{E}[L(m, D_{train}^S)] \tag{3.9}$$

where Div is the final calculated diversity, $\mathbb{E}$ refers to the expectation, and D_{train}^S refers to the augmented training set. The diversity can then be calculated as the average final training loss for each model, and is a measure of how difficult it is for a model to fit on the training set.

These two measures can be used to quantify the effect of an augmentation strategy on the training data.

3.3 Dataset

The clinical data used in this work consists of 135 routine second trimester fetal ultrasound scans acquired in a clinical setting. The gestational age of fetuses in scans varied from 18+0 to 20+6 weeks. The data was acquired using a GE Volusen E8 Scanner (General Electric, USA) by experienced clinicians. These scans were recorded as part of the PULSE (Perception Ultrasound by Learning Sonographer Experience) study[1], which was designed to learn how sonographers perform ultrasound clinical scans, and were collected at the John Radcliffe hospital in Oxford.

3.3.1 Pre-processing and Annotation

The classification task was to label fourteen key fetal plane images, listed in the table 3.3. These planes are defined by the NHS Fetal Anomaly Screening Programme [5].

As seen in the table, these biomedical scan planes included the planes which must be captured as part of the anomaly scan, as well as heart standard planes that are viewed as part of the scan. These classes were also used for ease of classification comparison with previous work [99, 137, 138], which the performance of our augmentation search strategy can be compared against.

Ultrasound scan planes captured by the sonographer during routine anomaly scanning were then manually classified into the fourteen classes. The annotations were subsequently validated by an experienced sonographer. The number of planes per class varied from 62 images of the kidneys plane to 148 images of the lips plane, as planes may not be represented in each of the 135 scans, giving a total of 1127 non-background video frames. A further 1127 background frames were sampled from each ultrasound video where no video was captured.

Scans were partitioned into three folds for three fold cross-validation during the experimental process, and scans were partitioned based on anonymized subject identity to prevent overestimation of classification accuracy.

[1] UK Research Ethics Committee Reference 18/WS/0051

Table 3.1: Fetal Anomaly Screening Programme (FASP) [5] derived planes for standard plane classification.

FASP Ultrasound Scan Base Menu	Description (and Key Usage)	FASP Capture Storage Requirement
Brain TC	Scan of the fetal brain at the level of the Cerebellum; used to measure transcerebellar diameter.	Yes
Brain TV	Scan of the fetal brain at the transventricular plane; used for atrium lateral ventrical measurement.	Yes
Lips	Coronal view of the lips and nasal tip; used for assessment for congenital abnormalities of the fetal face.	Yes
Abdominal	Scan of the abdomen; used to measure the abdominal circumference.	Yes
Kidneys	Scan of the kidneys in the axial view; used for assessment of anteroposterior (AP) renal pelvis diameter.	Yes (if AP renal pelvis diameter >7mm)
Femur	Scan of the femur plane; used for measurement of femur length.	Yes
Spine Sag.	Scan of the fetal spine in a sagittal plane; used for assessment of vertebrae and skin covering.	One of the sagittal or coronal plane
Spine Cor.	Scan of the spine in the coronal plane, used for assessment of vertebrae and skin covering.	
FASP Fetal Cardiac Protocol		
4CH	Four chamber view.	No
3VT	Three vessel and trachea view.	No
RVOT	Left ventricular outflow tract view.	No
LVOT	Right ventricular outflow tract view.	No
Other planes		
Profile	Facial profile in the median plane.	No
Background	Fetal scan images not in the above categories acquired during the scanning process.	No

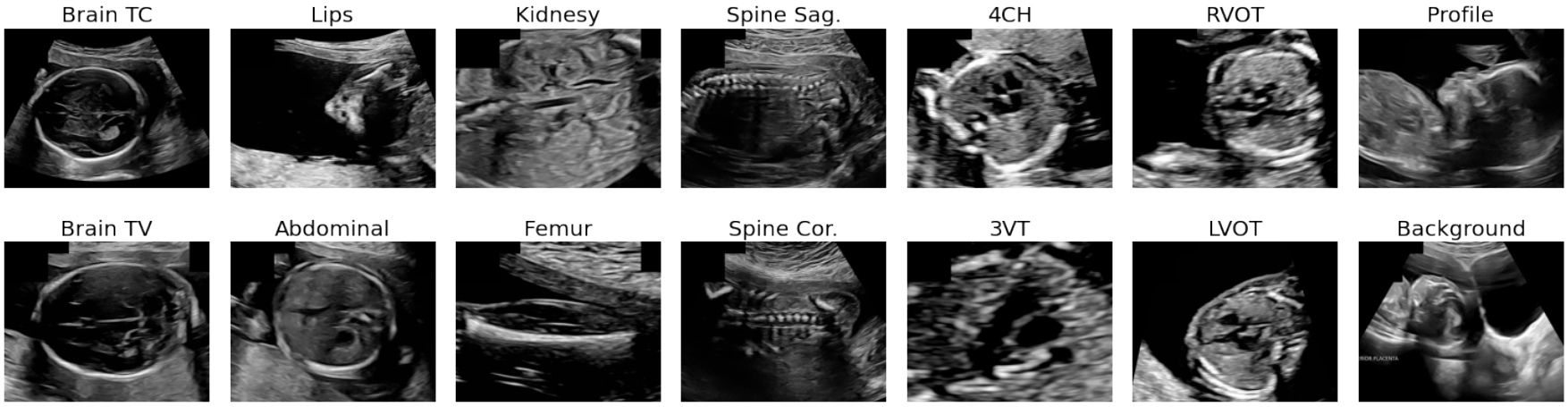


Figure 3.3: Example images of fetal planes used in fetal standard plane classification task

Plane images with sonographer labelling text or markers were removed. Images were downsampled to 224 by 224 pixels via bilinear resampling, and were intensity normalized before being used for training and testing.

3.4 Experiments and Results

Experiments were designed to evaluate the efficacy of training a SE-ResNeXt-50 [89] from scratch with various augmentation policies. The SE-ResNeXt-50 used in the standard plane classification had 16 channels in the final fully-connected layer for use in classification instead of the 1000 in the original implementation due to different classification requirements.

We try to answer the questions i) would a RandAugment based augmentation strategy search method improve test accuracy on biomedical imaging datasets, ii) would these augmentation strategy search methods work on the ultrasound imaging modality with significantly different imaging characteristics compared to natural images, and iii) can a mixed-example based augmentation policy be used with a RandAugment based augmentation strategy on ultrasound imaging?

Experiments were designed to evaluate the performance of standard plane classification on models trained with no augmentation, a standard hand-crafted augmentation policy commonly used in ultrasound image data analysis [99] (SN Pol.) and the RandAugment (RA) [97] augmentation search policy algorithm. We then evaluate the performance of a RandAugment inspired augmentation search policy with inclusions of grid and elastic distortions (see section 3.2.1), speckle noise augmentation, and include mixed-example data augmentation 3.2.2 in the policy search (Mix. RA).

3.4.1 Experiment Implementation Details

SE-ResNeXt-50 models were trained using SGD with Nesterov momentum with a batch size of 16, momentum of 0.9 and 1×10^{-5} weight decay. The learning rate was constant for every training parameter, and initial learning rates were set to 1×10^{-2}

and reduced by a factor of 10 up to three times during training loss stagnation, which was defined as 10 continuous epochs of training without validation loss improvement.

Mix. RA Implementation Details

The original RandAugment augmentation policy search algorithm simplifies the search space to two hyperparameters which can be learnt via simple grid search. The first hyperparameter n selects for the number of augmentation transformations that each image x undergoes, which is stochastically sampled from a predefined list of transformations. The second m selects for the magnitude of distortion for each augmentation transformation. More specifically:

$$\hat{x} = T_n(T_{n-1}(...T_2(T_1(x, m), m), m), ..., m) \tag{3.10}$$

$$\hat{y} = y \tag{3.11}$$

where x and $\hat{x}$ are the original training and augmented training image respectively, y and $\hat{y}$ are the original and augmented training objectives, and $T(x, m)$ represents the output of image x transformed with augmentation T with magnitude m. Due to the compression of the search space, the meaning of the magnitude of a transformation is necessarily different for each transformation.

Inspired by the simplification of the augmentation policy search space by RandAugment (RA) to two hyperparameters, we introduce the addition of mixed-example data augmentation to a RandAugment-like augmentation policy search algorithm tied to the same two hyperparameters Our Mix. RA augmentation policy then becomes:

$$\lambda_1, \lambda_2, \lambda_3 \sim Beta(m, m) \tag{3.12}$$

$$x_{1,2} = Mix(x_1, x_2, \lambda_{1,2,3}) \tag{3.13}$$

$$\hat{x}_{1,2} = T_n(T_{n-1}(...T_2(T_1(x_{1,2}, m), m), m), ..., m) \tag{3.14}$$

$$\hat{y}_{1,2} = (\lambda_3\lambda_1 + (1 - \lambda_3)\lambda_2)y_1 + (\lambda_3(1 - \lambda_1) + (1 - \lambda_3)(1 - \lambda_2))y_2 \tag{3.15}$$

where x_1, x_2 are the original training images, $beta(m, m)$ refers to the beta distribution, $Mix(x_1, x_2, \lambda_{1,2,3})$ is the mixed-example augmentation referred to in equation 3.7.

We also expand the list of potential transformations selected by the stochastic transformation sampler in the RA algorithm, and evaluate the performance of each included transformation. The full list of potential transformations T that can be selected as part of the augmentation search policy is as follows:

- Translations X/Y
- Shear X/Y
- Intensity Inversion
- Intensity Equalization
- Rotation
- Intensity Autocontrast
- Gamma Adjustment
- Posterize
- Cutout
- Sharpness
- Solarize
- Contrast Adjustment
- Brightness Adjustment
- Identity Transformation
- **Grid Distortion**
- **Elastic Distortion**
- **Speckle Distortion**

where the items in bold are extended over the original RandAugment implementation. We restrict m to be between $0 - 10$ and use the same range of image distortion per layer as the original RandAugment implementation. Examples of images through different m, n hyperparameters can be seen in figure 3.4. It can be seen that increasing either of m or n increases the degree of distortion for the final augmented image. It can be seen that a smaller degree of distortion at $n = 2$ and $m = 2$ tends to preserve geometric appearance of the entire image and image histogram distortion is minimal. Whereas at $n = 7$ and $m = 7$ key anatomical structures (such as the stomach bubble and spine for the abdominal

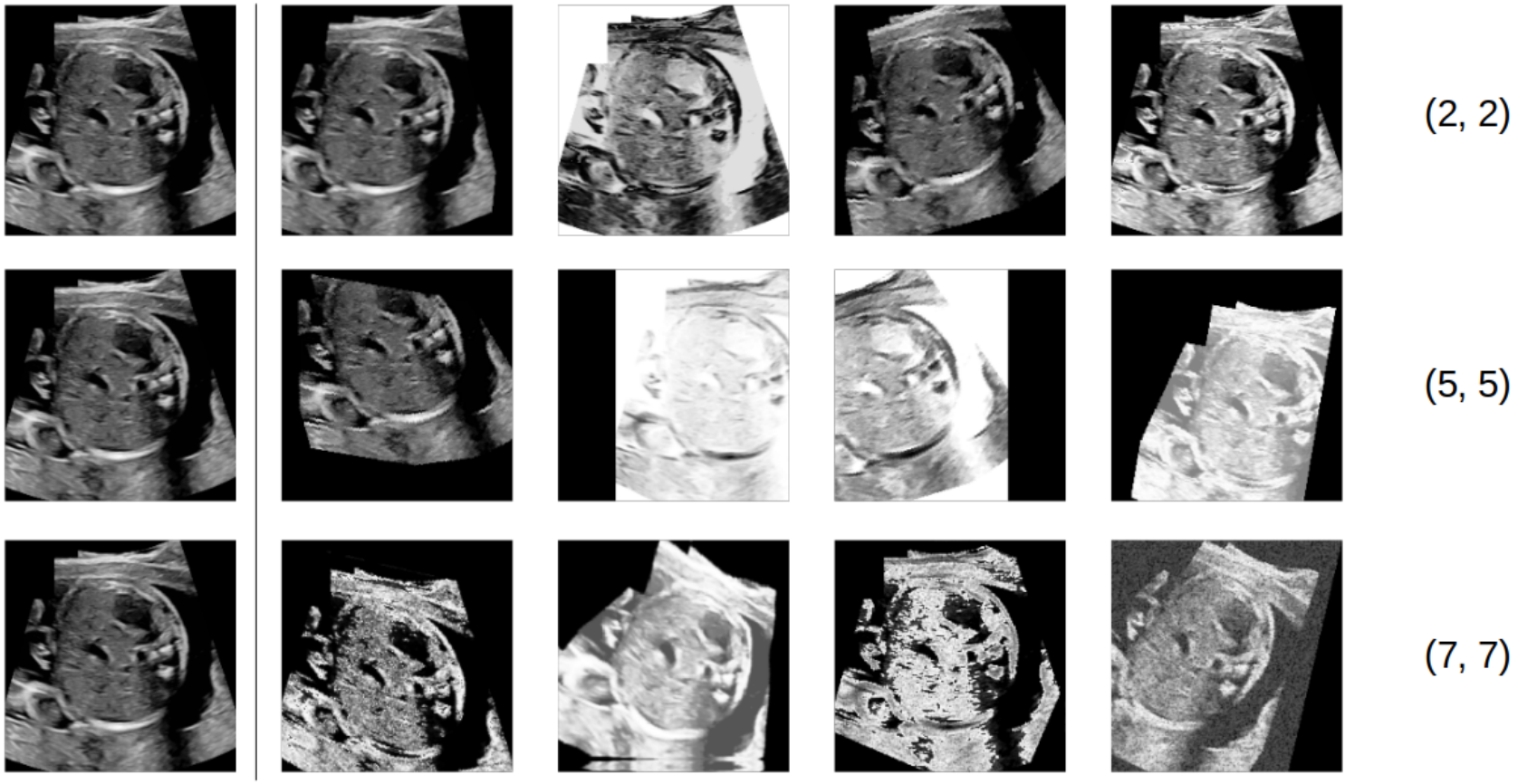


Figure 3.4: Augmented images through different hyperparameters of the Mix. RA algorithm The first column shows the original abdominal circumference plane image used for augmentation. The other columns show images augmented with (2, 2), (5, 5) and (7, 7) (m, n) respectively. As can be seen from the figure, increasing the magnitude of distortion for each augmentation and increasing the number of transformations selected drastically changes final augmented image distortion.

circumference plane) may be occluded/not visible and the distortion of the image intensity histogram takes the augmented image out of range for a "realistic" image.

Due to the limited number of hyperparameters in the augmentation policy definition with $n \in 0, 1, 2, 3, ..9$ and $m \in 0, 1, 2, 3, ..9$, we find the optimum (m, n) tuple by simple grid search evaluation. We train five models for each evaluated (m, n) tuple and use the best performing model in terms of accuracy to select the final (m, n) tuple used for model evaluation. We explore the classification performance of three RA based augmentation policies: the original RandAugment (RA) as proposed in [97], an extended RandAugment (Ext. RA) policy with additional transformations as proposed in section 3.4.1, and an RandAugment policy with mixed-example data augmentation (Mix. RA) followed by transformations selected from the selected list defined by equation 3.7. Random horizontal and vertical flipping was used in all RA-based augmentation policies as a baseline augmentation.

3.4.2 Quantitative Evaluation on Fetal Plane Classification

Grid-Search for Augmentation Policy Hyperparameter Optimization

The left of figures 3.5 and 3.6 shows the accuracy variation in final accuracy for $n \in 1, 3, 5, 7, 9$ and $m \in 1, 3, 5, 7, 9$ for RA and Mix. RA respectively, with the best performing hyperparameter tuple highlighted in red. The right of both figures show the standard deviation in accuracy for the final trained models. We find that (5, 3) and (3, 3) maximizes validation accuracy during training for each augmentation policy respectively, which stochastically generates augmented training images that can be seen in the bottom of both figures. Qualitatively, these images do not exhibit strong geometric distortions with images still identifiable as a fetal standard plane. It can also be seen that there is a trade off between m and n, values too high of either leading to reduced performance for final classification. We can also see that there is an increasing instability for final training accuracy with higher degrees of distortion. The SE-ResNeXt-50 models that were trained with the optimal (m, n) tuple had a variation in final validation accuracy of 0.6%, as compared to 2.2% for $(9, 9)$. This can be explained by augmentation policies

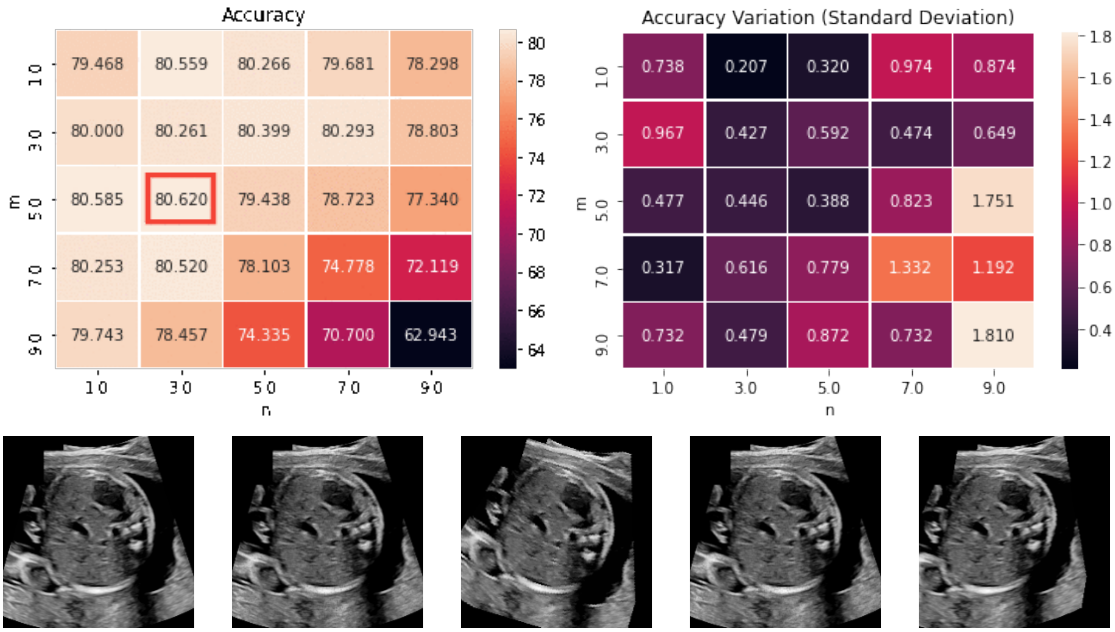


Figure 3.5: Final training classification accuracies for hyperparameter grid-search using RA for a SE-ResNeXt-50 model trained from random initialization Top Left: Best performing classification accuracy for various (m, n) tuples. **Top Right:** Variation in accuracy for evaluated models in standard deviation. **Bottom:** Bottom shows the example augmented training set generated with $(m, n) = (5, 3)$. The original ground truth image can be found in figure 3.4.

with higher stochasticity leading to different local minima during training due to the greater variation in the loss surface. We find that Mix. RA has a greater validation accuracy than RA. However, due to the large class imbalance, accuracy is not fully indicative of classification performance.

Quantitative Standard Plane Classification Performance

The final test classification error is therefore calculated in terms of macro-precision, recall and F1-scores due to the large class imbalance. The summary is found in table 3.2. SN Pol. [99] was included as an example of a standard hand-crafted augmentation policy, which consists of random scaling, horizontal flipping and rotations of $\pm$ 20 degrees.

The table reports macro-F1 scores due to the large sample imbalance between classes. We find that the RandAugment based augmentation policy searches (RA, Ext. RA, Mix. RA) outperform no augmentation (No Aug) and an example of a

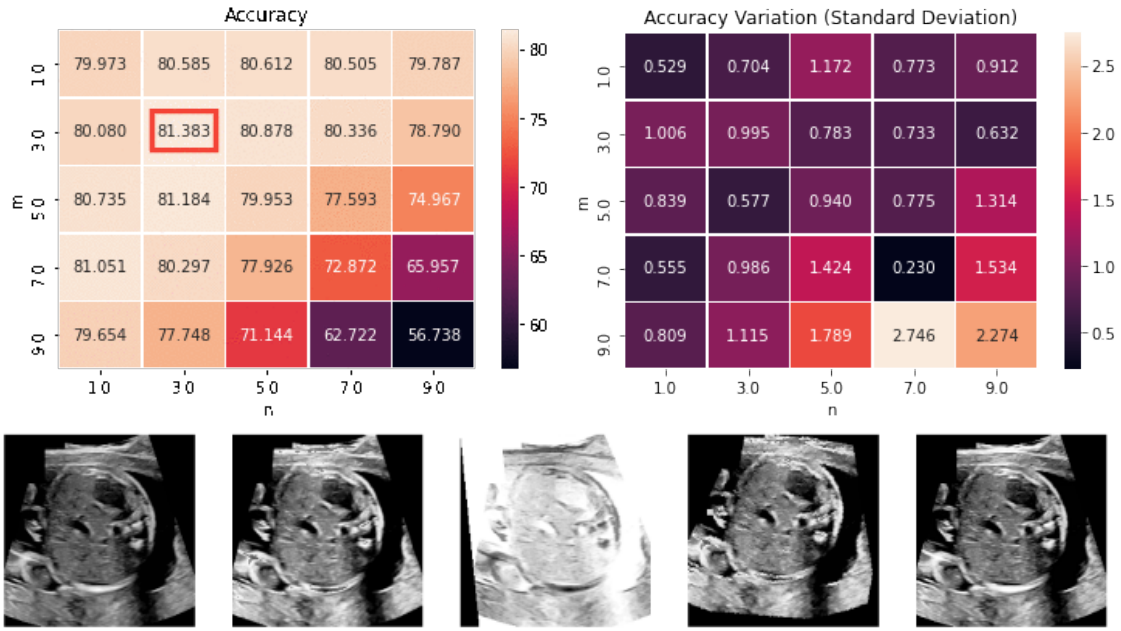


Figure 3.6: Final training accuracies for hyperparameter grid-search using Mix. RA for a SE-ResNeXt-50 model trained from random initialization Top Left: Best performing accuracy for various (m, n) tuples. **Top Right:** Variation in accuracy for evaluated models in standard deviation. **Bottom:** Bottom shows the example augmented training set generated with $(m, n) = (3, 3)$. The original ground truth image can be found in figure 3.4.

Table 3.2: Macro-precision, recall and F1-scores for evaluated augmentation policies compared to recent work. Table of final classification performance results. Results in grey include pre-training using external data sources such as gaze and video data, and are included as comparison. Results in black were trained from random initialization.

	No Aug.	SN Pol.	RA	Ext. RA (ours)	Mix. RA (ours)	Gaze [138]	Siam. Init. [137]	Saliency [138]
Precision	56.5±1.2	70.4±2.3	74.7±1.8	74.0±2.4	**75.1±1.8**	67.2±3.4	75.8±1.9	79.5±1.7
Recall	55.1±1.2	64.9±1.6	72.2±2.3	**74.5±1.3**	73.4±1.9	57.3±4.5	76.4±2.7	75.1±3.4
F1-Score	55.4±1.2	67.0±1.3	72.8±1.8	73.6±1.3	**74.0±1.8**	60.7±3.9	75.7±2.0	76.6±2.6

hand crafted augmentation policy (SN Pol.) significantly. The best performing Mix. RA in terms of overall F1-Score outperformed the SN Pol. by 6.0%. Furthermore, Mix. RA was able to outperform some previously reported pre-training methods including methods which required external gaze tracking data (Gaze) and pre-training using ultrasound video (Siam. Init.).

The classification error can be further broken down by class in the confusion matrix shown in figure 3.7 between the best performing Mix. RA and SN Pol.

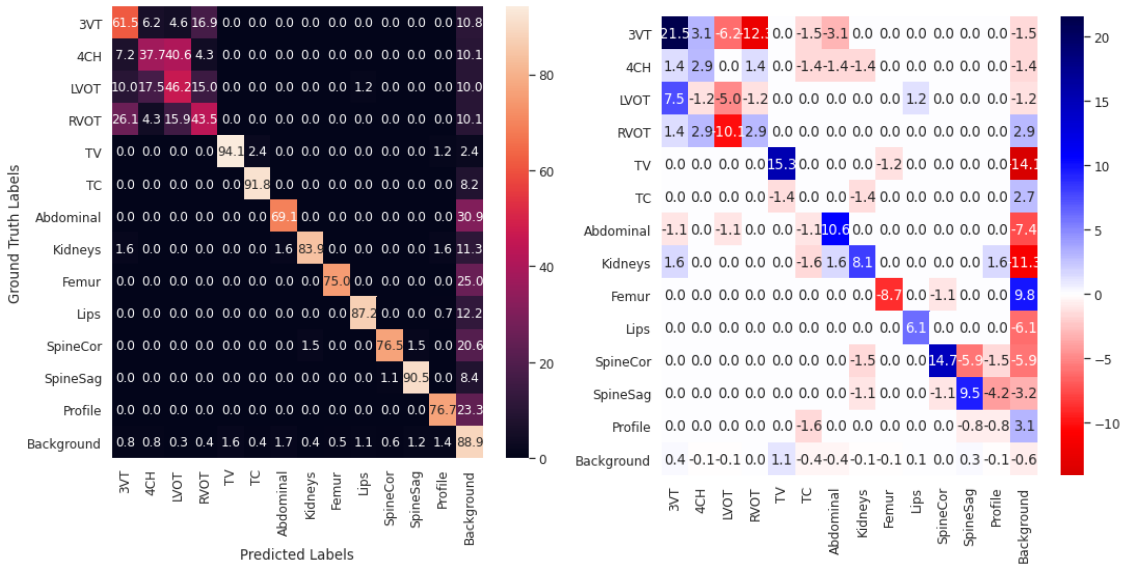


Figure 3.7: Class confusion matrix for the best performing augmentation policy found by Mix RA. augmentation policy search Left: Confusion matrix for accuracies broken down by class. **Right:** The difference in accuracy between a model trained with Mix. RA and a model trained with SN Pol. Both models were trained from random initialization.

We can see from figure 3.7 that a significant proportion of the performance increase was the correct classification of previously incorrectly classified background plane images. In particular, the kidney plane (11.3%), abdominal plane (7.4%), lips (6.1%) and coronal spine planes (5.9%) had incorrectly classified probability mass shifted to the correct classes. From the confusion matrix, we can also see that there is less-intra anatomical class confusion. The probability mass within the coronal and sagittal spine planes (SpineCor, SpineSag) are classified more accurately, with 5.9% of incorrectly classified spinal sagittal planes moving to the correctly classified spinal coronal plane, and 1.1% of the incorrectly classified spine coronal plane moving to the correctly classified sagittal plane. The confusion between the sagittal spine plane and the profile plane is also reduced by 4.2%.

Other sources of intra-anatomical confusion in the SN Pol. network can be seen between the abdominal and kidney planes. The majority of improvement in these planes is due to the subsequent correct classification of background planes, but there is some improvement from correctly differentiating between kidney and abdominal planes.

Table 3.3: Ablation study on the individual components of our *Mix. RA* policy search algorithm on training of a randomly initialized SE-ResNeXt-50 for ultrasound standard plane detection. All metrics are macro-averages due to the class imbalance. The *Linear Mix. RA* is included as a baseline mixed-example augmentation strategy.

	No Aug.	SN Pol.	RA	RA + Speckle ①	RA + Deform. ②	Ext. RA ① + ②	Linear Mix. RA ① + ②	Non-Linear Mix. RA ① + ②
Precision	56.5±1.2	70.4±2.3	73.9±1.7	74.6±1.7	74.6±1.7	74.0±2.4	74.6±1.6	**75.1±1.8**
Recall	55.1±1.2	64.9±1.6	72.9±2.0	72.1±1.8	72.9±1.8	**74.5±1.3**	73.2±1.5	73.4±1.9
F1-Score	55.4±1.2	67.0±1.3	72.8±1.8	72.9±1.7	73.2±1.2	73.6±1.3	73.7±1.6	**74.0±1.8**

However, the model still has difficulty in differentiating between the planes of the heart. From SN Pol. to Mix. Aug, the classification of heart planes is improved for the 3VT, 4CH and RVOT planes respectively. The RVOT sees improvement in classification from previously misclassified LVOT planes (10.1%), whilst the 3VT plane experiences the greatest improvement of 21.5% from correct classification of the LVOT and RVOT planes to 3VT (6.2% and 12.1%). However, this also leads to misclassification of LVOT planes from previously correctly classified LVOT planes to the 3VT classification (7.5%).

In general, the heart is poorly classified even in the best-performing Mix. Aug. model, with an average accuracy of 47.2% across all four planes. This is due to the similar appearance of the four heart planes which can be seen in figure 3.3. This difficulty in differentiating static images of heart planes is exhibited even by experienced sonographers.

The best performing planes are the fetal head trans-ventricular plane and the fetal head trans-cerebellar planes at 94.1% and 91.8% accuracy respectively. This may be due to the fact that a large part of the difficulty in classification of fetal planes is in the separation between the standard plane and the background class, and the fetal head is obviously characterized by the fetal skull.

We can also find the respective performance increase that each extension on the original RA augmentation gives in the table 3.3.

