Sonography Data Science

Clare Teng

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

Acknowledgements

Shhh... shhh... A quiet moment of reflection to thank the many individuals and communities who made Oxford a special place, and also specifically for their contributions towards this thesis. Thank you for your critiques, kindness, friendship and generosity, especially in times of adversity – a non-exhaustive list below. This DPhil was supported by the ERC Advanced Grant scheme led by Professor Alison Noble FRS (ERC-2015-AdG-694581) and the financial contributions of my parents.

DPhil Supervisor
Professor Alison Noble.

Examiners
Professor Min Chen.
Professor Frederick Shic.

Collaborators
Dr Lior Drukker.
Dr Lok Lee Hin.
Jayne Lander.
Professor Aris Papageorghiou.
Dr Harshita Sharma.

Et. al.
Dr Angelico F.
Dr Mohammad Alsharid.
Dr Tapabrata C.
Jamie B.
Sharon C.
The Wagners.
The Fields.
Juan T.

Finally, many thanks to my parents for their love and sacrifices. Discipline, curiosity and tenacity - I could not have finished this DPhil without the gifts that I was fortunate enough to have inherited from you.

And whatever you do, whether in word or deed, do it all in the name of the Lord Jesus, giving thanks to God the Father through him. Colossians 3:17

Abstract

Fetal sonography remains a highly specialised skill in spite of its necessity and importance. Because of differences in fetal and maternal anatomy, and human pyschomotor skills, there is an intra- and inter-sonographer variability amoungst expert sonographers. By understanding their similarities and differences, we want to build more interpretive models to assist a sonographer who is less experienced in scanning.

This thesis's contributions to the field of fetal sonography can be grouped into two themes. First I have used data visualisation and machine learning methods to show that a sonographer's search strategy is anatomical (plane) dependent. Second, I show that a sonographer's style and human skill of scanning is not easily disentangled.

We first examine task-specific spatio-temporal gaze behaviour through the use of *data visualisation*, where a *task* is defined as a specific anatomical plane the sonographer is searching for. The qualitative analysis is performed at both a *population* and *individual* level, where we show that the task being performed determines the sonographer's gaze behaviour.

In our *population*-level analysis, we use unsupervised methods to identify meaningful gaze patterns and visualise task-level differences. In our *individual*-level analysis, we use a deep learning model to provide context to the eye-tracking data with respect to the ultrasound image. We then use an event-based visualisation to understand differences between gaze patterns of sonographers performing the same task.

In some instances, sonographers adopt a different search strategy which is seen in the misclassified instances of an eye-tracking *task classification* model. Our *task classification model* supports the qualitative behaviour seen in our *population*-level analysis, where task-specific gaze behaviour is quantitatively distinct.

We also investigate the use of time-based skill definitions and their appropriateness in fetal ultrasound sonography; a time-based skill definition uses years of clinical experience as an indicator of skill. The developed task-agnostic *skill classification* model differentiates gaze behaviour between sonographers in training and fully qualified sonographers. The preliminary results also show that fetal sonography scanning remains an operator-dependent skill, where the notion of human skill and individual scanning stylistic differences cannot be easily disentangled.

Our work demonstrates how and where sonographers look at whilst scanning, which can be used as a stepping stone for building style-agnostic skill models.

Contents

List of Figures viii

List of Abbreviations xv

1 Introduction 1

1.1 Motivation . . . 1
1.2 Contributions . . . 3
1.3 Thesis Structure . . . 4
1.3.1 Publications . . . 5
1.4 Fetal Anomaly Screening Programme . . . 5
1.4.1 Standard Planes . . . 7
1.5 Perception Ultrasound by Learning Sonographic Experience (PULSE) 8
1.6 Definitions . . . 11
1.7 Acknowledgements . . . 12

2 Literature Review 13

2.1 Visualising Eye-Tracking Events for Videos . . . 13
2.1.1 2-Dimensional (2D) Eye-Tracking Visualisation for Videos . 15
2.1.2 Non-Parametric Classification of Eye Movements . . . 18
2.1.3 Challenges in Fetal Ultrasound . . . 19
2.2 Task Classification: Medical Applications . . . 21
2.2.1 On Surgical and Fetal Ultrasound Video Differences . . . 21
2.2.2 Classification of Medical Tasks using Gaze . . . 22
2.2.3 Challenges in Fetal Ultrasound . . . 24
2.3 Skill Classification . . . 25
2.3.1 Aggregated Eye Movement Characteristics . . . 27
2.3.2 Feature Engineered Eye-Tracking Data . . . 27
2.3.3 Pupillometry . . . 28
2.3.4 Challenges in Fetal Ultrasound . . . 29

3 Datasets and Pre-processing Methods 30
3.1 Software Packages . . . 31
3.2 A Brief Description of `pulsepytools` . . . 32
3.3 Datasets . . . 33
3.3.1 Manually Labelled Second Trimester Scans . . . 33
3.3.2 PULSENet Standard Planes . . . 36
3.3.3 Identification of Heart Standard Planes using Optical Character Recognition . . . 37
3.3.4 Trainer-Trainee Sessions . . . 40
3.3.5 Fully Qualified Sonographer Scan Sessions . . . 40
3.4 Pre-processing Methods . . . 41
3.4.1 Image Augmentation . . . 41
3.4.2 Eye-Tracking Data . . . 42
Zero Padding Eye-Tracking Data. . . . 43
Linear Interpolation of Eye-Tracking Data. . . . 43
Pupillary Pre-processing Method. . . . 44
Scale and Position Invariant Eye-Tracking Data (Feature Engineering). . . . 45
I-VT Algorithm. . . . 46

4 Population Level Visualisation of Spatial Temporal Gaze Characteristics of Sonographers 47
4.1 Introduction . . . 47
4.1.1 Example Visualisation of Gaze and Ultrasound Frames . . . 48
4.1.2 Contribution . . . 51
4.1.3 Data . . . 51
4.1.4 Definitions . . . 52
4.2 Methods . . . 52
4.2.1 Determining Areas-of-Interest (AOIs) using Unsupervised Clustering . . . 52
Clustering Distance Metrics. . . . 55
4.2.2 Visualising Scanning Characteristics in the Spatial and Temporal Domain . . . 57
Sonographer Visual Scanning Modes. . . . 58
Bi-variate Contour Plots. . . . 60
Baseline Comparisons. . . . 62
4.3 Results . . . 62
4.4 Discussion . . . 65
4.5 Summary . . . 66

5 Individual Level Visualisation of Spatial Temporal Gaze Characteristics of Sonographers 67
5.1 Introduction . . . 67
5.1.1 Contribution . . . 68
5.1.2 Data . . . 68
5.1.3 Definitions . . . 69
5.2 Methods . . . 70
5.2.1 Normalisation of Eye-Tracking Data by Localising Anatomy Circumference using Affine Transformer Networks . . . 71
Implementation. . . . 72
5.2.2 Visualisation of Eye-Tracking Data using Time Curves . . . 72
5.3 Results . . . 75
5.4 Summary . . . 79

6 Eye-Tracking Based Task Classification 81
6.1 Introduction . . . 81
6.1.1 Contribution . . . 82
6.1.2 Data . . . 82
6.1.3 Definitions . . . 83
6.2 Methods . . . 83
6.2.1 Normalisation of Gaze Data . . . 84
6.2.2 Time-series Classification Models . . . 85
Implementation. . . . 89
6.2.3 Qualitative Visualisation of Scanpath . . . 92
6.3 Results . . . 93
6.3.1 Task Classification Results . . . 93
6.3.2 Class Imbalance Models . . . 95
6.3.3 Qualitative Results . . . 96
6.4 Discussion . . . 101
6.5 Summary . . . 102

7 Skill and Style Classification 103
7.1 Introduction . . . 103
7.1.1 Contribution . . . 105
7.1.2 Data . . . 105
7.1.3 Definitions . . . 105
7.1.4 Software Packages . . . 106
7.2 Method . . . 106
7.2.1 Generating Features from Eye-Tracking Data . . . 107
7.2.2 Classification Models . . . 108

7.2.3 Skill and Style of Sonographer Gaze . . . 109
7.2.4 Class Imbalance Sampling . . . 111
7.2.5 Years of Scanning and Level of Expertise . . . 112
7.2.6 Implementation . . . 115
7.3 Results . . . 116
7.3.1 Skill Classification Models . . . 116
7.3.2 Qualitative Analysis of Trees . . . 117
7.3.3 Relationship between Years of Scanning and Level of Expertise 119
7.4 Discussion . . . 121
7.5 Summary . . . 123

8 Conclusion **124**
8.1 Summary . . . 124
8.1.1 Distinct Task-Specific Sonographer Gaze Patterns . . . 125
8.1.2 Human Skill and Style of Sonographer Scanning . . . 126
8.2 Limitations . . . 127
8.2.1 Hardware . . . 127
Eye-Tracking Accuracy and Precision Implications. . . . 128
8.2.2 Trimester Differences . . . 130
8.2.3 Implementation . . . 130
8.3 Future Avenues . . . 131
8.4 Conclusion . . . 132

List of Figures

1.1 An example timeline of ultrasound scanning. In this example timeline, the sonographer started performing their scan in B mode (between T_0 and T_1), before freezing the video. Before taking measurements, the sonographer rewinds the video through the buffered frames. These buffered frames are known as cinemode frames (between T_1 and T_2). Once they have found the frame they want to take measurements on, they begin the measurement phase (between T_2 and T_3). After measuring, they unfreeze the video and continue scanning (between T_3 and T_4). 6
1.2 Standard plane heart views. 8
1.3 Head and abdomen standard plane views. 8
1.4 PULSE equipment set-up in the John Radcliffe Hospital, Oxford. The eye-tracker is mounted underneath the ultrasound machine screen. The microphone was placed near/on the computer set-up. The probe motion was recorded using an inertial measurement unit (IMU) which was mounted onto the cable of the probe. The position of the phantom fetus is where the pregnant woman would lie down during the scan. 9

2.1 An example of the types of heatmap eye-tracking visualisations, using the abdomen plane as an example. In Figure 2.1a I show the original abdomen frame. In Figure 2.1b, I show a heatmap of the eye-tracking data superimposed on the original frame. The heatmap shows the concentration of the gaze points recorded by the eye-tracker around the stomach, umbilical vein and aorta. In Figure 2.1c, the areas-of-interest are numbered based on the order that the sonographer looked at the landmarks, and the size of the circle indicates how long they spent there. These figures are for illustration purposes only. . . 15
2.2 An example of scarf plot created using 3 areas-of-interest on the abdomen plane: stomach, aorta and umbilical vein. The colour identifies the anatomical landmark, and the length of the colour bar indicates how long was spent at that landmark. Illustration created in Microsoft PowerPoint using [92] as an example. 16

2.3 Example of a storyline visual. This is similar to the scarf plot in Figure 2.2, but the main idea behind a storyline visualisation is to compare the behaviour between participants. Here the white gaps indicate that the sonographer had not been looking at any of the areas-of-interest stomach, aorta or umbilical vein. PX refers to a unique participant. Illustration created in Microsoft PowerPoint using [71] as an example. 16
2.4 Example of spiral visualisation which uses 'slices' of the image. The left most image shows the abdomen plane, with a red box over a thin slice of the image. The red box's position is determined based on the eye-tracking x co-ordinate. The height of the box is arbitrary. In other similar visualisations, instead of a slice a cropped image can be used. The middle image shows several slices taken over a period of time. On the right most image, the slices are now collated to form a spiral, where the direction of time is anti-clockwise. Illustration created in Microsoft PowerPoint using [64] as an example. 17
2.5 Example of a space-time cube used to visualise eye-tracking data. In this illustration, the abdomen plane is used. In the space-time cube, both the spatial aspect and temporal aspect are displayed. The spatial element is meant to convey where the plane is with respect to the screen's location, and the temporal element shows how the image changes over time. Some key frames are also highlighted and identified (in blue, red and green), where the cumulative eye-tracking data until that point is shown as a heatmap. Illustration created in Microsoft PowerPoint using [95] as an example. 18
2.6 Example of gaze scanpath superimposed onto an image for task classification. X and Y represent the gaze x and y co-ordinates. The gaze points are connected using a lineplot. 23

3.1 Example of the GUI displayed to the annotator. The GUI was used to manually label anatomical planes. The GUI selects a frame with biometric measurements and 5 other frames based on the annotated frame using predefined offsets. The user is able to choose from a set of abbreviations which anatomical plane is displayed. If they are unsure, they can skip the frame and move to the next annotated frame. The example shown has been annoymised and displays the ultrasound image only without the frames corresponding clipboard annotations. 32
3.2 Overview of Anatomy Labels. Figure reproduced from [105]. 34

3.3 Comparison of a 3VV heart view (a) against a situs heart view (b), where the situs heart view shows the heart on the left side of image (b), and a corresponding abdomen plane on the right side of image (b). 36
3.4 Process of frames being pre-processed for maximum OCR. From left to right: (a) Original frame with text 'SITUS BREECH', (b) Non-green pixels converted to black pixels, (c) Green pixels converted to black pixels, and black pixels converted to white pixels. 38
3.5 Overview of Eye-Tracking Interpolation Methods using Linear Interpolation and Zero Padding. 44
3.6 An example showing how a raw gaze point (green) with co-ordinates G_x, G_y is normalised with respect to the hand drawn bounding box (yellow). The point of origin of the bounding box is given as the bottom left corner. 45

4.1 Gaze points from frame 16584 to 16684 plotted as a scatterplot and heatmap on the final frame 16684. 49
4.3 Gaze points from frame 16634 to 16684 plotted as a scatterplot, contour and heatmap on the last frame 16684. 50
4.2 Gaze points from frame 16584 to 16634 plotted as a scatterplot, contour and heatmap on the middle frame 16634. 50
4.4 First frame of the abdomen segment, frame 16584. 50
4.5 An example of the abdomen, brain and heart anatomical plane. . . 52
4.6 An example of determining the elbow of a clustering algorithm. The scoring metric used here is the Silhouette score discussed in Equation 4.9, and the number of clusters being tested ranges from 2 to 12. In this toy example, the optimal number of clusters is 7. 53
4.7 An example of globular and spherical data clustered using a k-means algorithm [160]. In this example, the k-means clustering algorithm has produced 3 different clusters. The x and y axis represent the numerical values of the dummy data points [Reproduced under the BSD License]. 54
4.8 Description of modes I, II, III and IV being considered for spatial and temporal analysis of sonographer visual scanning characteristics (top left, bottom left, top right, bottom right). 59
4.9 Example of bi-variate contour plot showing amount of temporal variance (x axis), and spatial variance (y axis) based on Figure 4.8. 61
4.10 Bi-variate contour plot of abdomen (**left**), brain (**middle**) and heart (**right**) AOIs calculated using HDBSCAN and I-VT and its AOIs landmarks. 64

4.11 Labelled anatomical landmarks.
Abdomen (stomach, aorta and umbilical vein).
Brain (anterior part of the falx cerebi, cavum speti pellucidi and choroid plexus) plane. 65

5.1 Selected brain and heart views from the PULSENet DS rescaled to 224x224 pixels. From left to right: (i) TVP, (ii) TCB, (iii) 3VT, (iv) 3VV, (v) 4CH, (vi) LVOT and (vii) RVOT. 69
5.2 Example of a drawn anatomy circumference (in yellow) of a brain plane. 70
5.3 Figure demonstrating the process of generating an event to represent a fixation from the original ultrasound frames.
1: Original ultrasound frames.
2: Frames transformed using an affine transformer network.
3: Selecting a representation of the event as the middle frame of the fixation.
4: Landmark focused on, defined as an *event*. 73
5.4 Figure summarising the process of developing the time curve (Illustrative purposes only, not to scale).
1: Creating events, here number of events n=9.
2: Calculating distance matrix between events, here numbered from 0 to 8.
3: Distance matrix reduced to 2D using multi-dimensional scaling (MDS).
4: Events connected using splines. 73
5.5 An example showing the performance of the ATN on standard plane images. **Left to right**: TVP, TCB, 3VT, 3VV, 4CH, LVOT, RVOT.
Top: Original images resized to 224x224 pixels.
tom: Images after ATN transformation, resized to 224x224 pixels. . 75
5.6 An example of several scanning patterns for brain plane TVP. . . . 76
5.7 An example of several scanning patterns for brain plane TCB. . . . 76
5.8 Time curves for the heart planes: 3VT, 3VV, 4CH, LVOT and RVOT. 77
5.9 Example of landmarks which were viewed while scanning for brain planes TVP (left) and TCB (right). To avoid visual clutter, event numbers are not shown in this figure. 78
5.10 Example of landmarks which were viewed while scanning for heart planes 3VT, 3VV, 4CH, LVOT, RVOT. To avoid visual clutter, event numbers are not shown in this figure. 79

6.1 A bounding box drawn (in yellow) around the (a) abdomen, (b) brain and (c) heart plane's circumference. To calculate the area occupied by the plane on the frame, the area of the yellow box is divided by the area of the frame. The frame has dimensions 1008x784 pixels. . 84
6.2 An example of calculating the distance between 2 time-series (in this instance the x-co-ordinate of the gaze data) using the dynamic time warping distance metric. The dotted lines represent the nearest match between two data points. 87
6.3 The cluster labels of gaze points after being classified by a k-means algorithm. There are 5 clusters, k, generated. These cluster labels are used as the states to train the hidden Markov models. 92
6.4 Contour density plot of abdomen gaze points normalised using the bounding box method, G_{xBB}, G_{yBB} (left), and an example abdomen scanpath (right). The dotted lines provide a 1:1 reference between the contour plot axes (left) and plotted scanpath (right). In this example, the abdomen scanpath on the left falls within the total distribution of abdomen gaze points on the right - below 0.5 on the y-axis. 93
6.5 Confusion matrix for the GRU(bb+A) model normalised with respect to total number of segments per anatomy plane in the test set (106 segments in total). 95
6.6 Contour plots of normalised eye-tracking data G_{xBB}, G_{yBB} of abdomen, brain and heart plane scanpaths. 97
6.7 Examples of individual abdomen scanpaths. The gaze points G_{xBB}, G_{yBB} are plotted as scatterpoints and joined with connecting lines in temporal order. In these examples, the sonographers have focused on the middle of the abdomen plane where the landmarks are located. In Figure 6.7a, the sonographer's gaze alternated between the different landmarks (up and down). 97
6.8 Examples of individual brain plane scanpaths. The gaze points G_{xBB}, G_{yBB} are plotted as scatterpoints and joined with connecting lines in temporal order. In these examples, the sonographers' gaze have traversed from the left to right representing their visual scanpath across the midline of the plane. 98
6.9 Examples of individual heart plane scanpaths. The gaze points G_{xBB}, G_{yBB} are plotted as scatterpoints and joined with connecting lines in temporal order.In these examples, the sonographers' gaze have focused near the center of the heart plane. In Figure 6.9b, the sonographer found the landmark they were interested in and quickly honed in. When they had finished scanning, their gaze quickly veered off towards a different point. 98

6.10 Abodomen scanpaths which were misclassified as brain and heart scanpaths. For scanpaths which were mislabelled as brain (Figures 6.10a and 6.10b), the sonographers' visual scanpath mimiced a brain scanpath, traversing 'across' the plane. For scanpaths which were mislabelled as heart (Figures 6.10c and 6.10d), the sonographer had focused in the middle of the plane. 99
6.11 Brain scanpaths which were misclassified as an abdomen scanpath. The sonographer had not scaled the brain plane to fit the screen fully, and did not focus along the midline horizontally but diagonally across the plane. 99
6.12 Heart scanpaths which were misclassified as an abdomen and brain scanpaths. For those mislabelled as abdomen scanpaths (Figures 6.12a and 6.12b), the sonographer had looked around the area of focus, as opposed to just focusing on the landmark and 'following' the landmark as they adjusted the probe. For the heart scanpath which was mislabelled as a brain scanpath, the sonographer was moving their gaze and the image simultaneously, resulting in a scanpath which was 'sweeping', rather than focused. 100
6.13 Labelled anatomical landmarks of the abdomen plane: stomach, aorta and umbilical vein. 101

7.1 An example lineplot of a positive monotonic relationship between years of scanning experience and percentage of segments predicted as expert. In this subsection, I have defined percentage of segments predicted as expert as *levels of expertise*. The legend '3 YEARS' refers to the years of experience the expert used to train the skill classification model. 115
7.2 The 50th boosted tree (100 trees were trained in total) trained. The tree was generated using the in-built LightGBM `plot_tree` function. The tree shows the percentage of data which falls into each leaf. The names of the features which were used at the splits are given as columns 26, 171, 158, 165, 10, 102, 81, 60, 52, 188. The list of features can be found in the list below. 118
7.3 Lineplots of models: $EX_{0,3}$, $EX_{10,11}$ and $EX_{14,15}$ which demonstrates a positive monotonic relationship. 121
7.4 Lineplots of models: $EX_{0,16}$, $EX_{1,2}$ and $EX_{2,3}$ which did not show a strong positive or negative relationship between years of experience and expertise. 121

8.1 Elbow method used to determine optimal number of states to use for training the Hidden Markov Models. Here the figure shows that the optimal number of clusters is 5. The scoring metric used here was `distortion` in the `YellowBrick` package [80], which is the sum of the squared distances between the gaze point and its assigned cluster center. 134

8.2 Elbow method used to determine optimal number of neighbours to train the k-nearest neighbour time-series classifier. Here the figure shows that the optimal number of clusters (k) is 2. The scoring metric used here was the accuracy of the tuned dataset using the `tslearn` package [122]. 135

List of Abbreviations

AOI . . . . . Area-of-interest.