In table 3.3, RA + Speckle refers to an RA based augmentation search policy with the list of transformations extended with speckle noise augmentation (see section 5), and RA + Deform refers to an RA based augmentation search policy

extended with both elastic and grid deformation transformations (see section 3.2.1). We can see that both the addition of speckle noise and deformation to the available list of transformations improves the performance of the augmentation policy search, which is further improved when the two are combined in Ext. RA. This performance is further enhanced with mixed-example data generation in Non-Linear Mix. RA, leading to a final F1-Score of 74.0%.

3.4.3 Qualitative Evaluation on Fetal Plane Classification

t-SNE Visualization

t-SNE [139] is a technique to aid visualization of high-dimensional data, and can be used to visualize the features extracted from trained models. The Kullbeck Leibler divergence between similarity measures for instances in the original high dimensional space and the low dimensional representation is minimized, which then allows for visualization of features in the original higher-dimensional manifold.

We extract the features from the penultimate layer of the SE-ResNeXt-50 layer to visualize the learned high-dimensional features using t-SNE.

Figure 3.8 shows the different t-SNE visualizations of the extracted high level features from the best performing Mix. RA augmentation policy and RA augmentation policy. We can see from the figure a clearer cluster delineation between anatomical planes (coloured) and background planes (grey), with fewer mis-classified anatomical planes in the background cluster. There is also continued entanglement between the heart standard planes, as is expected from the quantitative evaluation.

We can also visualize the improvement in classification by inspecting the frames which were previously misclassified using an RA trained model which are now correctly classified with a model trained with Mix. RA using integrated gradients.

Integrated Gradients for Axiomatic Attribution

Integrated gradients [140] is a technique that allows for attribution of the final classification to input features. The gradients of the final predicted output is calculated with respect to the final prediction. Integrated gradients satisfies two

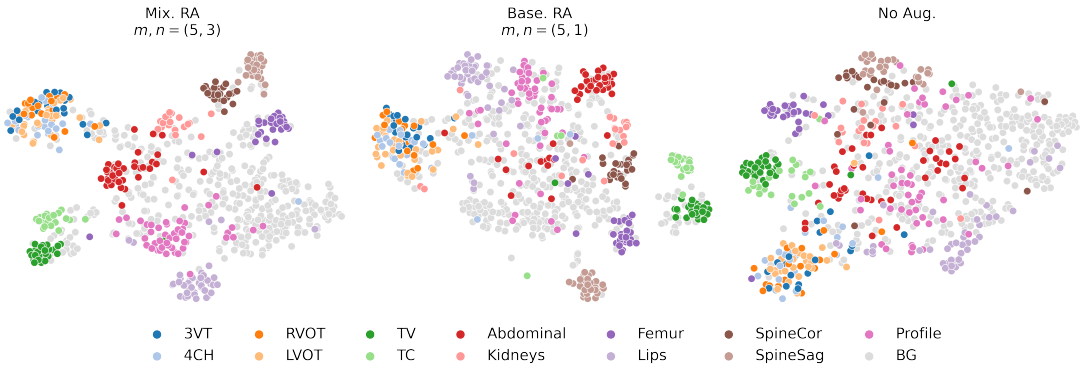


Figure 3.8: t-SNE visualizations of the extracted high level features from best performing RA and Mix. RA augmentation policies

axioms, sensitivity, and implementation invariance for attribution calculation, which makes it more robust than other commonly used attribution methods. An attribution method satisfies sensitivity if it gives non-zero attribution to features that generate outputs. An attribution method can be said to be invariant to implementation if two different functionally identical models with the same decision boundaries in the input space lead to the same attributions.

We can therefore use integrated gradients to visualize the attribution of pixels in the input image that lead to mis-classification in the original RA model and correct classification in Mix. RA. Examples of images which were previously mis-classified as background planes using a model trained with RA and correctly classified with Mix. RA can be seen in figure 3.9. These include images from classes which are commonly mis-classified when trained with RA, including the fetal brain transventricular plane, abdominal circumference plane, spine coronal plane, and kidneys plane (see figure 3.7).

In general, the RA model which misclassified head circumference planes as background planes failed to find the skull, with the classification based upon the inside of the skull. On the other hand, correctly classified head circumference planes have the skull clearly delineated This pattern is repeated for the abdominal plane, where the stomach bubble and spine is highlighted in the integrated gradients based attribution picture. For the coronal views of both the spine and kidneys, the fetus is more clearly localized and highlighted against the maternal anatomy. The spines

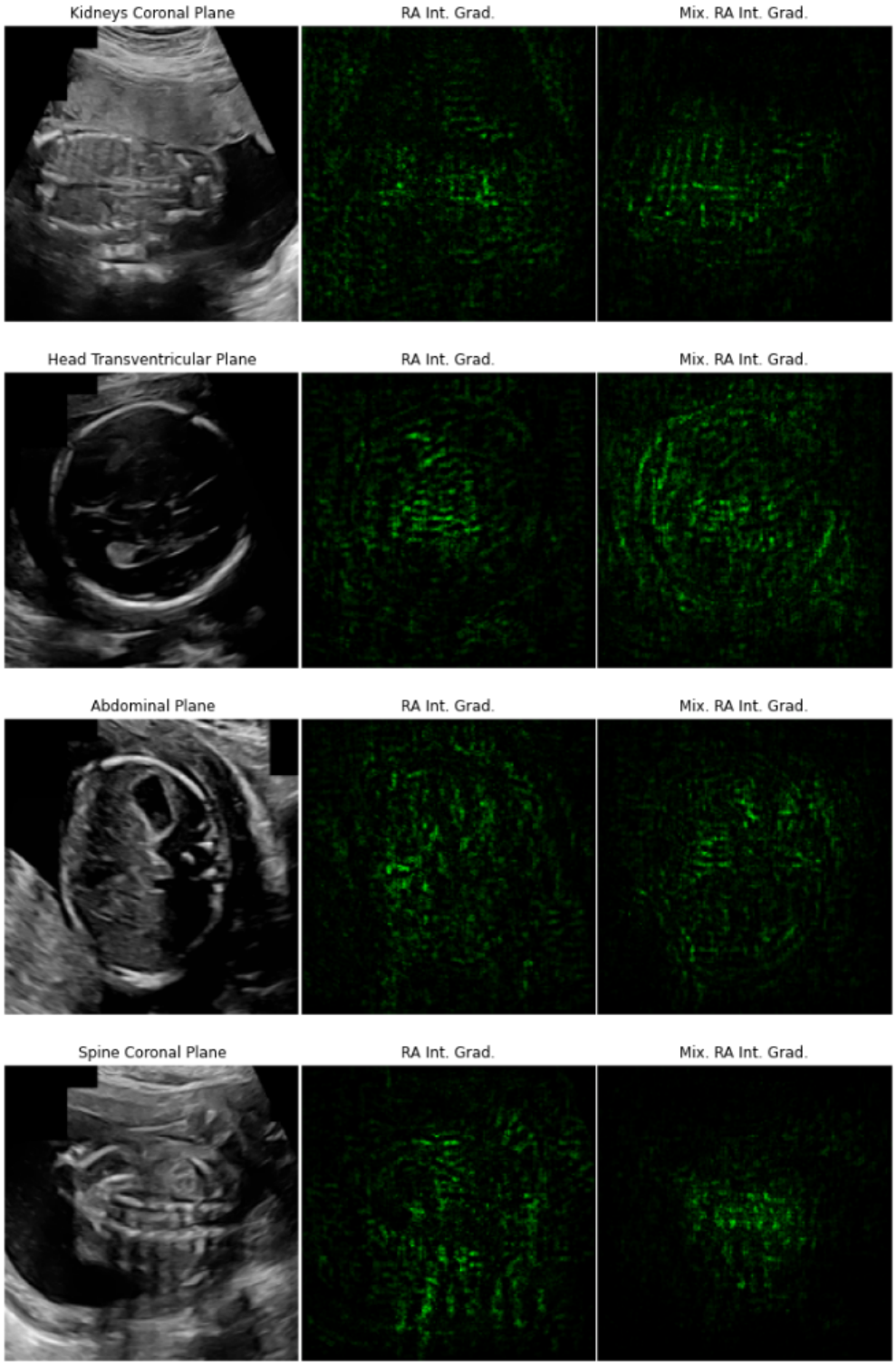


Figure 3.9: Integrated Gradients based attribution for image classification on RA and Mix. RA The model trained with the RA augmentation policy mis-classifies these images as background. These images are subsequently correctly classified by the Mix. RA model. The integrated gradients (Int. Grad.) based attribution assigns the attribution for the classification.

are clearly used in the decision for spinal classification in the spinal coronal view, which is in line with intuition. The above suggests that using Mix. RA allowed the model to learn stronger anatomical priors for what constitutes a standard plane compared to RA. This is likely due to the inclusion of non-linear deformation fields and speckle noise patterns in the augmentation transformation list.

3.4.4 Evaluation of Augmentation Policies using Affinity and Diversity

To examine the effects of the m and n hyperparameters in RA-based augmentation policy selection, we visualize the effects of the augmentation policy on affinity and diversity against final test accuracy in figure 3.10

There appears a sweet spot of affinity and diversity using non- mixed-class augmentation strategies at an affinity distance of 3.8 and diversity of 0.25 which maximized model performance. At high m, n values, whilst the diversity of generated images is increased, the augmented images are too out of distribution compared to the original training images which reduces model performance. At low m, n values, whilst there is a high affinity between the generated dataset and the original training dataset, the model sees too little variation in input data and model performance also is reduced. Using Mix. RA as the augmentation policy drastically increases the diversity of images with every given affinity due to the increased space of potential images generated, which could act to explain the final improved optimal model performance. This visualization of the diversity and affinity trade-off suggests that hand-crafted augmentation policies that lie on the same diversity/affinity curve as RA may lead to similar optimal performance, which may be easier to justify clinically when a model is trained with "realistic" images.

3.5 Discussion and Conclusion

To summarize, we have thoroughly experimented with automated augmentation policy search methods for fetal ultrasound plane classification. The results suggest that whilst the RandAugment augmentation policy search method is able to increase

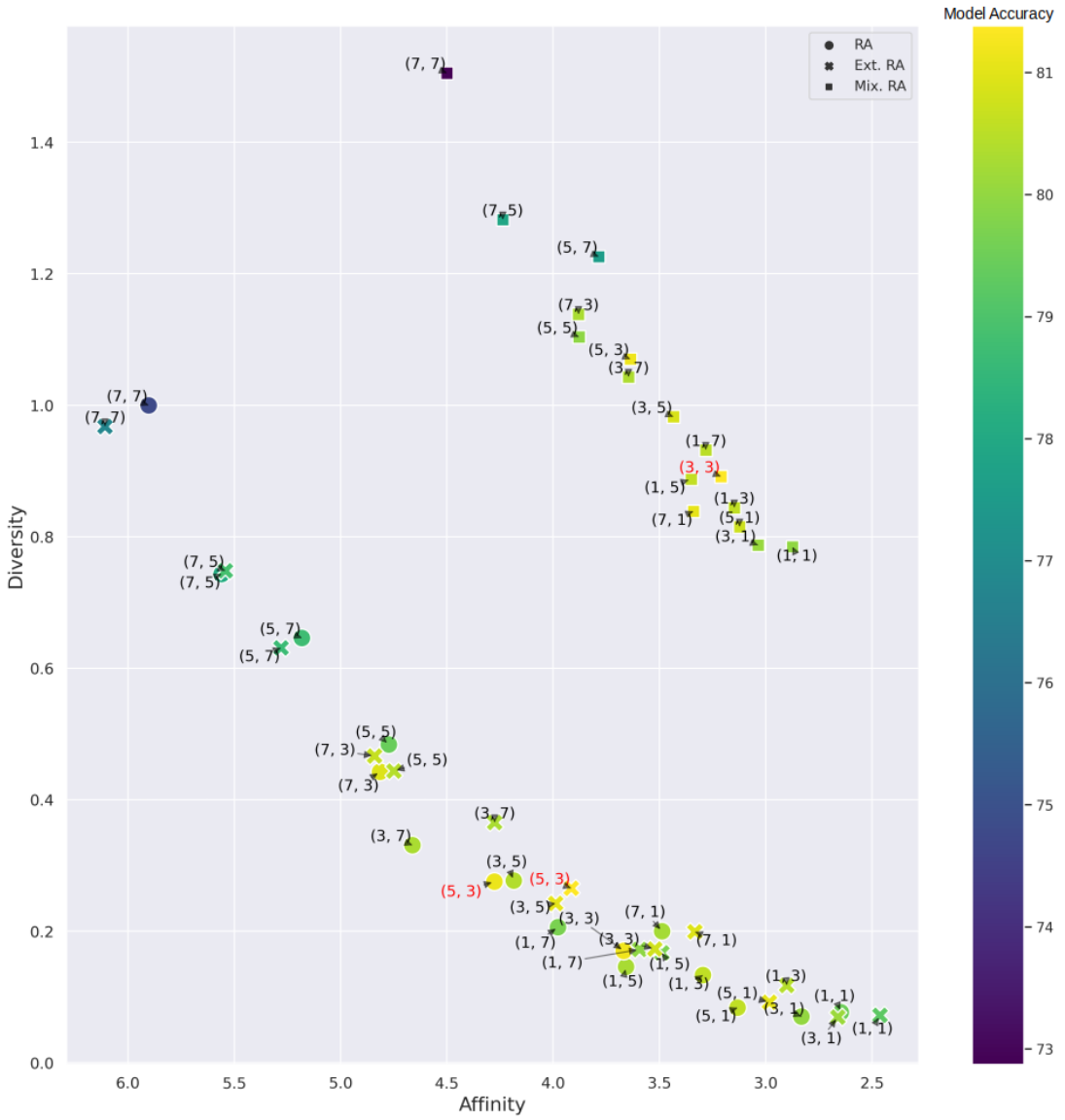


Figure 3.10: The effects of m and n on augmented image diversity and affinity Each point represents a (m, n) point in the hyperparameter search space. The best performing m and n hyperparameter for each augmentation policy search method is highlighted in red. Colours indicate final overall test accuracy.

performance over a common hand-crafted augmentation policy baseline, further improvement can be found by adding transformations based on image distortion and speckle noise. This is in-line with literature which finds neural network performance increases when elastic transformations [94, 141] and noise augmentation [129] are used in the biomedical imaging space. Generation of training images using mixed-example data generation further increases the diversity of training images seen by the neural network and increases the network performance.

The fetal plane classification task is characterized by the difficulty in separating background plane images and standard plane images, and background plane images may be taken from portions of the video where salient anatomies are visible in the image. The neural network is therefore forced to learn what views and anatomies constitutes a valid "standard plane". We find that this is difficult for static heart images even with optimal augmentation strategies, due to the similarity of the appearance of different heart planes. The classification of heart planes in general improved over hand-crafted augmentation engineering, although the accuracy of the LVOT plane decreased due to mis-classification as the 3VT plane. Other anatomical planes showed increases in performance, largely from correct classification of background planes. This has been qualitatively shown also in t-SNE visualizations which allow the high-dimensional feature space in the penultimate layer of trained models to be visualized in a 2-D chart.

We also visualize the attribution of images which were mis-classified with a model trained using RandAugment, and compare it to the attribution of images which were correctly classified with a model trained using our augmentation search policy which includes mixed-example augmentation. Integrated gradients was used to assign attribution to the input images. We found that the Mix. RA model correctly assigned attribution to salient anatomical structures such as the stomach bubble for the abdominal plane and the spinal column for the spinal coronal plane. The author argues that this is due to the increased elastic deformations that the model sees using Mix. RA, which forces greater reliance on anatomical layout.

We therefore find that i) unrealistic image augmentation with optimized affinity and diversities for a given dataset and model is able to improve fetal ultrasound standard plane classification despite the smaller datasets used than natural imaging and significantly different imaging modality characteristics, ii) improvements can be found by adding commonly used hand-crafted transformations in biomedical imaging to augmentation search policies.

It goes like this - a tav, a resh. A fearsome joy, a fervent wish.

— Scott Alexander's *Unsong*

4

Regression of Fetal Gestational Age using CNNs

Contents

4.1 Introduction 66
4.2 Dataset 67
4.2.1 Pre-processing 67
4.2.2 Pre-processing 68
4.2.3 Annotation 68
4.3 Supervised Regression using Neural Networks 72
4.3.1 The Classification-Based Binning Approach 73
4.3.2 Direct Deep Regression 75
4.3.3 Rank-consistent Ordinal Regression 75
4.4 Experiments and Results 76
4.4.1 SE-ResNeXt-50 76
4.4.2 Experiments 77
4.4.3 Implementation Details 78
4.4.4 Quantitative Evaluation 79
4.4.5 Qualitative Evaluation 87
4.4.6 Visualization of Anatomical Structures of Interest . . . 89
4.5 Discussion and Conclusion 93

In this chapter, we address the task of direct gestational age regression from obstetric ultrasound images using deep regression Convolutional Neural Networks (CNNs). The performance of current state-of-the-art CNN models are explored for this task, and multiple model variants for regression using CNNs where the loss

function is changed are explored. Moreover, we use image attribution to visualize the anatomical features of importance for regression at different stages of gestational development and find that this allows for insight into of CNN performance. We find that a consistent ordinal regression approach [142] allows for greater accuracy than direct regression of gestational age.

4.1 Introduction

Recent work has shown that neural networks can define a powerful way for image-based regression by letting the network directly predict regressed values. This has been demonstrated, for instance, for facial age estimation [104, 112] and pose estimation for natural images [106, 109] (see section 2.5 for comprehensive evaluation).

In this chapter, three different models of neural network regression are compared to estimate fetal gestational age directly from clinically acquired ultrasound images, i) classification and binning of the desired gestational age, ii) directly regressing the gestational age, and iii) ordinal regression. We explore the effectiveness of using different standard plane anatomies in gestational age regression, and evaluate the accuracy of each estimation. As the neural network is not provided with pixel size, the neural network is forced to use the appearance of anatomical features in each standard plane for regression. A number of feature aggregation methods are compared with the multiple plane model in terms of their final regression performance.

Integrated gradients [140] is also used to visualize the attribution of predicted gestational age, and the attributions of multiple images are registered to visualize the importance of different fetal anatomies in gestational age determination. This attribution allows us to visually confirm that the features and anatomical cues that the neural network learns can be clinically relevant.

4.2 Dataset

The training clinical data used in this work was provided from the International Fetal and Newborn Growth Consortium for the 21st Century Project (INTERGROWTH-21st) [4].

The INTERGROWTH-21st ultrasound image dataset consisted of a mixture of the three standard planes with no label of the plane acquired. Acquired standard plane scans were therefore initially classified into the three standard planes using a pre-commercial algorithm (ScanNav, Intelligent Ultrasound, United Kingdom). The performance of the classifier was then independently validated on 10% of the entire dataset by an experienced sonographer (Dr. Elizabeth Bradburn, EB), and the classifier achieved an accuracy of 99.7% on this subset. Due to the high accuracy of the preliminary classifier on the subset and the large size of the dataset making full manual classification unfeasibly time-consuming and expensive, the decision was made to use the ScanNav classified images as the initial sorting algorithm for each anatomical plane. Images which did not satisfy the scanning guidelines from the fetal anomaly scanning protocol [5] were discarded.

We investigate gestational age estimation from images of the femur plane, abdominal plane and head circumference plane. The dataset consisted of 62,664 femur plane images, 110,172 abdominal plane images and 120,972 head circumference plane images. Gestational ages are well distributed across the entire gestational age for all planes due to the acquisition protocol of the INTERGROWTH-21st study, as can be seen in figure 4.1. Gestational ages ranged from 13+0 weeks to 42+0 weeks. All images were acquired with the same ultrasound machine using the same protocol.

4.2.1 Pre-processing

Fetal images were taken from all three anatomies, as seen in Figure 4.2. As can be seen from the images, these were taken from clinical ultrasound scans and images therefore included common ultrasound imaging artefacts, such as acoustic shadowing and acoustic enhancement. These artefacts are commonly more pronounced in later trimesters as the fetus becomes larger and bones become more ossified.

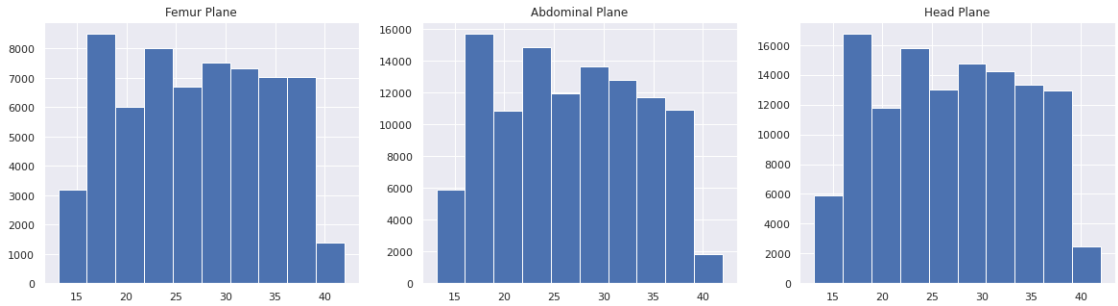


Figure 4.1: Ground truth gestational age distribution of the dataset of images from INTERGROWTH-21st. Ground truth gestational ages are approximately uniformly distributed across the entire gestational age range due to the acquisition protocol defined in the INTERGROWTH-21st study. **Left:** Femur length plane **Middle:** Abdomen circumference plane **Right:** Head circumference plane

4.2.2 Pre-processing

Fetal head images taken from the INTERGROWTH-21st study included sonographer markings, as seen in the original images shown in Figure 4.3. These included sonographer measurement markings and other extraneous text left on the image by the Phillips ultrasound machine. The locations of these artefacts were found using cross-correlation based template matching, and anomalous pixels were then smoothed with bilinear interpolation. This cleaned image size was at a resolution of 781x566 pixels. The final cleaned and resized images were then verified by a trained sonographer (Elizabeth Bradburn, EB) to be fetal head ultrasound images which still had the required clearly visible anatomical structures as defined in Figure 4.5.

Images were downsampled to 224x224 pixels using bilinear interpolation, with a subset (100 images from each plane) of the final cleaned and resized images verified by a trained sonographer (Elizabeth Bradburn, EB) to be of valid plane images. Images were then intensity normalized to have zero mean and unary standard deviation on a per individual image basis.

4.2.3 Annotation

The dataset included fetal planes of the fetal sagittal plane, abdominal circumference plane, fetal head circumference plane and fetal femur length standard planes. Fetal planes were defined using the ultrasound reference manual for pregnancy



Figure 4.2: Examples of the three planes (Abdominal, Femur, Head) used in the gestational age estimation work derived from INTERGROWTH-21st Images of each plane are taken from the first, second and third trimester in the first, second and third rows respectively. Obvious acoustic shadowing artefacts can be seen in the head and abdomen images for the second and third trimester, indicative of a common features of fetal ultrasound imaging.

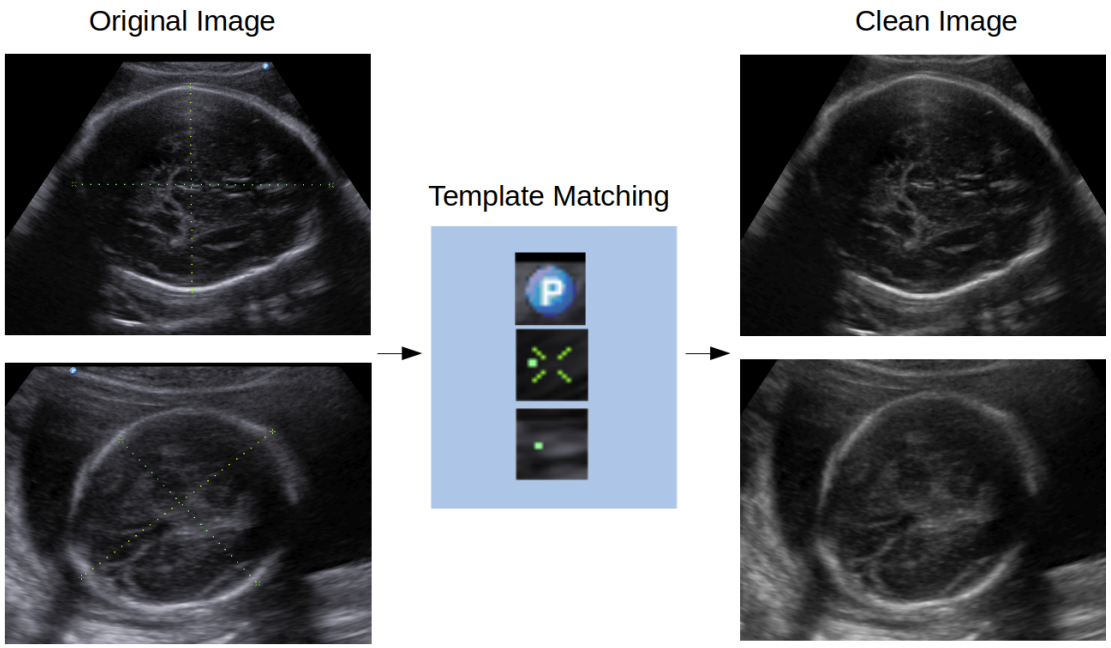


Figure 4.3: INTERGROWTH Standard Plane Image Cleaning and Pre-Processing. The standard plane images included text and sonographer markings. These markings were found with template matching and nearest-neighbour averaging. Images were then resized before input into the CNN due to memory limitations.

dating [5], and were defined from the visibility of specific fetal anatomies for each plane. In particular, the head plane is characterized by a visible cavum septum pellicidum, midline and thalami, the head should be centrally positioned, oval in shape, symmetrical, and as close as possible to horizontal, and the fetal head cross section should fill at least 30% of the visible ultrasound image. The abdominal plane is characterized by a circular presentation of the abdomen with the stomach bubble, ribs, spine and short segment of the umbilical vein visible. The abdomen should be as circular as possible, and the cross section of the abdomen should fill at least 30% of the visible ultrasound image. The femur plane visualizes the full length of the bone with no segment of the bone shadowed by adjacent ossified tissue. These anatomical requirements for the abdomen and femur can be seen in figure 4.4, and the head in figure 4.5. Additional requirements for the abdominal plane include the kidneys and the bladder not being visible, and the visible abdomen walls to be as circular as possible. Additional requirements for the femur image include the femur to be as horizontal as possible in the image.

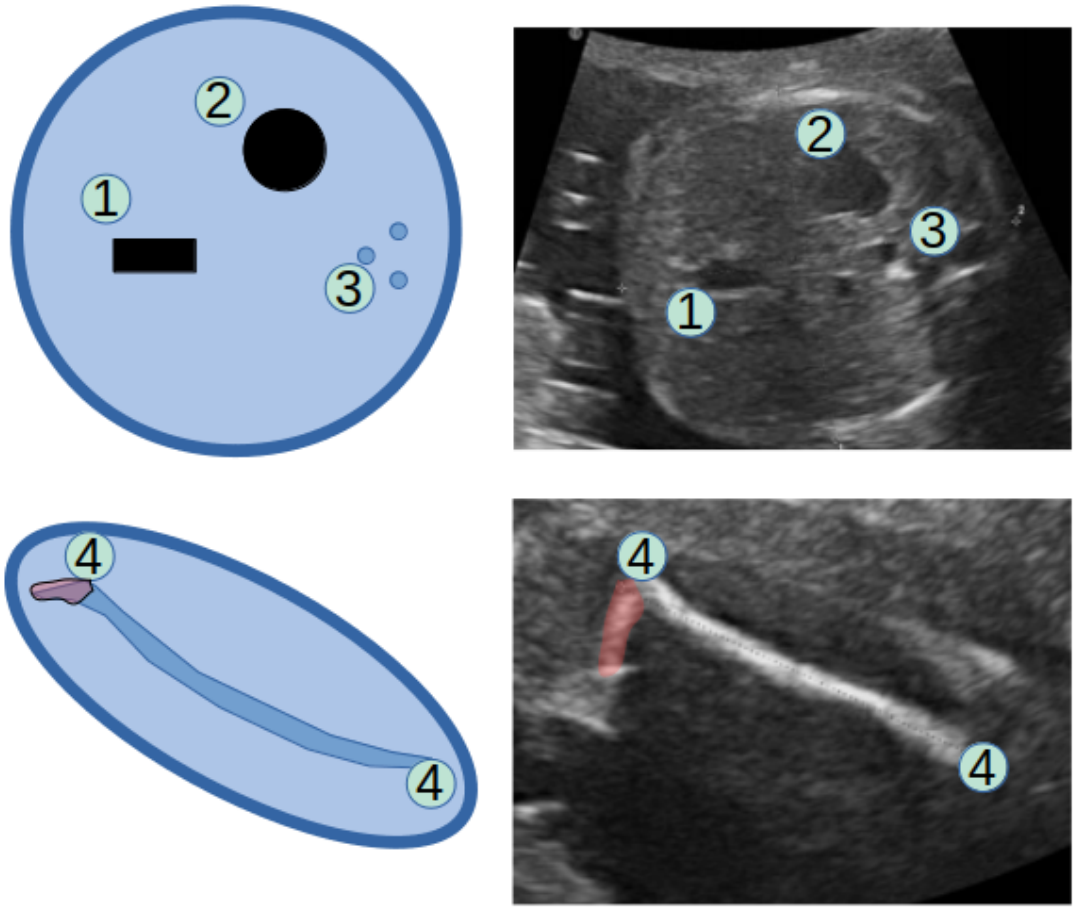


Figure 4.4: The anatomical definition of the femur and abdominal standard planes used in clinical gestational age estimation Top: The fetal anatomical structures used to define the abdominal plane include a segment of the umbilical vein (1), the stomach bubble (2) and the spine (3). **Bottom:** The femur length is measured between (4), with the greater trochanter (highlighted in red) not included in the femur measurement.

These images were classified using a pre-trained pre-commercial classification model (ScanNav, Intelligent Ultrasound, United Kingdom) into their respective planes. 10% of these images were initially manually classified into their respective anatomical classes by a trained sonographer (EB), and the performance of the ScanNav classification model was verified. An accuracy of 99.4% was achieved compared to the expert-classified images. Due to the large number of frames and high accuracy achieved by the model, the classification by the ScanNav model was used to separate the dataset into their respective classes.

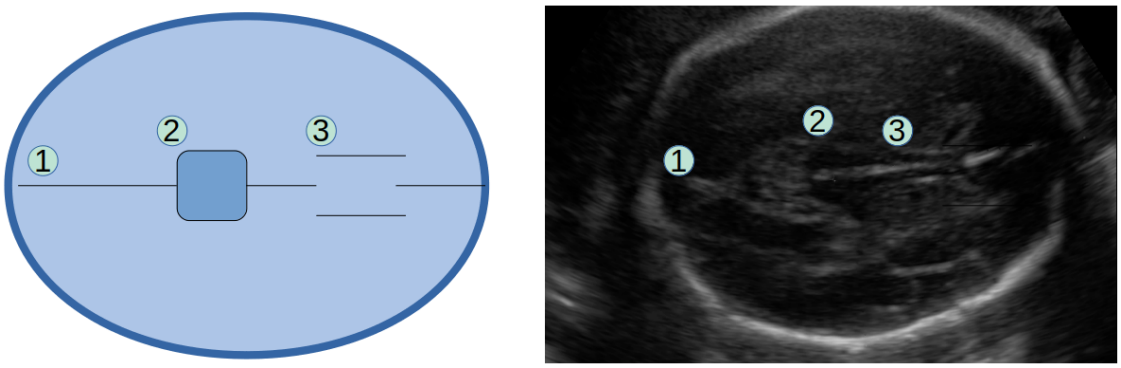


Figure 4.5: The anatomical definition of the head circumference standard planes used in clinical gestational age estimation. The appearance of each standard plane was defined by a central and continuous midline (1), with a symmetrical thalami around the midline (2) and the cavum septi pellucidi along 1/3 of the midline. The fetal head should also take up at least 30% of the image. This example fetal head standard plane has a gestational age of 26.3 weeks.

In-line with the rest of the INTERGROWTH-21st dataset, ground truth gestational age was dated in weeks from the last menstrual period (LMP), and validated with the initial crown rump length (CRL) measurement at the first scan. It should be noted that due to the acquisition protocol, ground truth gestational ages are spread uniformly across the entire gestational age range, which is a key strength in our work compared to previous literature.

4.3 Supervised Regression using Neural Networks

Neural networks have been shown to for effective in regression tasks in biomedical imaging. In this section, we explore the approaches that previous work have taken for regression tasks in biomedical imaging, where regression tasks are defined as the prediction of a variable y where y is an unknown continuous variable.

Xie and Chen [115, 143] both use a convolutional neural network for regressive tasks; Xie to count the number of cells on a microscopy slide and Chen to detect the occurrence of cell mitosis on a slide. However, both sidestep the regression problem by formulating the task as a segmentation task by quantizing the regression task and predicting the occurrences of the cells instead from the heatmap regression. This approach therefore requires post-processing after prediction by counting the

number of "hotspots" on the predicted heatmap. On the other hand, Gao [144] uses a convolutional neural network with a least squares loss to predict the occurrences of interstitial lung disease in computerised tomography (CT) scans. Stern [145] predicts biological age from MRI volumes of the hand by concatenating features extracted from multiple pre-defined locations of the cropped 3D bone volume. Namburete [126] use a three-dimensional convolutional regression network in order to predict a gestational age from 3D ultrasound scans of fetal heads.

We therefore explore the accuracy of different regression approaches for fetal gestational age estimation using deep convolutional neural networks with different loss functions. We explore the degree to which gestational age estimation can be performed on fetal images with no pixel size information and therefore no information on the absolute size of the imaged fetal anatomies. We also explore the degree to which the neural network learns anatomically relevant priors for gestational age estimation by visualizing the attribution that the network gives to different regions of an input image.

4.3.1 The Classification-Based Binning Approach

As seen in prior work [146], a potential approach to regression is to perform quantization of the final required variable and the conversion of the task into a classification task. More formally, given a dataset $D = (X, Y)$ with inputs $X = (x_i)$ and ground truths $Y = (Y_i)$ where Y is a continuous variable, a new dataset D_c is constructed instead where $D_c = (X, Y_c)$ and $Y_c \in (Y_1, Y_2, Y_3, Y_4, ..., Y_n)$, which is also known as feature discretization. The model can then be trained with a standard cross-entropy classification based loss for instance.

There therefore is an auxiliary task in the conversion of Y into Y_c. Common approaches for binning can be seen in table 4.1 [147].

This approach has been used in both the natural image space [146, 148] and in audio generation [149]. All three works use the equal-width binning approach to regression problem, and find that the performance of a model on a classification-based discrete cross entropy loss is improved over that of a regression based loss.

Table 4.1: Table of common approaches for bin selection for regression via classification

Binning Method	Description
Element-wise Binning	The width of each bin is selected in order to have an equal number of elements within each bin. This is equivalent to having an equal expectation of the final result in each bin during classification
Equal-width Binning	The width of each bin is selected based on equal widths of each bin.
Clustered Binning	Bins are selected based on pre-computed clustering (e.g. K-means), with a bin selected for each cluster.

Furthermore, a classification prediction can be arbitrarily multi-modal compared to a single-valued regression loss in conventional neural networks.

However, the process of binning and quantization leads to loss of information in the prediction. The width of the bins necessarily discards some of the information from the original continuous variable, with wider bins discarding more information. There therefore is an additional complexity-accuracy trade off that is introduced by the binning approach. Furthermore, by using a classification-based cross-entropy loss during neural network training, the loss from mis-predictions of adjacent bins and bins far away from the target value are the same, which is misrepresentative of continuous value regression.

Prior knowledge is therefore required in picking the bins. Bins are often picked based on existing quantization inherent in the continuous variable. In the case of generation of natural images, the quantization inherent in pixel values (0 - 255, [146]) leads to an easy decision of binning criteria. This can analogously be extended to fetal gestational age estimation. Due to the inherent uncertainty in gestational age derived from fetal growth charts [6] and the uniform distribution of gestational ages from the training dataset, we argue equal-width binning intervals of one week in width can be used during discretization of the desired gestational age for training.

4.3.2 Direct Deep Regression

Neural networks can also be used to directly regress a desired continuous variable. Oord [146] finds that convolutional neural networks trained to predict continuous variables perform worse than neural networks trained with cross-entropy loss. However, continuous information is not lost compared to the binning process, and this approach has therefore used in facial landmark detection [150] and human pose estimation problems [106]. Common loss functions used during prediction of continuous variables include the L_1 loss,

$$\mathcal{L}_1(\mathbf{r}) = \frac{1}{n}\sum_{i=1}^{n}|(\hat{y}_i - y_i)| \tag{4.1}$$

Huber loss [151],

$$\mathcal{L}_{hub}(\mathbf{r}) = \frac{1}{n}\sum_{i=1}^{n} k^2\sqrt{1 + \frac{(\hat{y}_i - y_i)^2}{k}} - 1 \tag{4.2}$$

and L_2 loss,

$$\mathcal{L}_2(\mathbf{r}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{4.3}$$

where y_i and $\hat{y}_i$ represent a sample of the ground truth and the prediction respectively, n is the index of the data, and k is the degree to which the Huber loss is sensitive to outlier predictions. Lathuiliere [108] finds that the optimal loss function is heavily dataset and model dependent, and that there is no best loss function that can be used for every regression task. We therefore investigate the performance of direct regression neural network models trained with all three loss functions.

4.3.3 Rank-consistent Ordinal Regression

Ordinal regression neural networks [142] attempt to predict ordinal labels. Ordinal labels are similar to binned classification, except the neural network is provided with information about the relative ranking for each bin. Given a dataset $D = (X, Y)$ with inputs $X = \{x_i\}$ and ground truths $Y = \{y_i\}$ where Y is a continuous

variable, a proxy target $Y_{ordinal} = \{y^i_{ordinal}\}$ is generated for inference with $y_k = (y_1 < y_2 < y_3 < ... < y_n)$ ordinal bins:

$$y^i_{ordinal} = \begin{cases} 1, & \text{if } y_i \geq y_k \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

The neural network can then output $n-1$ predictions of whether the predicted value is greater than each ordinal bin. The loss function is then

$$\begin{aligned} L(W,b) = -\sum_{i=1}^{N}\sum_{k=1}^{K-1} \Big[& \log\left(\sigma\left(g\left(x_i, W\right) + b_k\right)\right) y^i_n \\ & + \log\left(1 - \sigma\left(g\left(x_i, W\right) + b_k\right)\right) \left(1 - y^i_n\right)\Big] \end{aligned} \tag{4.5}$$

where $(g(x_i, W)$ are the inputs to the classifiers in the final layer for ordinal classification, σ is the logistic sigmoid function. This function can be interpreted as the weighted cross-entropy for $n-1$ binary classifiers used for ordinal regression, and encodes the ordinal information for each bin during the training of the neural network.

Some of the same drawbacks can be seen with this method as with classificiation-based binning methods. Bin widths and spacing need to be decided with prior knowledge and might lead to biases during prediction. However, the ordinal approach means that the bin proximity information is retained during training, which improves results in natural image regression [142].

4.4 Experiments and Results

4.4.1 SE-ResNeXt-50

The base structure of the SE-ResNeXt model can be see in figure 4.6. The SE-ResNeXt model therefore consists of each input feature travelling through multiple individual paths, which are subsequently summed together with the residual from the original input feature, similar to the ResNeXt model described in 2.4.6. The number of individual paths is controlled by an external hyperparameter C, also known as the cardinality of the ResNeXt model, and the dimension of the internal convolutional

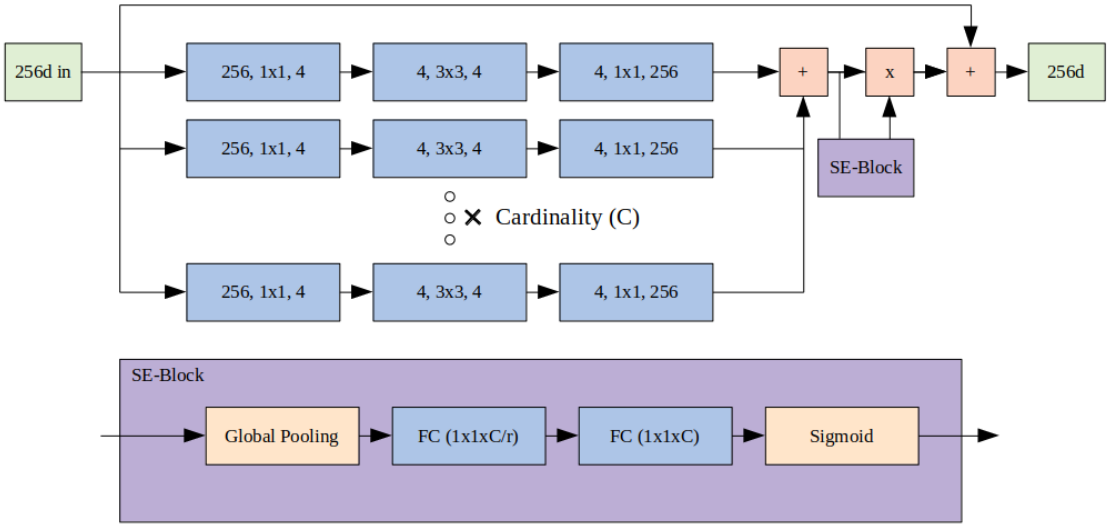


Figure 4.6: The base block of the SE-ResNeXt model [152] The figure shows the base building block of the SE-ResNeXt model. Each blue node represents a convolutional operation with parameters represented by (input channel number, kernel size, output channel number). The red nodes represent the summation operator, and the green nodes represent a 256-dimensional input feature and output feature respectively. The SE-Block includes a dimensionality reduction ratio r, which is set to 16 by default [152].

kernel d. A ResNeXt-50 model consists of an initial feature aggregating convolutional layer with 64 channels, a 7x7 kernel and stride 2, 48 of the ResNeXt blocks in figure 4.6, and an output layer canonically composed of a fully connected layer. The SE-ResNeXt-50 model [152] is a squeeze and excitation modified ResNeXt, which applies a channel-wise attention layer on each ResNeXt block which can be seen in the purple block in the figure. We choose this model due to its proven performance in the natural image space, and initialize the model with a Cardinality of 32, an internal dimension d of 4, and squeeze-excitation dimension reduction r of 16.

4.4.2 Experiments

Experiments were designed to train a SE-ResNeXt-50 with random initialization for regression of gestational age for three fetal standard planes, the head standard plane (HCP), abdominal standard plane (ACP) and the femur standard plane (FLP). We use the same architecture for all losses in order to limit the performance difference to solely the loss function used. We experiment with the different loss functions for

regression i) classification-based binned regression, ii) direct regression to desired gestational age and iii) rank-consistent ordinal regression for gestational age, and find the method that leads to the most consistently accurate results. For a Huber based regression loss (L_{hub}) we use $k = 1$ to approximate an in-between loss to outliers.

With regression-based losses, the final fully-connected layer in the SE-ResNeXt-50 is replaced with a tanh-activated single channel which represents the final predicted gestational age. With models trained with classification losses for binned regression, the final fully-connected layer is modified to include an output for each bin, with a softmax activation across the entire fully-connected layer.

We then also visualize the attribution that the image gives to each image, and the saliency information that can be achieved by inspecting the importance that the network gives each pixel in the input image for gestational age prediction using integrated gradients.

4.4.3 Implementation Details

All models were trained with the Adam optimizer [84, 153] with a batch size of 80 images[1]. A learning rate of 1×10^{-3} for all trainable weight parameters was used with AMSGrad weight decay. The learning rate was then decayed by 0.1 every 50 epochs until training termination. All weights in the SE-ResNeXt-50 network were randomly initialized. Validation accuracy was evaluated every epoch, and networks were trained until there were 10 consecutive epochs with no validation accuracy improvement. Networks were trained with Ext. RandAugment based data augmentation (see chapter 3) with hyperparameters $n = 3$ and $m = 1$. The hyperparameters for the Ext. RandAugment implementation was chosen based on grid search of $m \in (1, 2, 3, 4, 5), n \in (1, 2, 3, 4, 5)$ which was found to maximize network performance on the validation set. Mix. Aug was not used, as the ground truth for mixed-example data augmentation was ill-posed for regression problems. Models were trained on a Nvidia 1080Ti.

[1] a larger bach size was not possible due to computer GPU memory constraints

We split the original INTERGROWTH-21st dataset into a train, validation and test set to evaluate each model. The dataset was split 75/15/10 for the train, validation and test sets respectively on a patient basis to reduce data leakage between separate datasets. This led to (90,535, 82,454, 46,984) images in the training set, (18,165, 16,579, 9,428) images in the validation set, and (12,272, 11,139, 6,252) in the test set for the head, abdomen and femur planes respectively.

Networks trained with direct regression loss had a singleton final-fully connected layer at output network for continuous value prediction. Networks trained for binned-classification based regression and ordinal regression had bin widths corresponding to week of gestational age, as that is higher resolution than the maximum accuracy that can be achieved with clinical based gestational age regression methods in the first trimester (standard deviation of an estimated GA of 8 weeks from CRL based biometry is 2.6 weeks) [6]. Increasing bin widths would have led to unsatisfactory resolution in gestational age estimation. Ground truth gestational ages were normalized to between -1 and 1 for training neural networks with regressive losses, and were one-hot encoded for neural networks with classification losses.

4.4.4 Quantitative Evaluation

Table 4.2 shows the performance of each model on the hold-out test set, evaluated by mean squared error (MSE), mean absolute error (MAE) and explained variance. MSE and MAE are defined by $\mathcal{L}_2$ and $\mathcal{L}_1$ losses in section 4.3.2 respectively. Explained variance is defined by the equation:

$$\text{Explained Variance} = 1 - \frac{Var\{\hat{y} - y\}}{Var\{y\}} \tag{4.6}$$

where y is the ground truth, $\hat{y}$ is the prediction and Var is the variance.

Networks trained with L_2, L_{hub} and L_1 loss represent models trained with losses from equation 4.1, 4.2 and 4.3 respectively. We find that unlike in natural imaging and audio generation [149], models trained with regression loss consistently outperform models that were trained with binned-regression classification loss, with the best performing regression model outperforming the classification model by

Table 4.2: Results for gestational age regression on individual standard planes MSE/MAE/Explained Variance metrics are calculated based on maximum likelihood with gestational age estimation for models trained with a classification loss. Models in **bold** indicate the best performing model on all three metrics for each anatomy. The regression model highlighted in grey represents the best performing regression-based loss when the model is trained with for vanilla deep regression. Arrows indicate direction of better performance for each metric.

Model Training	MSE (Weeks^2, ↓)	MAE (Weeks, ↓)	Explained Variance (↑)
HCP SE-ResNeXt-50 Classification Loss	1.413	0.763	0.973
HCP SE-ResNeXt-50 Regression Loss (L_2)	1.004	0.722	0.981
HCP SE-ResNeXt-50 Regression Loss (L_{Hub})	0.927	0.698	0.982
HCP SE-ResNeXt-50 Regression Loss (L_1)	0.832	0.656	**0.984**
HCP SE-ResNeXt50 Ordinal Regression Loss	**0.823**	**0.615**	**0.984**
FLP SE-ResNeXt-50 Classification Loss	2.330	1.058	0.957
FLP SE-ResNeXt-50 Regression Loss (L_2)	1.701	0.984	0.970
FLP SE-ResNeXt-50 Regression Loss (L_{Hub})	1.630	0.966	0.969
FLP SE-ResNeXt-50 Regression Loss (L_1)	1.798	0.988	0.966
FLP SE-ResNeXt50 Ordinal Regression Loss	**1.371**	**0.831**	**0.974**
ACP SE-ResNeXt-50 Classification Loss	2.445	1.104	0.952
ACP SE-ResNeXt-50 Regression Loss (L_2)	1.802	1.023	0.965
ACP SE-ResNeXt-50 Regression Loss (L_{Hub})	2.053	1.045	0.960
ACP SE-ResNeXt-50 Regression Loss (L_1)	2.112	1.098	0.959
ACP SE-ResNeXt50 Ordinal Regression Loss	**1.279**	**0.806**	**0.975**

Table 4.3: The p values calculated using Welch's unequal variance t-test [154] to determine the probability that the mean absolute error samples are drawn from different populations. HCP, FLP and ACP represent the head circumference plane, femur length plane and abdominal circumference plane respectively. **Top:** Regression models trained on the HCP. **Middle:** Regression models trained on the FLP. **Bottom:** Regression models trained on the ACP

	HCP $\mathcal{L}_2$	HCP $\mathcal{L}_1$	HCP $\mathcal{L}_{Hub}$
HCP $\mathcal{L}_2$	1	-	-
HCP $\mathcal{L}_1$	1.24e-14	1	-
HCP $\mathcal{L}_{Hub}$	6.30e-03	4.42e-07	1
	FLP $\mathcal{L}_2$	FLP $\mathcal{L}_1$	FLP $\mathcal{L}_{Hub}$
FLP $\mathcal{L}_2$	1	-	-
FLP $\mathcal{L}_1$	7.94e-01	1	-
FLP $\mathcal{L}_{Hub}$	2.55e-01	1.72e-01	1
	ACP $\mathcal{L}_2$	ACP $\mathcal{L}_1$	ACP $\mathcal{L}_{Hub}$
ACP $\mathcal{L}_2$	1	-	-
ACP $\mathcal{L}_1$	6.65e-10	1	-
ACP $\mathcal{L}_{Hub}$	1.09e-01	9.95e-15	1

(0.11, 0.09, 0.08) weeks for the head, femur and abdomen planes on a mean absolute error basis.

We find that the optimal regression loss varies for each anatomical plane. Specifically, the neural network performed the best on the test set when trained with $\mathcal{L}_1$ loss with the head plane, $\mathcal{L}_2$ loss for the abdominal plane and $\mathcal{L}_{Hub}$ for the femur plane. We further evaluate the statistic significance of this finding by using Welch's unequal variance t-test [154] on the distribution of the mean absolute errors for each regression model loss function, seen in table 4.3. We use Welch's unequal variance t-tests to test the hypothesis that different regression loss functions $(\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_{Hub})$ lead to different mean absolute errors. The difference in regression performance is statistically significant for the fetal head plane ($p_{max} < 6.3 \times 10^{-3}$), but less so for the abdomen plane ($p_{max} < 1.1 \times 10^{-1}$) and the femur plane ($p_{max} < 7.9 \times 10^{-1}$). This suggests that the optimal loss function for a deep vanilla regression model has to be found by trial and error.

Overall, we find that gestational age regressed from the head circumference plane consistently outperforms gestational ages regressed from the abdominal

Table 4.4: Predictive error dependence on gestational age Results below show the mean absolute error in weeks for different sets of images for the best performing neural network on each anatomical plane. Percentage increase numbers are relative to the gestational age range column on the immediate left.

	<23 Weeks	>23 Weeks and <31 Weeks	>31 weeks
HCP	0.539	0.676 (+25.4%)	0.950 (+40.1%)
FLP	0.772	0.961 (+24.5%)	1.211 (+26.0%)
ACP	0.821	1.031 (+25.6%)	1.224 (+18.7%)

circumference plane and the femur length plane, and gestational ages regressed from the femur plane outperform that regressed from the abdominal plane.

We also find that consistent with expectations, ordinal regression leads to the lowest error in regression accuracy for all three planes. Models trained with ordinal regression achieved a mean absolute error of (0.615, 0.831, 0.806) weeks for the HCP, FLP and ACP respectively, compared to (0.656, 0.966, 1.023) weeks for models trained with regression loss and (0.763, 1.058, 1.104) weeks for models trained with classification loss. This may be due to ordinal regression models combining the ease of classification-based training for neural network models (discussed in section 2.5), whilst providing ordinal and therefore bin ordering information during training.

We also plot the predicted gestational ages against ground truth gestational ages for networks trained with regression loss and networks trained with a regression L_1 loss respectively in figure 4.7. We find that the predictions in aggregate follow the line of equality for all three standard planes with any training loss, with points closely following the optimal red line where the predicted gestational age is equal to ground truth gestational age.

The lack of bias in prediction is visualized in the Bland-Altman [155] plots in Fig. 4.8. It can be seen that the the distribution of differences is equally distributed across the mean for all three training methods and standard planes. However, the quantization of the continuos gestational age variable is obvious in neural networks trained with classification loss, which leads to loss in predictive resolution and increased error when the maximum predicted bin is used as the predicted gestational age.

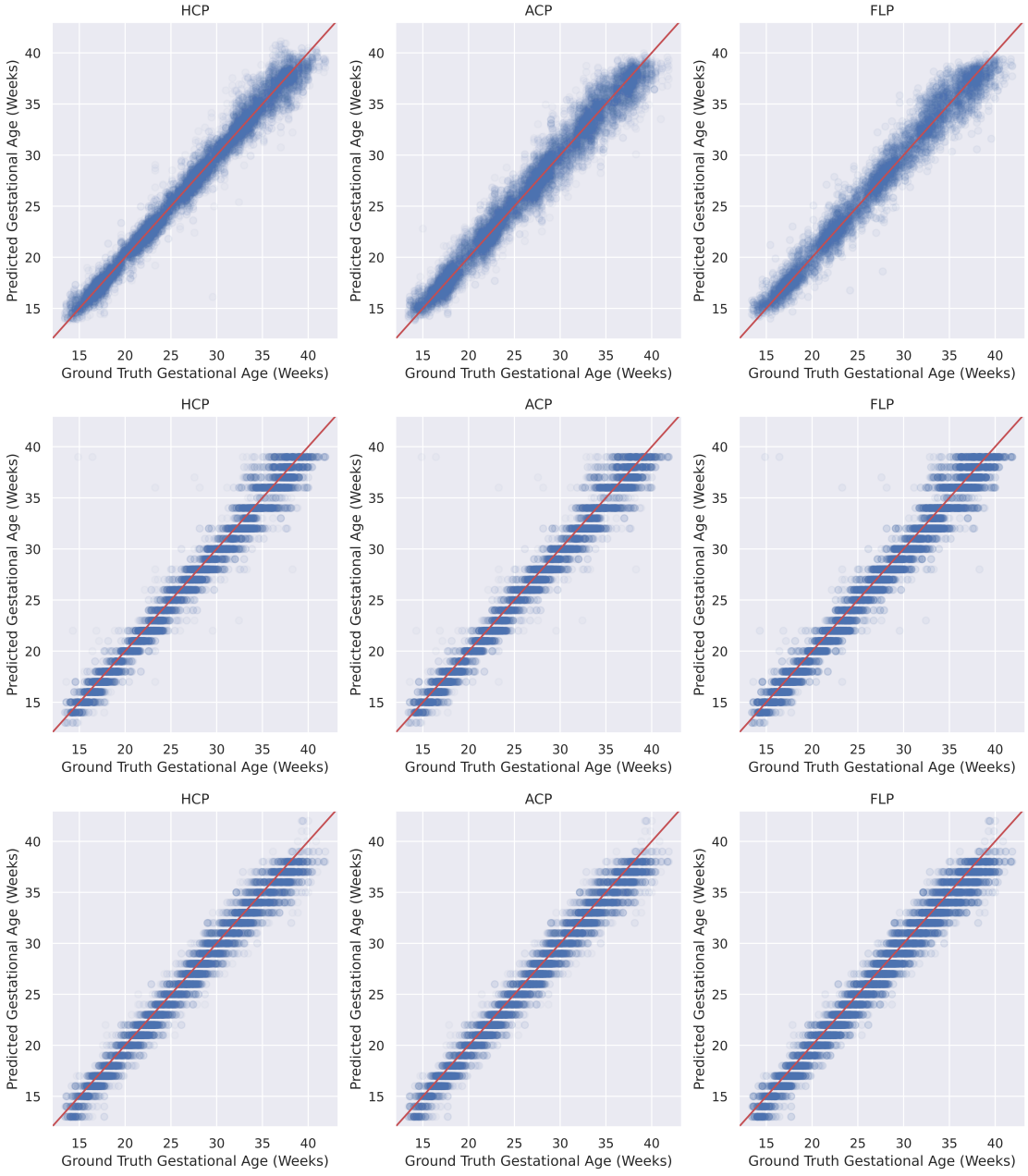


Figure 4.7: Predictions against ground truth gestational age for the best performing models trained with regression, classification and ordinal loss Top row: Results of the best models trained with regression loss. **Middle row:** Results of the best models trained with classification loss. **Bottom row:** Results of the best models trained with ordinal loss. **Left column:** Head circumference plane. **Middle column:** Abdominal circumference plane. **Right column:** Femur length plane. The red line shows the equality line where the predicted gestational age is equal to ground truth gestational age.

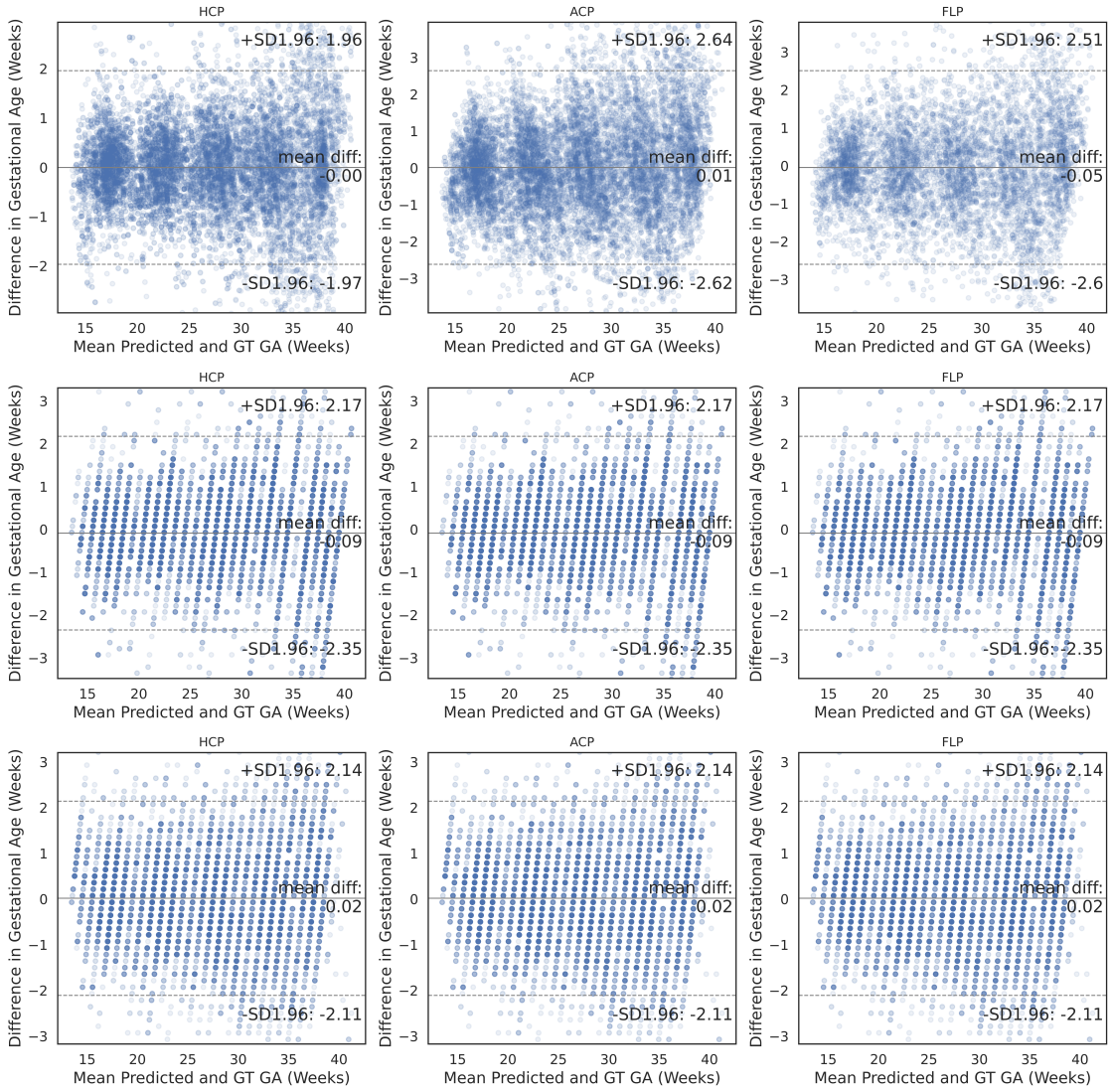


Figure 4.8: Bland-Altman plots [155] to compare the distribution of the predicted gestational ages and ground truth gestational ages. Each chart shows the Bland-Altman plots for the predicted gestational age and the ground truth gestational age **Top:** Models trained with regression loss. **Middle:** Models trained with classification loss. **Bottom:** Models trained with ordinal loss. HCP (left), ACP (middle) and FLP (right) represent the head circumference plane, femur length plane and abdominal circumference plane respectively.