AC . . . . . Fetal abdominal circumference measurement.

I-VT . . . . . An off-the-shelf velocity-threshold fixation identification algorithm by [38] that separates fixations and saccades.

LVOT . . . . . Fetal aorta or left ventricular outflow tract which shows the outflow tract of the left ventricle.

OCR . . . . . Optical character recognition.

RVOT . . . . . Fetal pulmonary/right ventricular outflow tract which shows the outflow tract of the right ventricle.

TVP . . . . . Fetal head circumference plane and atrium of the lateral ventricle also known as transventricular plane.

TCB . . . . . Fetal suboccipitobregmatic view demonstrating measurement of the transcerebellar diameter also known as TCP.

2D . . . . . Two dimensional.

3D . . . . . Three dimensional.

3VV . . . . . Fetal 3 vessel view which shows the outflow tract of the right ventricle including the pulmonary artery.

3VT . . . . . Fetal 3 vessel and trachea view which shows the main pulmonary artery in direct communication with the ductus arteriosus, the transverse aortic arch and the superior vena cava.

4CH . . . . . Fetal 4 chamber view showing the transverse section of the thorax including one complete rib and the crux of the heart.

1 Introduction

Contents

1.1 Motivation **1**
1.2 Contributions **3**
1.3 Thesis Structure **4**
1.3.1 Publications 5
1.4 Fetal Anomaly Screening Programme **5**
1.4.1 Standard Planes 7
1.5 Perception Ultrasound by Learning Sonographic Experience (PULSE) **8**
1.6 Definitions **11**
1.7 Acknowledgements **12**

1.1 Motivation

Fetal ultrasound screening is a service provided to pregnant women during their pregnancy to assess the health of their fetus and maternal well-being [90]. The scan is carried out by a qualified sonographer and for the second trimester scans, usually takes about 30 minutes to complete. In spite of its necessity and importance, fetal ultrasound screening is a highly specialised skill that takes several years to acquire because of differences in sonographer skill, fetal and maternal anatomy, and individual response to real-time visual feedback [127].

There is a need to better understand how a sonographer performs the scan. This is done through available data modalities, such as ultrasound videos and gaze behaviour (of sonographers) recorded using eye-trackers. Perception Ultrasound by Learning Sonographic Experience (PULSE) is a unique study (to date) that investigates sonographer scanning behaviour by analysing these data modalities. Sonographer gaze is currently used for saliency detection to describe the visual navigation process [109, 112], learn new tasks [133] and classifying skill [137]. However, less work has been done to analyse gaze behaviour of a population of sonographers. There is research interest for analysing spatio-temporal gaze characteristics to understand different searching strategies and quantification of sonographer expertise [126].

Prior work has shown that there is intra- and inter-sonographer variability amoung a group of experts [39]. Consequently, the searching strategy of one expert could differ from another expert. Analysing a large amount of data is not trivial because computational resources required scales with the size of the dataset. Hence, methods to understand sonographer search strategies need to be both efficient and informative. There is also room to extend current skill quantification methods to include sonographers who have not yet fully qualified, and consider approaches which do not arbitrarily quantify expertise by grouping fully qualified sonographers based on years of scanning experience.

This thesis is concerned with analysing spatio-temporal gaze characteristics of sonographers, and extending current definitions of skill classification and assessment that use eye-tracking data. Hence in the literature review (Chapter 2), I cover 3 different topics. The first topic investigated is the use of data visualisation to analyse eye-tracking data recorded. Data visualisation is a tool used to understand participant behaviour. Analysing eye-tracking data of sonographers is not trivial because of the amount of data collected (over 2-3 years in the PULSE project) and the number of distinct anatomical planes being searched for in a short period of time.

Data visualisation can be thought of loosely as a qualitative evaluation of gaze behaviour. Consequently, this thesis also aims to perform a quantitative evaluation

using task classification models that separates sonographer eye-tracking data when searching for different anatomical planes. Hence, the second topic is concerned with the use of eye-tracking to differentiate scanning tasks.

Finally, the third topic is concerned about current definitions of skill used in literature, and more specifically medical literature. I aim to extend present definitions of skill in fetal sonography.

1.2 Contributions

The main contribution I have made is with respect to the field of fetal sonography. Specifically, I have used data visualisation and machine learning methods to show how sonographers perform their search visually, and where sonographers look at when searching for 3 different planes: abdomen, brain and heart. Moreover, I have shown that the search strategy is (anatomical) plane dependent. I also extend current time-based and task-specific definitions of skill used in medical literature by building a general (task-agnostic) skill model which shows that the relationship between human skill and style are not easily disentangled.

First, I investigated spatio-temporal gaze characteristics of sonographers at the *population* and *individual* level using *data visualisation* methods. Then, following on from the *population* analysis, I built a *task classification model* using eye-tracking data of sonographers. Finally, I extend current definitions of sonographer skill by building a *skill classification model*, where current time-based definitions are over simplified and omit other important factors that contribute to skill level.

In the *population* level analysis, I use unsupervised methods to identify meaningful clusters of gaze points and then visualise the spatio-temporal characteristics of these clusters. The visualisation method also uses ultrasound images as an additional data modality to further separate the gaze clusters.

In the *individual* level analysis, I use a deep learning model to first localise the anatomical plane. Then I use an event-based visualisation to consider differences when performing tasks of different levels of difficulty.

Following on these qualitative analyses, I build an eye-tracking *task classification model*. The task classification model complements the *population* level analysis as it affirms that the distinct spatio-temporal gaze characteristics observed are also quantitatively distinct.

Finally, I extend current definitions of skill in fetal sonography by building an eye-tracking based *skill classification model*. Unlike prior methods [123, 137], I consider all fully qualified sonographers as experts and sonographers learning to scan as trainees.

1.3 Thesis Structure

In **Chapter 1**, I discuss the main contributions and global definitions used throughout the thesis. The Fetal Anomaly Screening Programme (FASP) and the project Perception Ultrasound by Learning Sonographic Experience (PULSE) are introduced. Relevant publications and acknowledgements of collaborators are described. In **Chapter 2**, I discusses current literature on 3 main topics: 1) the use of data visualisation methods to analyse eye-tracking data, 2) the use of eye-tracking to differentiate tasks and 3) the use of eye-tracking for skill classification of clinicians, including adjacent medical fields to that of fetal ultrasound. In **Chapter 3**, I provide an overview of the ultrasound video datasets which were used in this thesis. These include any pre-processing and feature engineering methods of the ultrasound video and eye-tracking data of sonographers. In **Chapters 4 and 5**, I present 2 applications of data visualisation methods on sonographer eye-tracking data to perform population and individual level analysis. In **Chapter 6**, I present a task classification model that was used to differentiate different anatomical planes using only sonographer eye-tracking data. In **Chapter 7**, I present a skill classification model that was used to correlate sonographer scanning skill and years of scanning experience.Finally, in **Chapter 8**, I summarising the contents of my thesis, its limitations and suggest some directions for future work.

1.3.1 Publications

Work from the following publications form the basis of sections of this thesis as detailed below:

- **Chapter 4** C. Teng, H. Sharma, L. Drukker, A. T. Papageorghiou, Alison J. Noble, 'Visualising Spatio-Temporal Gaze Characteristics for Exploratory Data Analysis,' In: *14th ACM Symposium on Eye Tracking Research and Applications (ETRA 2022), Poster Presentation* [149]
- **Chapter 5** C. Teng, L. H. Lee, J. Lander, L. Drukker, A. T. Papageorghiou, Alison J. Noble, 'Skill Characterisation of Sonographer Gaze Patterns during Second Trimester Clinical Fetal Ultrasounds using Time Curves,' In: *14th ACM Symposium on Eye Tracking Research and Applications (ETRA 2022), Poster Presentation* [148]
- **Chapter 6** C. Teng, H. Sharma, L. Drukker, A. T. Papageorghiou, Alison J. Noble, 'Towards Scale and Position Invariant Task Classification using Normalised Visual Scanpaths in Clinical Fetal Ultrasound,' In: *2nd International Workshop of Advances in Simplifying Medical UltraSound (ASMUS 2021) at MICCAI, Oral Presentation* [140]
  Reproduced with permission from Springer Nature
- **Chapter 7** C. Teng, L. Drukker, A. T. Papageorghiou, Alison J. Noble, 'Skill, or Style? Classification of Fetal Sonography Eye-Tracking Data' In: *Gaze Meets ML Workshop at NeurIPS 2022, Poster Presentation* [153]

1.4 Fetal Anomaly Screening Programme

The National Health Service (NHS) offers a second trimester ultrasound scan to pregnant women through the Fetal Anomaly Screening Programme (FASP) [90]. The purpose of the scan is to check the health of the fetus who is usually between 18^{0} to 20^{+6} weeks old. During the scan, anatomical structures of the fetus are assessed by accredited sonographers who are required to measure and capture specific anatomical planes. Typically, 30 minutes is allocated for the scan. These

standard planes are also sometimes referred to as the gold standard, where specific anatomical structures must be visible to be considered as a 'textbook' image. In practice, sonographers may not always capture a gold standard plane because of differences in fetal position, maternal anatomy and time constraints. However, the captured plane is sufficient for diagnostic purposes.

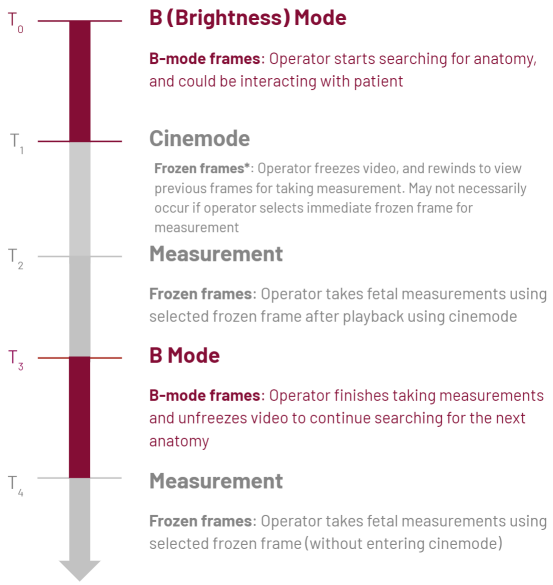


Figure 1.1: An example timeline of ultrasound scanning. In this example timeline, the sonographer started performing their scan in B mode (between T_0 and T_1), before freezing the video. Before taking measurements, the sonographer rewinds the video through the buffered frames. These buffered frames are known as cinemode frames (between T_1 and T_2). Once they have found the frame they want to take measurements on, they begin the measurement phase (between T_2 and T_3). After measuring, they unfreeze the video and continue scanning (between T_3 and T_4).

There are several different modes that the ultrasound video can be used in. This thesis used the brightness mode (B-mode) frames (Figure 1.1 between T_0 and T_1, T_4 onwards) which consist of both unfrozen and frozen frames; frozen frames can be measurement and or cinemode frames (Figure 1.1 between T_1 and T_3). When the video is not frozen, sonographers are navigating or searching for the required anatomical plane and adjusts the ultrasound probe based on what they see on the display of the ultrasound screen (Figure 1.1 between T_0 and T_1, and, T_3 and T_4). Frozen frames occur when sonographers have frozen the video on the ultrasound machine to take a measurement of the anatomical plane (Figure 1.1 between T_2

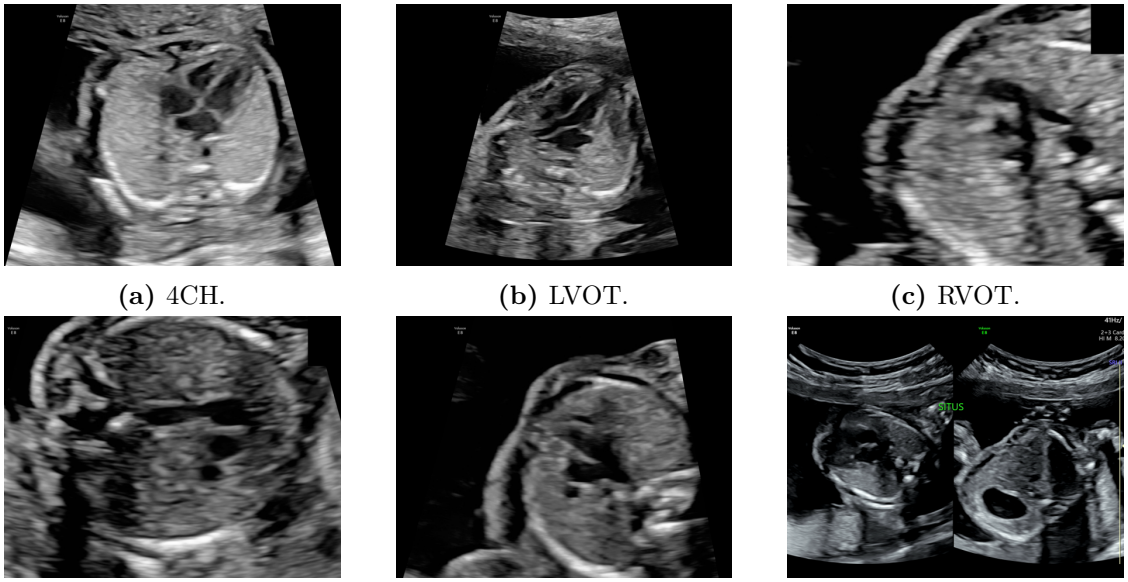
and T_3). If the sonographer rewinds the video to obtain a higher quality image, cinemode (buffered) frames will be generated(Figure 1.1 between T_1 and T_2). An example of the scanning process is shown in Figure 1.1.

1.4.1 Standard Planes

During second trimester ultrasound scans, several standard imaging planes are captured and measured by sonographers [90]. I mention those which are relevant to this thesis, which are the abdomen, brain and heart plane views (Figure 1.2).

According to [90], the 6 heart views which are assessed include the situs, the 4 chamber view (4CH), the aorta/left ventricular outflow tract (LVOT), the pulmonary/right ventricular outflow tract (RVOT), the 3 vessel view (3VV) and the 3 vessel and trachea view (3VT). The situs is a heart view used to determine whether the fetus is cephalic (head facing upwards with respect to the cervix) or breech (head facing towards the cervix). Typically, the sonographer will take a view of the heart (left of the image) and move the probe towards the abdomen (right of the image) to determine the orientation of the fetal organs with respect to the maternal orientation. I specifically mention the situs because the situs view presents itself differently (Figure 1.2f) from other heart views 4CH, LVOT, RVOT, 3VV and 3VT on the ultrasound machine.

The 2-head views which are assessed are the suboccipitobregmatic view demonstrating measurement of the transcerebellar diameter (TCB) (Figure 1.3b) and the head circumference and atrium of the lateral ventricle, also known as the transventricular plane (TVP) (Figure 1.3a). The abdominal view which is assessed is the abdominal circumference (AC) (Figure 1.3c).



(a) 4CH. **(b)** LVOT. **(c)** RVOT.

(d) 3VV. **(e)** 3VT. **(f)** Situs

Figure 1.2: Standard plane heart views.

(a) TVP. **(b)** TCB. **(c)** AC.

Figure 1.3: Head and abdomen standard plane views.

1.5 Perception Ultrasound by Learning Sonographic Experience (PULSE)

The Perception Ultrasound by Learning Sonographic Experience (PULSE) project (ERC-2015-AdG-694581) started in 2016 with data collection approved by the UK Research Ethics Committee (Reference 18/WS/0051). Written informed consent was given by all pregnant women who participated. The women were at least 19 years old when they came in for their routine ultrasound scan, and up to 20 different sonographers participated in the study [127].

The project aims included to build assistive technology through more powerful

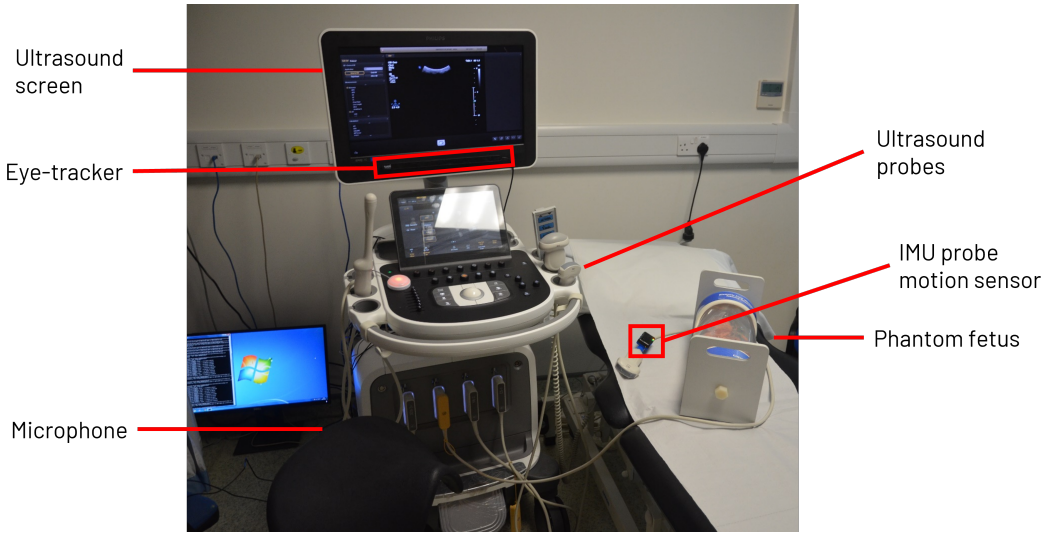


Figure 1.4: PULSE equipment set-up in the John Radcliffe Hospital, Oxford. The eye-tracker is mounted underneath the ultrasound machine screen. The microphone was placed near/on the computer set-up. The probe motion was recorded using an inertial measurement unit (IMU) which was mounted onto the cable of the probe. The position of the phantom fetus is where the pregnant woman would lie down during the scan.

interpretive models than was previously possible by only analysing ultrasound videos and images. The project recorded 4 synchronised data sources: ultrasound videos of first, second and third trimester routine clinical fetal ultrasound scans, sonographer gaze using an eye-tracker, sonographer probe motion using an inertial measurement unit (IMU) which was mounted onto the ultrasound probe, and audio recordings of sonographers providing qualitative descriptions of fetal images during the scans [127, 94]. Figure 1.4 shows the set-up of the PULSE equipment, where the probe motion was recorded using a mounted IMU, the eye-tracker mounted underneath the screen, and the microphone placed within appropriate vicinity of the sonographer scanning [125].