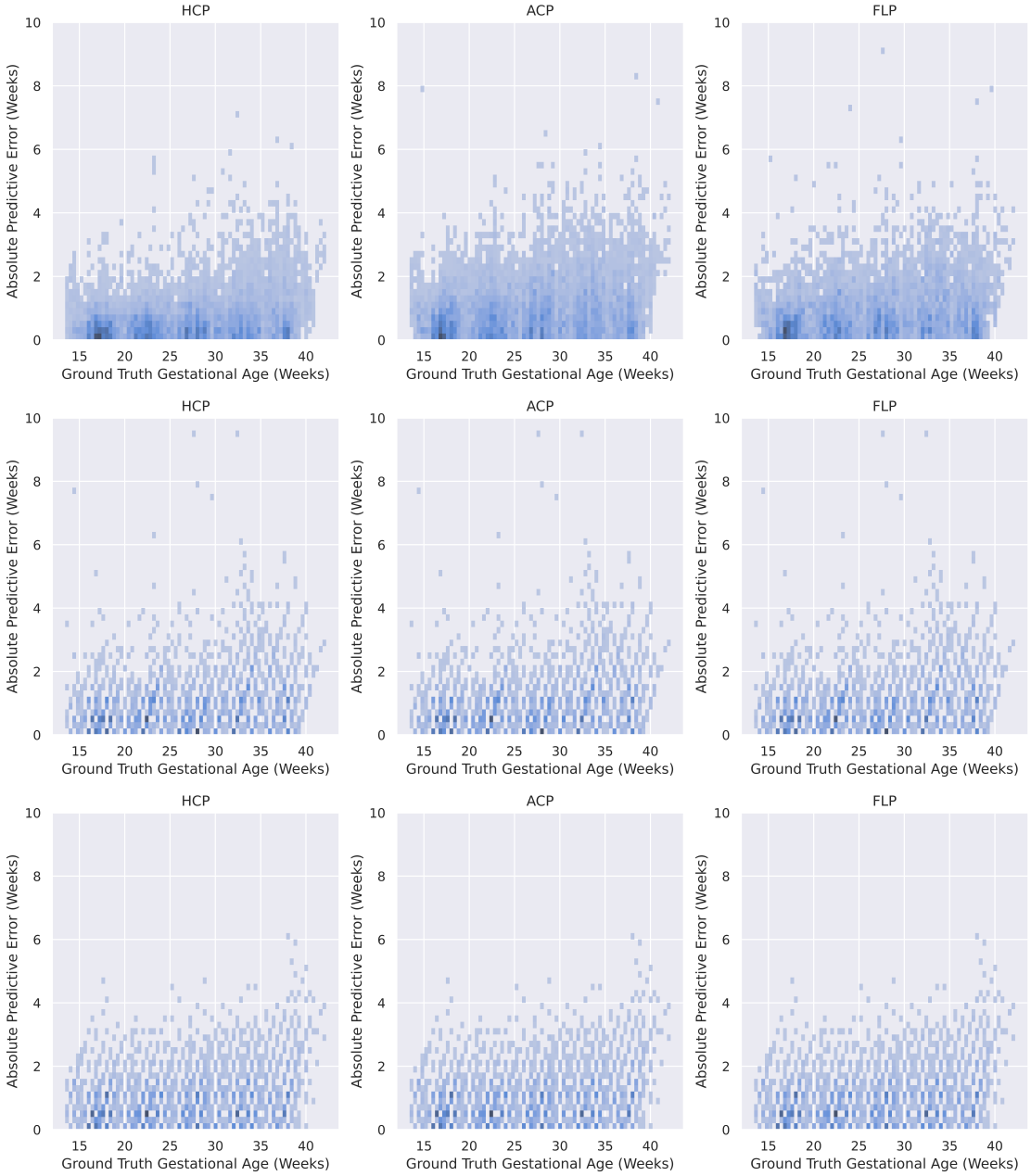


Figure 4.9: Absolute predictive error for trained models against ground truth gestational age Each chart shows the distribution of absolute error against ground truth gestational age. **Top:** Error distribution of models trained with regression loss. **Middle:** Error distribution of models trained with classification loss. **Bottom:** Error distribution of models trained with ordinal loss. HCP (left), ACP (middle) and FLP (right) represent the head circumference plane, femur length plane and abdominal circumference plane respectively.

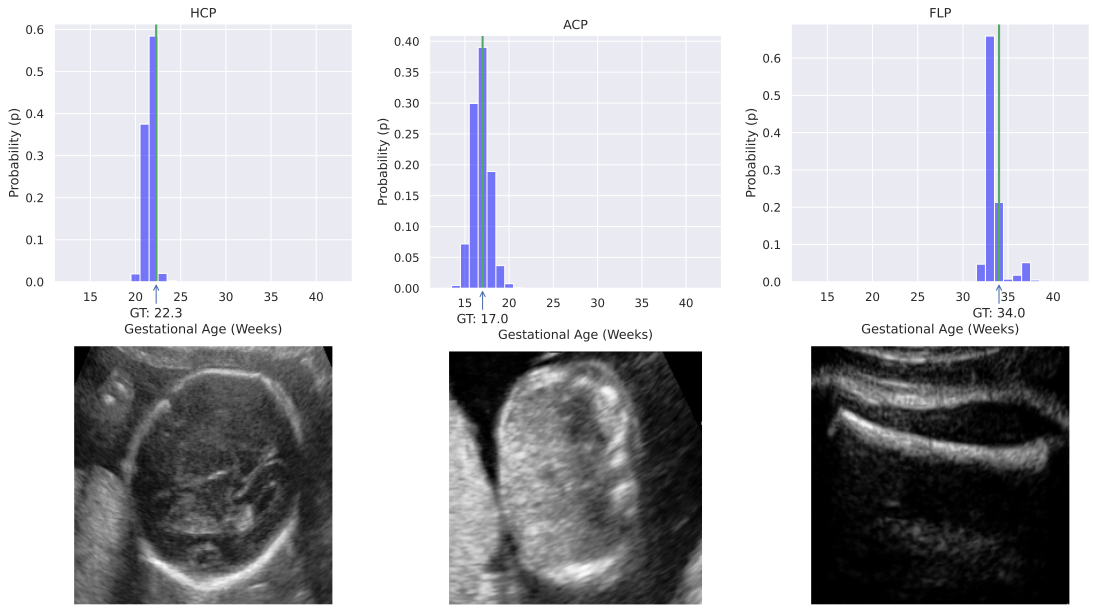


Figure 4.10: Successful examples of gestational age prediction for each fetal standard plane. Top: The output of neural networks trained to predict gestational age. The bar chart shows the distribution of the final predictions of a model trained with a classification loss. The green line is the prediction of the model trained with the best performing regression loss. The ground truth is indicated by the blue arrow on the x-axis. **Bottom:** Images of the head circumference plane (HCP), abdominal plane (ACP) and femur length plant (FLP) respectively used to generate the regression results.

Furthermore, it appears that predictive error increases with increased gestational age. This can more clearly be seen in Fig. 4.9, which shows the variation in absolute predictive error against gestational age. We find that there is a positive correlation with predictive error as gestational age increase, see from the increased absolute predictive error with increased ground truth gestational age. This is further highlighted in Table 4.4, where we find that increasing the gestational age band increases the predictive error in general for all three planes. Note that, the predictive error for the femur head increases the most for gestational ages greater than 31 weeks. This may be due reduced anatomically salient information from the femur long bone in late pregnancy. We can also see from Fig. 4.9 that some of the performance of the model trained with ordinal regression can be explained by the reduction of outlier prediction errors.

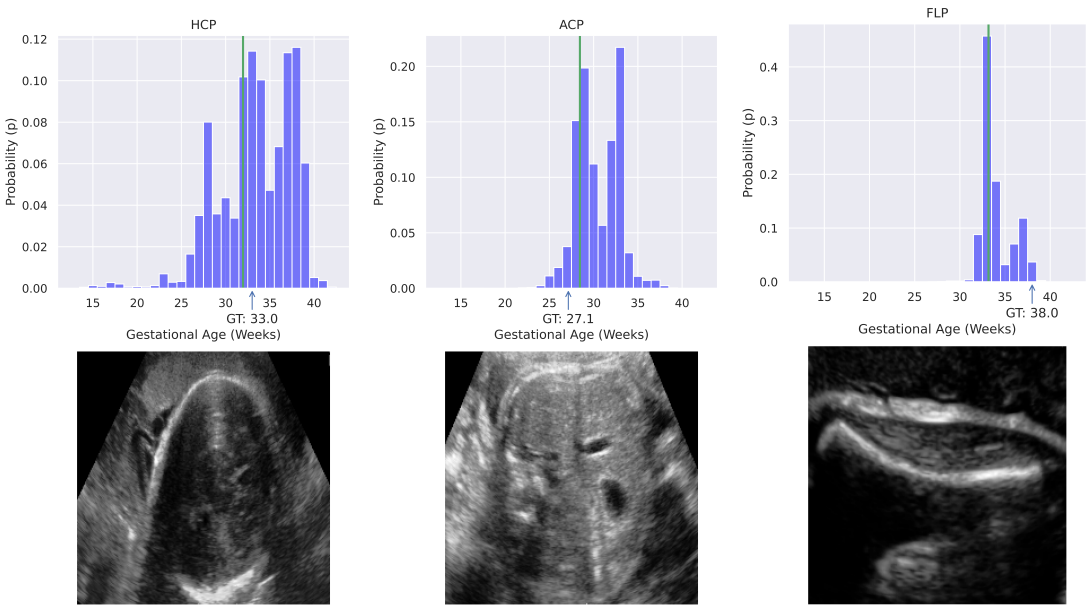


Figure 4.11: Multi-modal results from classification model gestational age estimation Top: The output of neural networks trained to predict gestational age. The bar chart shows the distribution of the final predictions of a model trained with a classification loss. The green line is the prediction of the model trained with a regressive L_1 loss. The red line is the ground truth. **Bottom:** Images of the head circumference plane (HCP), abdominal plane (ACP) and femur length plant (FLP) respectively.

4.4.5 Qualitative Evaluation

Evaluation of successful gestational age prediction (success defined as gestational age estimation < 1 week from ground truth gestational age with either networks trained with regressive and classification losses) show a successful prediction by a network trained with a regressive loss is likely to be successfully predicted by a network trained with classification loss ($p = 0.96$). Figure 4.10 shows examples of successfully predicted gestational age for the three different anatomical planes for models trained with regression and classification loss. In general, results with high accuracy show strongly clustered classification probabilities despite a networks not knowing the ordinal nature of the bins.

However, due to the more expressive nature of classification models where predictions are arbitrarily multi-modal (see section 4.3.1) the predictions from the neural network trained with a classification loss can be used to qualitatively evaluate images which have high gestational age estimation errors.

One failure mode can be qualitatively evaluated from the results of the classification model, seen in figure 4.11. The figure shows the final predicted distribution that the gestational age has with a SE-ResNeXt-50 model trained with a cross-entropy classification loss with bin widths of one week. We find that the classification model is better able to explain the error in prediction of the model compared to the ground truth. Predictions in figure 4.11 appear to be multi-modal, which suggests that the neural network is uncertain of which gestational age to predict. This can be seen in the output of the classification model, but this nuance cannot be shown in a model trained with a regressive loss despite networks trained with regressive loss being more accurate overall. This can be seen most clearly in the HCP in figure 4.11, where there are two strong responses at the 15 week bin and the 38 week bin respectively. Assigning the predicted gestational age to the bin with the highest probability then discounts the other modes and leads to a larger prediction error. This can also be seen in the prediction of the regression model, where it appears that the single-valued regression prediction is pulled between modes, which leads to greater error. However, the multi-modal nuance cannot be seen in a regression prediction, as the model only outputs a single number.

Another source of high gestational age prediction error appeared to come from images which were erroneously included in the INTERGROWTH-21st dataset despite failing the acquisition criteria. Due to the size of the dataset, manual validation of every image by an experienced sonographer was impossible. Furthermore, as INTERGROWTH-21st is multi-site and multi-sonographer based, some images may have been included into the dataset which did not completely fulfil the imaging and anatomical criteria for a standard plane which are outlined in section 4.2.3. Examples of these images for each plane can be seen in figure 4.12. These include head images which are partially occluded either due to being out of frame or significant acoustic shadowing/echoing artefacts, or heads images which are not of the trans-ventricular plane as defined in the acquisition protocol. For the ACP, these may include abdominal planes which mistakenly include the kidneys as part of the acquisition plane, or do not include the stomach bubble or spine. For the

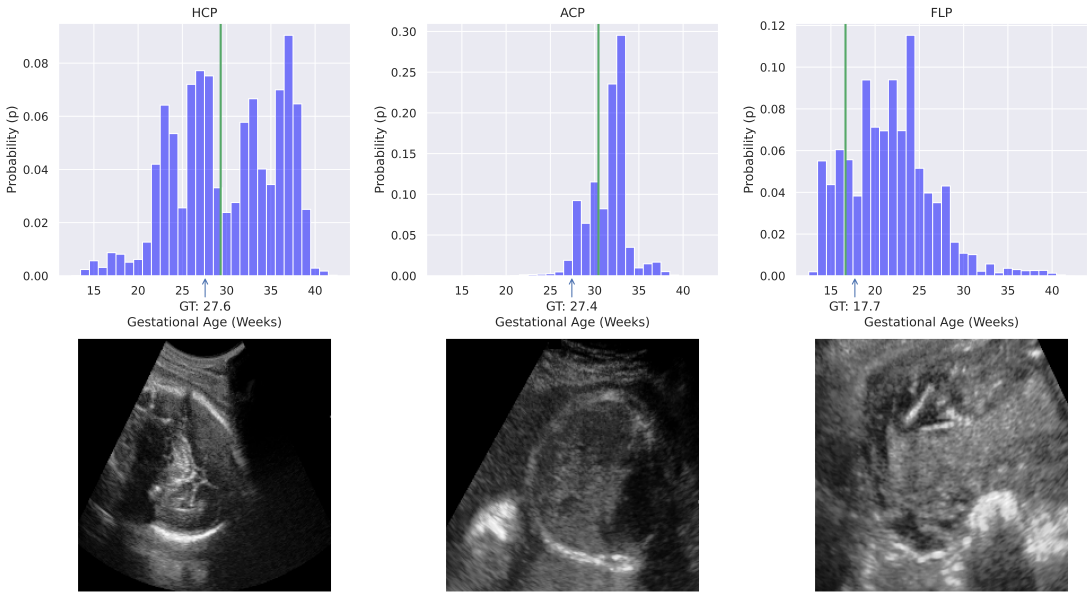


Figure 4.12: Images with high gestational age prediction error due to not adhering to acquisition protocols Left: The HCP image shows significant shadowing and the fetal head is not fully in view. **Middle:** The abdominal plane does not have a clear umbilical vein section and stomach bubble, and is not fully in view with a portion of the ribs hidden by acoustic shadowing. **Right:** The femur is not clearly visible in the image, and is not situated horizontally. The femur does not constitute 30% of the image.

FLP, these include images of the femur which are a small part of the image, or femurs that are partially occluded by either the frame or angle of acquisition. This shifts the test data out of distribution compared to the training set, which therefore increases gestational age prediction error.

4.4.6 Visualization of Anatomical Structures of Interest

In this section, we use integrated gradients [140] to visualize the anatomies that are regressed for each trimester in order to determine the anatomies of importance to the neural network in gestational age estimation. We separate each anatomy into the gestational age ranges in Table 4.4 in order to better visualize the difference in anatomical presentation with changes in gestational age.

Less than 23 weeks gestational age

The HCP plane fetal head is clearly delineated in the attribution image as can be seen in figure 4.13. Attribution is clearly given to the visible ventricles and the

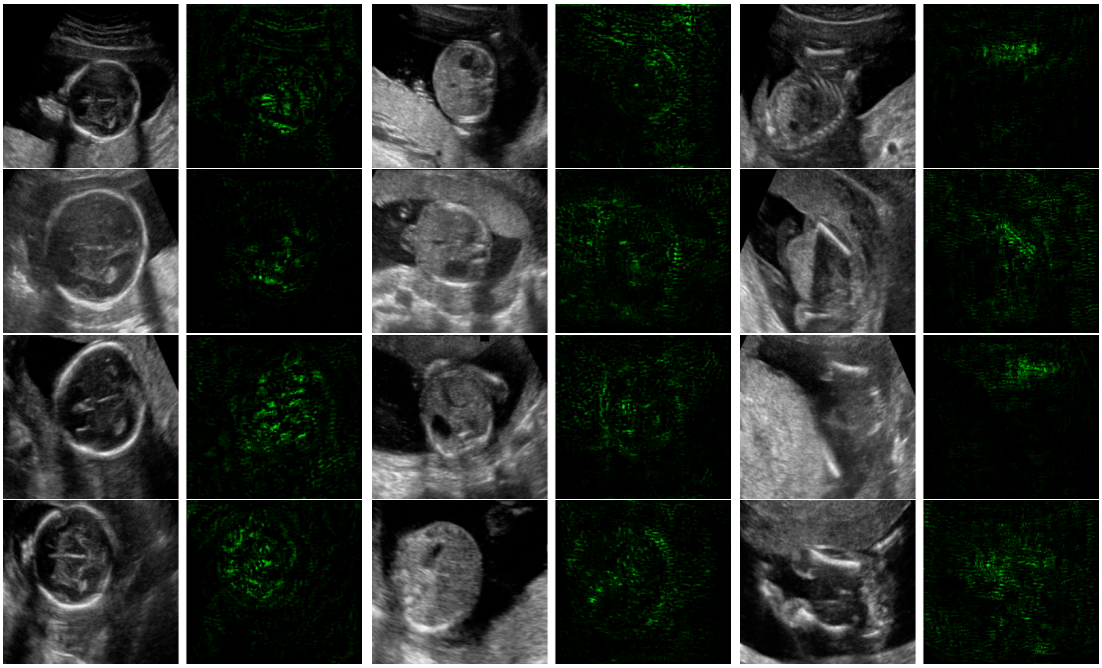


Figure 4.13: Attributions to input images of fetal planes less than 23 weeks of gestational age Images are from left to right columns the HCP, ACP and FLP planes respectively. The green indicates the degree of attribution to each individual pixel for gestational age estimation. Images are best viewed digitally.

CSP, with some residual going to the midline. This is expected, as these anatomical structures are well defined in the acquisition protocol for the head plane, and visibly change with gestational age. This is also the case for the abdominal plane, with most stomach bubbles and umbilical veins highlighted. The attribution images for the femur plane show that the neural network is able to segregate the femur from the rest of the ultrasound image during gestational age estimation despite segmentation maps not provided to the neural network, which suggests that the neural network is able to learn the anatomies that are relevant to gestational age estimation without explicit prior knowledge being provided. However, it appears that when two potential femurs are visible, the attribution is spread across both instead of the neural network focusing on the one visible femur, which can be seen on the bottom femur image of figure 4.13.

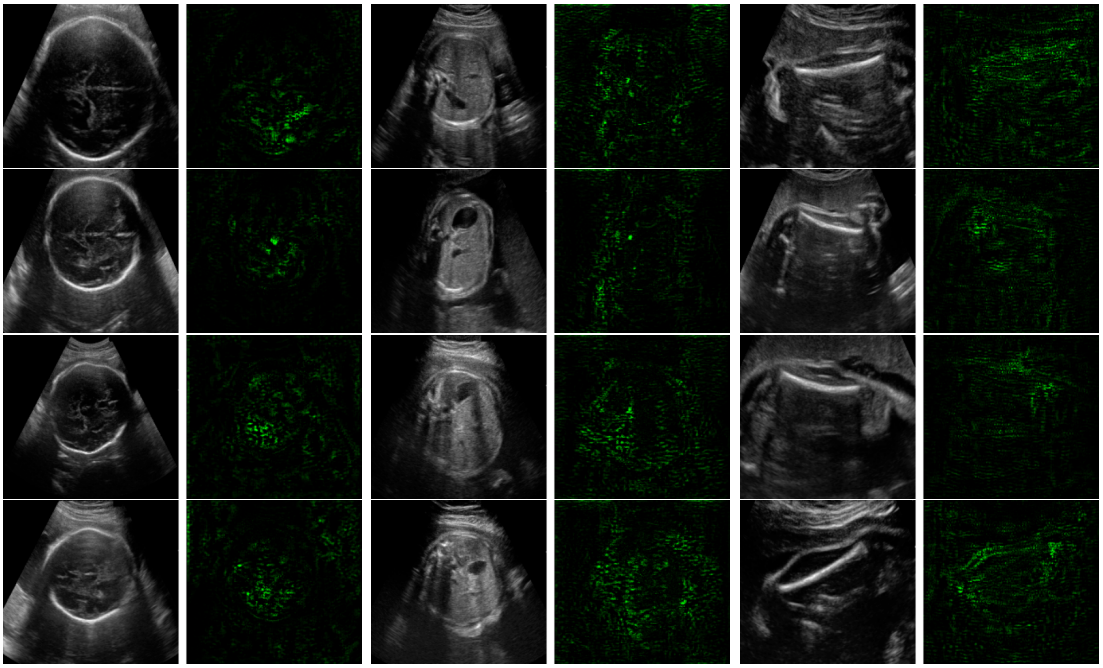


Figure 4.14: Attributions to input images of fetal planes between 23 weeks and 31 weeks gestational age. Images are from left to right columns the HCP, ACP and FLP planes respectively. The green indicates the degree of attribution to each individual pixel for gestational age estimation. Images are best viewed digitally

Between 23 weeks and 31 weeks gestational age

This pattern is repeated for fetal images with a gestational age between 23 weeks and 31 weeks, which can be seen in figure 4.14. Localization of the anatomical structure in question is less evident for the skull, as the larger size of the fetus means that anatomical structures are larger in the image plane. Similar fetal head structures are still highlighted in such as the thalami and the cavum septum pellucidi in HCP planes and the stomach bubble in ACP planes. However, it is harder to qualitatively inspect regions of high attribution due to the attribution being more spread out over the input image. This is also the case for the femur plane. When the femur remains relatively well-localized by the attribution calculated by integrated gradients, gestational age estimation accuracy remains high. However, as can be seen on the bottom two femur images on Fig. 4.14, other femur-like structures in the image with a strong acoustic response can be mistakenly attributed to during gestational age estimation, which might have contributed to the reduced

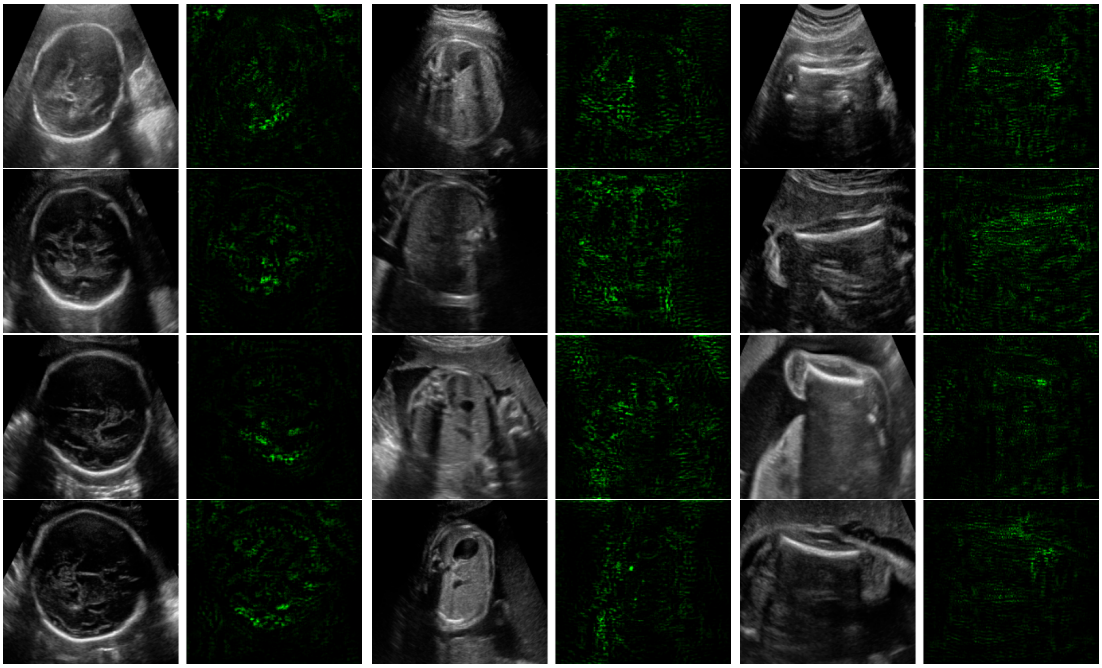


Figure 4.15: Attributions to input images of fetal planes greater than 31 weeks gestational age. Images are from left to right columns the HCP, ACP and FLP planes respectively. The green indicates the degree of attribution to each individual pixel for gestational age estimation. Images are is best viewed digitally

accuracy that gestational age based on femur plane has compared to gestational age estimation from the head plane.

Greater than 31 weeks gestational age

The head continues to be well-captured in head circumference images with gestational ages greater than 31 weeks (figure 4.15). However, the network appears to be focusing on the appearance of the ventricles for gestational age prediction. This is in line with prior knowledge, as the posterior and anterior ventricles are strongly age-discrminating. The network also appears to localize the stomach bubble for the abdominal plane, but it appears that the presence of acoustic shadowing artefacts are also used for gestational age estimation. This might be due to images from higher gestational ages having more shadowing artefacts, especially for the abdomen plane, from acoustically opaque objects such as the ossified rib. This also appears to be the case in femur plane images. Whilst the femur continues to be well localized with significant attribution to each end of the femur, pixels corresponding

to acoustic shadowing from the femur bone appear to contribute to gestational age estimation. This suggests that the network has learned that additional acoustic shadowing is correlated with increased gestational age, and the degree of shadowing can be used to infer gestational age.

4.5 Discussion and Conclusion

To summarize, we have demonstrated that it is possible to regress gestational age from fetal standard planes of the femur head, abdomen and femur respectively, despite a lack of pixel size information. In this chapter, a SE-ResNeXt-50 model was trained with various regression and classification based losses to determine the optimal loss for gestational age regression. We find that models trained with ordinal regression and regression loss consistently outperform classification loss for every anatomy, but models trained with a classification loss are able to show the uncertainty and distribution of predictions which may be more clinically useful and useful to evaluate the performance of the model.

Furthermore, we find that errors in gestational age prediction increase with increased gestational age, which is in line with current clinical chart regression. This might be due to the increased variation in fetal anatomy in late pregnancy, as well as increased acoustic artefacting such as acoustic attenuation and acoustic shadowing. We observed two ways that the neural network fails to infer the gestational age. The first appears to be when there are two candidates for possible gestational ages, which can only be seen when the network is trained with a classifier loss. The other is when the anatomy violates the fetal plane acquisition requirements, including images which occlude the anatomy of interest in some way, or are not of the correct anatomically defined plane.

Integrated gradients was used to visualize the regions of each input image which were used to infer gestational age. We find that structures of the fetal head are well localized, which implies that the head is used for fetal head estimation through all gestational ages. Structures which are of interest to the neural network include the ventricles and cavum septi pellucidi, the appearance of which are known to vary

with gestational age [74]. The stomach bubble and umbilical cord appear to be used for ACP gestational age estimation for smaller gestational ages, but the network appears to focus on acoustic artefacts such as shadowing at higher gestational ages. This is repeated for femur images, with a well localized femur at lower gestational ages and more varied attributions at higher gestational ages. This suggests that the network learns to use acoustic abberations for gestational age estimation at higher gestational ages instead of anatomical presentation.

A potential improvement would therefore be to be train a neural network to use all three standard plane of a fetus simultaneously for gestational age estimation. This would allow the network to focus on more salient features on every gestational age, and might lead to greater accuracy. However, we find that we are not able to get uncertainty distribution information from a network trained with regressive loss, despite the higher accuracy using conventional convolutional neural networks Therefore, in the next chapter, we propose the usage of Bayesian neural networks in order to model the prior distribution over neural network weights, which can then be used to give predictive uncertainties.

Even Leonard Cohen writes, in his Anthem, "There is a kraken: everything".

— Scott Alexander's *Unsong*

5

Uncertain Gestational Age Estimation of Fetal Head Images with Bayesian Neural Networks

Contents

5.1 Introduction 96
5.2 Dataset 97
5.2.1 Pre-processing 98
5.2.2 Validation Datasets 99
5.3 Training Bayesian Neural Networks 100
5.3.1 Variational Inference 102
5.3.2 Bayes by Backpropagation 102
5.3.3 Minimizing Gradient Variance using Flipout Monte Carlo Estimator 103
5.3.4 Kullbeck-Leibler Loss Annealing 104
5.4 Aleatoric and Epistemic Uncertainty 105
5.5 Accuracy of Uncertainty Estimates 106
5.5.1 Uncertainty Calibration 108
5.6 Experiments and Results 109
5.6.1 Implementation Details 111
5.6.2 Quantitative Evaluation 111
5.6.3 Visualization of Uncertainty Calibration 114
5.6.4 Evaluation of Aleatoric and Epistemic Uncertainties . . 115
5.6.5 Out of Distribution Data Detection 120
5.7 Discussion and Conclusion 123

In this chapter, we consider the problem of how to estimate the gestational

age of a fetus from fetal standard plane images using a Bayesian Neural Network (BNN). The usage of a Bayesian neural network model quantifies the uncertainties of the estimation to be predicted even if the model was trained with a regressive loss. This may be attractive in a clinical setting, as uncertain gestational estimates can be further evaluated with expert guidance. We use a VGG-16 network, a tried and tested neural network architecture in the natural image space. Estimated uncertainties are evaluated and calibrated so that predicted uncertainties are in-line with actual model accuracies. We show that the BNN approach is useful in predicting network uncertainties and can be used to find input domain shifts and invalid anatomies entered as test input. We train the BNN using Bayes-by-backprop, which learns the probability distribution on the weights of the neural network by standard gradient backpropagation. An early version of the work presented in this chapter has been reported in:

Lee L.H., Bradburn E., Papageorghiou A.T., Noble J.A. (2020) Calibrated Bayesian Neural Networks to Estimate Gestational Age and Its Uncertainty on Fetal Brain Ultrasound Images. Proceedings of the International Workshop on Advances in Simplifying Medical Ultrasound (MICCAI ASMUS) 2020 [15].

5.1 Introduction

Convolutional neural networks have proven to be a powerful tool in medical image based regression [145, 156]. Cao uses a deep regressive neural network to predict a deformation field for registration of pairs of images in medical MRI images, and Stern predicts biological age from 3D hand MRI volumes. However, predictions are deterministic and without nuance; and therefore require careful consideration during use in a clinical setting. It would therefore be more powerful if we could use a BNN to predict both the required unknown variable and the uncertainty associated with the variable. A BNN that is capable of predicting both would be much more useful in a clinical setting, where deep learning model interpretability and trust is important [157, 158].

In this chapter, we train both a deterministic CNN and a BNN on a dataset of fetal head images acquired in a multi-country, multi-centre and multi-ethnic project where gestational age is known and over a wide range of gestational age. In line with the previous chapter, both neural networks are used to estimate gestational age directly without accompanying pixel size information, which forces each neural network to predict gestational age based on anatomical structure appearance only. We find that the uncertainty predicted by a BNN may not be accurate, which is then calibrated using an auxiliary isotonic regression model in section 5.5.1. The BNN is then used to estimate the uncertainty of a prediction, which can then be used to determine input data domain shift as we show later in section 5.6.5.

5.2 Dataset

The clinical training data used in this work was taken from the International Fetal and Newborn Growth Consortium for the 21st Century Project (INTERGROWTH) [4]. The INTERGROWTH-21st study was prescribed optimal fetal and infant growth standards based on preterm postnatal growth and infant developmental markers [4, 6, 159]. As part of the INTERGROWTH-21st study, ultrasound fetal images were taken of healthy women with low-risk pregnancies. As this was a longitudinal study, scans were then taken every 5 weeks ($\pm$ 1 week), which led to a dataset that provided accurate gestational age coverage across the entire pregnancy. The subset of data used in this chapter is summarised as can be seen in Figure 5.1.

All scans were taken using the Philips HD-9 (Philips Ultrasound, Bothell, WA, USA) with curvilinear abdominal transducers (C5-2, C6-3, V6-3). Fetal head images were captured in the same anatomically defined standard plane as described in Figure 4.5.

Gestational ages of these fetal images were labelled using the first day of the last menstrual period (LMP), further verified with crown-rump length (CRL) measurement during the first ultrasound scan. If both figures matched within $<$7 days, the LMP was used. Otherwise, the ultrasound scans were discarded.

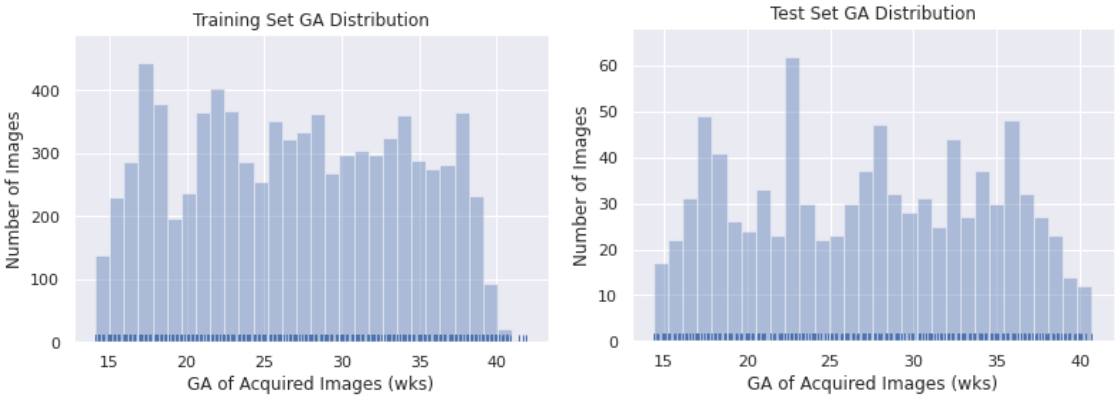


Figure 5.1: Ground Truth Gestational Age Distribution of Fetal Head Images from INTERGROWTH. Due to the longitudinal study formulation, ground truth gestational ages were well distributed across the entire gestatational age range. Pregnant women included in the study were invited back to acquire additional ultrasound scans every 5 weeks. Instances of fetal head images are therefore distributed in all three trimesters, with the youngest fetal head images at 13 weeks and the oldest at 42 weeks.

The full study methodology and low-risk criteria are described in the methodological paper [159].

5.2.1 Pre-processing

Fetal head images taken from the INTERGROWTH-21st study included sonographer markings. These fetal images included sonographer markings, which were removed in the same process used in section 4.2.2. This cleaned image size was at a resolution of 781x566 pixels, which was then downsampled to 224x224 pixels using bilinear interpolation with the tradeoff in losing finer image details. Image downsampling was performed due to the limited GPU memory available (1.1×10^4 megabytes on a Nvidia GTX 1080 Ti) when training a BNN.

The final cleaned and resized images were then verified by a trained sonographer (EB) to be fetal head ultrasound images which still had the required clearly visible anatomical structures as defined in Figure 4.5.

The dataset was then split on a subject by subject basis, as there could be multiple images in the same gestational age range from the same patient which would reduce generalization and overestimate performance. To fully validate the training of the Bayesian Neural Network, all images were individually scrutinized

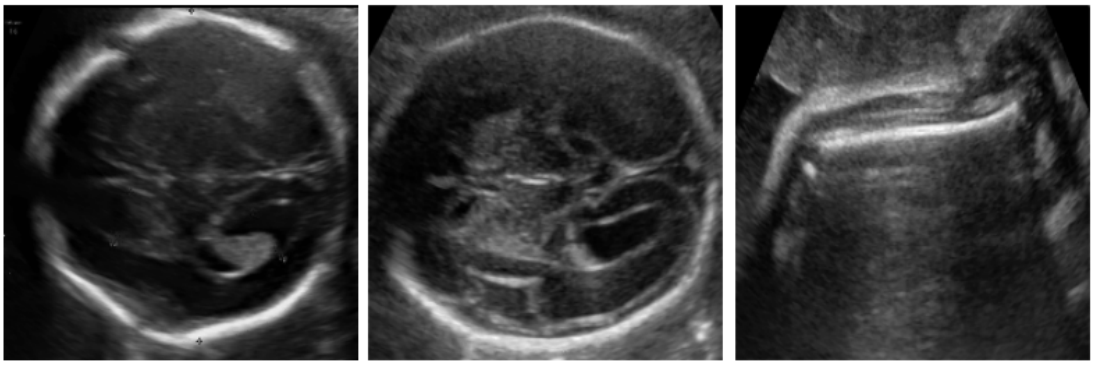


Figure 5.2: Examples of the INTERGROWTH-21st and PULSE datasets used in Bayesian neural network training and testing. Left: A picture of a fetal head with GA of 20+2 weeks from the PULSE dataset, used only for validation. **Middle:** A picture of a fetal head drawn from the INTERGROWTH-21st dataset with GA of 21+2 weeks. **Right:** A picture of a fetal femur drawn from the INTERGROWTH-21st dataset with GA 21+2 weeks.

and validated to be valid fetal head transthalamic standard planes according to the guidelines set out in the fetal anomaly screening programme [5] by an experienced sonographer. A training set of (8369/3016 images/patients), validation set (91/35) and test set (849/300) was generated respectively and used to train and test the Bayesian neural network.

5.2.2 Validation Datasets

In addition to the INTERGROWTH-21st fetal head images dataset, two additional datasets were used for validation of the Bayesian neural network. Examples of all three datasets used in this work can be seen in Figure 5.2

To test the performance of the neural network on out-of-domain data, an additional fetal head dataset was used (PULSE). These scans were taken in a clinical setting in a different protocol to the INTERGROWTH-21st dataset. Ultrasound scans were acquired using a Voluson E8 version BT18 (General Electric Healthcare, Zipf, Austria), which had different imaging characteristics and in-machine image post-processing compared to the machine for the INTERGROWTH-21st dataset. A total of 20 fetal head planes were extracted from routine clinical recordings, with gestational ages ranging from 18+6 weeks to 22+1 weeks. However, compared

to the LMP-based gold standard ground truth used for the INTERGROWTH-21st dataset, the gestational ages had to be inferred based on existing clinical fetal growth charts alone.

To test the performance of the Bayesian neural network on anatomies that the neural network was not trained on, femur standard plane images were taken from the INTERGROWTH-21st dataset. These were acquired using the same machine and operators as described in Section 5.2. However, images of the femur standard planes were used, with different plane requirements that can be seen in figure 4.4.

5.3 Training Bayesian Neural Networks

Conventional convolutional neural networks (CNNs) are composed of an input layer, multiple hidden layers of point trainable and untrainable weights, and an output layer. Each layer is composed of a linear transformation of the previous layer, followed by an activation function. These linear transformations have learnable parameters, also known as weights W, which are tuned during training to optimize for training loss.

More formally, given a training dataset $D = X, Y$ with inputs $X = x_i$ and ground truths $Y = y_i$, a neural network can be viewed in a probabilistic frame as $p(Y|X, W)$. In standard maximum likelihood estimation and routine neural network training, we minimize a loss function (conventionally cross-entropy for classification tasks and mean squared error for regression) and maximize $p(D|W)$ to find the maximum likelihood point estimates of the final neural network parameterized by W.

However, the maximum likelihood method of finding model parameters can lead to problems with generalization outside the training dataset, limited clinical confidence in "black-box" type deep learning predictions means that uncertainties associated with regression predictions would be more useful in that setting.

Bayesian neural networks aim to find a posterior distribution over the weights, $p(W|D)$ by first using a prior distribution $p(W)$ that allows us to enforce our beliefs on the distribution of the weights, as illustrated in figure 5.3 We can then use the posterior predictive distribution

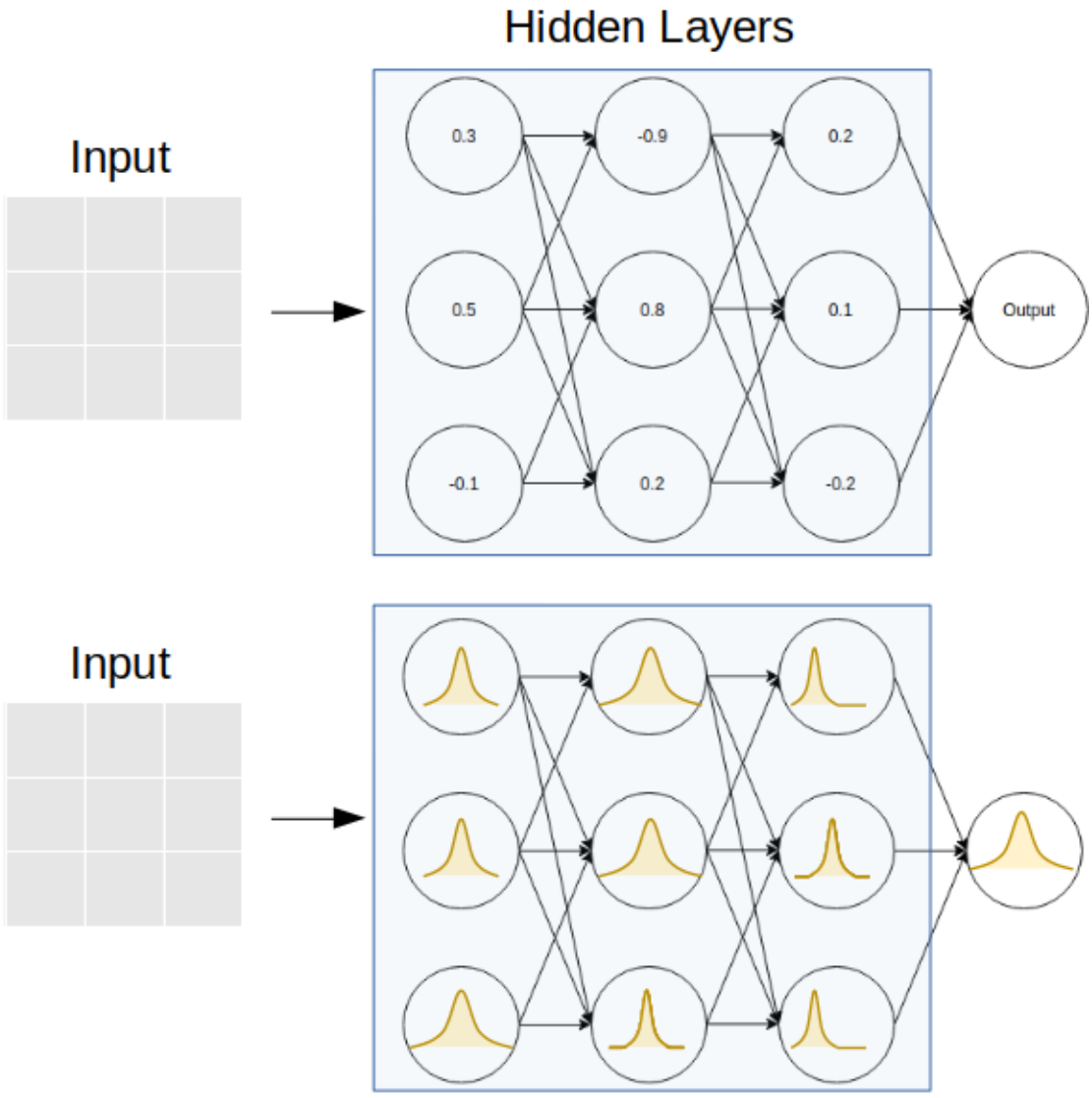


Figure 5.3: Differences Between Conventional Convolutional Neural Networks and Bayesian Neural Networks. Top: In a conventional neural network, each weight in the hidden layer is characterized by a single point value. Bottom: in a Bayesian neural network, probability distributions over weights are learned.

$$p(y|x, D) = \int p(y|x, W)p(W|D)dW \tag{5.1}$$

to perform inference on unknown samples of x.

In practice, $p(W|D)$ found in Equation 5.1 cannot be solved analytically, and must be approximated. The rest of this section is then dedicated to variational inference, which is the method we use to approximate the posterior $p(W|D)$, and methods used to optimize the approximation to the posterior during neural

network training.

5.3.1 Variational Inference

Variational Inference approximates the posterior $p(W|D)$ with a more tractable variational distribution $q(W, \theta)$, where θ represents the parameters which characterize the variational distribution. We can then minimize the difference between $q(W, \theta)$ and $p(W|D)$ which is characterized by the Kullback-Leibler divergence between $p(W|D)$ and $q(W, \theta)$, which can be minimized by varying θ during model training.

The Kullback-Leibler divergence between $q(W, \theta)$ and $p(W|D)$ is defined as:

$$KL(q(W, \theta)||p(W|D)) = \int q(W, \theta) \log(\frac{q(W, \theta)}{p(W|D)})dW \tag{5.2}$$

Using Jensen's inequality [160] and the fact that the Kullbeck-Leibler divergence is always a positive value, it can be shown [161] that that maximizing the evidence lower bound (ELBO) minimizes the Kullbeck-Leibler divergence between the variational distribution $q(W, \theta)$ and the true posterior $p(W|D)$.

$$\text{ELBO} = KL(q(W|\theta)||p(w)) - \mathbb{E}_{q(W,\theta)} \log(p(D|Q)) \tag{5.3}$$

This expected lower bound can then be approximated through a Monte Carlo sampling procedure by drawing samples of $\mathbf{w}$ from the variational distribution $q(W, \theta)$.

$$F(\theta) = \frac{1}{M} \sum_{i=0}^{M} \log q(\mathbf{w}|\theta) - \log p(\mathbf{w}) - \log p(D|\mathbf{w}) \tag{5.4}$$

where M represents a batch and F represents the final cost function.

5.3.2 Bayes by Backpropagation

Optimizing equation 5.4 can be done via backpropagation, in a manner analogous to conventional neural network backpropagation in a method named Bayes-by-backprop [162]. In the forward pass, weights are stochastically "drawn" from the variational distribution $q(W, \theta)$ using an external independent noise variable $\epsilon_k \sim p(\epsilon)$, where each ϵ_k is redrawn for each k mini-batch. A specific posterior sample of the weights $\mathbf{w}$

can then be drawn from the variational distribution, and backpropagated gradients can then be defined to this specific posterior sample for each mini-batch. More specifically, the Bayes-by-Backprop algorithm goes:

Algorithm 5: Algorithm for updating the variational distribution for a Bayesian neural network using Bayes-by-Backprop. In this case, the variational distribution is assumed to be Gaussian.

for $i \leftarrow 0$ **to** n **do**
  Sample $\epsilon_k \sim p(\epsilon)$
  Set current batch network weights $\mathbf{w} = \mu_q + \sigma_q \epsilon_k$
  Calculate loss $\Delta F(\theta)$ from Equation 5.4 and backpropagate gradients g with respect to current batch weights $\mathbf{w}$
  $\theta = \theta - \alpha \Delta F(\theta)$
end

where α is the final learning rate and n is the number of batches. Bayes-by-backprop therefore acts as an unbiased estimator of the gradient for variational inference. However, due to the reparamterization trick [161], the calculated gradient would depend on ϵ_k and the convergence will therefore be noisy.

5.3.3 Minimizing Gradient Variance using Flipout Monte Carlo Estimator

Unfortunately, the stochasticity of the Bayes-by-backprop Algorithm 5 means that convergence is slow for Bayesian neural networks, and may be untenable for large deep neural networks. This is in part due to the same ϵ_k that is drawn for every minibatch n. This same perturbation during the sampling of $\mathbf{w} = \mu_q + \sigma_q \epsilon_k$ for every weight means that there would be gradient correlations within each mini-batch, leading to variance that cannot be reduced by the averaging operation in Equation 5.4. This variance in gradient estimation during stochastic variation inference reduces convergence speed and increases overall computational requirements.

The Flipout Monte Carlo Estimator [163] reduces the correlation between the gradients by randomly negating each sampled weight perturbation $\sigma_q \epsilon_k$ for each sample within a mini-batch. As the random negation is independent of the sampled weight, the final randomly negated perturbation has the same distribution as the

original perturbation, and the final loss gradients are therefore also distributed the same as the original perturbation.

The Bayes-by-Backprop algorithm augmented by the Flipout Monte Carlo Estimator then becomes:

Algorithm 6: Algorithm for updating the variational distribution for a Bayesian neural network using Bayes-by-backprop, augmented with Flipout Monte Carlo Estimation which reduces gradient correlation and therefore improves convergence. In this case, the variational distribution is assumed to be Gaussian. $s_j \in (-1, 1)$ is a rank-one matrix sampled from the Rademacher [164] distribution.

for $i \leftarrow 0$ **to** n **do**
  Sample $\epsilon_k \sim p(\epsilon)$
  for $j \leftarrow 0$ **to** k **do**
    Sample $s_j \sim$ Rademacher distribution
    $\epsilon_{k,j} = \epsilon_k s_j$
    Set current sample network weights $\mathbf{w} = \mu_q + \sigma_q \epsilon_{k,j}$
    Calculate loss $\Delta F(\theta)$ from Equation 5.4 and backpropagate gradients g with respect to current batch weights $\mathbf{w}$
  end
  $\theta = \theta - \alpha \Delta F(\theta)$
end

The random negation as supplied by sampling from the Rademacher [164] distribution decorrelates the per-sample gradients within each mini-batch with a low computational cost. This reduces gradient estimation variances without introducing bias and therefore increases convergence rate.

5.3.4 Kullbeck-Leibler Loss Annealing

The final loss function for gestational age regression therefore takes the form:

$$\mathcal{L}_{total} = \frac{1}{N} \sum_{n} -\frac{1}{2} (\frac{y - \mu_i}{\sigma_i})^2 + \alpha \mathcal{L}_{KL} \tag{5.5}$$

Where L_{total} is the final overall loss, N is the number of samples in the batch, μ_i and σ_i represent the predicted mean gestational age and scale of the aleatoric uncertainty (see section 5.4), $\mathcal{L}_{KL}$ is the Kullbeck-Leibler loss as described in equation 5.4, and α is the scale of the Kullbeck-Leibler loss and the regression loss.

Empirically it is found that the network does successfully converage with $\alpha = k$ if k is constant. We therefore also explore two separate α schedules that vary with each epoch, such that $\alpha(epoch)$ is not constant. We explore two recent training schedules as described in literature [165, 166]. [166] uses a linear annealing scheme where $\alpha epoch$ is gradually linearly increased from zero to one by the hundredth epoch. This can inituitively be analogous to stochastically training a deterministic regression CNN before priors are variationally constrained to the variational distribution. Alternatively, [165] proposes a cyclical annealing schedule with a sawtooth profile. $\alpha epoch$ linearly increases from zero to one every hundred epochs, but regularly drops to zero every 150 epochs. The efficacy of both methods are explored in section 5.6.2.

5.4 Aleatoric and Epistemic Uncertainty

The overall uncertainty of an estimation can be decomposed into two sources - **Epistemic Uncertainty** and **Aleatoric Uncertainty** [12].

Epistemic uncertainty is uncertainty in the definition of our model parameters, which can be captured in the final θ standard distribution in the variational posterior distribution $q(W, \theta)$. Epistemic uncertainty can therefore be reduced by training the model with more data, which will marginalize over the variational posterior distribution.

Aleatoric uncertainty refers to uncertainty that cannot be reduced with additional data. This is uncertainty that is inherent during the generation of the data, which may include sensor noise, or unwanted movement.

The delineation between the two uncertainties is useful from both a clinical and a deep learning standpoint. Variations in epistemic uncertainty can inform further model training of deficiencies in the amount of model data, or when the model is operating on data that is out-of-distribution compared to input training data, whilst aleatoric uncertainties can give a baseline "minimum" uncertainty that the current neural network architecture and training method.

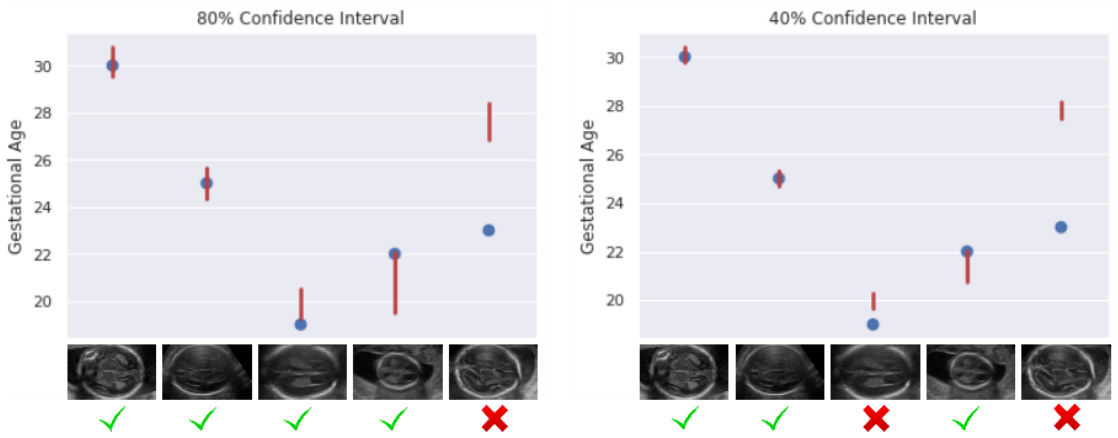


Figure 5.4: Confidence Interval Accuracy for a given Neural Network. The above figures show confidence intervals (in red) for each prediction (in blue) from the same neural network. **Left (Calibrated):** A well-calibrated 80% confidence interval would correctly predict confidence intervals for 80% inferences, which is the case here with five examples. **Right (Mis-Calibrated):** A well-calibrated 40% confidence interval would have an expectation that the ground truth lies within the confidence interval 40% of the time. However, 60% of the predictions lie within the confidence interval, implying that the right figure is mis-calibrated.

5.5 Accuracy of Uncertainty Estimates

Despite the above training methods, recent research has found that Bayesian Neural Networks trained with variational methods can have miscalibrated uncertainty estimates [167, 168]. An uncertainty estimate can be said to be miscalibrated if there is a mistmatch between the predicted credibility interval and the true observed outcome. An example of confidence interval calibration quality is demonstrated in Figure 5.4. A Bayesian neural network predicts gestational age for a series of images in the figure, as well as 80% confidence intervals on the left and 40% confidence intervals on the right. The BNN is well-calibrated on the predictions with the 80% confidence interval, as 80% of the predictions lie within the estimate confidence interval. However, the BNN is not well-calibrated on the 40% confidence interval, as 60% of predictions lie within the confidence interval.

This can result from two main sources of error. Firstly, Bayesian Neural Networks trained with Variational Bayes (See Section 5.3.1) and a variational approximation of the posterior may fail to capture the true posterior distribution even if the

Kullbeck-Leibler distance between the two ais minimalized. Vladimirova [169] finds that the weights of a trained fully-connected feed-forward neural network has weights that are distributed differently depending on their location in the neural network and layer depth. The distribution of weights on the top layer are Gaussian, but deeper layers are drawn from a sub-Weibull distribution [170] with different distribution parameters θ depending on layer depth, with weights being increasingly heavy-tailed compared to an equivalent Gaussian distribution. The equivalent weight distribution is as yet unknown for convolutional neural networks with more complicated architectures, and a Gaussian posterior may therefore be unable to accurately capture the entirety of the posterior probability mass. Secondly, this uncertainty miscalibration may arise from a limited model capability. The model may not be able to assign accurate probabilities to every confidence interval, and therefore mis-predict confidences for different confidence intervals.

Estimated confidence intervals can therefore be said to be calibrated if empirically observed accuracy is well-matched to predicted uncertainty intervals for all confidence probabilities. Figure 5.4 shows an example of a well-calibrated 80% confidence interval, but a mis-calibrated 40% interval.

More generally, we can define the calibration of a model by its accuracy across every confidence interval. Figure 5.5 shows the difference between well-calibrated uncertainties and uncalibrated uncertainties for a given model. For every confidence interval between 0% and 100%, we should expect the accuracy to match the confidence interval. Quantitatively, we can quantify uncertainty calibration error [171] as:

$$\text{calibration} = \sum_{i=0}^{m} (p_i - \hat{p}_i)^2 \tag{5.6}$$

where we choose m confidence intervals to calibrate our chosen model on distribution, p_i refers to the specific chosen confidence interval and $\hat{p}_i$ refers to the empirical accuracy as evaluated on the test dataset. In Figure 5.5, $m = 10$, and equally spaced confidence levels ($p_0 = 0.0, p_1 = 0.1, ..., p_{10} = 1$) were used to assess calibration accuracy.