The processing of obtaining fetal images during the ultrasound scan can be described as follows [127]. The sonographer adjusts the probe whilst looking for the standard plane, while receiving real-time visual feedback from the video on the screen. They re-position the transducer based on what they see, which changes the video on the ultrasound machine screen. This thesis uses only second trimester fetal ultrasound video scans and the sonographers' corresponding eye-tracking data, focusing on the visual aspect of the sonographer 'loop'.

The ultrasound machine used was a commercial General Electric (GE) Healthcare Voluson E8 or E10 with a sampling frequency of 30Hz, or frames, per second [127]. The video files were recorded using lossless compression [126]. An eye-tracker (Tobii Eyetracking Eye Tracker 4C, Danderyd, Sweden) was mounted under the ultrasound machine screen, and the calibration procedure is described in [110]. Sonographers do not have any knowledge of whether the eye-tracker is functioning [127]. The eye-tracker samples gaze data at 90Hz. The eye-tracker is used to record the sonographer's gaze, as on the ultrasound machine screen, while they are scanning.

The eye-tracker equipment was set up and calibrated prior to the start of the doctorate. Since this thesis uses the eye-tracking data as the primary data modality, a brief description of the calibration parameters are listed here. Where possible, I have included the corresponding details below.

- Eye tracker model/specifications: *Tobii Eyetracking Eye Tracker 4C, Danderyd, Sweden.*
- Sampling frequency, either of eye tracker itself, or for analogue systems the sampling frequency of any AD conversion (or some such): *90Hz.*
- A description of the setup and geometry: *Mounted at the bottom of a 1920x1080 pixel ultrasound machine screen using a magnetic mount and silicon adhesive.*
- Calibration procedure: *9-point calibration per sonographer [110] who participated in the study.*
- Environmental conditions: *Eye-tracking data was recorded in a hospital room at the John Radcliffe Hospital, where the lighting conditions (due to curtains drawn during the scanning procedure) remain largely unchanged during and between scans.*

The accuracy and precision of the eye-tracker was investigated in [110]. In their work, they performed an in-situ study with 3 sonographers. I report their metrics below.

- Median accuracy: *0.65 degrees, (30.1 pixels).*

- Median precision: *0.09 degrees, (4.5 pixels).*

In [110], the authors found that there was no effect on the precision of the eye-tracker, while they reported a loss of accuracy of 0.16 degrees between calibration and later use. Note that the eye-tracker was calibrated for each sonographer who participated in the study [136]. In this thesis, the main assumption is that the eye-tracker's accuracy validated using the procedure described in [110, 127] is accurate for the analysis performed.

1.6 Definitions

A list of definitions which are used globally in the thesis are given below. These definitions are specific to the methods presented in this thesis.

- Segment: A group of adjacent and consecutive frames in the fetal ultrasound video between time t and $t+n$.
- `PULSENet`: A fetal ultrasound frame classification model which was trained on the PULSE ultrasound videos [142].
- `pulsepytools`: A `Python` toolbox built by Dr. Richard Droste which extracts and pre-processes the PULSE project data collected between 2018 - present [126].
- Standard plane: A standard plane view of an anatomy must contain key anatomical structures specified by [90].

A list of eye-tracking specific definitions which are used globally in the thesis are given below.

- Gaze point: The position of a user's attention on a 2-dimensional screen captured using an eye-tracker. The gaze point is recorded using x and y co-ordinates abbreviated as G_x, G_y.
- Gaze velocity: The difference between G_x, G_y at time t and $t + \delta t$ divided by δt, where δt is the inverse of the sampling frequency of the eye-tracker.
- Fixation: The eye is focused on a particular object or area [161].

- Saccade: The eye is moving rapidly from one point of interest to another [161].
- Post saccadic oscillation: Ocular instability just after a saccade which occurs before reaching a steady-state value [6, 45].
- Smooth pursuit: The fovea is actively following a moving object [161].
- Areas-of-interest (AOI): an AOI is a specific anatomical landmark that the sonographer has looked at while performing the scan.

1.7 Acknowledgements

I am grateful to my collaborators and colleagues who have contributed to the work in this thesis. With thanks to the following colleagues:

- Dr. Lior Drukker and Jayne Lander for their guidance in navigating the clinical interpretation of the results.
- Professor Aris T. Papageorghiou who was the co-PI of the PULSE project.
- Dr. Richard Droste whose work in building `pulsepytools` was instrumental in helping me pre-process the eye-tracking data.
- Dr. Harshita Sharma, Dr. Lior Drukker and Dr. Pierre Chatelain whose labelled data was used in Chapter 4.
- Dr. Richard Droste, Dr. Jianbo Jiao, Dr. Lok Lee Hin and Dr. Zeyu Fu for providing the data which was used in Chapter 5.
- For their contributions to Chapters 4 and 5, Dr. Lok Lee Hin, Jayne Lander, Dr. Harshita Sharma. For their time and initial guidance that eventually formed Chapter 5, Professor Min Chen.
- For their contributions in Chapter 6, Dr. Harshita Sharma.
- For proof-reading the publication which formed Chapter 7, Dr. Qianhui Men and Dr. Mohammad Alsharid.

2
Literature Review

Contents

2.1 Visualising Eye-Tracking Events for Videos 13
2.1.1 2-Dimensional (2D) Eye-Tracking Visualisation for Videos 15
2.1.2 Non-Parametric Classification of Eye Movements 18
2.1.3 Challenges in Fetal Ultrasound 19
2.2 Task Classification: Medical Applications 21
2.2.1 On Surgical and Fetal Ultrasound Video Differences . . 21
2.2.2 Classification of Medical Tasks using Gaze 22
2.2.3 Challenges in Fetal Ultrasound 24
2.3 Skill Classification 25
2.3.1 Aggregated Eye Movement Characteristics 27
2.3.2 Feature Engineered Eye-Tracking Data 27
2.3.3 Pupillometry 28
2.3.4 Challenges in Fetal Ultrasound 29

In this chapter we review literature relevant to the thesis.

2.1 Visualising Eye-Tracking Events for Videos

The purpose of video visualisation is to provide a visual representation of important events and features that occurred over time in the video. It is used as a tool to summarise meaningful information for the end user using graphical representations, one example being glyph-based visualisation [34]. In eye-tracking specifically,

visualisation systems can be used to identify unique user scanpaths, user specific fixation/saccadic behaviour and areas-of-interest [47].

There are many factors to consider when designing an eye-tracking visualisation as outlined in [47]. Some that are relevant to fetal ultrasound are:

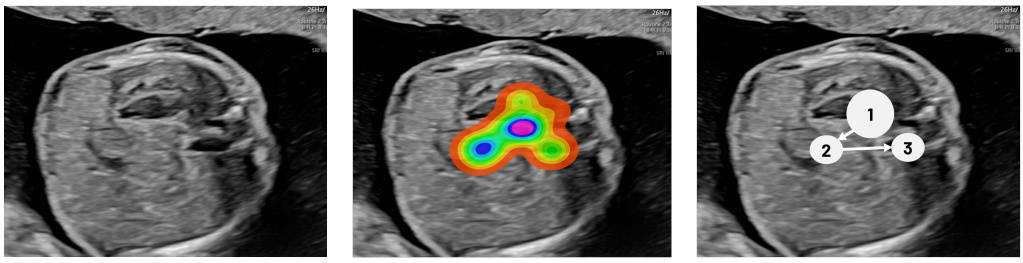
- the nature of the stimuli - static or dynamic. Examples of static stimuli are images, and dynamic stimuli appear when watching a video. In fetal ultrasound, the fetus moves based on the probe movement and the nature of the stimuli is said to be dynamic.
- the desired representation of recorded gaze data in the constructed visualisation. In fetal ultrasound, both spatial and temporal visual representation of gaze characteristics are desirable.
- the dimensions of recorded data. In the PULSE project, the gaze data is recorded in 2-dimensions - an x and y co-ordinate.
- the number of participants. In the PULSE project, there are multiple sonographers of different years of experience whose data was collected.

Eye-tracking visualisation methods follow a general pipeline of 1) classification of eye movements, 2) calculation of eye movement characteristics and 3) visualisation of aggregated gaze behaviour. Once raw eye-tracking data is separated into their respective eye movements, a set of aggregated metrics is calculated. Examples include, length of fixations, time to completion of task, average radius of fixation (in degrees). These metrics are collated to produce a suitable visualisation method, depending on what the user is interested in. For example, the transition between fixations or spatio-temporal order of areas-of-interest visited. The final visuals are used to inform the user of differences between participants, such as classification of skill. Because of the complexity involved across applications, there is no one-size-fits-all data visualisation solution [75, 72, 31].

Eye-tracking visualisations are usually presented in 2D or 3D. As 3D systems are not trivial to implement, in this literature review I consider current 2D visualisation methods which use video (dynamic) stimuli (Section 2.1.1). These methods also

typically rely on the successful classification of the different types of eye movements fixations/saccades/smooth pursuits, or a suitable expert who can label them. Hence, I also consider non-parametric methods of classifying eye movements (Section 2.1.2), since using an off-the-shelf algorithm such as Tobii's algorithm [38] might not always be suited for the specific application, and expert labellers might not be available for every study.

2.1.1 2-Dimensional (2D) Eye-Tracking Visualisation for Videos



(a) Frame of an abdomen plane.

(b) Heatmap superimposed on the abdomen plane in Figure 2.1a

.

(c) Order of areas-of-interest visited, labelled using the numbers. The size of the circle indicates how long was spent at each area-of-interest.

.

Figure 2.1: An example of the types of heatmap eye-tracking visualisations, using the abdomen plane as an example. In Figure 2.1a I show the original abdomen frame. In Figure 2.1b, I show a heatmap of the eye-tracking data superimposed on the original frame. The heatmap shows the concentration of the gaze points recorded by the eye-tracker around the stomach, umbilical vein and aorta. In Figure 2.1c, the areas-of-interest are numbered based on the order that the sonographer looked at the landmarks, and the size of the circle indicates how long they spent there. These figures are for illustration purposes only.

There are several popular methods which are used to display eye movement characteristics in 2D such as the attention heatmap [32] (Figure 2.1) and scarf plots (Figure 2.2). Some of these methods focus on spatial characteristics, while others incorporate both spatial and temporal information. Location of fixation and saccade is represented in x and y co-ordinates, while time can be represented with a different attribute such as colour. When creating a visualisation for videos, there are added complexities such as dynamic stimuli, where there can be multiple

objects of interest which change in size and location over time. Plotting multiple heatmaps could be used for multi-participant videos. Videos could be broken down into different tasks being performed to perform a between participant comparison. However, the spatial changes in gaze might not be easily captured for long videos where tasks are not easily defined by a fixed time period. In those instances, a suitable temporal measure such as length of time is required.

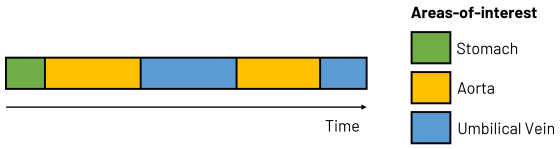


Figure 2.2: An example of scarf plot created using 3 areas-of-interest on the abdomen plane: stomach, aorta and umbilical vein. The colour identifies the anatomical landmark, and the length of the colour bar indicates how long was spent at that landmark. Illustration created in Microsoft PowerPoint using [92] as an example.

For certain applications, the temporal order of objects of interest visited (Figure 2.1c) is important information to compare between groups of participants. Another key element of visualisation design to consider is how to display multiple participant characteristics while avoiding visual design clutter which make interpretation difficult.

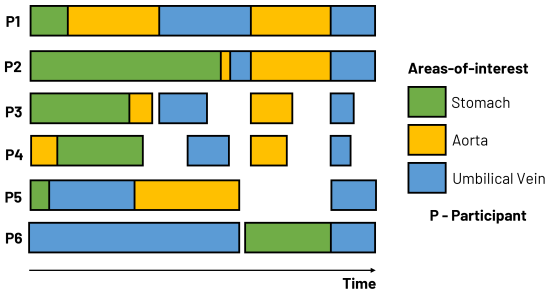


Figure 2.3: Example of a storyline visual. This is similar to the scarf plot in Figure 2.2, but the main idea behind a storyline visualisation is to compare the behaviour between participants. Here the white gaps indicate that the sonographer had not been looking at any of the areas-of-interest stomach, aorta or umbilical vein. PX refers to a unique participant. Illustration created in Microsoft PowerPoint using [71] as an example.

Methods which calculate areas-of-interest display the temporal order of areas visited using timeline / storyline visuals (Figure 2.3) [63, 71, 50], nodes [66] and

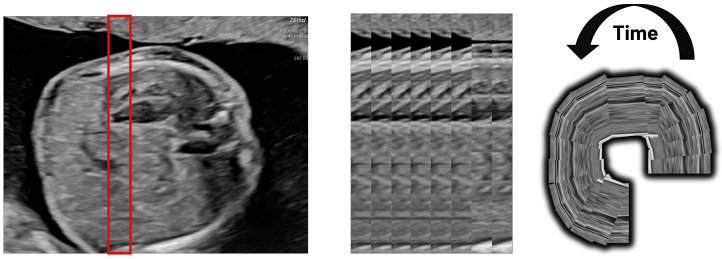


Figure 2.4: Example of spiral visualisation which uses 'slices' of the image. The left most image shows the abdomen plane, with a red box over a thin slice of the image. The red box's position is determined based on the eye-tracking x co-ordinate. The height of the box is arbitrary. In other similar visualisations, instead of a slice a cropped image can be used. The middle image shows several slices taken over a period of time. On the right most image, the slices are now collated to form a spiral, where the direction of time is anti-clockwise. Illustration created in Microsoft PowerPoint using [64] as an example.

scarf plots [41] (Figure 2.2). Other methods are annotation free and do not identify specific areas-of-interest, such as [64, 145, 64, 61, 79] which represents images as thumbnails, or uses clustering-based methods to identify areas-of-interests [102, 57]. More labour intensive methods design dedicated platforms which are integrated with the eye-tracker and computer [118, 50].

[79] uses Hilbert maps to visualise spatial and temporal characteristics of gaze behaviour. [64, 145, 64] represents each frame as a small slit (horizontal or vertical) taken at the x (or y) co-ordinate of the recorded gaze point. [145] uses a spiral visualisation and does not require any specific annotations. Temporal order is represented in the anti-clockwise direction (Figure 2.4). [71] uses dynamic time warping to represent similarities between areas-of-interest and clusters the scanpaths into a 'storyline' visualisation. In storyline visualisations, time is represented on the x-axis, and an ordered preference of objects viewed is plotted on the y-axis.

[50] uses the 3D space-time cube [44] to produce a timeline based visualisation of areas-of-interest visited. [41] uses a Sankey diagram to represent multiple areas-of-interest and the behavior of participants over time.

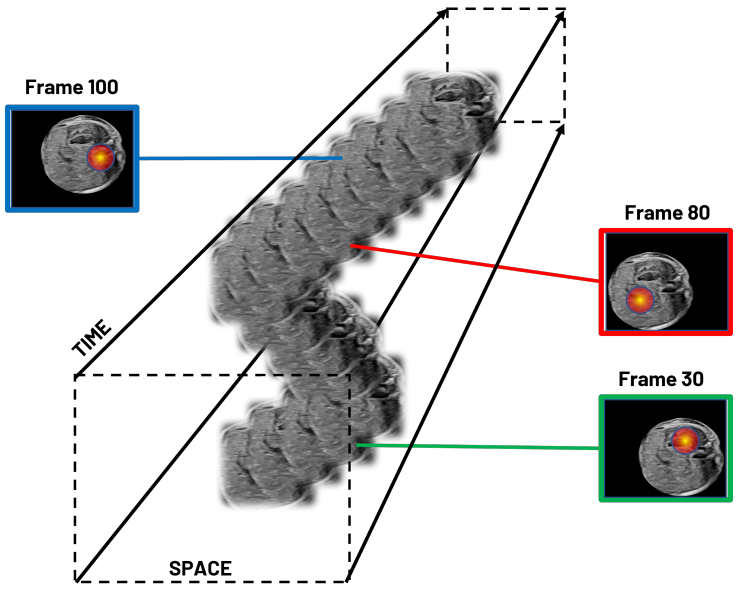


Figure 2.5: Example of a space-time cube used to visualise eye-tracking data. In this illustration, the abdomen plane is used. In the space-time cube, both the spatial aspect and temporal aspect are displayed. The spatial element is meant to convey where the plane is with respect to the screen's location, and the temporal element shows how the image changes over time. Some key frames are also highlighted and identified (in blue, red and green), where the cumulative eye-tracking data until that point is shown as a heatmap. Illustration created in Microsoft PowerPoint using [95] as an example.

2.1.2 Non-Parametric Classification of Eye Movements

Some of the eye-tracking visualisation literature discussed relies on having a suitable eye movement classification algorithm. These algorithms classify different eye movements based on whether a threshold, either gaze distance or velocity, is exceeded. [13] outlines 5 eye movement classification algorithms which form the building blocks of many other threshold-based algorithms [67]. Briefly, these are usually written as 'I-XX', where I stands for identification, and XX represents the type of algorithm being used. In this thesis, I refer to Tobii's [38] Velocity-Threshold Identification (I-VT) most frequently. Their algorithm considers gaze points corresponding to ≤ 30 degrees/s as a fixation, otherwise as a saccade.

Suitable thresholds may not always be available based on the study's requirements. Hence researchers have implemented non-parametric methods for separating fixations/saccades/smooth pursuits from raw eye-tracking data [49, 37, 43]. In their methods [49, 37, 43], they do not choose a specific threshold based on expert

knowledge or rely on in-built eye-tracker algorithms [38]. [28, 70] provide an overview of how choosing the threshold metrics can return vastly different results. Being able to choose thresholds or parameters which are informative and accurate for the application is challenging.

Current methods include custom quantitative metrics [28, 43], distribution based methods [37, 55, 22], probabilistic methods [67] and machine learning based methods [106, 93, 49]. Quantitative metrics, distribution based methods and probabilistic methods aim for a 'sweet spot', where the metric or distribution value reaches a steady state value. Machine-learning based methods use random forest [93], convolutional neural networks [106] and clustering to separate eye movements.

[28] created a group of behaviour scores which allowed a set of optimal threshold values to be determined when fixations and saccades are present, which was followed by [43] who used [28]'s existing metrics and additional scores to include smooth pursuits. [43]'s customised metrics combined of classical eye movement methods. [67] uses Bayesian decision theory to automate classification of fixation/saccades/smooth pursuits in real-time from low frequency eye-trackers. Unlike classical threshold-based methods, their work uses a probabilistic approach based on eye speed. [55] uses a 3-step segmentation algorithm to identify fixations and smooth pursuit events. Their work uses the Rayleigh test which tests the hypothesis that the direction vectors between samples are uniformly distributed around a unit circle. [93] uses a random forest to classify fixations/saccades/post-saccadic oscillations, and performs post-processing heuristics to return the final labels. More recent work uses advanced machine learning techniques such as deep learning. For example, [106] who used convolutional neural networks to classify fixations/smooth pursuit/noise using an open source labelled dataset [27].

2.1.3 Challenges in Fetal Ultrasound

Eye-tracking visualisation that are designed specifically for videos usually involve a complex set-up. These equipment are not trivial to implement in a clinical setting where the participants can be patients, clinicians or both. In fetal sonographer,

to utilise methods like the space-time cube [50, 44] which allow for post-analysis of gaze behaviour, the video that participants watch have to be identical so that their behaviour can be compared. However, the fetus presents itself differently in each scan. No two fetal videos will look exactly alike.

Applications which visualise gaze behaviour while watching videos used benchmark datasets like that of [51] use coloured videos with defined objects of interest for the participant to follow. For example, a car driving along the road or a group of people play a game of UNO (which have colours) [51]. Fetal ultrasound images are recorded in greyscale. This makes annotation free methods like [145] difficult to use because each slit of the frame would look similar to the other slits. The slit also needs to be wide enough to provide overall context of the anatomical landmark that the sonographer is looking at. For planes like the heart plane where the aortas and pulmonary arteries are spatially close, segmentation of these landmarks is not trivial. Other methods that require labelling of areas-of-interest would be labour intensive.

To reduce the eye-tracking labelling efforts involved for a fetal ultrasound video, non-parametric methods could be used to separate the eye tracking data into their respective eye movements. However, the current non-parametric methods are not easily extendable to the PULSE data because of sampling frequency differences and length of data available on a task-basis. This is because the fetal ultrasound videos are recorded 'in-the-wild' with pregnant women and sonographers, and there are also more than 20 standard planes which need to be captured over a 30-40 minute period. As a result, the time of capturing a specific plane could be as short as a few seconds [135]. Conversely, methods like that of [37, 49] require a reasonable amount of data being available per participant, per task.

There is a need to find a method that can analyse the gaze behaviour of sonographers, taking into account the complexities such as grey-scaled images, fast transition between landmarks of interest and the low sampling frequency of the eye-tracker. This question is explored in this thesis (Chapters 4 and 5).

2.2 Task Classification: Medical Applications

The literature on task classification is usually closely intertwined with skill assessment, so I will discuss task classification and skill assessment in separate sections. Task classification is used to identify the specific task being performed. In surgical applications this could be tying a knot whilst closing a suture [68]. Whilst using a sewing machine [111], this could be adjusting the knob's settings to increase tension in the thread. In fetal ultrasound, a task is defined as the sonographer searching for a specific anatomy plane (for e.g. TCB or TVP). The literature around task classification using eye-tracking can be segregated based on the use of a) images or videos, and b) the use of either eye-tracking (raw/feature engineered/calculated eye movements) or pupillometry [138].

I focus my literature review on medical applications using eye-tracking in videos, such as fetal ultrasound [126] or surgery [26], and images such as radiology [108], for example chest X-rays [18, 21] and breast mammograms [10].

2.2.1 On Surgical and Fetal Ultrasound Video Differences

Surgical skill data science is an example of a field in parallel to that of sonography data science [147]. I briefly comment on the similarities and differences between surgery and fetal ultrasound and why the methods used in surgical research cannot necessarily be applied directly to analyse fetal ultrasound videos.

Surgery and ultrasound are similar in the following ways. Surgical skill motions can be broken down into smaller gestures such as passing the needle through the suture, knot tying and clipping [68, 20]. Similarly, fetal ultrasound probe motion can be described as 6 general motions slide, rock, sweep, fan, pressure/compression and rotation [60]. Both use hand-eye co-ordination, requiring interpretation of medical images while manipulating their hand/probe to achieve the desired view. For example, the use of psychomotor skills [52] involving the use of visual attention (eyes) and hand motion to manipulate the tools (scalpel and probe) and reading of medical videos. There are instances where the sonographer/surgeon looks away from the video, for example when changing instruments [20] or when the sonographer

interacts with the patient [127]. Finally, simulators (or phantoms [58] in fetal ultrasound and cadavers [26] in surgery) are also widely used for training and experimental studies to ensure patient safety.

In spite of these similarities, there are also some key differences. The process of carrying out surgery is more structured than that of fetal ultrasound. Each surgical task is well defined and considers a specific point of entry when performing the surgery e.g. the nose in sinus surgery [26]. Beginning and ending points of fetal ultrasound is dependent on sonographer skill, maternal anatomy and fetal anatomy. Movements of the object of interest (fetus in fetal ultrasound, and anatomy in surgery) also differ: a fetus can be actively moving in response to the probe, in surgery the patient is typically under anesthetic and hence still. Consequently, it is not easily replicable to use eye-tracking methods for surgery in fetal ultrasound research. The fetus is also small in comparison to an adult patient, and correspondingly their organs are also much smaller than an average adult. In surgeries which use endoscopes, optic cameras are attached to obtain a clear view of the patient's anatomy whilst performing the surgery [26, 20]. Ultrasound imaging has its own challenges such as acoustic shadows [119]. For example, during fetal ultrasound scanning, the fetal bones can cast shadows onto the region of interest. In those cases, it may not be possible to get a good standard view of the fetus and in this case the sonographer would capture a view that is sufficient for diagnostic purposes.

2.2.2 Classification of Medical Tasks using Gaze

Images and Pupillometry. In some studies, the scanpath is superimposed onto an image [137, 144, 96] (Figure 2.6) . Gaze points can also be convolved with a Gaussian kernel to return a visual attention map or saliency maps [137, 109, 113, 69, 139].

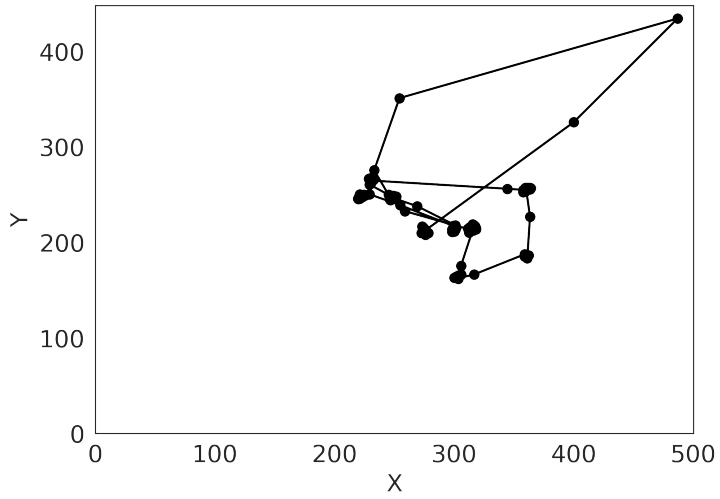


Figure 2.6: Example of gaze scanpath superimposed onto an image for task classification. X and Y represent the gaze x and y co-ordinates. The gaze points are connected using a lineplot.

Pupillometry, the study of the measurement of pupil size and reactivity, has been used to understand cognitive load or operator work load [146]. Larger changes in pupil diameter are associated or indicative of a higher cognitive load. Pupillary activity can be calculated using the raw pupil diameter to compute metrics such as Low/High Index of Pupillary Activity [114] (LHIPA), Index of Pupillary Activity [85] (IPA) or Task Evolved Pupillary Response [5] (TEPR). [116] uses a combination of pupillometry and raw eye-tracking to generated hand-crafted and engineered features to discriminate patients with and without dementia. Their work shows that using eye-tracking can be used to detect significant biomarkers of cognitive differences between patients with and without dementia. [136] uses both pupillometry and eye-tracking to classify different fetal ultrasound tasks, brain and heart planes, and found statistically significant variation between sonographer expertise groupings and their pupillary responses. Their results suggest that more experienced sonographers have lower cognitive workloads. The use of pupillometry has been studied in real-world ultrasound scanning conditions [136]. The authors have noted that it is challenging to control environmental factors which can lead to errors in the observed data, specifically in the context of pupil data because the pupil diameter is sensitive to changes in environmental conditions such as light.

Raw and Feature Engineered Gaze Points. Gaze data can be used in its raw form or feature engineered into useful features for task classification. For example, gaze points can be discretised [58, 26] to return a sequence. The gaze data can be discretised into a sequence of numbers. For example, raw gaze points can be clustered using k-means and return a string of label clusters [26]. Another example of is computing a sequence of labelled states, where the states are identified by unique fixations visited in temporal order [58]. Gaze points are also commonly used to calculate eye movements such as fixations, saccades and smooth pursuits [130, 129, 58, 73, 96], and correspondingly metrics such as time spent per type of eye movement such as fixations, and transition between fixations [134]. This can be done using off-the-shelf algorithms such as Tobii's I-VT algorithm [38] (defined in Section 3.4.2) or non-parametric methods [37, 28]. In some studies, this is also done manually by suitable experts. Some medical areas which have been researched include radiographs [73], autism [96], fetal ultrasound [58], dementia [116], schizophrenia [33] and porcine laparoscopic cholecystectomy [20].

2.2.3 Challenges in Fetal Ultrasound

The works discussed above are not applicable or easily extendable to fetal ultrasound for the following reasons. They largely depend on having a suitable algorithm or expert to separate different types of eye movements, namely fixations, saccades and smooth pursuits. This can be very labour intensive because no two fetal ultrasound videos present in the same manner. Studies which superimpose the scanpath onto the image being read rely on the image/object-of-interest being fixed in time and space, as opposed to reading a video. Typically, these studies use simulated environments or are conducted in highly controlled environments. There are opportunities to repeat the experiment if they are unable to collect the data properly due to equipment malfunction, and also do not necessarily account for external influencing factors from their surroundings such as patient-clinician interaction. In the PULSE project, data is being collected 'in-the-wild' whilst the sonographer is performing the ultrasound scan. Prior work has been carried out

using image-based methods [137] but images are computationally expensive to use for classification compared to time-series data. There is a need to find a method that can classify eye-tracking behaviour of sonographers when searching for different anatomical planes considering only eye-tracking data, in its raw form or feature engineered form. This question is explored in this thesis (Chapter 6).

2.3 Skill Classification

There are 2 different aspects of skill that I will discuss in this subsection. First, the definition of skill used across studies. Secondly, the types of different methods used to classify skill.

Defining Skill. The definition of skill in the medical literature is most often quantified by the number of years of experience the clinician has been practicing for. In fetal ultrasound, this typically corresponds to the number of years after qualification. For example, if a sonographer has been scanning for 2 years or less [123, 137, 131], then they are considered newly qualified, while someone who has been scanning for 10 years is considered an expert. [35, 29, 99] considers experts as those who performed a minimum number of surgeries, and novices as those who did not meet the threshold. [128, 82, 141, 23, 25] considers those at a specific stage of their career to be experts. For example, [82, 141] considers expertise based on number of semesters completed, with fellows as experts. [23] considers those with no prior experience as novices, those with 1 year of training as intermediate and others as experienced readers. In other medical fields such as surgery, where possible, studies quantify skill based on the number of instances the specific surgery is performed. [117, 86]. In dentistry, the number of semesters completed by the students is used as a measure of skill [82, 141]. This time-based definition of skill can also be found in surgical literature [115, 91, 98, 26, 124].

This time-based measure of skill does not necessarily take into account other factors that can contribute to skill. For example, the frequency of scanning could decrease over time as the sonographer takes on additional responsibilities later in

their career. Another limitation of this time-based definition is that it does not consider the differences (and hence difficulty) between types of task. A 2-year and 10-year experienced sonographer could be equally skilled in searching for the head or abdomen plane, since it is considered an easier plane to search for. However, the differences in skill could be more separable when searching for the heart plane, where multiple views of the heart need to be examined and the probe movement to reach these different slices of the heart involves fine movements.

Methodologies. In medical studies where eye-trackers have been used, researchers use metrics such as number of fixations, time taken to complete the task, number of saccades to differentiate between groupings of clinicians [91, 141]. They also use statistical properties of the distribution of fixation and saccadic properties (e.g. mean length of fixation, median length of fixation) to determine if the two groups (newly qualified and expert) are significantly different [98, 16]. I discuss studies which use pre-calculated metrics in Section 2.3.1. These studies have the unique characteristic of using a predefined eye movement classification algorithm or suitable experts to separate eye-tracking data into fixations, saccades, smooth pursuits and areas-of-interest (defined in Section 1.6).

Studies which utilise raw eye-tracking data, or some form of feature engineered eye-tracking data (not including the pre-calculated metrics above) are discussed in Section 2.3.2. These studies do not separate eye-tracking data into separate eye movement types before classification.

Eye-trackers are also able to record the pupil diameter, which is referred to as pupillary data and the study of which is referred to is pupillometry. Pupillary data is more commonly used to assess cognitive workload in participants, where a larger dilation in pupil diameter is an indicator of higher cognitive workload. It is used to compare differences between varying levels of difficult tasks and consequently, an indirect measure of skill level of clinicians [117]. I discuss studies which utilise pupillary data only in Section 2.3.3.

2.3.1 Aggregated Eye Movement Characteristics

There are many works which consider skill using aggregated eye movements to characterise skill. For example, [128] consider arthroscopic surgery using cadavers, dentistry expertise in reading orthopantomograms [82, 141], detecting lesions [23], visual search patterns in colonoscopists [25], different phases of micro-neurosurgery [35] and laparoscopic cases [29, 99]. In their work they use a combination of eye movement characteristics such as blinks, fixations, saccades and pupil diameter to separate experts and non-experts at performing the clinical task. [82] compares the reading scanpaths of dentists and trainees, using pre-labelled stimuli to measure instances of participants observing the correct area of interest.

There are several works which use statistical modelling methods such as [35, 99, 25, 23], who use statistical models such as analysis of variance (ANOVA) to discriminate skill levels between groups of clinicians. There are also other works which use machine learning models such as support vector machines [128], long-short term memory deep learning models [82], linear discriminate function analysis [141] and nonlinear neural network analysis [29] to classify skill groupings. In [136], authors use both statistical models and machine learning models to classify skill.

2.3.2 Feature Engineered Eye-Tracking Data

Unlike previous works mentioned in Section 2.3.1, there are several studies which use raw eye-tracking data for skill classification as opposed to aggregated eye movement characteristics. In some studies, the eye-tracking data is combined with other data such as tool motion data recorded using sensors. These have been researched in applications such as surgery and fetal ultrasound [17, 19, 123].

[26] uses a Hidden Markov model to separate 95 experts and 139 novices using both their eye-tracking and tool motion data. [30] fits a statistical model to eye-tracking and tool motion data for separating 7 experts and 13 novices in endoscopic sinus surgery. [137] uses a combination of raw eye-tracking, pupillary data and image data to classify newly qualified and expert sonographers. They use convolutional neural networks and consider uni-modal and multi-modal data for 2 different tasks.

Their results show that majority of their multi-modal models outperform uni-modal models. [26] considers an expert as a surgeon who has knowledge of sinus anatomy structure and operation of the endoscope. Novices are those without prior endoscopic experience. [137] separates skill using a 2 year experience threshold.

2.3.3 Pupillometry

In fetal ultrasound, [137, 136] uses gaze data to compare different years of experience of sonographers for two different tasks, searching for the brain and heart anatomical plane. [136] suggests, as measured by pupillary data, that different expertise and tasks in fetal ultrasound show significant statistical variation. In parallel, in surgical studies, [146, 117, 86, 46] use pupillary data to assess differences in skill for thoracostomy, laparoscopy, laparoscopic Roux-en-Y gastric bypass and ophthalmoscopy skills respectively. In [146], they found that when the trainees' required help, their pupillary response showed a difference compared to when they were 'performing normally'. Their results suggest that pupillary response is an important indicator of when a trainee is struggling with a particular task. In [117], they found that using metrics such as larger pupil size, indicating a greater cognitive workload, could be used objectively to label the difficulty of surgical tasks. In [86], experts focused more on important areas-of-interest while having a reduced cognitive workload. Finally, in [46] experience was a significant factor in differences in pupillary response.

[117, 86] consider different skill levels of surgeons based on the number of procedures they performed. [117] define experts as having >50 procedures, and novices without any experience. In their study, they used 16 surgeons (5 experts, 11 novices) watched surgical videos with 8 different steps. [86] consider experts as clinicians who completed >75 procedures and junior as those without. In total, they had 12 junior, 8 expert surgeons and they watched 20 procedures. [117] show that novice surgeons have higher average pupil diameter during the duration of the video compared to expert surgeons. [86] show that experts have a smaller maximum pupil size during the operation segments, a reduced mental workload and increased

concentration. [46] considered medical students, residents and attending physicians (75 in total, but specific breakdown not specified) and 3 different factors which could affect performance: experience, frequency of cases and viewable fundus field range. Their work suggests that experience is a factor which affected performance. [146] consider only a single group of trainees for easy and difficult tasks and their results show differences in behaviour depending on task type. In their work they did not specify the number of trainees who participated. These studies focused specifically on participants watching surgical videos [117] or using simulated data [46]. In [46] there were a large number of unique participants, but they did not specify the percentage breakdown of the number of students, residents and physicians. In [86] the surgeons watched 20 procedural videos.

2.3.4 Challenges in Fetal Ultrasound

The literature I have described in this section on skill classification relies on several assumptions which are not necessarily available in fetal ultrasound. Eye-tracking studies separate raw eye-tracking data into different eye movements. However, research has shown that results are parameter dependent [13]. In the studies cited above, simulations or videos are also often used to quantify differences between expert and non expert. However, due to the differences (between fetal ultrasound videos and other similar medical domains) I mentioned in Section 2.2.3, it is not always possible to consider a like-for-like comparison between an expert sonographer's gaze behaviour and a non-expert.

There is also an open question around the time-based definition of skill. In fetal ultrasound, skill result in events such as fast probe movement and transitions between anatomical planes. They also depend on sonographer experience, and maternal and fetal anatomy [127]. In fetal ultrasound, there is an open question on how skill is defined, and whether time-based measures are sufficient. Following on, whether it is possible to use eye-tracking data to quantify skill without making prior assumptions about the groupings of expertise. This question is explored in this thesis (Chapter 7).

3

Datasets and Pre-processing Methods

Contents

3.1 Software Packages 31
3.2 A Brief Description of `pulsepytools` 32
3.3 Datasets 33
3.3.1 Manually Labelled Second Trimester Scans 33
3.3.2 PULSENet Standard Planes 36
3.3.3 Identification of Heart Standard Planes using Optical Character Recognition 37
3.3.4 Trainer-Trainee Sessions 40
3.3.5 Fully Qualified Sonographer Scan Sessions 40
3.4 Pre-processing Methods 41
3.4.1 Image Augmentation 41
3.4.2 Eye-Tracking Data 42

In this chapter, I introduce the software packages used to pre-process the PULSE data. The data collected are 'in-the-wild', therefore the ultrasound videos need to be annotated at an anatomical plane level for analysis, and eye-tracking data processed to account for any tracking errors that occur during the scan; tracking errors occur when the eye-tracker did not record any gaze points at a specific point in time whilst the sonographer was scanning. My colleague Dr. Richard Droste (also part of the PULSE project) created a toolbox `pulsepytools` to process the data; the functionality relevant to my thesis will be described.

The ultrasound videos which were used for the methods presented in Chapters 4 to 7 are described. Finally, the methods used to pre-process ultrasound images and sonographer eye-tracking data are discussed. The recorded ultrasound videos were annotated using different labelling methods. The data was not labelled by the same group of annotators each time, and the final quantity of labelled data was also driven by the project's need of what labels were required then. Hence, the number of videos which were annotated differs. The labelled datasets also did not have an equal number of anatomical planes for each anatomy which were captured during the scan. For example, the abdomen has a single anatomical plane, the brain has 2 planes and the heart has 5 planes (Section 1.4.1). To increase the dataset size and improve robustness for image-based models, image augmentation methods were used which are described in Section 3.4.1.

3.1 Software Packages

The software packages that were used to process the data is listed below. The packages `COBYLA` and `pytesseract` were chosen by [126] who built `pulsepytools` to analyse the ultrasound videos and eye-tracking data collected in PULSE. A detailed explanation of how these packages were used is found in Section 3.2. `difflib` was used in Section 3.3.3 as a supplement to the OCR algorithm that I built to detect sonographer text written on the images.