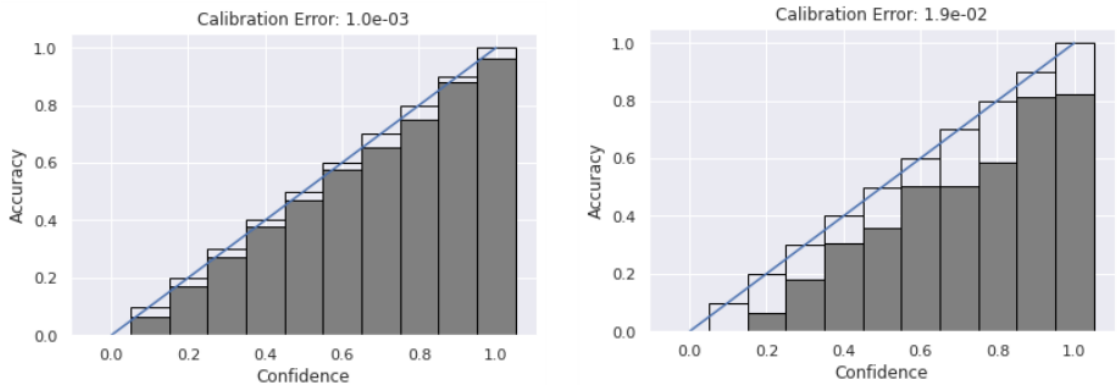


Figure 5.5: Quantifying Uncertainty Calibration in Deep Learning. Left: A well-calibrated uncertainty estimate would have the accuracy of prediction similar to the confidence of predictions for every given confidence interval. **Right** Mis-calibrated uncertainty estimates will have accuracies different from estimated confidence intervals. In this case, the model is consistently under-forecasting and estimating uncertainties that are too low. Calibration errors are calculated with Equation 5.6. For illustrative purposes, this figure is shown with $m = 10$ with equally spaced confidence intervals for evaluation.

5.5.1 Uncertainty Calibration

This suggests that naive usage of uncertainties predicted by a Bayesian Neural Network would lead to inaccurate and under/overconfident predictions. We therefore also use a calibration of uncertainty approach [167, 171], where predicted uncertainties from the Bayesian Neural Network are passed through an additional learned model in order to produce more accurate uncertainty estimates. More formally, an auxiliary model $R : [0, 1] \leftarrow [0, 1]$ can be trained to take uncertainty estimates from the Bayesian Neural Network and would transform the estimates to more accurate estimates.

A method proposed by Kuleshov [171] was used for this purpose in our work. We first generate a series of uncertainty calibration intervals ($p_0 = 0 < p_1 < p_2 < ... < p_m = 1$), where the greater the m the greater the number of calibration intervals used for final auxiliary model training. In our case, we use $m = 10$ with p_m equally spaced between 0 and 1.

For every p_m value and (x, y) in the training set, an auxiliary dataset

$$D = (H(y_i), H(\hat{y}_i))_{i=1}^{m} \tag{5.7}$$

is generated, where $H(y_i)$ represents the estimated confidence intervals for the the interval i for each y value, and $H(\hat{y_i})$ is the actual empirical accuracy of the predictions from the network evaluated with the ground truth.

A regression algorithm is then trained to predict the empirical accuracy from the estimated confidence intervals as an auxiliary model. We use an isotonic regression model [172] as the auxiliary model, as the estimated probability should be monotonically increasing. As the isotonic regression model is non-parametric, an arbitrarily accurate R auxiliary model can be learnt with sufficient data. To avoid overfitting, we train R on a holdout set separate from both the training and test set. During inference and usage, the predicted uncertainty estimates from the Bayesian Neural Network are then calibrated through the auxiliary network R before being used.

5.6 Experiments and Results

The proposed Bayesian neural network was trained for gestational age estimation by directly regressing to the known gestational age. In the following experiments, a conventional deterministic convolutional neural network and a Bayesian neural network were trained with a common architecture. The training of the Bayesian neural network was further investigated with two different loss functions, as described in section 5.3.4.

Through our experiments, we evaluate the performance of a Bayesian neural network compared to a convolutional neural network for a gestational age regression task, given the same dataset limitations and the same architecture. The accuracy of the uncertainty estimates from Bayesian neural networks were evaluated on a hold-out test set. We then train an auxiliary isotonic regression model in order to calibrate uncertainty estimates from the bayesian neural networks, and investigate the use of a Bayesian neural network to discern between in training distribution and out of training distribution input data in the test set.

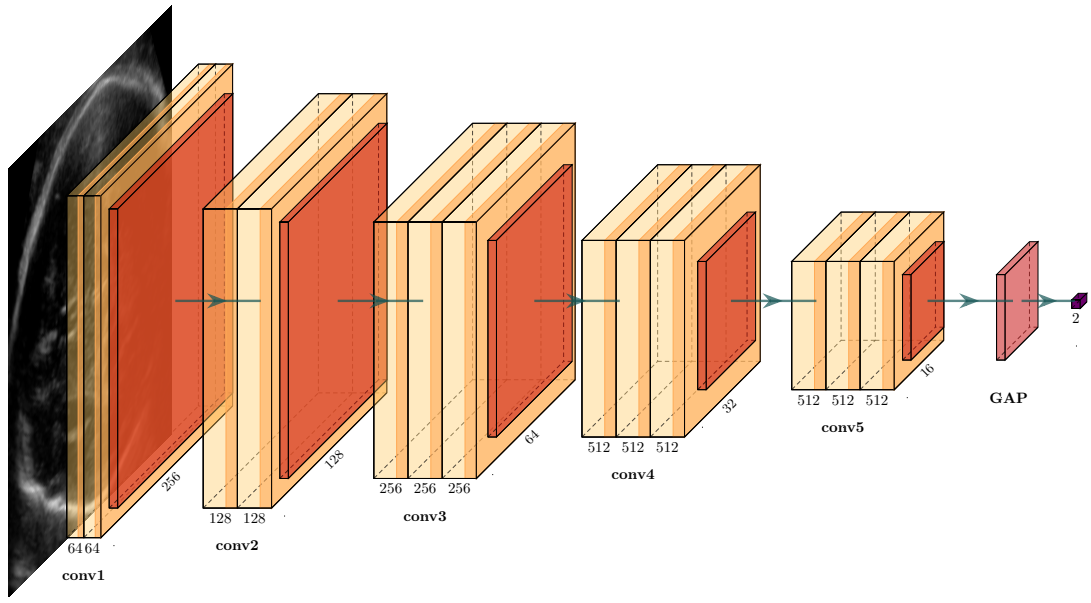


Figure 5.6: Schematic of the VGG-16 neural network architecture used in this chapter. The neural network architecture used in this chapter for both Bayesian neural networks and convolutional neural networks. Each **conv** block consists of multiple rectified convolutional layers (yellow) followed by a maxpooling layer (orange). In contrast to the VGG-16 network as described in [79], a general average pooling (GAP) layer is used at the top for feature aggregation before the final aleatoric uncertainty and gestational age estimation. Due to the regression objective, the final layer consists of fully connected layer with a linear activation.

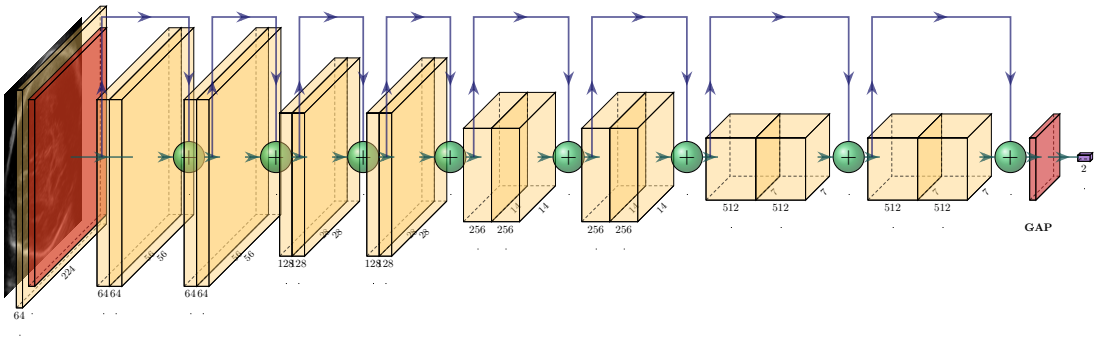


Figure 5.7: Schematic of the ResNet-18 neural network architecture used in this chapter. The ResNet-18 neural network is defined in line with [89]. A general average pooling (GAP) layer is used at the top for feature aggregation before the final aleatoric uncertainty and gestational age estimation. Blue arrows represent skip connections, and a convolutional layer with a (1x1) kernel is used in the skip connection when there is a mismatch between feature sizes in the input and the output. The + symbol represents concatenation in the network diagram.

5.6.1 Implementation Details

We use tried-and-tested neural network architectures VGG-16 [79] and ResNet-18 [89] as the neural network architecture is not the focus of this work. The VGG-16 and ResNet-18 neural network architectures are summarized in figure 5.6 and figure 5.7 respectively. The neural network size is constrained due to computational and memory limitations, as a variational Bayesian neural network with a Gaussian variational prior has double the number of trainable parameters compared to a convolutional neural network.

All models were trained using the Adaptive Moment Estimation (Adam) optimizer (described in Section 2.4.5) with a learning rate of 0.001, β_1 of 0.9 and β_2 of 0.999. All weights were randomly initialized. All networks were trained with a batch size of 50 due to memory constraints of the Bayesian neural network architecture. Networks were trained until validation loss does not decrease for ten continuos epochs at the above learning rate. During the validation pass, each network was sampled stochastically fifty times to compute epistemic uncertainties and mean regressed gestational ages.

The range of valid gestational ages was re-scaled to be between [-1, 1] to assist network optimization. Input pixel values were also re-scaled to between [-1, 1] before neural network training. Images were manually cropped such that the fetal skull defined the boundaries of the image. During training, the training dataset was augmented with random horizontal flipping, random resizing of $\pm 10\%$ and random shifting of up to 20 pixels vertically and horizontally. These ranges were found by manually inspecting the variation in TV plane image appearance in the INTERGROWTH-21st dataset. Of key note is that no pixel size information is given to the network, which would force the network to rely solely on image appearance for gestational age estimation.

5.6.2 Quantitative Evaluation

The performance of both the convolutional neural network and the Bayesian neural network trained with the methods outlined in section 5.3 is reported in this section.

Table 5.1: Quantitative performance for gestational age estimation for both CNNs and BNNs on the INTERGROWTH dataset. Best results for each metric are bolded. RMSE and MSE Metrics are calculated with the maximum likelihood for both aleatoric and epistemic uncertainties. The "Calibration Error" column shows uncertainty calibration error pre- and post- calibration respectively.

	μ RMSE (days)	μ MAE (days)	Calibration Error	No. of Parameters
VGG-16 with KL Cycling (BNN)	**9.6**	**12.5**	0.24/**7.4e-4**	18.8M
VGG-16 with KL Annealing (BNN)	12.2	15.4	**0.21**/1.2e-3	18.8M
ResNet-18 with KL Cycling (BNN)	14.5	18.1	0.25/3.5e-3	11.1M
ResNet-18 with KL Annealing (BNN)	17.1	21.8	0.22/2.5e-3	11.1M
VGG-16 (CNN)	11.9	14.5	N/A	9.4M
ResNet-18 (CNN)	10.6	12.6	N/A	5.6M

We summarise the overall results of both neural networks in the table 5.1. The author reports the root mean square error (RMSE) and mean absolute error (MAE) of the estimates respectively on the test dataset for both baseline neural network architectures (VGG-16 and ResNet-18). Two methods of Bayesian neural network training were experimented with for the Bayesian neural networks; Kullbeck Leibler loss cycling and Kullbeck Leibler loss annealing. The performance of the Bayesian neural networks were compared with the deterministic conventional neural network counterparts. We also investigate the pre- and post- calibration error for each Bayesian neural network with calibrated with an auxiliary isotonic regression model in this section.

Referring to table 5.1 we find that the best performing neural network architecture and training method is the VGG-16 network architecture with Kullbeck Leibler loss cycling.

For the VGG-16 based neural network architecture, the Bayesian Neural Network ("VGG-16 with KL Cycling", "VGG with KL Annealing") appears to outperform the conventional convolutional neural network formulation of VGG-16 ("VGG-16 (Deterministic)"). It should be noted that the Bayesian neural networks had double the parameters of the conventional convolutional neural network, as each weight was parameterized by both the variational mean and standard deviation respectively. The best performing VGG-16 network reduced mean absolute error in days by 2.0 days compared to the deterministic convolutional neural network, and could predict gestational age of a fetal head image to a RMSE of 9.6 and MAE of 12.5 days

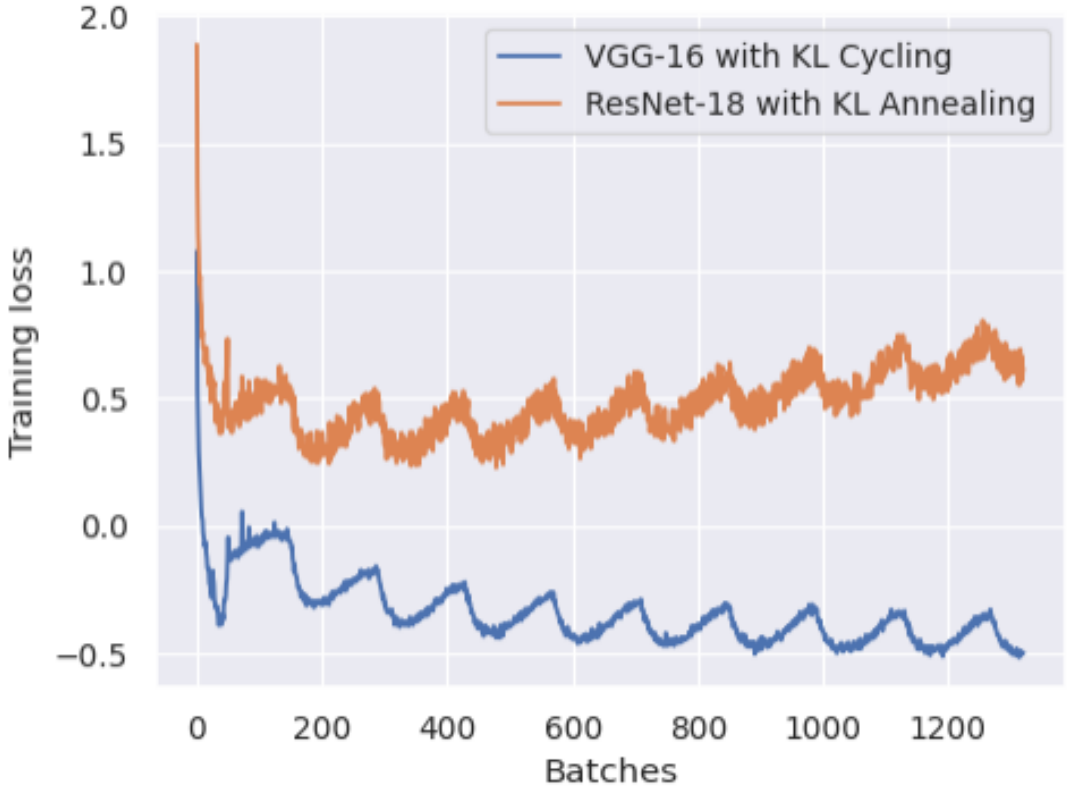


Figure 5.8: Training loss progression of the Bayesian Neural Networks architectures The figure shows the overall training loss of both neural network architectures when trained with Kullbeck Leibler loss cycling. The cycling rate can be easily visible from cyclical nature of the training loss. The VGG-16 network continues to converge with subsequent training batches with this training regime, whilst the ResNet-18 architecture diverges with each subsequent Kullbeck Leibler loss cycle.

respectively across the entire gestational age range without pixel size information. The outperformance of the Bayesian neural networks might be explained by the greater generalization power of Bayesian neural networks compared to conventional convolutional neural networks due to greater parameter count. Bayesian neural networks also act in concert during inference as an ensemble network, which may also account for the increased accuracy.

For both Bayesian neural network architectures, a constant α in the loss (Equation 5.5) leads to unsuccessful loss convergence, results of which were therefore not reported in table 5.1. However, it can be seen that Kullbeck Leibler loss cycling consistently outperformed Kullbeck Leibler loss annealing, which may have been due to more consistent alteration of the loss surface.

The ResNet-18 architectures underperformed when trained with a Bayesian Neural Network ("ResNet-18 with KL Cycling", "ResNet-18 with KL Annealing") compared to the deterministic convolutional neural network ("ResNet-18 (Deterministic)"). This may be due to the more complex architecture leading to a more complex loss surface that is less conducive to a modified loss function. This could be seen in the quantitative evaluation of the final training loss of the ResNet variants of the Bayesian neural networks in figure 5.8. The Kullbeck Leibler loss term did not decrease significantly beyond network initialization, and during cycling appeared to worsen the performance of the neural network regression task. This also suggests that the true posterior distribution of the weights in the ResNet-18 network has a greater Kullbeck Leibler distance from the Gaussian variational posterior than the true posterior distribution of the VGG-16 based architecture, as ADAM was successfully able to jointly optimize both the Kullbeck Leibler distance and the regression loss in the VGG-16 based architecture. This can be seen in the training loss in figure 5.8, and is also suggested by the greater variance that the overall loss for the ResNet-18 architecture experiences during training.

5.6.3 Visualization of Uncertainty Calibration

Figure 5.9 shows the calibration of the proposed Bayesian VGG-16 network trained with Kullbeck Leibler cycling as described in section 5.5. An ideally calibrated Bayesian neural network would have the accuracy-confidence line lying on the $x = y$ axis line, depicted by the black diagonal line. However, the network appears to systematically overestimate uncertainty during inference across every confidence interval range, as can be seen on the left of figure 5.9. This suggests that the Gaussian distribution used as the posterior prior is insufficiently capturing the true prior posterior probability distribution. However, using the procedure described in 5.5, an isotonic regression model can be trained in order to calibrate uncertainty estimates. The uncalibrated error for the VGG-16 neural network model trained with Kullbeck Leibler cycling is 0.22, but drops to $7.4e-4$ when calibrated on the training set as can be seen on the right of figure 5.9.

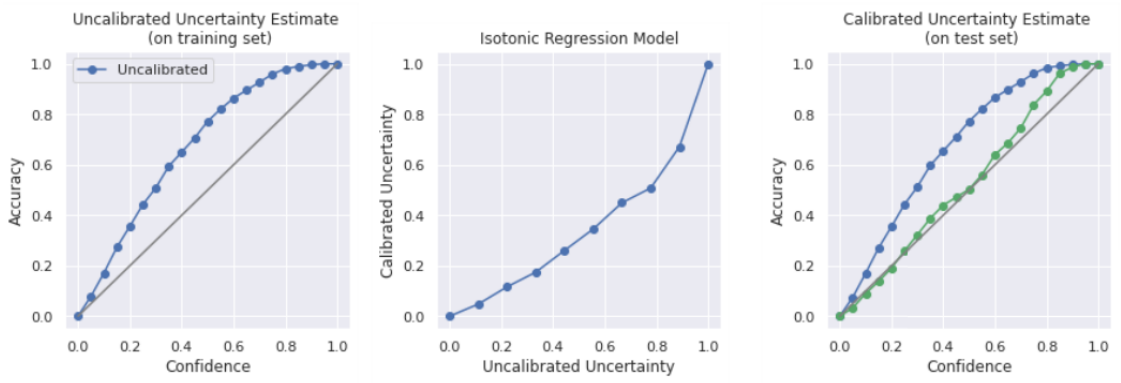


Figure 5.9: Calibration accuracy and auxiliary isotonic regression model for a VGG-16 based Bayesian neural network. Left: Uncalibrated overall uncertainties from the VGG-16 based Bayesian neural network trained with Kullbeck Leibler loss cycling on the uncertainty calibration training set. **Middle:** The auxiliary isotonic regression middle learnt from the training set. **Right:** The uncalibrated and calibrated uncertainty estimates on the test set after uncertainty estimates are put through the auxiliary regression model. The uncalibrated blue line represents a calibration error of 0.21, and the calibrated green line has a calibration error of 1.2×10^{-3}.

However, there appears not be sufficient differentiation between uncertainty intervals with confidence estimates > 0.8, as the true accuracy of these estimates are near 1 within the confidence interval. This leads to the isotonic regression model failing to calibrate uncertainty estimates with high confidence, as can be seen from the deviation from the ideal accuracy-confidence line on the right of the figure.

5.6.4 Evaluation of Aleatoric and Epistemic Uncertainties

Figure 5.10 shows the distribution of uncertainty estimates with gestational age. It should be noted that whilst both uncertainty estimates increase with gestational ages with both best-fit gradient values $m > 0$, the gradient for aleatoric uncertainty increases greater than that of epistemic uncertainty (5.68e-02 versus 1.90e-02). This indicates that the uncertainty inherent in the data (aleatoric) increases with gestational age. This is supported by the fact that the variation in fetal head images appearance increases with increased in gestational age, which would lead to greater variations in data that may not be able to be modelled by the Bayesian neural network. We also find that epistemic uncertainty increases with gestational

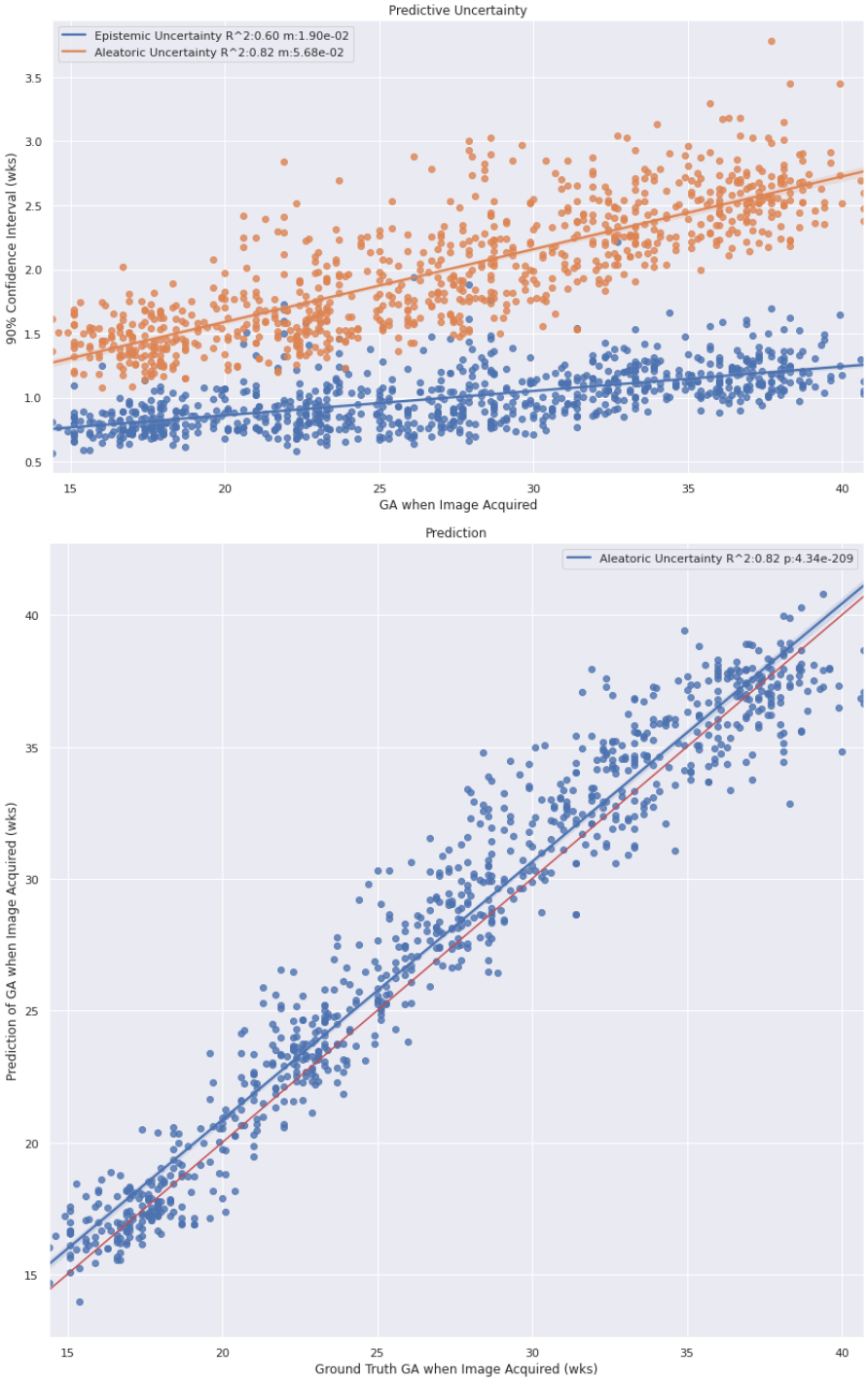


Figure 5.10: Aleatoric and Epistemic Uncertainty Change with Gestational Age. Top: Aleatoric and Epistemic uncertainty variation with differences in gestational age. **Bottom:** Prediction against ground truth gestational ages for the hold-out INTERGROWTH training set. The red line shows the ideal line where predicted gestational age is equal to the ground truth gestational age.

age, with average epistemic uncertainty at 0.76 weeks at early gestational ages, increasing to 1.3 weeks at late gestational ages. This suggests that there continues to be uncertainty that availability of more data may reduce. Furthermore, as the imaged gestational ages are uniformly distributed across the fetal age range, but there is variation in epistemic uncertainty across gestational age, this suggests that gestational age estimation at higher gestational ages requires more data to concentrate the variational prior, which is also in line with the fact that fetal head presentation variation increases with gestational age. We can also see from the bottom of the figure that the predicted gestational age is consistently greater than the ground truth gestational age on average of 0.7 weeks, with the gap decreasing as gestational age increases.

A further breakdown on the relationship between ground truth gestational ages, validation error and uncertainty can be seen in figure 5.11. The categories in the plot are designed to split the validation ground truth gestational age into three equal quantiles. It can be seen that the ground truth gestational ages are well distributed across the entire ground truth range. We can also see that the validation error distribution in images of gestational ages < 22 (blue) is more skewed towards the left than the other two quantiles, which is reflected in the different aleatoric and epistemic distributions for each gestational age quantile, with each successive gestational age quantile moving the median aleatoric uncertainty and epistemic uncertainty further to the right. The modal epistemic uncertainty for each quantile is (0.78, 0.88, 1.12) for each ground truth range respectively, and the modal aleatoric uncertainty for each quantile is (1.46, 1.79, 2.51) respectively. Furthermore, it can be seen that there is a positive relationship between both uncertainties and errors, though the Bayesian neural network defaults to underestimating confidence in the prediction results due to the use of the Gaussian variational posterior.

The variation in image appearance for different aleatoric uncertainties can be seen in figures 5.12 and 5.13. It can be clearly seen from figure 5.13 that images with a higher aleatoric uncertainty have brain anatomical structures that are occluded or not visible. The ventricles and midline structures in the images of the first quantile

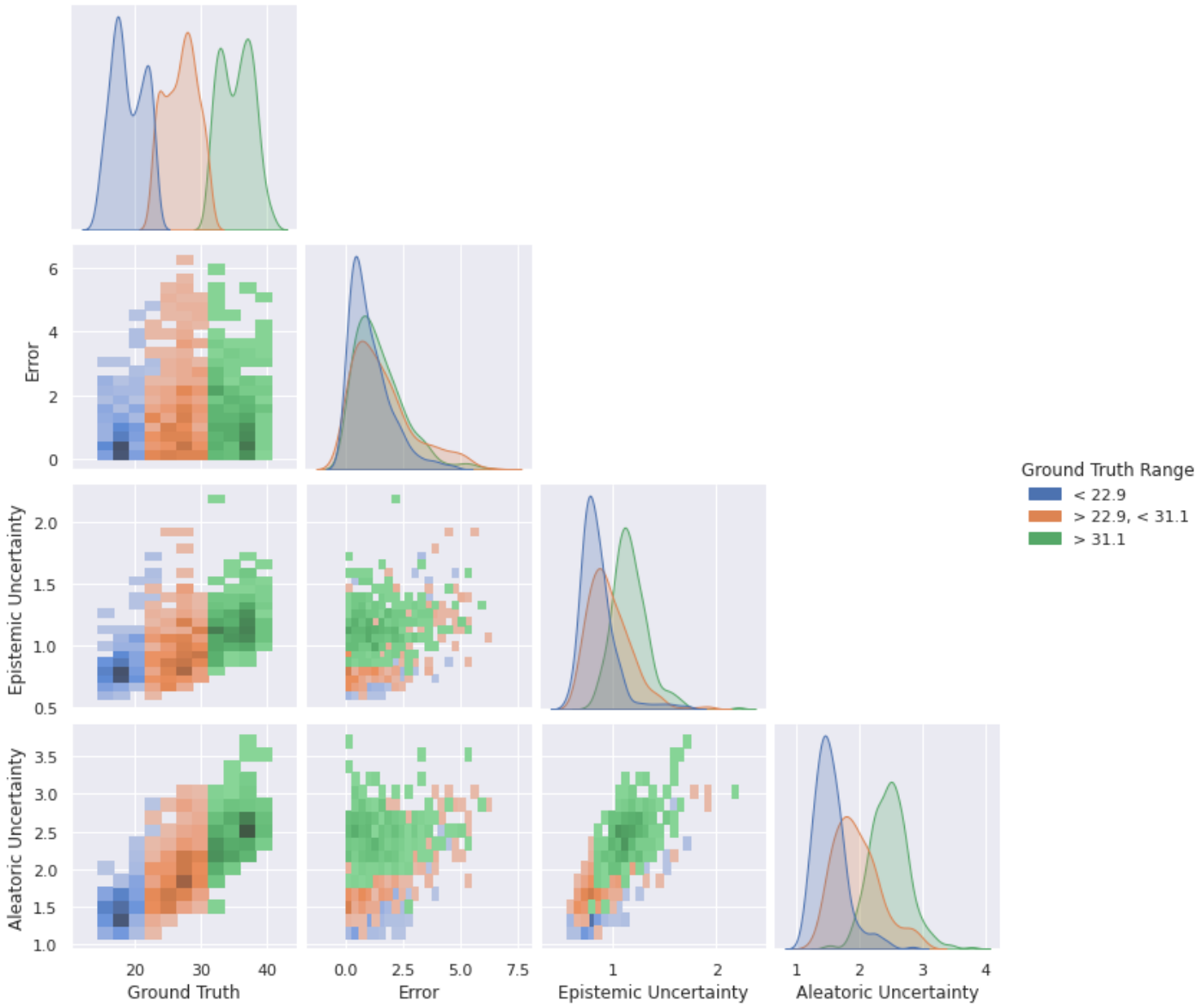


Figure 5.11: Relationship Plot between Gestational Age Error and Uncertainty. All values in both x- and y- axis are in gestational age weeks. The ground truth ranges are selected based on gestational age quantile ranges, and the validation set is divided into three equal sets for easier visualization.