- `COBYLA` [8] A numerical optimisation method where the gradient of the objective function is unknown. The parent package of `COBYLA` is `SciPy` (version 1.7.0).
- `difflib` [158] A `Python` package that compares two different words and returns a score metric that measures the degree to which the two words are similar (version 0.18.0).
- `pytesseract` [157] An OCR algorithm that was first developed at Hewlett-Packard Laboratories Bristol UK and at Hewlett-Packard Co, Greeley Colorado USA. Since November 2018, `pytesseract` has been developed and maintained by Google (version 0.3.9).

3.2 A Brief Description of `pulsepytools`

I give a brief overview of the functionality in `pulsepytools` that was used in this thesis. `pulsepytools` was a toolbox built by Dr. Richard Droste [126] during his doctorate and was used to extract and process the PULSE data from the data storage server.

Resampling 90Hz Eye-Tracking Data to match 30Hz Video Frequency. The eye-tracking and video sampling frequency were 90Hz and 30Hz respectively. `COBYLA` [159, 9] was used to estimate the geometric median of these 3 gaze points to return a single gaze point per video frame.

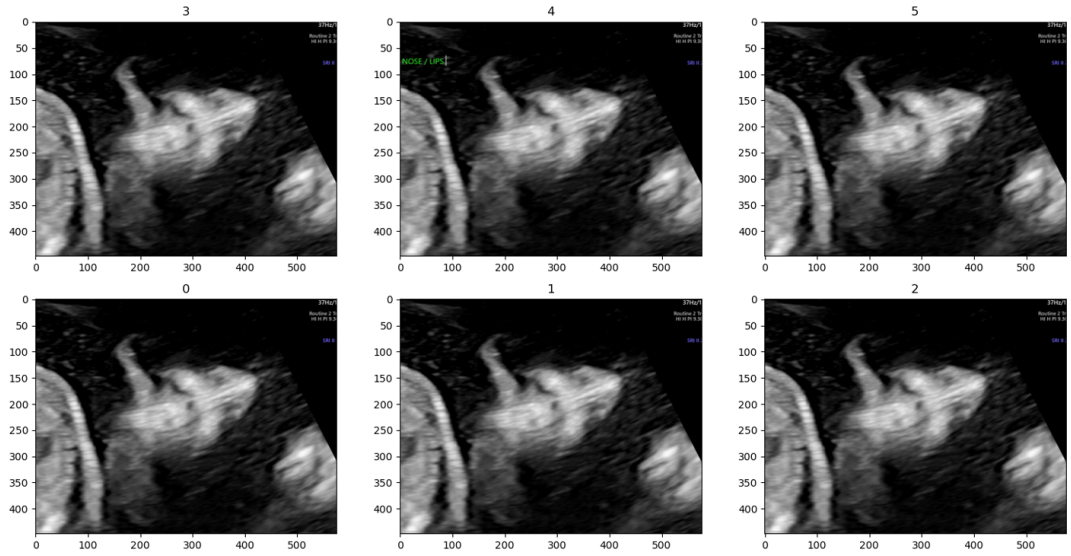


Figure 3.1: Example of the GUI displayed to the annotator. The GUI was used to manually label anatomical planes. The GUI selects a frame with biometric measurements and 5 other frames based on the annotated frame using predefined offsets. The user is able to choose from a set of abbreviations which anatomical plane is displayed. If they are unsure, they can skip the frame and move to the next annotated frame. The example shown has been annoymised and displays the ultrasound image only without the frames corresponding clipboard annotations.

Graphical User Interface for Extracting Standard Planes. A graphical user interface (GUI) (Figure 3.1) was built by Dr. Richard Droste [126] to extract and label frames of full-length ultrasound scans. The GUI first selects frames which

have been annotated with biometric measurements by the sonographer. For each measurement segment, (n, m), where n and m are the start and end frame of the measurement segment, the GUI selects 5 other frames based on predefined offsets. The offsets are $n-60$, $n-9$, $n-8$, $n-7$, $n+30$, $m-30$. At a 30Hz video sampling frequency, these offsets correspond to 2 seconds before the measurement segment ($n-60$), and just before freezing ($n-9$, $n-8$, $n-7$). The frozen frame 1 second after freezing the video ($n+30$) and 1 second before unfreezing the video ($m-30$) is also displayed. These specific frames would be able to show which anatomy the sonographer was searching for just before freezing. Since the frozen frame does not change unless the sonographer rewinds the video after, the frozen frame just after freezing and just before unfreezing is displayed. The GUI displays these 6 frames and allows the user to choose a label based on the appearance of these 6 frames using specified abbreviations e.g. *kk* for kidneys. The full list of abbreviations can be found in Table 3.1.

3.3 Datasets

3.3.1 Manually Labelled Second Trimester Scans

The dataset presented in this section was used in Chapters 4 and 6. The research presented in Chapters 4 and 6 was carried out in late 2020 and early 2021. This labelled datasets was made available from earlier works of colleagues who wanted to investigate the clinical workflow of sonographers [104]. The workflow analysis included which anatomical planes were searched for the most frequently, and whether there was any specific ordering to the planes sonographers were looking for. Subsequently, full-length second trimester scans were manually labelled at an anatomy level, where the labels did not distinguish between different views of the anatomy. For example, the heart was labelled as 'Heart' and was not further separated into the different views RVOT, LVOT, 3VV, 3VT and 4CH (Section 1.4.1).

The manually labelled segments contained 150 frames, corresponding to 100 frames before freezing and 50 frames after freezing. These were labelled using the procedure set out in [121], and performed by 2 engineering researchers and 1 clinical

research fellow using the GUI presented in Section 3.2. In total, there were 25 different labels (Table 3.1). A sample of labelled images are shown in Figure 3.2. The overall inter-annotator agreement for labelling is 79.7% [121].

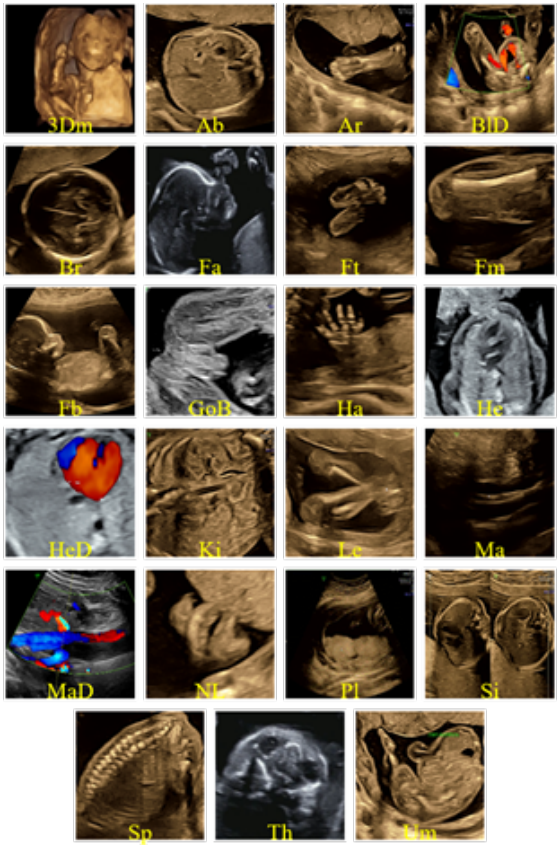
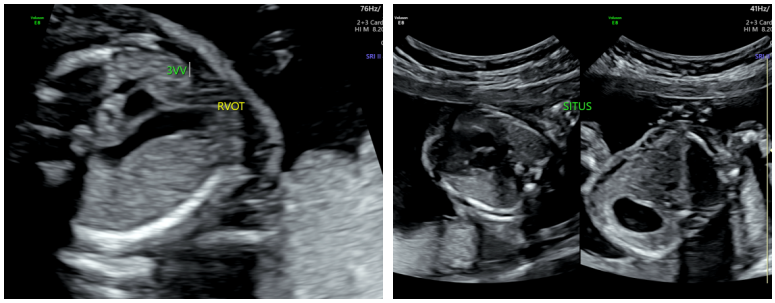


Figure 3.2: Overview of Anatomy Labels. Figure reproduced from [105].

When I qualitatively investigated the appearance of these segments, I found several images which contained a heart and abdomen view side by side. In the later stage of my doctorate, I learnt that these views are the situs view (Figure 3.3b, Section 1.4.1). Since the sonographer has only displayed the heart view on half of the screen, and the abdomen on the other half, the heart view was half the scale of the other heart views (Figure 3.3). For this dataset, the situs views were not included since they presented differently from other heart views and the corresponding sonographer's gaze would present differently as well. For example, the sonographer would focus their gaze on one half of the screen at any time in comparison to reading the image at its full size. There were several labelled segments

Label name	Abbrv.	Comments
3D and 3D+t mode	3Dm	Views taken in static or real-time 3D mode, showing surface rendering of the fetal head and face.
Abdomen	Ab	Fetal abdomen (with biometric measurements).
Arms	Ar	Fetal arms.
Background search	Bk	Operator quickly froze-unfroze as they did not finalise the frozen (or saved) frame as a standard view during search.
Bladder with Doppler	BlD	Fetal bladder (including Doppler mode).
Brain with skull head and neck	Br	Fetal brain (with biometric measurements).
Face side profile	Fa	Side view of the fetal face.
Feet	Ft	Fetal feet.
Femur	Fm	Fetal femur (with biometric measurements).
Full body side profile	Fb	Full-body side views of the fetus. May include face, hands, heart, ribs, spine, diaphragm.
Girl or boy	GoB	Views to determine fetal sex.
Hands	Ha	Fetal hands.
Heart	He	Fetal heart without Doppler mode.
Heart with Doppler	HeD	Fetal heart with Doppler mode.
Kidneys	Ki	Fetal kidney (including Doppler mode).
Legs	Le	Fetal lower legs.
MiscellaneousMaternal anatomy	Ma	Maternal uterine artery without Doppler mode.
MiscellaneousMaternal anatomy with Doppler	MaD	Maternal uterine artery with (pulse) Doppler mode.
Mixed	Mx	Clip containing views (frames) of more than one annotation label, representing abrupt scene changes.
Front face with nose and lips	NL	Fetal front face showing nose or lips or both.
Placenta	Pl	Placenta (with biometric measurements).
Situs	Si	Situs
Spine	Sp	Fetal spine (may be full spine or part of spine).
Top head with eyes and nose	Th	Top of the fetal head showing eye sockets and/or nose.
Umbilical cord insertion	Um	Insertion of the umbilical cord.

Table 3.1: Description of Manual Labels. Reproduced from [105].



(a) 3VV **(b)** Situs

Figure 3.3: Comparison of a 3VV heart view (a) against a situs heart view (b), where the situs heart view shows the heart on the left side of image (b), and a corresponding abdomen plane on the right side of image (b).

which did not have 150 frames as set out in [104]. This could have been due to an error in the pre-processing pipeline, or that the time period between frozen frames were shorter than 100 frames. For consistency, segments which had 100 and 50 unfrozen and frozen frames labelled respectively were included.

3.3.2 PULSENet Standard Planes

The dataset in this section was used in Chapter 5. The dataset was created because my colleagues [142] wanted to build a deep-learning model (`PULSENet`) to classify fetal standard planes using domain-specific knowledge and characteristics of the fetus. Having a fetal ultrasound standard plane classification model meant that tasks such as clinical workflow analysis and identification of the desired standard plane is made much faster and easier. In a fetal ultrasound scan video, this is not insignificant as each video can be up to 30-40 minutes long. At a 30Hz sampling frequency, this corresponds to at least 54,000 frames to filter through. My colleagues used `pulsepytools`'s GUI to extract the anatomical planes [126]. The research presented in Chapter 5 was carried out in late 2021 and early 2022.

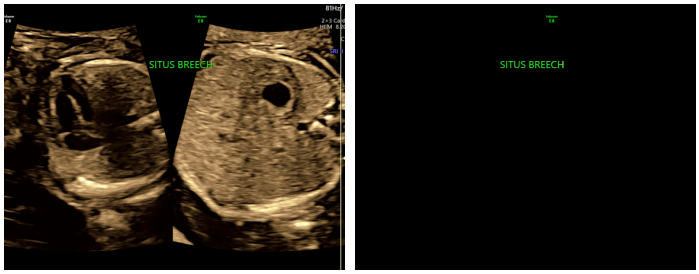
The labels generated for the manually labelled second trimester scans (Section 3.3.1) did not separate the different planes for each anatomy; for example, the 2 head planes were labelled as 'Brain' (Table 3.1). To create view-specific labels, [142] annotated a selection of standardised anatomical views based on prior work [97] which included: 3VT, 4CH, RVOT, LVOT, TCP, TVP, two views of the spine

(coronal and sagittal), abdomen, femur, kidneys, lips, profile and background. They selected frames that did not contain any sonographer annotation on the ultrasound frame so that the network would not learn the text on the image that contained the abbreviation annotated by the sonographer. An example of sonographer annotation on the frame can be seen in yellow and green text in Figure 3.3.

3.3.3 Identification of Heart Standard Planes using Optical Character Recognition

Whilst verifying the labels generated by the `PULSENet` team, it was found that some of the heart view labels were noisy. Some of the labels were incorrect and some of the key anatomical structures required to achieve a standard plane (Section 1.4.1) were not present in several frames. The latter means that the frame does not qualify as a clinical standard plane. The sonographer can easily capture multiple heart views in the same video segment by manipulating the probe with fine movements. By selecting the heart view a few frames before the annotated frame, it is not surprising that some frames were mislabelled. For example, if the sequence of heart views captured by the sonographer was 3VV, LVOT and RVOT, a frame labelled as an LVOT could be a 3VV view. Instead of manually labelling the frames (Section 3.3.1), or using the GUI presented in Section 3.2, I used OCR to identify the different heart views: 3VV, 3VT, 3CH, LVOT, RVOT and Situs. This is possible because sonographers can label the frozen segments of videos using text (Figure 3.3a). The OCR aims to read these sonographer-annotated labels. The dataset presented in this section was used in Chapter 7.

The Python package used for OCR is `pytesseract` [155] (Section 3.1). The training data for `pytesseract` used images with black text on a white background which made it challenging to use the model on fetal ultrasound frames directly; ultrasound images have low contrast and are majority grey pixels, and the text is in a shade of yellow or green. Some of the suggestions in [157] were used to improve the visibility of the text on the ultrasound frame to maximise the performance of the `pytesseract` model on the frame.



(a) Original frame **(b)** Green text on black background **(c)** Black text on white background

Figure 3.4: Process of frames being pre-processed for maximum OCR. From left to right: (a) Original frame with text 'SITUS BREECH', (b) Non-green pixels converted to black pixels, (c) Green pixels converted to black pixels, and black pixels converted to white pixels.

Pre-processing Frames. To improve the accuracy of the text detected by `pytesseract`, the frames were pre-processed before running the OCR algorithm on the frame. The original text abbreviation on the frame is green or yellow. First, a green/yellow mask was created to detect any green/yellow pixels in the frame (Figure 3.4, b). Then, any pixels which were not green/yellow were converted to black, and green/yellow pixels were converted to white. This step was carried out because it was empirically found that `pytesseract` performed better with black text on a white background compared to white text on a black background. Finally, the white and black pixels were inverted; any white pixels were converted to black and black pixels to white (Figure 3.4, c).

Curating Frames to Label. The ultrasound videos that need to be processed have a mean of 12983 frames and a median of 12667 frames in length. To reduce the computational resources required to run the OCR model on every frozen frame, every 45th frame blue (corresponding to 1.5 seconds at 30Hz sampling frequency) of each frozen segment in a full-length scan was labelled. For example, if the frozen segment corresponds to frame numbers (100, 200) inclusive, the 100th, 145th, 190th frame were labelled. 45 was chosen as a suitable interval because the video sampling frequency is 30Hz. If the sonographer annotates a frame, it would be present for at least 1 second but could be less than 2 seconds. By sampling every 1.5 seconds,

at least 2 out of 3 seconds of the video will pass through the labelling process, ensuring that any sonographer written text is picked up.

Rule-Based Filtering. Empirically, it was found that using the pre-processing steps was not sufficient to identify the text on the frames correctly and several frames were still mislabelled. One reason was because there was other text present in the frame such as the measurements, and removing such text is not a trivial task even after removing the clipboard annotations; the clipboard annotations are located on the left side of the screen where sonographers can save frames.

In particular, '3VV' and '3VT' were often mistaken for each other. `pytesseract` was also not able to correctly identify 'VV' as 2 separate V's and returned 'W'. `pytesseract` also often detected 'SL' which was a half-written 'SITUS' label ('SI'), i.e. the sonographer had not completed the text. 'LVOT' was also often picked up as 'WOT'. Some miscellaneous text such as 'WW' was also detected by the OCR, which was misidentified as '3VV' or '3VT'.

To increase the accuracy of identifying the views correctly, a rule-based filter was implemented after the text was read by `pytesseract`. These rules were based on empirical observation, where `pytesseract` was unable to read the characters properly. Before implementing these rules, special characters (e.g. %, #) that appeared in the read text were removed, and all alpha numeric characters converted to uppercase. The rules can be found in the Appendix 8.4.

The final frame label was chosen as the closest matching label, with a probability of at least 80%, compared to a list of pre-defined labels using the Python package `difflib` [158]. The probability threshold was manually selected. The pre-defined labels were 4CH, Situs, LVOT, RVOT, 3VV and 3VT. As a final check, a manual inspection of all frames was performed to ensure that the final label corresponded to the sonographer's annotation.

3.3.4 Trainer-Trainee Sessions

To analyse task-agnostic differences in skill, where a task is a specific anatomical plane being searched for, data were collected from not yet fully qualified sonographers learning how to scan with a fully qualified sonographer. The dataset was used in Chapter 7. This data presented in this section was collected from March 2022 until July 2022, after the PULSE project was able to resume data collection due to Covid-19 restrictions.

The videos in Table 3.2 were scans performed by a fully qualified sonographer training a trainee (not-yet fully qualified) sonographer how to scan. This involved a teacher (fully qualified sonographer) and the trainee present during the scan, where the teacher would show the trainee how to scan. During the training sessions, due to time constraints, the trainee does not necessarily perform the scan but is instead given opportunities to try searching for planes with some guidance from the teacher. In total, there were 4 unique trainees (Table 3.2) and a single teacher with 5 years of scanning experience at the time of these sessions.

	Trainee 1	Trainee 2	Trainee 3	Trainee 4	Teacher
# of pregnant women	6	6	1	1	14

Table 3.2: Number of unique pregnant women scanned by a teacher training a trainee sonographer.

3.3.5 Fully Qualified Sonographer Scan Sessions

To complement the trainer-trainee dataset, the fully qualified sonographer scan sessions dataset was curated to serve as the expert population of sonographers for skill analysis. This dataset was necessary because the fully qualified sonographer in the trainer-trainee dataset was accompanied by a trainee sonographer. The presence of the trainee can affect the way that a teacher would perform the scan since the sonographer is also interacting with the trainee for teaching purposes. In contrast, the fully qualified sonographers in the second trimester scan performed the scan individually and reflect more typical behaviour of a sonographer whilst scanning. The dataset is used in Chapter 7.

The total number of second trimester scans performed by the cohort of PULSE sonographers are shown in Table 3.3. Table 3.3 does not include the teacher-trainee videos in Table 3.2 and were performed by fully qualified sonographers individually.

Years of experience	0	1	2	3	5	6	7	8	10	11	14	15	16
# of pregnant women	136	115	33	8	22	16	5	4	18	13	104	39	2

Table 3.3: Number of unique pregnant women scanned by fully qualified sonographers.