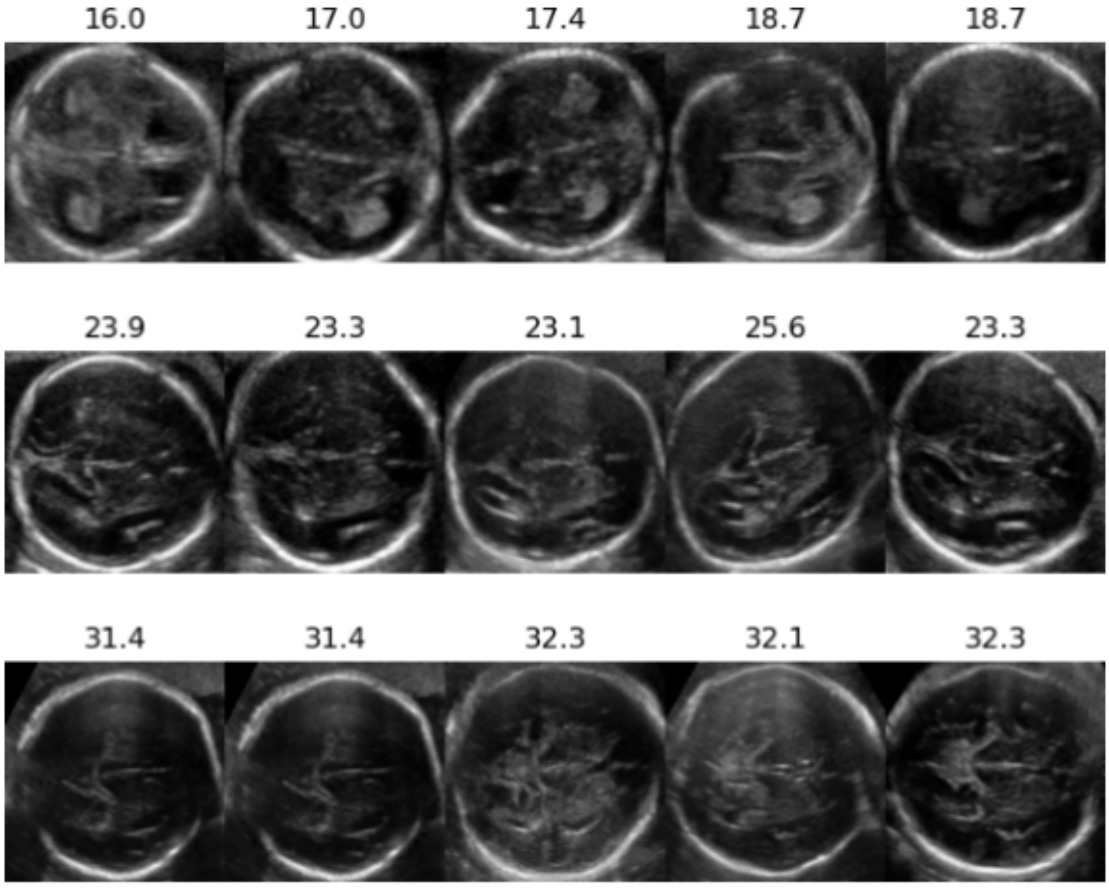


Figure 5.12: Visualization of Fetal Head Images with Low Aleatoric Uncertainty. Numbers above each image indicate ground truth gestational age. Figure shows fetal head images with the lowest aleatoric uncertainties in each quantile.

are clearly visible in all of the images in figure 5.12, whilst the ventricle is only visible for one image in the first quantile in 5.12. This means that some of the anatomical criteria for a valid fetal head plane (see figure 4.5) is not found in the image. This pattern is repeated in the other two quantiles, with the Corpus Callosum and Cavum Septum Pellucidum clearly visible in all images in figure 5.12 but not in figure 5.13. This also acts to suggest that the neural network is successfully using anatomical structural appearance to regress gestational age, images with low visibility of internal head anatomical structures have an associated high aleatoric uncertainty.

However, epistemic uncertainty qualitatively appeared less correlated with fetal brain anatomical structure visibility, as seen in figure 5.14. However, it appears that fetal head images that are rotated away from the horizontal plane and partially

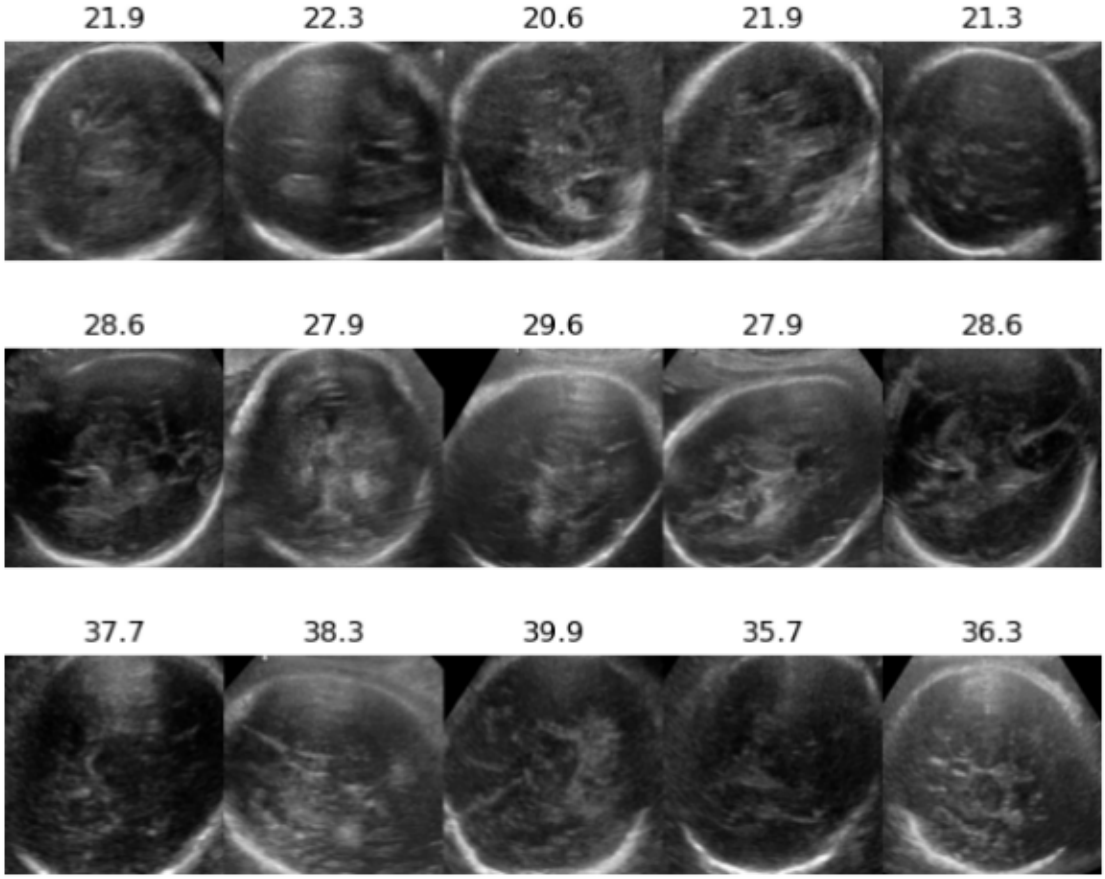


Figure 5.13: Visualization of Fetal Head Images with High Aleatoric Uncertainty. Numbers above each image indicate ground truth gestational age. Figure shows fetal head images with the highest aleatoric uncertainties in each quantile.

occluded fetal heads have a higher epistemic uncertainty. This may be because a large proportion of the dataset consists of fetal heads that are in the left-right orientation. It also appears that increasing gestational age is a factor in increasing epistemic uncertainty, as seen in Figure 5.11. This may be explained by the increasing visible variation in anatomical presentation with increasing gestational age requiring more data to decrease epistemic uncertainty, which is not reflected in the uniform distribution of gestational age in the INTERGROWTH-21st dataset used for training.

5.6.5 Out of Distribution Data Detection

The VGG-16 model trained with Kullbeck Leibler cycling was then evaluated on the two holdout datasets described in section 5.2.2. The results can be seen in Figure 5.15

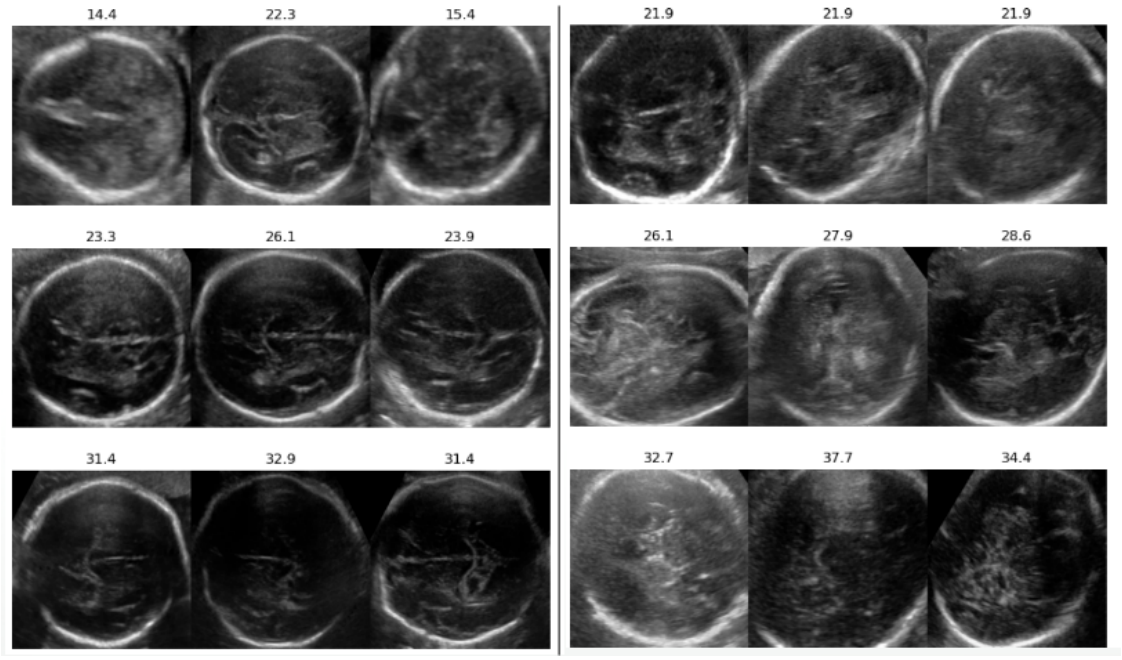


Figure 5.14: Visualization of Fetal Head Images and Epistemic Uncertainties. Left: Head images in each quantile with the lowest epistemic uncertainty. **Right:** Head images in each quantile with the highest epistemic uncertainty.

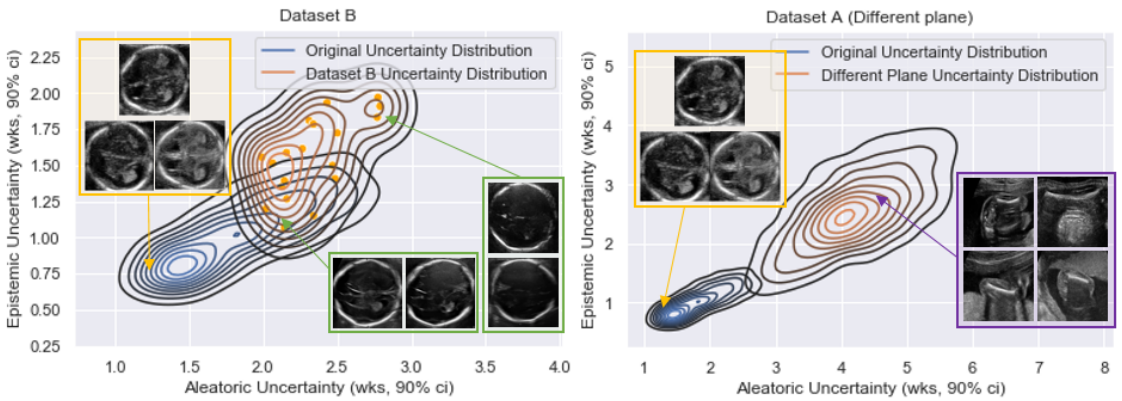


Figure 5.15: Kernel density estimates of uncertainty estimates on out-of-distribution datasets. Images in yellow are from the original INTERGROWTH fetal head dataset. Images in green are fetal head images from the PULSE dataset. Images in purple are fetal femur images from the INTERGROWTH-21st dataset. **Left:** Uncertainty distribution for the Bayesian neural network evaluated on the same anatomical plane but with a different acquisition protocol **Right:** Uncertainty distribution for the Bayesian neural network evaluated on a different anatomical plane with the same protocol.

PULSE Head Scan Plane: Same Anatomy, Different Protocol

As described in section 5.2.2, we validate the performance of the best performing Bayesian neural network on the fetal head anatomical plane but acquired with a different scanning protocol and machine. N=20 images were used in this validation. We find that the neural network is able to predict gestational age in this context to a mean absolute error of 19.6 days and a root mean squared error of 18.1 days respectively. However, this is evaluated based off of ground truth that was regressed from routine fetal growth charts, which have their own associated uncertainty compared to the gold-standard LMP (see section 5.2). As an example, a fetus with a clinically-regressed gestational age will have a clinical 90% confidence interval of 1.4 weeks [26], which will affect the final MAE and RMSE calculation. We also find that evaluated epistemic and aleatoric uncertainties are increased beyond that of the validation set, as can be seen in the left of figure 5.15. This is expected due to the different data distributions compared to the training set; qualitatively, images from the PULSE dataset have a higher contrast than INTERGROWTH-21st image and show greater variation in probe placement. This increased uncertainty may therefore be used to judge neural network confidence during gestational age estimation, which could help increase clinical confidence in the prediction.

INTERGROWTH-21st Femur Scan Plane: Different Anatomy, Same Protocol

We continue to evaluate the ability for the Bayesian neural network to reject out-of-distribution data by using images taken with the same imaging distribution but of different anatomical structures. As expected, both the epistemic and aleatoric uncertainties for femur plane images are much greater than that of the validation set. This results in two distinctly separate distributions in the right of figure 5.15. The convolutional neural network when tested with femur scan planes continued to output valid gestational ages, as the output of all models were constrained to [-1, 1.]. A convolutional neural network would have therefore produced a gestational age despite the input image being of an invalid anatomical plane.

5.7 Discussion and Conclusion

To summarize, we have evaluated thoroughly the performance of a Bayesian neural network as compared to a conventional neural network. The performance of these two regimes were evaluated on two neural network architectures, the VGG-16 and ResNet-18. The results suggest that Bayesian neural networks can be used for gestational age estimation, and both epistemic and aleatoric estimates of uncertainties can be generated through such a network. We achieve results comparable to clinically regressed gestational age estimates without pixel size, solely regressing based on the information inherent in the appearance of the image. We also find that, perhaps not unexpectedly, images with occluded anatomical structures or significant shadowing lead to increased aleatoric uncertainty. These uncertainties in the regression cannot be reduced with more data.

Furthermore, these uncertainties may be mis-calibrated due to the un-reducable difference in the true posterior and the variational posterior, and therefore an auxiliary isotonic regression model was used to calibrate uncertainty estimates. The auxiliary isotonic model re-calibrated uncertainty estimates for all trained Bayesian neural networks, as shown in reduction in calibration error in Table 5.1. However, the use of Kullbeck Leibler loss cycling allowed for better regression estimates and convergence for the best performing network compared to both a constant Kullbeck Leibler loss and a linearly annealing Kullbeck Leibler loss.

The results also suggest that Bayesian neural networks can be used for out-of-distribution data detection, whether it is for invalid anatomies or different imaging characteristics compared to the original dataset which may be useful in a clinical setting.

6
Summary and Future Work

Contents

6.1 Main Thesis Contributions 125
6.2 Future Work 128
6.2.1 Gestational Age Estimation from Concurrent Multi-Anatomy Standard Planes 128
6.2.2 Mixed Example Extended RandAugment on other Ultrasound Tasks 130
6.2.3 End-to-end Solution for Gestational Age Estimation in a Clinical Setting 131

This thesis has presented a method for automated data augmentation strategy selection and proposed a series of models to address the challenge of gestational age estimation from standard plane images without reliance on fetal biometry across the entirety of the second and third trimesters. In this chapter, we summarise the main contributions of this thesis, discuss the strengths and limitations in the work we have presented and outline potential research avenues for future exploration.

6.1 Main Thesis Contributions

Below we list the contributions of this thesis:

1. **Mixed-Example Extended RandAugment: An extended automated**

framework for automatic optimization of augmentation strategy for ultrasound medical image analysis.

We presented a novel approach for automated augmentation strategy search to train convolutional neural networks, called mixed-example extended RandAugment (Mixed RA). Mixed RA is an augmentation search policy that requires of two hyperparameters to be learnt, and can be implemented via simple grid search. We labelled a dataset consisting of clinically acquired second trimester ultrasound scans. A convolutional neural network model was designed to classify planes selected from the fetal anomaly scanning protocol [5], including fetal biometry standard planes. We trained the convolutional neural network model with different augmentation strategies to evaluate the performance of Mixed RA on this task. It was found that meta-learning for augmentation policy optimization lead a higher performance compared to traditional hand-crafted augmentation policies. A CNN trained with an optimized augmentation policy achieved recall and F1-scores of 75.1%, 73.4% and 74.0% on standard plane classification compared to 70.4%, 64.9% and 67.0% for a CNN trained on a commonly used hand-crafted augmentation policy. We also used diversity and affinity as metrics to quantify the effect of augmentation policy on network performance, and find that there is a sweet spot where the diversity of the augmented data was maximized and affinity of the augmented data was minimized to maximize network performance. It was also found that non-realistic looking images generated by an optimized augmentation policy may act to increase network performance over generating images with a realistic appearance, which is contrary to how hand-crafted augmentation policies were chosen in previous work [99, 138, 173].

2. **An evaluation of methods using convolutional neural networks for regression of gestational age without fetal biometry.**

   We described the potential methods of gestational age regression using convolutional neural networks, including regression with different losses, binned

classification for regression, and ordinal regression. We trained a convolutional neural network on the novel task of direct gestational age estimation across both the second and third trimesters from standard plane ultrasound images without using clinical biometry growth charts. The best performing model achieved mean absolute error, mean squared error and explained variance of 0.615 weeks, 0.823 weeks^2 and 0.984 respectively across a gestational age of 13+0 weeks to 42+0 weeks, which outperforms current state-of-the-art from clinical growth charts. We found that direct gestational age estimation is less accurate in late pregnancy, and we used integrated gradients, a method for attribution assignment in order to visualize the areas of input images that the neural network was focused on. The visualization allowed us to explain the lower accuracy at late pregnancy due to shadowing and anatomical variation. As the convolutional neural network learnt features automatically, we were also able to visualize the image features that were used for gestational age estimation, and found that both anatomical features in standard planes (such as the cavum septi pellucidi in the head circumference plane, stomach bubble in the abdominal circumference plane, and femur in the femur length plane) and ultrasound acoustic features (such as acoustic shadowing) were used to assist in gestational age estimation.

3. **Bayesian neural network for uncertain regression of gestational age from fetal head standard planes.**

   Deep vanilla regression neural networks are only able to predict point values during inference. To overcome this limitation, we used Bayesian Neural Networks to quantify the uncertainty of the regressed gestational age. We found that more complex neural networks such as the ResNet-18 could not converge in a Bayesian Neural Network context, whereas a simpler feed forward non-residual network such as the VGG-16 converges. We calibrated both the epistemic and aleatoric uncertainties using an isotonic regression model, and found that images with occluded anatomical structure or otherwise

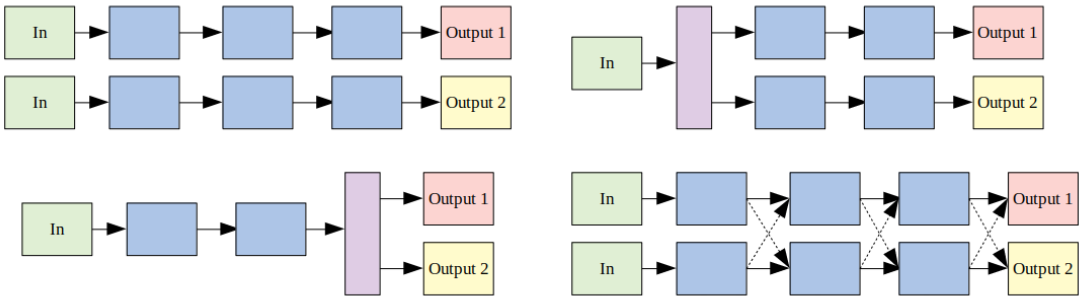


Figure 6.1: Examples of neural network architectures used for Multi-Task Learning (MTL) in previous literature. Convolutional layers are depicted in blue, and aggregation layers are depicted in purple. **Top left:** Two separate neural networks for two different tasks on the same input image. **Top right:** Neural network with early division of convolutional layers for the two tasks. **Bottom left:** Neural network with late division of features for the two tasks. **Bottom right:** Cross-stitch network [174] for self-learnt feature aggregation for tasks. Dashed arrows represent learnt cross-stitch attention modules.

violations of fetal anomaly scanning protocol [5] guidelines increased aleatoric uncertainty. Two separate datasets were used to validate the performance of the network, and the Bayesian neural network was found to be able to detect out-of-distribution input images from images of different anatomies and different scanning protocols.

6.2 Future Work

There are some interesting possible avenues of future research extending on the results as shown in this thesis. This includes 1) Evaluation of Mix. RA on other fetal ultrasound tasks; 2) The integration of multiple anatomical planes of the same fetus into one convolutional neural network for concurrent estimation of gestational age; and 3) Combination of end-to-end plane extraction and gestational age estimation to progress towards a solution for gestational age estimation in the clinical setting.

6.2.1 Gestational Age Estimation from Concurrent Multi-Anatomy Standard Planes

Chapters 4 and 5 presented methods of fetal gestational age estimation using frames from single standard planes. However, it was found that whilst there was

different accuracies in gestational age estimation for different standard planes, neural networks were in general successful ($<$ 1 week gestational age estimation error across second and third trimesters) in learning anatomical cues for regression from different standard planes.

There may therefore be potential in utilizing multiple standard planes for more accurate gestational age estimation. In the fetal anomaly scanning protocol guidelines [5], the standard planes are all routinely taken in the same ultrasound scan. Papageorghiou et al [7] find that gestational age estimation in late pregnancy in a clinical setting is more accurate when based on both femur length and head circumference measurements together, compared with regression based on any single biometry. This suggests that there is potential improvement in accuracy if a neural network was trained to regress gestational age from multiple standard plane images of the same fetus at the same time. However, it would also be interesting to understand the method of data fusion in the neural network to incorporate information from all three standard plane images. This ties into recent research into Multi-Task Learning (MTL) neural networks, which learn to predict multiple values from multiple inputs. Recent research MTL [174, 175] research has found that a single network trained with multiple targets on the same image perform better than separate networks trained separately with example architectures seen in Fig. 6.1, where different tasks benefit from feature aggregation at different representation levels. This is also found in medical ultrasound research [176] where the combination of late and early fusion networks for input information led to greater outperformance in standard plane classification. We may therefore be able to use a neural network architecture similar to the cross-stitch network [174], depicted in Fig. 6.1, but with multiple inputs and a single output for gestational age regression. This would allow the neural network to self-learn the levels of abstraction required for gestational age regression across different anatomies, and may therefore lead to higher accuracy in estimation.

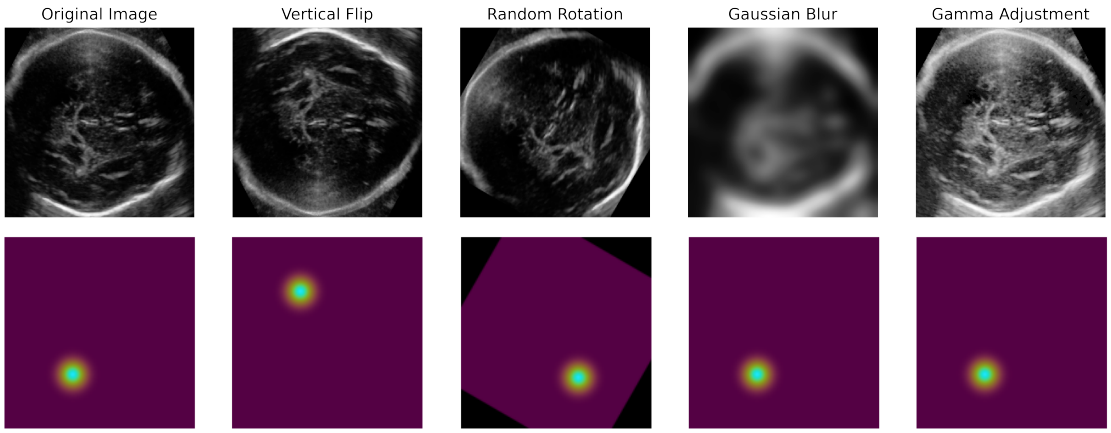


Figure 6.2: Examples of common augmentation transformations and the effect of the transformation on a landmark detection task. The landmark detection task used for illustration here is the identification of the transcerebellar boundary. Affine transformations such as vertical flipping and rotation lead to corresponding changes in the training target landmark, whereas image intensity histogram manipulation and blurring such as Gaussian blur and Gamma adjustments do not.

6.2.2 Mixed Example Extended RandAugment on other Ultrasound Tasks

We have evaluated the performance of our augmentation policy search method on the classification task of standard plane detection and the regression task of gestational age estimation. However, the performance of such augmentation policy search methods have not been evaluated on other medical imaging tasks, such as segmentation or landmark detection.

The application of such an augmentation policy search may not be straightforward, as augmentation transformations might need to be differentially applied to both input images and target features. This is unlike classification or regression, where the optimized augmentation policy can be unilaterally applied to the training dataset. This is shown in figure 6.2, where we show that some transformations have to be applied to both the final heatmap regression as well as the original image. It is an open research question as to whether or not meta-learning based augmentation policy search methods can be applied to segmentation problems, and the efficacy of such methods compared to hand-crafted augmentation

6.2.3 End-to-end Solution for Gestational Age Estimation in a Clinical Setting

Recent work has shown the efficacy of point-of-care low-cost probes in an obstetric setting [177–179], even with simplified ultrasound sweep protocols compared to freehand scanning used by trained sonographers. An end-to-end solution for gestational age estimation as described in the thesis would potentially be useful in an low-and-medium country context where a woman might present at a clinic with unknown dating of the pregnancy. In this case we need to determine the gestational age which may be at any point in pregnancy, to determine the correct patient care. The accuracy required will depend on the scenario, and feedback on quality and uncertainty of the estimate (as per chapter 5) may be more important than accuracy (chapter 4). A potential clinical workflow may be the extraction of standard planes from a fixed-protocol sweep of a pregnant mother, and the estimation of gestational age from standard planes extracted from this upstream algorithm. Future work could also then focus on the translational impact of trained machine learning algorithms in the clinical setting as a direction to explore.

References

[1] S Campbell et al. "Routine ultrasound screening for the prediction of gestational age". In: *Obstetrics and gynecology* 65.5 (1985), 613—620.

[2] Kimberly Butt et al. "Determination of gestational age by ultrasound". In: *Journal of Obstetrics and Gynaecology Canada* 36.2 (2014), pp. 171–181.

[3] C Ioannou et al. "Systematic review of methodology used in ultrasound studies aimed at creating charts of fetal size". In: *BJOG: An International Journal of Obstetrics & Gynaecology* 119.12 (2012), pp. 1425–1439.

[4] Aris T Papageorghiou et al. "The INTERGROWTH-21st fetal growth standards: toward the global integration of pregnancy and pediatric care". In: *American journal of obstetrics and gynecology* 218.2 (2018), S630–S640.