3.4 Pre-processing Methods

3.4.1 Image Augmentation

Image augmentation is a technique that is used to increase the size of a dataset available to train a machine learning model. Its purpose is to increase the robustness of the trained model by ensuring that the model is able to learn invariant features. For example, the model should be able to correctly classify two ultrasound images that have the same anatomical landmarks but different orientations. Given that the number of unique views of an anatomy is different (Chapter 1.4.1), for example, the brain has 2 views to be scanned and the heart has 5, the number of planes available for each anatomy is imbalanced. To analyse differences in sonographer behaviour when scanning for the abdomen vs. heart, for example, the dataset needs to be balanced first. As such, in this section, I describe the image augmentation methods which were applied on manually labelled second trimester scan dataset presented in Chapters 5 and 6 on the `PULSENet` standard plane dataset, as set out by [132].

Augmenting Images by Flipping. To increase the size of the manually labelled dataset in Chapter 6, I augmented the images by flipping the images about the horizontal, vertical, and horizontal and vertical axis.

The method presented in Chapter 6 uses only eye-tracking data to train the model, and not images. Consequently, the orientation of the augmented image is important since sonographers have to capture the anatomical plane in a specific orientation [90], and other augmentations such as rotation or translation would

not have been appropriate.For example, the TVP and TCB plane needs to be captured horizontally and centered on the ultrasound machine to ensure that the head circumference can be measured accurately. Hence these specific types of augmentations were chosen for the dataset in Chapter 6.

Principled Data Augmentation of Images. In Chapter 5, the `PULSENet` standard plane dataset was augmented using principled data augmentation [132] to increase the size of the PULSENet standard plane dataset and ensure robustness for the image-based model presented. In [132], authors found that using 3 random augmentations of scale 3 outperformed conventional augmentation methods like that of flipping. The magnitude of the augmentation depends on the type of augmentation that is performed. For example, a translation of magnitude 3 means that the image is translated by 3 pixels (in the positive or negative direction).

3.4.2 Eye-Tracking Data

The recorded eye-tracking data needs to be pre-processed before being used to build any models. The methods below describe industry standard methods which are used to process raw eye-tracking data. Eye-tracking data is usually used in terms of gaze angle, the change in angle between two gaze points. Subsequently, the change in gaze angle over a unit of time is used as gaze velocity. In the PULSE data, one unit of time δ_t is used as $\frac{1}{90}$ seconds (at a 90Hz sampling frequency).

In this section I describe how the change in gaze angle between time t and $t + \delta_t$ is calculated, as the gaze angle is used frequently in this thesis. I also describe the pre-processing and feature engineering methods used to generate normalised eye-tracking data, and interpolation methods for filling in missing values.

Gaze Angle. This is defined as difference between a gaze point at time t and $t + \delta_t$, calculated in degrees or radians. Gaze angles are used to calculate gaze velocity, which is typically used to separate different types of eye movements. The method of calculating the change in gaze angle between time t and $t + \delta_t$ is given

in Algorithm 1. δ below refers to the phrase 'change in' for e.g. δ pixel per mm refers to change in pixel per mm.

Algorithm 1 Pseudocode showing how to calculate the gaze angle between two gaze points at time t and $t + \delta_t$.

```
x_pixel_mm ←w / w_res                                          ▷ Constant
y_pixel_mm ←h / h_res                                          ▷ Constant
unit_pixel_mm ←sqrt(x_pixel_mm^2 + y_pixel_mm^2)               ▷ Constant
δd_pixels ←sqrt(δx_pixel^2 + δy_pixel^2)
δd_mm ←unit_pixel_mm × δd_pixels
Gaze angle ←tan^−1(0.5 * δd_mm / d)
```

- d: distance from screen to user (mm)
- δx_pixel: difference in G_x between time t and $t + \delta_t$
- δy_pixel: difference in G_y between time t and $t + \delta_t$
- h: height of screen (207 mm)
- w: width of screen (332 mm)
- h_res: screen height in pixels (1080 pixels)
- w_res: screen width in pixels (1920 pixels)
- unit_pixel_mm: change in mm per pixel (mm/pixel)

Zero Padding Eye-Tracking Data.

In Chapter 6, the presented method requires that the time-series data are of equal length. To satisfy this constraint, any eye-tracking segments which are missing gaze points just before freezing, for example, the 100th unfrozen frame in the segment, or due to natural variable length of sequences [137], were zero-padded to create equal-length time-series. For a visual example of where this method is used, see Figure 3.5.

Linear Interpolation of Eye-Tracking Data.

Zero padding should not be used for missing gaze data where tracking errors occurred whilst the video is unfrozen, since the missing gaze was not due to natural variable length data. The missing gaze points are filled using linear interpolation on [38], and is used in Chapters 4 to 7.

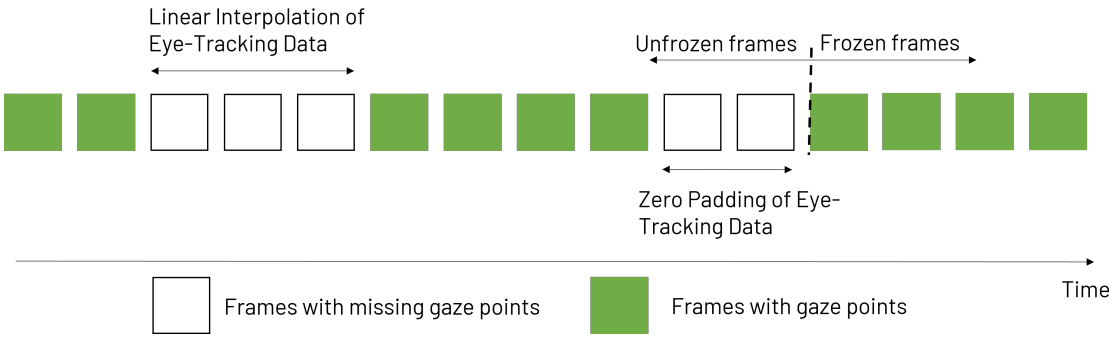


Figure 3.5: Overview of Eye-Tracking Interpolation Methods using Linear Interpolation and Zero Padding.

First, a step size is calculated as the difference between time t and $t+i$. The inverse of the step size, $\frac{1}{t-t_i}$, is multiplied by the difference in raw gaze point $G_t - G_{t_i}$. This is multiplied by the number of steps, $(n-i+1)$ and summed with the gaze point at t_i to return the interpolated gaze point value at t_{i+n}. n is the index where the gaze point is missing.

$$G_{t_{i+n}} = \frac{1}{t_i - t} \times (G_{t_i} - G_t) \times (n - i + 1) + G_{t_i} \tag{3.1}$$

where $t <$ n $< t_i$. For a visual example of where this method is used, see Figure 3.5.

Pupillary Pre-processing Method.

Tracking errors of an eye-tracker would also affect any pupillary data which was recorded. However, any recorded pupil diameter that is out of range of the human eye needs to be discarded before linearly interpolating any missing values. The pre-processing method is used in Chapter 7.

The pupillary data pre-processing method used is outlined in [136]. Any pupil diameters <1.5mm and >9.0mm was discarded, and any missing values were linearly interpolated. For the interpolation, only gaps <210ms (or 7 frames at 30Hz) were interpolated. The final pupil diameter was smoothed using a Gaussian window with a standard deviation of 1.

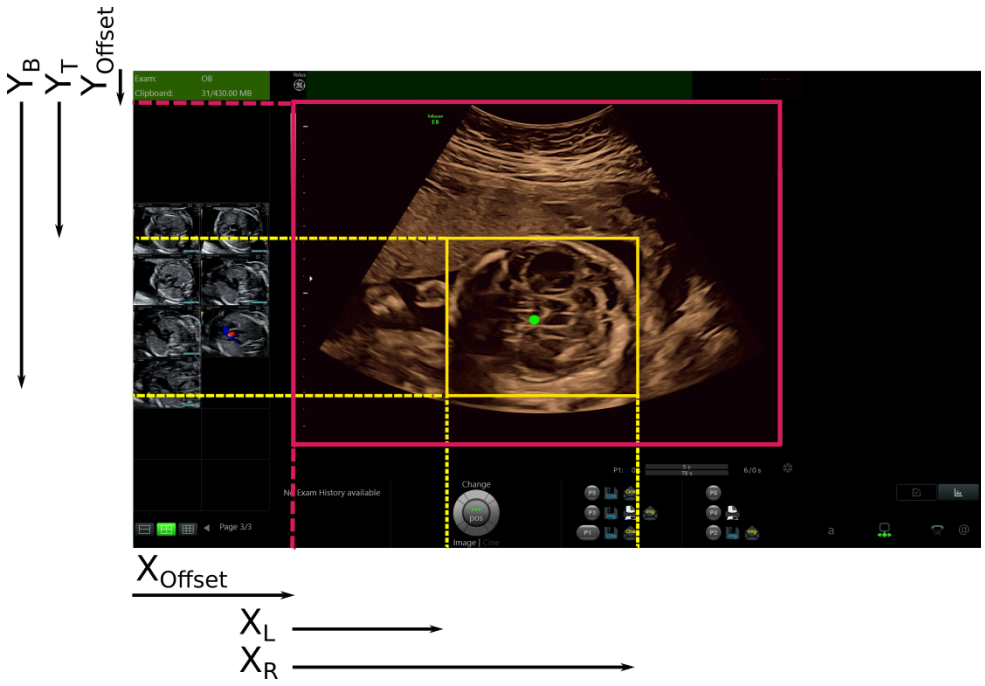


Figure 3.6: An example showing how a raw gaze point (green) with co-ordinates G_x, G_y is normalised with respect to the hand drawn bounding box (yellow). The point of origin of the bounding box is given as the bottom left corner.

Scale and Position Invariant Eye-Tracking Data (Feature Engineering).

Whilst scanning, a sonographer spends the majority of their time looking specifically at the anatomical view and landmarks that they are interested in. This view is normally centered on the screen, but can also be off-center. The same landmark, for example the aorta, could be at different positions of the screen depending on the scan. There is a need to normalise the gaze data such that observing the same landmark, even when the view of the plane does not occupy the same space on the screen, returns the same co-ordinates. To provide contextual information to the gaze data recorded by the eye-tracker, eye-tracking data is normalised with respect to the anatomy to return scale and position invariant gaze points between multiple scans. This eye-tracking processing method is used in Chapter 5 and Chapter 6.

Manual bounding boxes using `OpenCV` [156] are drawn around the circumference of the anatomy plane on a cropped (1008×784 pixels) image (shown as the red box Figure 3.6). All text and clipboard images are excluded to view the circumference

clearly. Then, gaze points are normalised with respect to the corner positions of the bounding box along the x and y axis: X_L, X_R, Y_T, Y_B where L, R, T and B represent left, right, top and bottom. The X and Y offsets (shown as X_{offset}, Y_{offset} in Figure 3.6) is 427 and 66 pixels respectively. An example of this normalisation process is shown in Figure 3.6.

Raw gaze points recorded by the eye-tracker along the x and y axis with respect to the screen dimensions of 1920×1080 pixels are defined as G_x, G_y. Raw gaze points normalised by co-ordinates of a hand drawn bounding box on the image are given as G_{xBB}, G_{yBB} (Equation 3.2).

$$G_{xBB} = \frac{G_x - X_L - X_{offset}}{X_R - X_L} \quad \text{and} \quad G_{yBB} = \frac{G_y - Y_B - Y_{offset}}{Y_T - Y_B} \tag{3.2}$$

I-VT Algorithm.

Tobii's velocity-threshold fixation identification (I-VT) algorithm [38] is the in-built algorithm in Tobii's eye-trackers used to separate fixations and saccades. It has also been used in previous fetal ultrasound eye-tracking studies and therefore serves as a baseline for the method in Chapter 4. Their algorithm to separate eye movements works as follows. Any eye movement greater than 30 degrees/s will be classified as a saccade, and less than that will be classified as a fixation. Any missing data is interpolated using the procedure described in Section 3.4.2. Tobii's velocity-threshold fixation identification (I-VT) algorithm is referred to in Chapter 4.

4

Population Level Visualisation of Spatial Temporal Gaze Characteristics of Sonographers

Contents

4.1 Introduction 47
4.1.1 Example Visualisation of Gaze and Ultrasound Frames . 48
4.1.2 Contribution 51
4.1.3 Data 51
4.1.4 Definitions 52
4.2 Methods 52
4.2.1 Determining Areas-of-Interest (AOIs) using Unsupervised Clustering 52
4.2.2 Visualising Scanning Characteristics in the Spatial and Temporal Domain 57
4.3 Results 62
4.4 Discussion 65
4.5 Summary 66

4.1 Introduction

Data visualisation provides a way to understand what story the data is telling – it can be used to spot data trends and subsequently anomalies. Eye-tracking visualisations help us understand what was observed when a participant was reading

an image or watching a video. These visualisations potentially assist to differentiate participant behaviour when performing different tasks, and in clinical applications help differentiate more experienced clinicians from less experienced clinicians. The added complexity of analysing gaze behaviour of sonographers is that the fetus may move and change its position during the scan. In addition, even though the sonographer is observing the same anatomy structure, for example the brain, there are several different parts of the brain that they observe. Taking these factors into account, we want to come up with a way to understand both spatial and temporal characteristics of where and how sonographers look when they perform their visual search. Broadly, this could be useful when training new sonographers to help them locate important landmarks that will help them reach a desired standard plane.

Visualising longitudinal data collected from projects such as PULSE adds an additional challenge because of the amount of data available for analysis. Extracting meaningful summary statistics whilst maintaining the level of granularity of data desired is a trade-off between how simple, and easy to implement, and complex the final visual is. In this chapter, I aim to find a balance to understand gaze behaviour of sonographers at a population level. Population-level analysis enables us to learn general gaze characteristics demonstrated by the sonographers while searching for a specific anatomical plane.

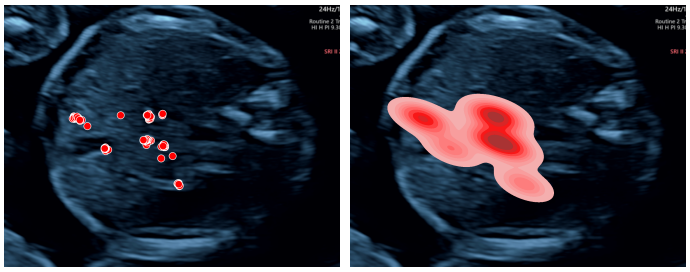
Current challenges to visualising 'eye-tracking data' (discussed in Section 2.1.3) include: the implementation complexities involved in using visualisation methods that analyse videos, and the requirement of annotated eye movements. In this chapter I present a visualisation method that helps us overcome these constraints.

4.1.1 Example Visualisation of Gaze and Ultrasound Frames

Before introducing the method, I present some examples here to show some of the current challenges in applying visualisation techniques to gaze and ultrasound frames. In this example, I have used an abdomen gaze segment of 100 frames and plotted the gaze points as scatterpoints and heatmaps. The first frame in this segment is frame 16584, and the last frame is 16684.

A more general note on the differences between a point (or binary) representation (scatterpoint) versus a distribution representation (heatmap). The scatterpoints allow the user to see precisely where the sonographer has looked at, where the color of the point represents a specific position of the image the sonographer's gaze focused on. The heatmap provides an approximate distribution of the sonographer's attention, where the colours represent the different probability levels of the sonographer looking at a certain area of the image.

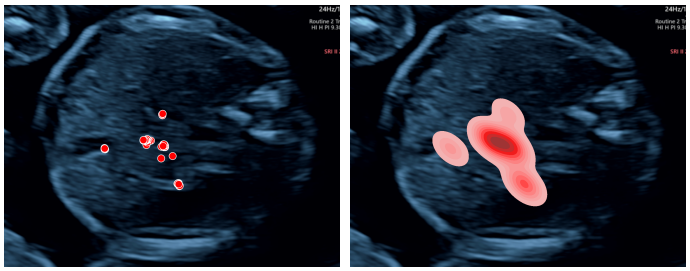
The first example in Figure 4.1 I present is a summary of all gaze points from frames 16584 to 16684 which were superimposed on the last frame 16684. The gaze points are represented as scatterpoints and a heatmap. The plots show some of the areas of the abdomen which the sonographer had looked at whilst scanning. Spatially, these visualisations appear to have provided a reasonable estimation of what the sonographer had looked at.



(a) Scatter: Frame 16584. **(b)** Heatmap: Frame 16684.

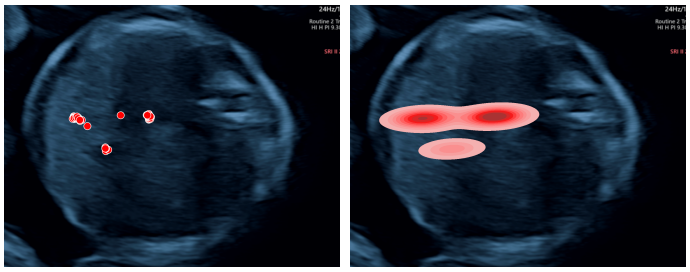
Figure 4.1: Gaze points from frame 16584 to 16684 plotted as a scatterplot and heatmap on the final frame 16684.

I also investigate the temporal gaze patterns to determine whether the representation in Figure 4.1 is sufficient. I plot the middle frame 16634, with gaze points recorded from frames 16584 to 16634 in Figure 4.2. Some limitations of using the presented visualisation methods for fetal ultrasound videos start to be seen here. The image itself has changed over time, where the landmarks are not as clear, since the sonographer is still finishing their search. Also, the gaze pattern presents differently in Figure 4.2, where the gaze attention in Figure 4.1 is clustered on the left side but the attention in Figure 4.2 is horizontal.



(a) Frame 16684. **(b)** Frame 16684.

Figure 4.3: Gaze points from frame 16634 to 16684 plotted as a scatterplot, contour and heatmap on the last frame 16684.



(a) Scatter: Frame 16634. **(b)** Heatmap: Frame 16634.

Figure 4.2: Gaze points from frame 16584 to 16634 plotted as a scatterplot, contour and heatmap on the middle frame 16634.

I also plot the last frame 16684, with gaze points recorded from frames 16634 to 16684. Instead of a horizontal gaze sweeping on the left side of the frame (Figure 4.2), the gaze pattern is slightly diagonal, veering towards the right side of the frame (Figure 4.2b).

Finally, I plot the first frame of the abdomen segment frame 16584 (Figure 4.4) to investigate whether the abdomen plane had changed. Here, another challenge of using the visualisation method shown in Figure 4.3 can be seen - the sonographer had scaled the abdomen plane to fit the screen before freezing.

Figure 4.4: First frame of the abdomen segment, frame 16584.

Some of the general challenges associated with using current methods (Section 2.1.3) on fetal ultrasound videos can be seen here. For example, the complexity of the video segment. The size of the abdomen changed during the video segment (Figure 4.4, 4.3). The temporal gaze patterns also show variation over time (Figure 4.2, 4.3), depending on which landmark the sonographer was observing.

4.1.2 Contribution

The method presented in this chapter proposes an unsupervised clustering method using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [42] to cluster raw eye-tracking data to first locate areas-of-interest (AOIs) of fetal ultrasound scan videos. Their corresponding images are used to capture granular changes within AOIs. The transitions within and between AOIs both spatially and temporally are visualised using a two-dimensional contour plot. The presented method presents both spatial and temporal information which distinguish between gaze patterns when searching for different anatomical planes. The proposed method combines unsupervised machine learning and data visualisation which is suitable for exploratory data analysis when analysing a large eye-tracking dataset consisting of several participants performing different tasks.

4.1.3 Data

The dataset used in this chapter was manually labelled second trimester scans (Section 3.3.1). To compare task differences, anatomical planes which sonographers spent the most time on whilst scanning were used. These planes were *Ab*, *Br* and *Ht* which represent the abdomen, brain and heart respectively. An example of each of these planes is shown in Figure 4.5. In total, there are 84, 160, 122 abdomen, brain and heart plane clips respectively. These were acquired by 10 fully qualified sonographers on 76 unique pregnant women. Eye-tracking data corresponding to the live-B mode video frames was used. The sonographer is actively searching for the anatomical plane during this time and differentiating gaze

behaviour could be present. Any gaze data which was missing was interpolated using the method described in Section 3.4.2.

(a) Abdomen. **(b)** Brain. **(c)** Heart.

Figure 4.5: An example of the abdomen, brain and heart anatomical plane.

4.1.4 Definitions

Before presenting the method, I define terminology that is used in this chapter. Depending on the application, an area-of-interest (AOI) can be defined as a fixation or as gaze points which fall within an area specified by the user. In this chapter, an area-of-interest (AOI) is defined as a *specific anatomical landmark that the sonographer has looked at while performing the scan.*

4.2 Methods

In this work, I use unsupervised methods to reduce the labelling effort required to identify meaningful eye movement events. Specifically, unsupervised clustering methods allows the user to identify meaningful clusters of sonographer recorded gaze points. Next, I use 2D visualisation methods to make useful inferences about the spatial and temporal characteristics of the gaze data; 2D methods are chosen over 3D because there are fewer dimensions to visualise whilst presenting meaningful information about the data.

4.2.1 Determining Areas-of-Interest (AOIs) using Unsupervised Clustering

Unsupervised clustering algorithms were used to reduce the need for manually annotating different eye movements and AOIs from raw eye-tracking data. Clustering

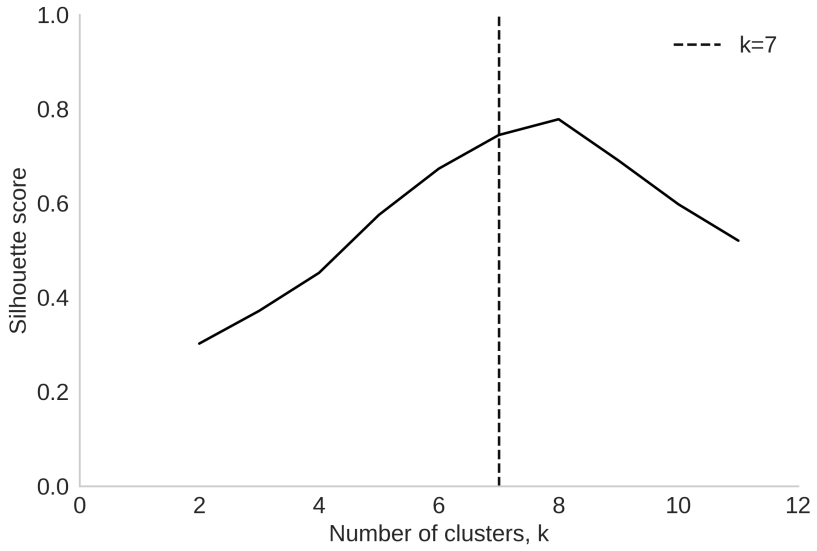


Figure 4.6: An example of determining the elbow of a clustering algorithm. The scoring metric used here is the Silhouette score discussed in Equation 4.9, and the number of clusters being tested ranges from 2 to 12. In this toy example, the optimal number of clusters is 7.

will identify gaze points which are spatially close - a first step towards identifying meaningful AOIs. Eye-tracking data includes different eye movements, such as saccadic movement, resulting in gaze data that is outside a specified radius of a fixation. These gaze data would need to be classified into their own cluster, and the clustering algorithm would need to capture single (or several) gaze points which represent saccadic movement.

k-means clustering. The first algorithm, k-means clustering, was chosen because it has been used in previous eye-tracking studies to separate fixations and saccades [49]. k-means clustering is an unsupervised clustering algorithm that partitions the data into k clusters. A new data point is assigned to the nearest cluster using a specified distance metric, typically the Euclidean distance.

The optimal number of clusters for a dataset is determined using the elbow method. The elbow method [1] is a heuristic method used to determine the optimal number of clusters in a dataset for a given clustering algorithm. To determine the elbow, a graph of the selected scoring metric (e.g. Euclidean distance) against the

number of clusters is plotted. The elbow is the point in the graph which returns the maximum curvature. An example is given in Figure 4.6.

One of the main assumptions of using k-means clustering for any data is that the algorithm assumes that the underlying data distribution is globular and spherical (example in Figure 4.7). This means that clusters with irregular shapes are not likely to be separated well via the k-means clustering algorithm. Isolated gaze points are more likely to be assigned to the nearest cluster as opposed to being considered as a single cluster.

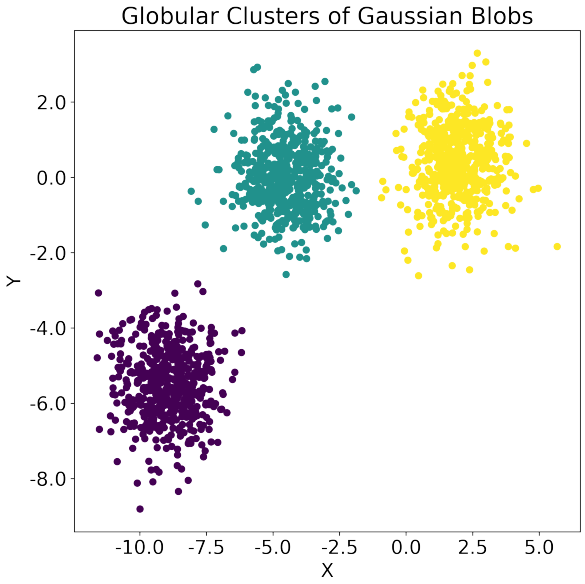


Figure 4.7: An example of globular and spherical data clustered using a k-means algorithm [160]. In this example, the k-means clustering algorithm has produced 3 different clusters. The x and y axis represent the numerical values of the dummy data points [Reproduced under the BSD License].

Hierarchical Density-Based Spatial Clustering of Applications with Noise. Conversely, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm is able to capture and classify single gaze points in their own cluster. HDBSCAN is a density-based clustering algorithm that uses a minimum spanning tree[1] to construct a cluster hierarchy (given a minimum cluster size). HDBSCAN [42, 78] does not assume spherical clusters and considers varying

[1]A graph that connects all the vertices using the minimum possible edge weights.

cluster densities and accounts for signal noise. This allows points which are far away from the AOI to be classified as a single cluster, which satisfies the constraint described above of isolating saccadic eye movement from fixations.

Clustering Distance Metrics.

To assess how well the clustering algorithm performs, metrics compute how well the points in a cluster are tightly packed, and how far apart each cluster is from the next cluster. If the clusters are tightly packed and far apart, then that is usually considered as a 'good' clustering performance. However, the nature of unsupervised clustering is such that unlike supervised methods, where the aim is to minimize (or maximise) the value of a specified loss function, there are no strict labels in unsupervised clustering. Therefore defining a 'good' performance is usually based on comparing the value of a calculated metric across different clustering algorithms. In this chapter, the metrics used to evaluate the clustering algorithm performance are: Davies-Bouldin [4], Calinski-Harabasz [3] and the Silhouette coefficient [7] . These 3 metrics compute the inter-and-intra similarity of clusters.

Davies-Bouldin (DB) Index.

$$S_i = \left\{ \frac{\sum_{j=1}^{N_i} ||d_j - c_i||^2}{N_i} \right\}^{\frac{1}{2}} \tag{4.1}$$

$$R_{ij} = \frac{S_i + S_j}{||c_i - c_j||_2} \tag{4.2}$$

$$DB = \frac{\sum_{i=1}^{K} max(R_{ij})}{K} \tag{4.3}$$

- c_i cluster center of cluster i
- K total number of clusters
- d_j data points assigned to cluster i
- N_i total number of points in cluster i

The DB index of a dataset (Equation 4.3) is calculated as the average ratio of the distances between each cluster center and all other points assigned to the same cluster (Equation 4.1), and the distance between cluster centers (Equation 4.2). A lower DB value is returned if the distance between points assigned to each cluster is small, and the distance between different cluster centers is large. Hence, a low DB value is indicative of a better clustering algorithm.

Calinski-Harabasz (CH) Index.

$$WG_k = \frac{\sum_{k=1}^{K} \sum_{j=1}^{N_i} \|d_j - c_k\|^2}{N - K} \tag{4.4}$$

$$BG_k = \frac{\sum_{i=1}^{K} N_i \|c_i - c\|^2}{K - 1} \tag{4.5}$$

$$CH = \frac{BG_k}{WG_k} \tag{4.6}$$

- c_i cluster center of cluster i
- K total number of clusters
- d_j data points assigned to cluster i
- N_i total number of points in cluster i
- c centroid of the dataset
- N total number of points in the dataset

BG_k calculates the 'between groups' distance between cluster centers c_i and centroid of the dataset c given by Equation 4.5. WG_k calculates the total 'within groups' distance between each cluster center c_k and all points assigned to the same cluster d_j.

The CH index of a dataset with N number of observations is calculated as the ratio between 2 variables defined as BG_k and WG_k, abbreviated for 'between-group' and 'within-group' respectively. A larger CH value indicates a large between group distance and small within group distance value; clusters are well-separated and individual clusters are densely packed. Hence, a larger CH value is indicative of a better clustering algorithm.

Silhouette coefficient.

$$\mu_{d_I} = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j) \tag{4.7}$$

$$\mu_{d_J} = \min_{J \neq I} \frac{1}{|C_J|} \sum_{k \in C_J} d(i, k) \tag{4.8}$$

$$s_i = \frac{\mu_{d_J} - \mu_{d_I}}{max(\mu_{d_I}, \mu_{d_J})} \tag{4.9}$$

The Silhouette coefficient of a data point is calculated as shown in Equation 4.9. If a data point s_i is assigned to cluster I, then the Silhouette coefficient of s_i is calculated as $\mu_{d_I} - \mu_{d_J}$ divided by the maximum value of μ_{d_I}, μ_{d_J}. The Silhouette coefficient falls within a range of (-1, 1). The higher the coefficient, the better the clustering algorithm.

μ_{d_I} (Equation 4.7) calculates the average distance between s_i and all other points j assigned to cluster I. This measures how densely packed a cluster is. μ_{d_J} (Equation 4.8) calculates the average distance between s_i and all other points k assigned to the nearest cluster J; this measures how far apart the next nearest cluster J is to s_i.

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) was used to identify AOIs after pre-processing using the methods described in Section 3.4.2. Only valid AOIs are used in this work; a valid AOI is a gaze segment which has ≥ 7 identical consecutive cluster labels.

4.2.2 Visualising Scanning Characteristics in the Spatial and Temporal Domain

In this subsection, I present a 2D visualisation method that was used to visualise meaningful clusters of gaze data. Common industry measures of describing participant behaviour using eye-tracking data are used to create this visualisation - time spent at each AOI and the change in gaze angle between AOIs. Instead of assuming that a sonographer's visual attention remains fixed within each AOI, I aim to capture local changes within each AOI. This step was performed to also consider changes in the video frames over time.

To capture these local changes, a snapshot of the sonographer's focus is taken by cropping a bounding box using the gaze point as the center. The size of the bounding box spans ±1.5 degrees in the x and y direction, where 1.5 degrees is a typical range of the fovea human field of view [36, 81] around a fixation point. Then, the cosine similarity is calculated between adjacent cropped images. The cosine similarity is a distance metric that measures how far apart two vectors are in Euclidean space, calculated using the dot product between two vectors $\vec{A}$ and $\vec{B}$ given in Equation 4.10.

$$\vec{A}.\vec{B} = ||A|| \times ||B|| cos\theta_{AB} \tag{4.10}$$

Where the cosine similarity is less than a threshold, chosen as two standard deviations less than the mean μ_{CS} of all anatomy specific cosine similarity values: $\mu_{CS} - 2\sigma_{CS}$, where σ_{CS} is the standard deviation at the population level, the sonographer is considered to have looked at a different AOI on the image. The distance between AOIs is defined by the change in gaze angle between the centroids of each AOI. The time taken between AOIs provides a quantitative measure of temporal variance, while the change in gaze angle between AOIs is a measure of spatial variance.

Sonographer Visual Scanning Modes.

Four visual scanning modes are considered, and they aim to capture the most common sonographer gaze behaviour observed in the data. They are presented in Figure 4.8 and described in Table 4.1. Two factors being considered are: how long the sonographer looks at a AOI, calculated in number of frames, at time t before moving onto the next AOI at $t + n$, and how far the gaze has travelled from one AOI to another.

In Table 4.1, large spatial variance is defined as >3 degrees, twice that of the typical range of the human field of view. Accordingly, a small spatial variance is ≤3 degrees. A small temporal variance is defined as ≤30 frames (1 second at a 30Hz video sampling frequency), and a large temporal variance as >30 frames.

Qualitative Description of Scanning Modes

Modes	Spatial variance	Temporal variance	Example action of sonographer
I	Large	Small	Shifts focus quickly between landmarks which are far apart in space.
II	Small	Small	Following a landmark as a guide to refine the selected anatomy plane.
III	Large	Large	Transitioning between landmarks to perform final adjustments to the image.
IV	Small	Large	Focusing on a specific anatomical landmark.

Table 4.1: Description of modes I, II, III and IV being considered for spatial and temporal analysis of sonographer visual scanning characteristics (top left, bottom left, top right, bottom right).

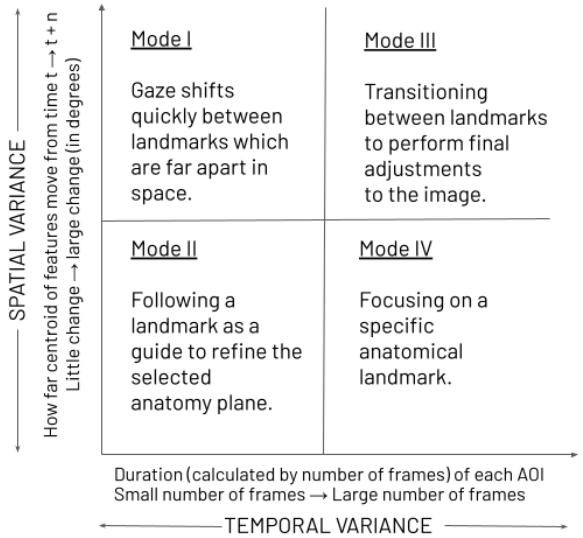


Figure 4.8: Description of modes I, II, III and IV being considered for spatial and temporal analysis of sonographer visual scanning characteristics (top left, bottom left, top right, bottom right).

Referring to Figure 4.8 and Table 4.1, Mode I considers the sonographer looking at different positions in space and at different AOIs. Mode II considers the sonographer focusing on the same position in space while changing the image quickly. Mode III considers the sonographer changing their focus at a specific position in space. Last, mode IV considers the sonographer focusing at a specific position in space and at a specific AOI.

Figure 4.8 shows a qualitative description of how these modes can be thought of, and is used to present the results. The modes are described in detail in Table 4.1. In the context of eye-movements, the spatial variation indicates the size of the area-of-interest, where the larger landmarks, the greater the spatial variation. For temporal variation, more time spent at an area-of-interest means a larger temporal variance.

Bi-variate Contour Plots.

To create the visualisation, a bi-variate distribution is calculated by superimposing a Gaussian kernel on each data point and returning a normalised density mass function. These data points represent AOI characteristics - length of AOI calculated by number of frames (x axis), and distance from previous AOI to following AOI (y axis) (Figure 4.8). Figure 4.9 shows an example of how to read the visualisation, where the contour lines represent cumulative density masses at 10 equally spaced levels from 0 to 1, where 0.1 is the outer-most contour and 1 is the inner-most contour. To avoid cluttering the contour plots, these contour level labels are not shown in the result plots.

I provide a brief description on how to interpret the visualisation in Figure 4.9 as an example of how to interpret the contour plots. An AOI in Figure 4.9 is defined using the definition above, where each AOI is a unique part of the image that the sonographer has looked at while scanning. The specific example shown in Figure 4.9 shows that the sonographer spent a short time looking at different AOIs, where on average they looked at each AOI for between 20-40 frames (bottom left of the figure). At a sampling frequency of 30Hz, that corresponds to approximately 1 second. There are several (20%) data points which lie on the

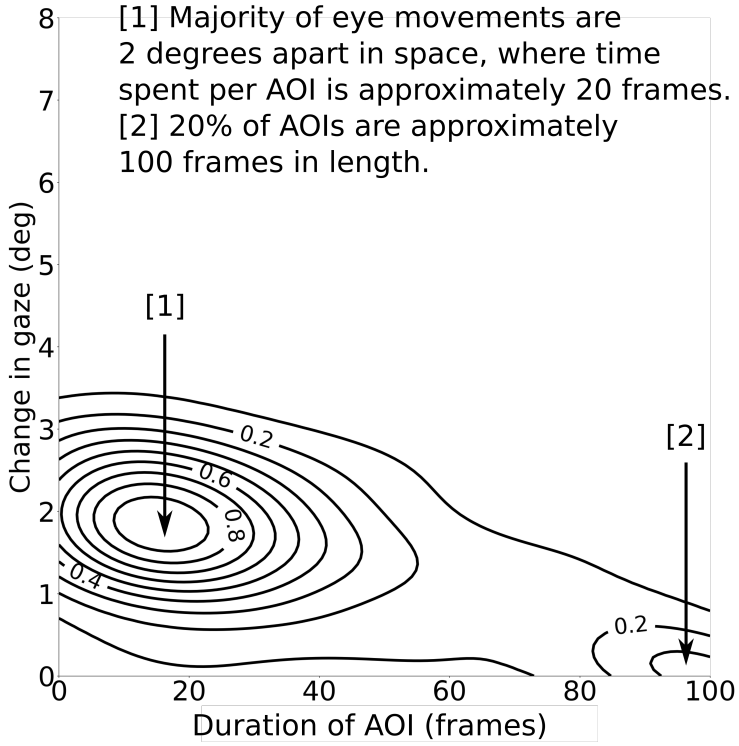


Figure 4.9: Example of bi-variate contour plot showing amount of temporal variance (x axis), and spatial variance (y axis) based on Figure 4.8.

bottom right side, which indicates that these AOIs were looked at for longer periods of time, approximately 100 frames ($\approx$ 3 seconds). Both these observations are related to the temporal aspect of how long the AOIs were observed for before the sonographer moved on to the adjacent AOI. For the spatial aspect, I observe the plot characteristics along the y-axis. As seen in Figure 4.9, the majority of the AOIs which were observed were not far in distance (degrees). Therefore the sonographer looked at AOIs which were close together on the screen. Finally, to relate it back to the visual scanning modes discussed above, the sonographer would have been considered to be scanning in mode II and IV, where the AOI data points are located mostly in the bottom left, and right of the plot. Mode II and IV occurs when the sonographer's gaze has not changed much during the scan (less spatial variance, as seen in Figure 4.9 along the y axis), while the image changed quickly for most of the scan (concentration of gaze points on the left side of the x-axis

representing temporal variance). In some parts of the scan, the image remained constant (concentration of gaze points on the right side of the x-axis).

Baseline Comparisons.

As a baseline technique to compare with, the I-VT algorithm [38] was chosen since one or more fetal ultrasound studies have also used I-VT for detection of standard fetal imaging planes [81]. The AOIs using the I-VT algorithm are calculated using the procedure described by [38]. The calculated AOIs are represented using the visualisation in Figure 4.9. A second baseline is chosen using the method proposed by [49], where they used k-means to separate eye-tracking data into fixations and saccades. Since I-VT was used specifically in a previous study [81], a visual inspection of qualitative differences was performed. Generated contour plots and visual scanning modes were compared for the output from I-VT and HDBSCAN.