[5] NHS. "Fetal Anomaly Screening Programme Handbook". 2018.

[6] AT Papageorghiou et al. "International standards for early fetal size and pregnancy dating based on ultrasound measurement of crown–rump length in the first trimester of pregnancy". In: *Ultrasound in Obstetrics & Gynecology* 44.6 (2014), pp. 641–648.

[7] A. T. Papageorghiou et al. "Ultrasound-based gestational-age estimation in late pregnancy". In: *Ultrasound in Obstetrics & Gynecology* 48.6 (Dec. 2016), pp. 719–726. URL: https://doi.org/10.1002/uog.15894.

[8] Uganda Bureau of Statistics. *Uganda Demographic and Health Survey 2016*. 2018.

[9] Saikat Deb et al. "Performance of late pregnancy biometry for gestational age dating in low-income and middle-income countries: a prospective, multicountry, population-based cohort study from the WHO Alliance for Maternal and Newborn Health Improvement (AMANHI) Study Group". In: *The Lancet Global Health* 8.4 (2020), e545–e554.

[10] M Mongelli et al. "Third-trimester ultrasound dating algorithms derived from pregnancies conceived with artificial reproductive techniques". In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 26.2 (2005), pp. 129–131.

[11] Stephan Karl et al. "Preterm or Not – An Evaluation of Estimates of Gestational Age in a Cohort of Women from Rural Papua New Guinea". In: *PLOS ONE* 10.5 (May 2015). Ed. by Claudio Romero Farias Marinho, e0124286. URL: https://doi.org/10.1371/journal.pone.0124286.

[12] Alex Kendall and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?" In: *arXiv preprint arXiv:1703.04977* (2017).

[13] Lok Hin Lee, Yuan Gao, and Julia Alison Noble. "Principled Ultrasound Data Augmentation for Classification of Standard Planes". In: *Information Processing in Medical Imaging (IPMI)*. Springer, 2021.

[14] Lok Hin Lee and Julia Alison Noble. "Generating Controllable Ultrasound Images of the Fetal Head". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020, pp. 1761–1764.

[15] Lok Hin Lee et al. "Calibrated Bayesian Neural Networks to Estimate Gestational Age and Its Uncertainty on Fetal Brain Ultrasound Images". en. In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Ed. by Yipeng Hu et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 13–22.

[16] Lok Hin Lee and J Alison Noble. "Automatic Determination of the Fetal Cardiac Cycle in Ultrasound Using Spatio-Temporal Neural Networks". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1937–1940.

[17] Yangdi Xu et al. "Simulating realistic fetal neurosonography images with appearance and growth change using cycle-consistent adversarial networks and an evaluation". In: *Journal of Medical Imaging* 7.5 (2020), p. 057001.

[18] Howard T Sharp et al. "Executive summary of the reVITALize initiative: standardizing gynecologic data definitions". In: *Obstetrics & Gynecology* 129.4 (2017), pp. 603–607.

[19] American College of Obstetricians, Gynecologists, Committee on Obstetric Practice, et al. "Methods for estimating the due date". In: *Obstet Gynecol* 129 (2017), pp. 959–960.

[20] Roger H Hertz et al. "Clinical estimation of gestational age: Rules for avoiding preterm delivery". In: *American journal of obstetrics and gynecology* 131.4 (1978), pp. 395–402.

[21] John T Queenan, Catherine Y Spong, and Charles J Lockwood. *Management of high-risk pregnancy*. Wiley Online Library, 1999.

[22] David A Savitz et al. "Comparison of pregnancy dating by last menstrual period, ultrasound scanning, and their combination". In: *American journal of obstetrics and gynecology* 187.6 (2002), pp. 1660–1666.

[23] Ganesa Wegienka and Donna Day Baird. "A comparison of recalled date of last menstrual period with prospectively recorded dates". In: *Journal of Women's Health* 14.3 (2005), pp. 248–252.

[24] WB Barr and CC Pecci. "Last menstrual period versus ultrasound for pregnancy dating". In: *International journal of gynaecology and obstetrics* 87.1 (2004), pp. 38–39.

[25] Kelly A Bennett et al. "First trimester ultrasound screening is effective in reducing postterm labor induction rates: a randomized controlled trial". In: *American journal of obstetrics and gynecology* 190.4 (2004), pp. 1077–1081.

[26] Aris T Papageorghiou et al. “International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project”. In: *The Lancet* 384.9946 (2014), pp. 869–879. URL: https://www.sciencedirect.com/science/article/pii/S0140673614614902.

[27] Torvid Kiserud et al. “The World Health Organization fetal growth charts: concept, findings, interpretation, and application”. In: *American journal of obstetrics and gynecology* 218.2 (2018), S619–S629.

[28] Pekka Taipale and Vilho Hiilesmaa. “Predicting delivery date by ultrasound and last menstrual period in early gestation”. In: *Obstetrics & Gynecology* 97.2 (2001), pp. 189–194.

[29] HP Robinson and JEE Fleming. “A critical evaluation of sonar “crown-rump length” measurements”. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 82.9 (1975), pp. 702–710.

[30] FP Hadlock et al. “Fetal crown-rump length: reevaluation of relation to menstrual age (5-18 weeks) with high-resolution real-time US.” In: *Radiology* 182.2 (1992), pp. 501–505.

[31] A Pexsters et al. “Clinical implications of intra-and interobserver reproducibility of transvaginal sonographic measurement of gestational sac and crown–rump length at 6–9 weeks’ gestation”. In: *Ultrasound in obstetrics & gynecology* 38.5 (2011), pp. 510–515.

[32] R Napolitano et al. “Pregnancy dating by fetal crown–rump length: a systematic review of charts”. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 121.5 (2014), pp. 556–565.

[33] Eric O Ohuma et al. “Estimation of gestational age in early pregnancy from crown-rump length when gestational age range is truncated: the case study of the INTERGROWTH-21 st Project”. In: *BMC medical research methodology* 13.1 (2013), pp. 1–14.

[34] S MacGregor and R Sabbagha. “Assessment of Gestational Age by Ultrasound”. In: *Global Library of Women’s Medicine’s* (2008).

[35] James R Shields, Arnold L Medearis, and Moraye B Bear. “Fetal head and abdominal circumferences: ellipse calculations versus planimetry”. In: *Journal of clinical ultrasound* 15.4 (1987), pp. 237–239.

[36] Frank P Hadlock et al. “Estimating fetal age: effect of head shape on BPD”. In: *American Journal of Roentgenology* 137.1 (1981), pp. 83–85.

[37] CB Benson and PM Doubilet. “Sonographic prediction of gestational age: accuracy of second-and third-trimester fetal measurements.” In: *AJR. American journal of roentgenology* 157.6 (1991), pp. 1275–1277.

[38] AP Souka et al. “Performance of third-trimester ultrasound for prediction of small-for-gestational-age neonates and evaluation of contingency screening policies”. In: *Ultrasound in obstetrics & gynecology* 39.5 (2012), pp. 535–542.

[39] TC Chang et al. “Prediction of the small for gestational age infant: which ultrasonic measurement is best?” In: *Obstetrics and gynecology* 80.6 (1992), pp. 1030–1038.

[40] Frank P Hadlock et al. “A date-independent predictor of intrauterine growth retardation: femur length/abdominal circumference ratio”. In: *American journal of roentgenology* 141.5 (1983), pp. 979–984.

[41] Anthony M Vintzileos et al. “Fetal weight estimation formulas with head, abdominal, femur, and thigh circumference measurements”. In: *American journal of obstetrics and gynecology* 157.2 (1987), pp. 410–414.

[42] Charles W Hohler. “Ultrasound estimation of gestational age.” In: *Clinical obstetrics and gynecology* 27.2 (1984), pp. 314–326.

[43] JAMES P Crane and Mazie M Kopta. “Prediction of intrauterine growth retardation via ultrasonically measured head/abdominal circumference ratios.” In: *Obstetrics and gynecology* 54.5 (1979), pp. 597–601.

[44] Mark B Landon, Marshall C Mintz, and Steven G Gabbe. “Sonographic evaluation of fetal abdominal growth: predictor of the large-for-gestational-age infant in pregnancies complicated by diabetes mellitus”. In: *American journal of obstetrics and gynecology* 160.1 (1989), pp. 115–121.

[45] Jorge L Gamba et al. “Accuracy of ultrasound in fetal femur length determination. Ultrasound phantom study.” In: *Investigative radiology* 20.3 (1985), pp. 316–323.

[46] RN Wolfson et al. “Comparison of biparietal diameter and femur length in the third trimester: effects of gestational age and variation in fetal growth.” In: *Journal of ultrasound in medicine* 5.3 (1986), pp. 145–149.

[47] Philippe Jeanty et al. “Ultrasonic evaluation of fetal ventricular growth”. In: *Neuroradiology* 21.3 (1981), pp. 127–131.

[48] RB Goldstein, RA Filly, and G Simpson. “Pitfalls in femur length measurements.” In: *Journal of ultrasound in medicine* 6.4 (1987), pp. 203–207.

[49] ISRAEL Goldstein, CHARLES Lockwood, and JOHN C Hobbins. “Ultrasound assessment of fetal intestinal development in the evaluation of gestational age.” In: *Obstetrics and gynecology* 70.5 (1987), pp. 682–686.

[50] JC Konje et al. “Determination of gestational age after the 24th week of gestation from fetal kidney length measurements”. In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 19.6 (2002), pp. 592–597.

[51] J Sagi et al. “Fetal kidney size related to gestational age”. In: *Gynecologic and obstetric investigation* 23.1 (1987), pp. 1–4.

[52] Philippe Jeanty et al. “Estimation of gestational age from measurements of fetal long bones.” In: *Journal of ultrasound in medicine* 3.2 (1984), pp. 75–79.

[53] Shaul Yarkoni et al. “Clavicular measurement: a new biometric parameter for fetal evaluation.” In: *Journal of Ultrasound in Medicine* 4.9 (1985), pp. 467–470.

[54] DM Sherer et al. “Fetal clavicle length throughout gestation: a nomogram”. In: *Ultrasound in obstetrics & gynecology* 27.3 (2006), pp. 306–310.

[55] Gülçin Dilmen et al. “Scapula length measurement for assessment of fetal growth and development”. In: *Ultrasound in medicine & biology* 21.2 (1995), pp. 139–142.

[56] Lyndon M Hill et al. “Fetal heart circumference as a predictor of menstrual in fetuses affected by disturbances in growth”. In: *American journal of obstetrics and gynecology* 169.2 (1993), pp. 347–351.

[57] Brian M Mercer et al. “Fetal foot length as a predictor of gestational age”. In: *American journal of obstetrics and gynecology* 156.2 (1987), pp. 350–355.

[58] Katelyn Stevens et al. “Updating fetal foot length to gestational age references: a chart review of abortion cases from 2012 to 2014”. In: *Contraception* 101.1 (2020), pp. 10–13.

[59] Eleanor A Drey et al. “Improving the accuracy of fetal foot length to confirm gestational duration”. In: *Obstetrics & Gynecology* 105.4 (2005), pp. 773–778.

[60] M Ahmed. “Eye Tracking to Aid Fetal Ultrasound Image Analysis”. In: *PhD Thesis* (2017).

[61] Sonia H Contreras Ortiz, Tsuicheng Chiu, and Martin D Fox. “Ultrasound image enhancement: A review”. In: *Biomedical Signal Processing and Control* 7.5 (2012), pp. 419–428.

[62] C Shobana Nageswari and K Helen Prabha. “Despeckle process in ultrasound fetal image using hybrid spatial filters”. In: *2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE)*. IEEE. 2013, pp. 174–179.

[63] J Alison Noble. *Reflections on ultrasound image analysis*. 2016.

[64] Bero O Verburg et al. “Intra-and interobserver reproducibility study of early fetal growth parameters”. In: *Prenatal Diagnosis: Published in Affiliation With the International Society for Prenatal Diagnosis* 28.4 (2008), pp. 323–331.

[65] SC Perni et al. “Intraobserver and interobserver reproducibility of fetal biometry”. In: *Ultrasound in obstetrics & gynecology* 24.6 (2004), pp. 654–658.

[66] F Yang et al. “Fetal biometry by an inexperienced operator using two-and three-dimensional ultrasound”. In: *Ultrasound in Obstetrics and Gynecology* 35.5 (2010), pp. 566–571.

[67] I Sarris et al. “Intra-and interobserver variability in fetal ultrasound measurements”. In: *Ultrasound in obstetrics & gynecology* 39.3 (2012), pp. 266–273.

[68] Snehal Kolapkar and Asmita Wakankar. “Automated gestational age estimation for monitoring fetal growth”. In: *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*. IEEE. 2014, pp. 1104–1109.

[69] Christine W Hanna and Abou Bakr M Youssef. “Automated measurements in obstetric ultrasound images”. In: *Proceedings of International Conference on Image Processing*. Vol. 3. IEEE. 1997, pp. 504–507.

[70] Gustavo Carneiro et al. “Automatic fetal measurements in ultrasound using constrained probabilistic boosting tree”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2007, pp. 571–579.

[71] Wei Lu, Jinglu Tan, and Randall Floyd. "Automated fetal head detection and measurement in ultrasound images by iterative randomized Hough transform". In: *Ultrasound in medicine & biology* 31.7 (2005), pp. 929–936.

[72] Anca Ciurte et al. "Semi-supervised segmentation of ultrasound images based on patch representation and continuous min cut". In: *PloS one* 9.7 (2014), e100972.

[73] Sylvia Rueda et al. "Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge". In: *IEEE Transactions on medical imaging* 33.4 (2013), pp. 797–813.

[74] Ana IL Namburete et al. "Learning-based prediction of gestational age from ultrasound images of the fetal brain". In: *Medical image analysis* 21.1 (2015), pp. 72–86.

[75] Dinggang Shen, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis". In: *Annual review of biomedical engineering* 19 (2017), pp. 221–248.

[76] Alexander Selvikvåg Lundervold and Arvid Lundervold. "An overview of deep learning in medical imaging focusing on MRI". In: *Zeitschrift für Medizinische Physik* 29.2 (2019), pp. 102–127.

[77] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[78] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[79] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[80] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Deep sparse rectifier neural networks". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 315–323.

[81] Min Lin, Qiang Chen, and Shuicheng Yan. "Network in network". In: *arXiv preprint arXiv:1312.4400* (2013).

[82] Dominik Scherer, Andreas Müller, and Sven Behnke. "Evaluation of pooling operations in convolutional architectures for object recognition". In: *International conference on artificial neural networks*. Springer. 2010, pp. 92–101.

[83] Jack Kiefer, Jacob Wolfowitz, et al. "Stochastic estimation of the maximum of a regression function". In: *The Annals of Mathematical Statistics* 23.3 (1952), pp. 462–466.

[84] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[85] Xiaoxia Wu, Rachel Ward, and Léon Bottou. "Wngrad: Learn the learning rate in gradient descent". In: *arXiv preprint arXiv:1803.02865* (2018).

[86] Rong Ge et al. "The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares". In: *arXiv preprint arXiv:1904.12838* (2019).

[87] Samuel L Smith et al. "Don't decay the learning rate, increase the batch size". In: *arXiv preprint arXiv:1711.00489* (2017).

[88] Rong Ge et al. "Rethinking learning rate schedules for stochastic optimization". In: (2018).

[89] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[90] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[91] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult". In: *IEEE transactions on neural networks* 5.2 (1994), pp. 157–166.

[92] Alon Halevy, Peter Norvig, and Fernando Pereira. "The unreasonable effectiveness of data". In: *IEEE Intelligent Systems* 24.2 (2009), pp. 8–12.

[93] Chen Sun et al. "Revisiting unreasonable effectiveness of data in deep learning era". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 843–852.

[94] Jakub Nalepa, Michal Marcinkiewicz, and Michal Kawulok. "Data augmentation for brain-tumor segmentation: a review". In: *Frontiers in computational neuroscience* 13 (2019), p. 83.

[95] Zeshan Hussain et al. "Differential data augmentation techniques for medical imaging classification tasks". In: *AMIA Annual Symposium Proceedings*. Vol. 2017. American Medical Informatics Association. 2017, p. 979.

[96] Ekin D Cubuk et al. "Autoaugment: Learning augmentation policies from data". In: *arXiv preprint arXiv:1805.09501* (2018).

[97] Ekin D Cubuk et al. "Randaugment: Practical automated data augmentation with a reduced search space". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 702–703.

[98] Jianpeng Zhang et al. "Medical image classification using synergic deep learning". In: *Medical image analysis* 54 (2019), pp. 10–19.

[99] Christian F Baumgartner et al. "SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound". In: *IEEE transactions on medical imaging* 36.11 (2017), pp. 2204–2215.

[100] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639 (2017), pp. 115–118.

[101] Emran Mohammad Abu Anas, Parvin Mousavi, and Purang Abolmaesumi. "A deep learning approach for real time prostate segmentation in freehand ultrasound guided biopsy". In: *Medical image analysis* 48 (2018), pp. 107–116.

[102] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[103] Zeynettin Akkus et al. “Deep learning for brain MRI segmentation: state of the art and future directions”. In: *Journal of digital imaging* 30.4 (2017), pp. 449–459.

[104] Yi Sun, Xiaogang Wang, and Xiaoou Tang. “Deep convolutional network cascade for facial point detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 3476–3483.

[105] Vasileios Belagiannis et al. “Robust optimization for deep regression”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2830–2838.

[106] Alexander Toshev and Christian Szegedy. “Deeppose: Human pose estimation via deep neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1653–1660.

[107] Xiabing Liu et al. “3D head pose estimation with convolutional neural network trained on synthetic images”. In: *2016 IEEE international conference on image processing (ICIP)*. IEEE. 2016, pp. 1289–1293.

[108] Stéphane Lathuilière et al. “A comprehensive analysis of deep regression”. In: *IEEE transactions on pattern analysis and machine intelligence* 42.9 (2019), pp. 2065–2081.

[109] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked hourglass networks for human pose estimation”. In: *European conference on computer vision*. Springer. 2016, pp. 483–499.

[110] Jun Zhang et al. “Breast mass detection in mammography and tomosynthesis via fully convolutional network-based heatmap regression”. In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Vol. 10575. International Society for Optics and Photonics. 2018, p. 1057525.

[111] Vasileios Belagiannis and Andrew Zisserman. “Recurrent human pose estimation”. In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE. 2017, pp. 468–475.

[112] Prachi Punyani, Rashmi Gupta, and Ashwani Kumar. “Neural networks for facial age estimation: a survey on recent advances”. In: *Artificial Intelligence Review* 53.5 (2020), pp. 3299–3347.

[113] Ryan Cunningham, Peter Harding, and Ian Loram. “Deep residual networks for quantification of muscle fiber orientation and curvature from ultrasound images”. In: *Annual Conference on Medical Image Understanding and Analysis*. Springer. 2017, pp. 63–73.

[114] Yufan He et al. “Fully Convolutional Boundary Regression for Retina OCT Segmentation”. In: *Lecture Notes in Computer Science*. Springer International Publishing, 2019, pp. 120–128. URL: https://doi.org/10.1007/978-3-030-32239-7_14.

[115] Weidi Xie, J Alison Noble, and Andrew Zisserman. “Microscopy cell counting and detection with fully convolutional regression networks”. In: *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization* 6.3 (2018), pp. 283–292.

[116] Shun Miao, Z Jane Wang, and Rui Liao. "A CNN regression approach for real-time 2D/3D registration". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1352–1363.

[117] Kh Tohidul Islam, Sudanthi Wijewickrema, and Stephen O'Leary. "A deep learning based framework for the registration of three dimensional multi-modal medical images of the head". In: *Scientific Reports* 11.1 (2021), pp. 1–13.

[118] Bob D de Vos et al. "A deep learning framework for unsupervised affine and deformable image registration". In: *Medical image analysis* 52 (2019), pp. 128–143.

[119] Xuhua Ren et al. "Regression Convolutional Neural Network for Automated Pediatric Bone Age Assessment From Hand Radiograph". In: *IEEE Journal of Biomedical and Health Informatics* 23.5 (Sept. 2019), pp. 2030–2038. URL: https://doi.org/10.1109/jbhi.2018.2876916.

[120] Vladimir I. Iglovikov et al. "Paediatric Bone Age Assessment Using Deep Convolutional Neural Networks". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer International Publishing, 2018, pp. 300–308. URL: https://doi.org/10.1007/978-3-030-00889-5_34.

[121] Lok Hin Lee and J. Alison Noble. "Automatic Determination of the Fetal Cardiac Cycle in Ultrasound Using Spatio-Temporal Neural Networks". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, Apr. 2020. URL: https://doi.org/10.1109/isbi45749.2020.9098613.

[122] Zahra Sobhaninia et al. "Fetal Ultrasound Image Segmentation for Measuring Biometric Parameters Using Multi-Task Deep Learning". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2019. URL: https://doi.org/10.1109/embc.2019.8856981.

[123] Hariharan Ravishankar et al. "Hybrid approach for automatic segmentation of fetal abdomen from ultrasound images using deep learning". In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, Apr. 2016. URL: https://doi.org/10.1109/isbi.2016.7493382.

[124] Mohammed A. Maraci et al. "Toward point-of-care ultrasound estimation of fetal gestational age from the trans-cerebellar diameter using CNN-based ultrasound image analysis". In: *Journal of Medical Imaging* 7.01 (Jan. 2020), p. 1. URL: https://doi.org/10.1117/1.jmi.7.1.014501.

[125] Thomas L.A. van den Heuvel et al. "Automated Fetal Head Detection and Circumference Estimation from Free-Hand Ultrasound Sweeps Using Deep Learning in Resource-Limited Countries". In: *Ultrasound in Medicine & Biology* 45.3 (Mar. 2019), pp. 773–785. URL: https://doi.org/10.1016/j.ultrasmedbio.2018.09.015.

[126] Ana IL Namburete, Weidi Xie, and J Alison Noble. "Robust regression of brain maturation from 3D fetal neurosonography using CRNs". In: *Fetal, Infant and Ophthalmic Medical Image Analysis*. Springer, 2017, pp. 73–80.

[127] PY Simard, D Steinkraus, and JC Platt. "Best practices for convolutional neural networks applied to visual document analysis". In: *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* IEEE. 2003, pp. 958–963.

[128] Alexander Buslaev et al. "Albumentations: fast and flexible image augmentations". In: *Information* 11.2 (2020), p. 125.

[129] Lennart Bargsten and Alexander Schlaefer. "SpeckleGAN: a generative adversarial network with an adaptive speckle layer to augment limited training data for ultrasound image processing". In: *International journal of computer assisted radiology and surgery* 15.9 (2020), pp. 1427–1436.

[130] Hoo-Chang Shin et al. "Medical image synthesis for data augmentation and anonymization using generative adversarial networks". In: *International workshop on simulation and synthesis in medical imaging*. Springer. 2018, pp. 1–11.

[131] Cecilia Summers and Michael J Dinneen. "Improved mixed-example data augmentation". In: *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2019, pp. 1262–1270.

[132] Hongyi Zhang et al. "mixup: Beyond empirical risk minimization". In: *arXiv preprint arXiv:1710.09412* (2017).

[133] Tong Che et al. "Mode regularized generative adversarial networks". In: *arXiv preprint arXiv:1612.02136* (2016).

[134] Naveen Kodali et al. "On convergence and stability of gans". In: *arXiv preprint arXiv:1705.07215* (2017).

[135] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. "Which training methods for GANs do actually converge?" In: *International conference on machine learning*. PMLR. 2018, pp. 3481–3490.

[136] Raphael Gontijo-Lopes et al. "Affinity and diversity: Quantifying mechanisms of data augmentation". In: *arXiv preprint arXiv:2002.08973* (2020).

[137] Jianbo Jiao et al. "Self-supervised representation learning for ultrasound video". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1847–1850.

[138] Richard Droste et al. "Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention". In: *Information Processing in Medical Imaging (IPMI)*. Vol. 11492. LNCS. Springer, 2019.

[139] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[140] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3319–3328.

[141] Eduardo Castro, Jaime S Cardoso, and Jose Costa Pereira. "Elastic deformations for data augmentation in breast cancer mass detection". In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE. 2018, pp. 230–234.

[142] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. "Rank-consistent ordinal regression for neural networks". In: *arXiv preprint arXiv:1901.07884* (2019).

[143] Hao Chen, Xi Wang, and Pheng Ann Heng. "Automated mitosis detection with deep regression networks". In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2016, pp. 1204–1207.

[144] Mingchen Gao et al. "Multi-label deep regression and unordered pooling for holistic interstitial lung disease pattern detection". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2016, pp. 147–155.

[145] Darko Štern et al. "Automated age estimation from hand MRI volumes using deep learning". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 194–202.

[146] Aaron van den Oord et al. "Conditional image generation with pixelcnn decoders". In: *arXiv preprint arXiv:1606.05328* (2016).

[147] Luís Torgo and João Gama. "Regression by classification". In: *Brazilian symposium on artificial intelligence*. Springer. 1996, pp. 51–60.

[148] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. "Pixel recurrent neural networks". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 1747–1756.

[149] Aäron van den Oord et al. "WaveNet: A Generative Model for Raw Audio". In: (2016).

[150] Zhen-Hua Feng et al. "Rectified Wing Loss for Efficient and Robust Facial Landmark Localisation with Convolutional Neural Networks". In: *International Journal of Computer Vision* 128.8-9 (Dec. 2019), pp. 2126–2145. URL: https://doi.org/10.1007/s11263-019-01275-0.

[151] Peter J Huber. "Robust estimation of a location parameter: Annals Mathematics Statistics, 35". In: (1964).

[152] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[153] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. "On the convergence of adam and beyond". In: *arXiv preprint arXiv:1904.09237* (2019).

[154] B. L. Welch. "The Generalization of 'Student's' Problem When Several Different Population Variances are Involved". In: *Biometrika* 34.1-2 (Jan. 1947), pp. 28–35.

[155] J. Martin Bland and DouglasG. Altman. "STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT". In: *The Lancet* 327.8476 (Feb. 1986), pp. 307–310. URL: https://doi.org/10.1016/s0140-6736(86)90837-8.

[156] Xiaohuan Cao et al. "Deformable image registration using a cue-aware deep regression network". In: *IEEE Transactions on Biomedical Engineering* 65.9 (2018), pp. 1900–1911.

[157] Gregor Stiglic et al. "Interpretability of machine learning-based prediction models in healthcare". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.5 (2020), e1379.

[158] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. "Explainable deep learning models in medical image analysis". In: *Journal of Imaging* 6.6 (2020), p. 52.

[159] J Villar et al. “The objectives, design and implementation of the INTERGROWTH-21st Project”. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 120 (2013), pp. 9–26.

[160] Johan Ludwig William Valdemar Jensen et al. “Sur les fonctions convexes et les inégalités entre les valeurs moyennes”. In: *Acta mathematica* 30 (1906), pp. 175–193.

[161] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. “Variational inference: A review for statisticians”. In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[162] Charles Blundell et al. “Weight uncertainty in neural network”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1613–1622.

[163] Yeming Wen et al. “Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches”. In: *International Conference on Learning Representations*. 2018.

[164] Paweł Hitczenko and Stanisław Kwapień. “On the Rademacher series”. In: *Probability in Banach Spaces, 9*. Springer, 1994, pp. 31–36.

[165] Hao Fu et al. “Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 240–250.

[166] Samuel Bowman et al. “Generating Sentences from a Continuous Space”. In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 2016, pp. 10–21.

[167] Chuan Guo et al. “On calibration of modern neural networks”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1321–1330.

[168] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. In: *NIPS*. 2017.

[169] Mariia Vladimirova et al. “Understanding Priors in Bayesian Neural Networks at the Unit Level”. In: *ICML 2019-36th International Conference on Machine Learning*. Vol. 97. 2019, pp. 6458–6467.

[170] Maurice Fr ’e chet. “On the law of probability ’e of the é maximum cart”. In: *Ann. Soc. Math. Polon.* 6 (1927), pp. 93–116.

[171] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. “Accurate uncertainties for deep learning using calibrated regression”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2796–2804.

[172] Alexandru Niculescu-Mizil and Rich Caruana. “Predicting good probabilities with supervised learning”. In: *ICML*. 2005.

[173] Yuan Gao and J. Alison Noble. “Learning and Understanding Deep Spatio-Temporal Representations from Free-Hand Fetal Ultrasound Sweeps”. In: *Lecture Notes in Computer Science*. Springer International Publishing, 2019, pp. 299–308. URL: https://doi.org/10.1007/978-3-030-32254-0_34.

[174] Ishan Misra et al. "Cross-stitch networks for multi-task learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3994–4003.

[175] Sebastian Ruder et al. "Sluice networks: Learning what to share between loosely related tasks". In: *arXiv preprint arXiv:1705.08142* 2 (2017).

[176] Yifan Cai et al. "SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 1475–1478.

[177] Yuan Gao et al. "Label Efficient Localization of Fetal Brain Biometry Planes in Ultrasound Through Metric Learning". In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Springer, 2020, pp. 126–135.

[178] Qingchao Chen et al. "Cross-Device Cross-Anatomy Adaptation Network for Ultrasound Video Analysis". In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Springer, 2020, pp. 42–51.

[179] A Self et al. "OC10. 03: Computer-assisted low-cost point of care ultrasound: an intelligent image analysis algorithm for diagnosis of malpresentation". In: *Ultrasound in Obstetrics & Gynecology* 56 (2020), pp. 28–28.