4.3 Results

The quality of the identified clusters of AOIs are assessed by using standardised clustering validity measures which calculate performance based on the distance between points within the cluster, and the distance between cluster centers. These measures calculate the compactness and separability of clusters, where a densely packed cluster far away from other clusters is considered to be compact and separable. Three metrics which assess unsupervised clustering are used: the Davies-Bouldin (DB) index [4], Calinski-Harabasz [3] (CH) index scores and the Silhouette coefficient [7]. These are defined in Section 4.1.4.

	Silhouette [7]	DB [4]	CH [3]
I-VT	0.43±0.28	1.25±1.47	684.11±1354.32
k-means	0.74±0.09	0.35±0.13	1602.47±2427.37
HDBSCAN	**0.83±0.08**	**0.22±0.10**	**3660.03±4781.0**

Table 4.2: Average and standard deviation of Silhouette, DB and CH scores using I-VT, k-means and HDBSCAN. A higher score for CH and Silhouette and a lower score for DB indicates a better clustering performance. In bold, the best performing clustering algorithm assessed against each of the metrics.

	Abdomen	Brain	Heart
I-VT	32	22	112
k-means	3	4	7
HDBSCAN	5	12	27

Table 4.3: Number of abdomen, brain and heart segments which returned a single AOI using I-VT, k-means and HDBSCAN.

Table 4.2 show the mean scores of Silhouette, DB and CH, which return a better score when using HDBSCAN compared to I-VT and k-means to cluster raw eye-tracking data. Clips which only had one predicted cluster label, corresponding to a single AOI, were not included in Table 4.2 as this would return a null score. The number of clips which returned a null score is shown in Table 4.3. The scores in Table 4.2 were for 76, 146 and 29 abdomen, brain and heart plane clips which returned more than 1 cluster.

For a qualitative assessment, results using HDBSCAN and I-VT are compared since the I-VT algorithm has been used and tested on one or more ultrasound fetal studies [81]. The visualisations for abdomen, brain and heart planes are shown in Figure 4.10. A comparison across tasks using HDBSCAN (red line) shows that heart planes use the least amount of spatial information while brain planes use the most. However, heart planes use more temporal information, where the contour of Figure 4.10 (right) is stretched in the temporal direction, while the contours of brain and abdomen (Figure 4.10, left and middle) are stretched in the spatial direction. These results show that the searching process is task dependent.

A visual comparison shows that HDBSCAN returns more granular and anatomy specific results compared to I-VT. For HDBSCAN, abdomen and brain planes show similar temporal characteristics (Figure 4.10 left, middle) but show more spatial variance for the brain compared to the abdomen. Using I-VT returns similar spatial-temporal gaze characteristics for abdomen and brain planes (Figure 4.10 left, middle). For heart planes, Figure 4.10 (right) shows opposite characteristics for the HDBSCAN and I-VT algorithm, where HDBSCAN returns mode II AOIs while I-VT returns mode IV. This is explained by the algorithm I-VT clustering 94 heart scans as a single cluster indicating no change in image or gaze over time.

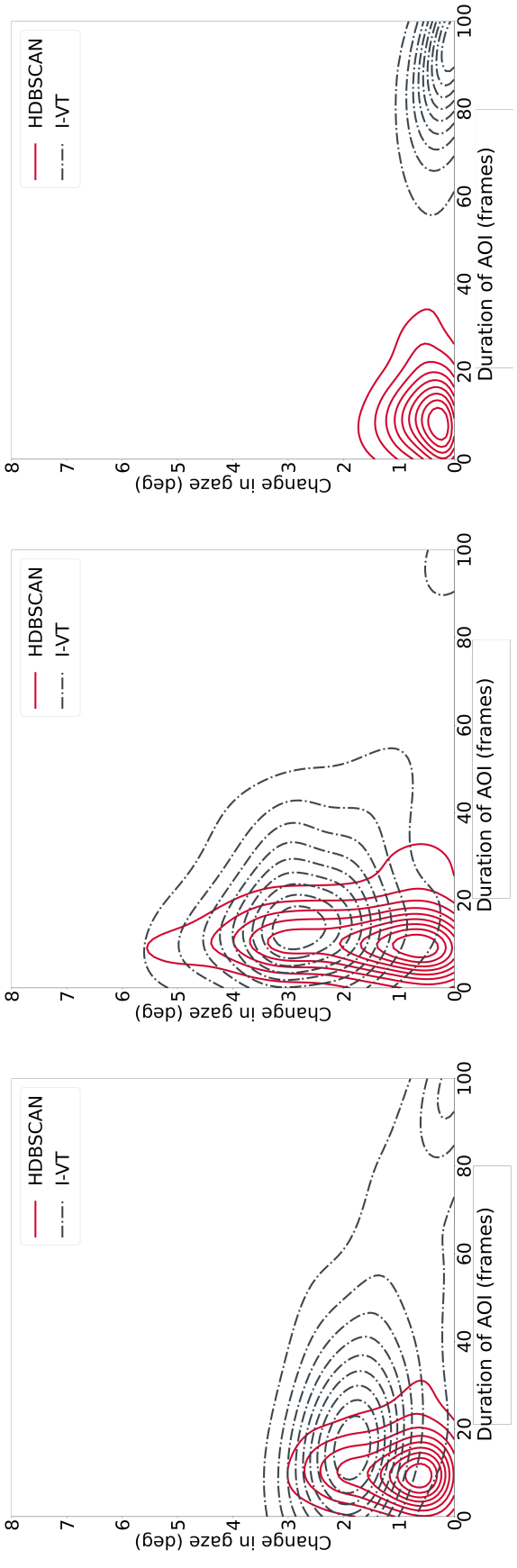


Figure 4.10: Bi-variate contour plot of abdomen (**left**), brain (**middle**) and heart (**right**) AOIs calculated using HDBSCAN and I-VT and its AOIs landmarks.

4.4 Discussion

Figure 4.10 (left) shows that whilst scanning for abdomen planes, sonographers use both spatial and temporal information. When empirically investigating the data, this was found to be true as sonographers focus on the areas within the three anatomical landmarks, stomach, aorta and umbilical vein (Figure 4.11), resulting in a general viewing area around the center of the image. In some cases, sonographers focused at the center of the image whilst refining the plane. I-VT captures this as mode IV (Figure 4.8). However, since image changes have been accounted for, these instances are now captured correctly as mode II (Figure 4.10, left) using HDBSCAN.

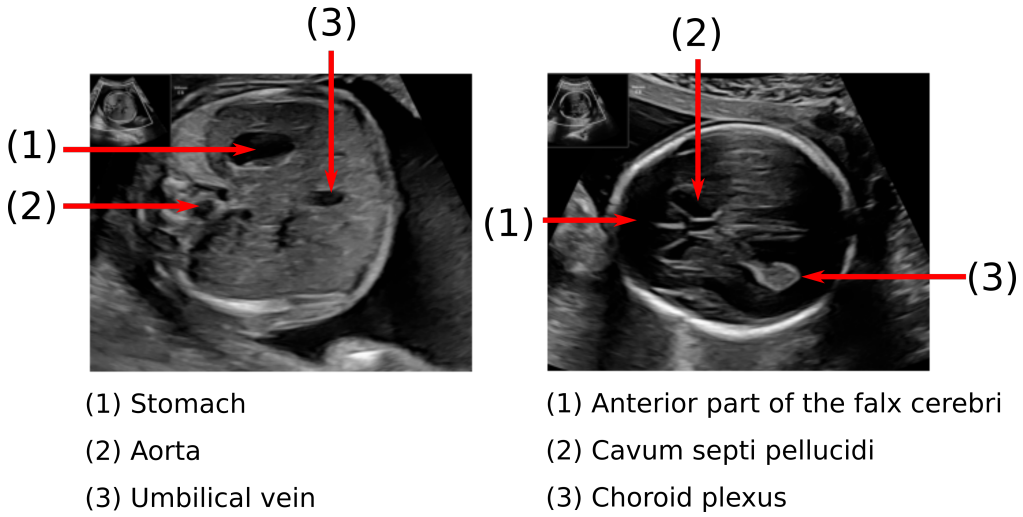


Figure 4.11: Labelled anatomical landmarks.
Left: Abdomen (stomach, aorta and umbilical vein).
Right: Brain (anterior part of the falx cerebi, cavum speti pellucidi and choroid plexus) plane.

For brain planes, sonographers often move between the anterior part of the falx cerebri, cavum septi pellucidi and the choroid plexus where measurement of the ventricular atrium occurs (Figure 4.11), and these two anatomical landmarks are at opposite sides of the head. This behaviour is reflected by the large change in spatial variance over short periods of time seen in Figure 4.10 (left), not captured by the I-VT algorithm.

For heart planes, it was observed that sonographers follow landmarks such as the septum or aorta closely to transition between heart planes; this is reflected

by the concentration of AOIs in mode II (Figure 4.10, right). However, the I-VT algorithm classified these instead as mode IV not having taken the change in image into account. This is a misleading result since the sonographer has in fact refined their image over time while following their chosen landmark closely showing little to no change in gaze.

Note that mode III is empty for all the anatomy planes. This is reasonable since such behaviour is more likely to be observed in a global static image, such as a sonographer reading an image. An alternative scenario is when an involuntary/voluntary saccade has occurred, but in this instance AOIs are being compared.

4.5 Summary

In this chapter, I used unsupervised methods to reduce the labelling effort required to identify meaningful eye movement events. Specifically, unsupervised clustering methods allows the user to identify meaningful clusters of sonographer recorded gaze points. I also use the ultrasound image to determine how similar the adjacent frames were. If they exceeded a certain threshold similarity, then the sonographer is considered to have moved onto the next meaningful event. By using a combination of unsupervised clustering and measuring image similarity metrics, eye movement classification algorithms were not required. Next, I used a simple 2D method to visualise patterns in sonographer eye-tracking data recorded during a longitudinal study which is easy to interpret and able to capture task specific characteristics when considering multiple viewing references and moving images such as videos.

Several types of scanning behaviour for abdomen, brain and heart planes were presented. Most noticeably, there are distinct behaviours for each type of task. HDBSCAN was used to cluster raw eye-tracking data and returned a more informative and meaningful eye-tracking data visualisation than using the established I-VT algorithm. The method presented in this chapter does not require manual labelling of AOIs or hand selecting threshold parameters for separating eye movements which is expensive and not always possible in large scale studies.

5

Individual Level Visualisation of Spatial Temporal Gaze Characteristics of Sonographers

Contents

5.1 Introduction 67
5.1.1 Contribution 68
5.1.2 Data 68
5.1.3 Definitions 69
5.2 Methods 70
5.2.1 Normalisation of Eye-Tracking Data by Localising Anatomy Circumference using Affine Transformer Networks . . . 71
5.2.2 Visualisation of Eye-Tracking Data using Time Curves . 72
5.3 Results 75
5.4 Summary 79

5.1 Introduction

In Chapter 4, I explained how data visualisation can be used to understand where and how sonographers have looked at whilst scanning for particular anatomical planes. The results of Chapter 4 were specifically for analysing the gaze behaviour of a population of sonographers. It is expected that on average a population of

sonographers search for the same anatomical plane in similar ways. This assumption means that studies such as [109, 113] are able to use gaze to build saliency prediction methods that describe the visual search of sonographers. Naturally, this leads to the question: are we able to visualise gaze behaviour on a per-scan basis? Broadly, a deeper understanding of how gaze patterns can differ between scans helps us discover new (or different) ways in which sonographers perform their scan which may be helpful for characterising sonographer skill.

To answer this, we want to visualise gaze behaviour using a concise and informative method. In this chapter, I present a method that is concise and informative of individual gaze behaviour. Individual-level analysis enables us to learn where and what sonographers have looked at across different scans. Performing the analysis on a per-scan basis also means that we can analyse the differences between tasks that have different levels of difficulty. This chapter complements Chapter 4, in that, alongside understanding a population of sonographers' gaze behaviour, we aim to 'dig deeper' and understand the gaze behaviour on an individual level.

5.1.1 Contribution

The approach I describe combines a deep learning model to normalise the eye-tracking data and an event-based visualisation method to characterise the normalised gaze data. The deep learning model is an affine transformer network [54] for localising the anatomy circumference of a fetus ultrasound video which allows for scale and position invariance of the fetal image between scans. Then, a data visualisation methodology, time curves [59], is used to characterise sonographer scanning patterns for different tasks. The proposed method allows us to analyse sonographer gaze behaviour on a per-scan basis and compare different gaze characteristics when a population of sonographers are searching for the same plane.

5.1.2 Data

The dataset used was the `PULSENet` standard planes described in Section 3.3.2. The subset of anatomies chosen were heart and brain planes to compare differences

in sonographer skill. Brain planes are considered easier to search for compared to the heart due to differences in anatomy size, and [104] showed that operators spend the most time searching for these anatomy planes during a routine clinical second trimester scan. The planes which were used are TVP, TCB, 3VT, 3VV, 4CH, LVOT and RVOT. Figure 5.1 shows examples of the 5 heart and 2 brain views respectively. In total, there were 185 TVP, 188 TCB, 65 4CH, 50 LVOT, 61 RVOT, 57 3VV and 65 3VT planes. 10 fully qualified sonographers performed the scans, and there were 250 unique pregnant women. Any gaze data which was missing was interpolated using the method described in Section 3.4.2.

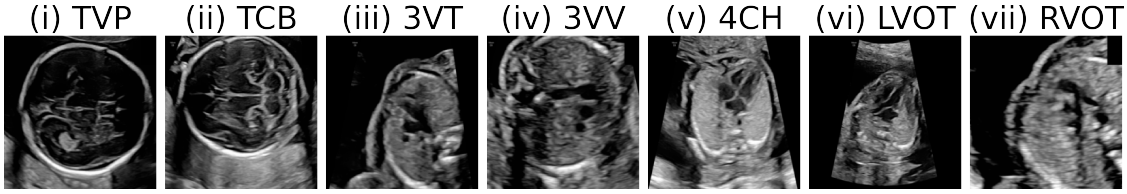


Figure 5.1: Selected brain and heart views from the PULSENet DS rescaled to 224x224 pixels. From left to right: (i) TVP, (ii) TCB, (iii) 3VT, (iv) 3VV, (v) 4CH, (vi) LVOT and (vii) RVOT.

5.1.3 Definitions

Before describing the method used to analyse individual-level gaze characteristics, I define terminology that is used in this subsection.

The deep learning model presented in this chapter builds on the scanpath normalisation method presented in Section 3.4.2. In Section 3.4.2, the eye-tracking data was normalised manually using bounding boxes. To reduce the effort required to draw bounding boxes, an automatic way of normalising the eye-tracking data is desired. Instead, the *anatomy circumference* is identified using a deep learning model. The anatomy circumference is defined as *the visible outer bounds of the fetus's anatomy displayed on the ultrasound machine screen* (also described in Section 3.4.2). An example of the brain's circumference is shown in Figure 5.2.

The visualisation method presented is an *event-based* visualisation - *time curves*. An event is defined as a period of time when a meaningful activity has taken place. These activities are application dependent [40] and study specific. In this fetal

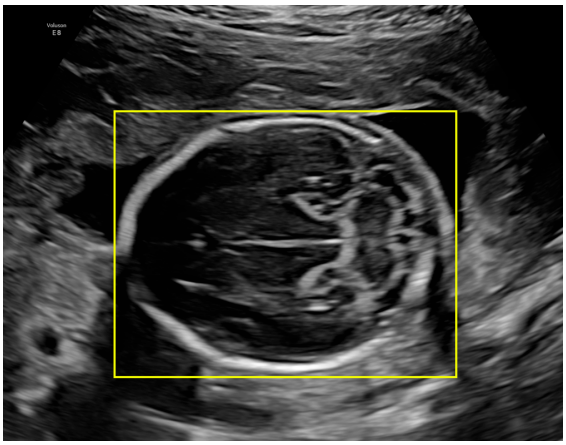


Figure 5.2: Example of a drawn anatomy circumference (in yellow) of a brain plane.

ultrasound application, an event is defined as *a unique anatomical landmark that the sonographer has looked at whilst scanning.*

Since each scan presents differently, assessing each scan individually is required to analyse differences between sonographer gaze behaviour. Event-based visualisations are useful for highlighting different events that occurred during each scan and are a useful way of providing summary information to the user quickly. As shown in Chapter 4, the population of sonographer gaze characteristics have different spatial and temporal characteristics. The chosen event-based visualisation must be able to account for spatio-temporal differences.

Time curves [59] was chosen because it is a 2D visualisation method that preserves temporal order of defined events and displays spatial similarity between events. First a linear timeline of events is plotted along the x axis, time. A distance matrix is used to represent the similarity between events. In general, the distance metric chosen is application dependent. Events are mapped onto a lower dimensional space (2D) using the distance matrix. Finally, a suitable interpolation curve, for example, splines, is used to connect event points. An example is shown in Figure 5.4.

5.2 Methods

A method is presented for localising the anatomy circumference in ultrasound frames using an affine transformer network inspired by spatial transformer networks [54].

Time curves [59] are then used to visualise sonographer recorded eye-tracking data. A variety of scanning styles is observed while searching for heart and brain anatomical planes.

5.2.1 Normalisation of Eye-Tracking Data by Localising Anatomy Circumference using Affine Transformer Networks

Spatial transformer networks (STNs) [54] are differentiable modules that can be used within deep learning network architectures to remove spatial and position variance between images for a downstream task such as classification or object detection. They have been used for ultrasound image registration [100, 120]. The affine transformer network (ATN) architecture proposed by [54] is used in this work. The network localises the fetal anatomy circumference to ensure that the fetus is the object of reference for the raw eye-tracking data. The AOI is the anatomy plane that the sonographer is searching for.

Normalising eye-tracking data with respect to the anatomy circumference has been shown to improve the performance of using eye-tracking data for anatomy plane classification (Chapter 6). However, for that work the circumference was manually labelled for >300 segments (Chapter 6). To reduce the manual labour needed to label a large-scale dataset, the proposed method uses an ATN to localise the anatomy circumference. The estimated affine transformation contains 6 parameters (Equation 5.1) and is able to account for scale, translation, shear and rotation [54].

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \tag{5.1}$$

The learnt transformation (Equation 5.1) is applied to the eye-tracking data for normalisation.

Implementation.

Following their proposed architecture, the backbone of the ATN uses GoogleNet [53]. The last average pooling layer and last fully connected layer are removed. On top of this, the following layers are stacked: (i) a 1 x 1 convolutional layer to reshape the AOI channels from 1024 to 128, (ii) a fully connected layer with 128x7x7 output and (iii) a fully connected layer with 6 outputs representing the affine transformation parameters [54]. The dataset was augmented using the principled data augmentation method described in Section 3.4.1.

The AdamW optimizer [77] is used, and the base learning rate starts at 1e-4 and reduces by a factor of 10 after 25, 50, 75 and 100 epochs. The loss is calculated as the mean squared error between the transformed image and the ground truth at the pixel level [103]. The model is allowed to train for 150 epochs and early stopping is implemented if the validation loss does not decrease by 0.001 after 10 consecutive epochs.

5.2.2 Visualisation of Eye-Tracking Data using Time Curves

An event-based visualisation method was chosen because this class of methods inform the user of what event has occurred and its characteristics. Each event must depict attributes of the data independently, similar to a glyph-based visualisation [40]. Time curves [59] were used because they are informative and simple to implement and interpret. To construct the time curve two variables are required: time spent at an event, and a similarity matrix to define how similar events are.

Fixations and saccades are calculated using the standardised I-VT algorithm [38] since a previous study [81] used the I-VT algorithm.

An *event* is defined as a snapshot of what landmark the sonographer is focusing on during a fixation. The position of a fixation is typically calculated as the average gaze point [38] during the fixation. Likewise, the middle time stamp (Figure 5.3, step 2) of the fixation is used as the average frame representing the AOI. A segment (20x20 pixels) of the image is cropped around the gaze point to create an event (Figure 5.3, step 4). Since the images are resized to 224x224, a 20x20 crop was

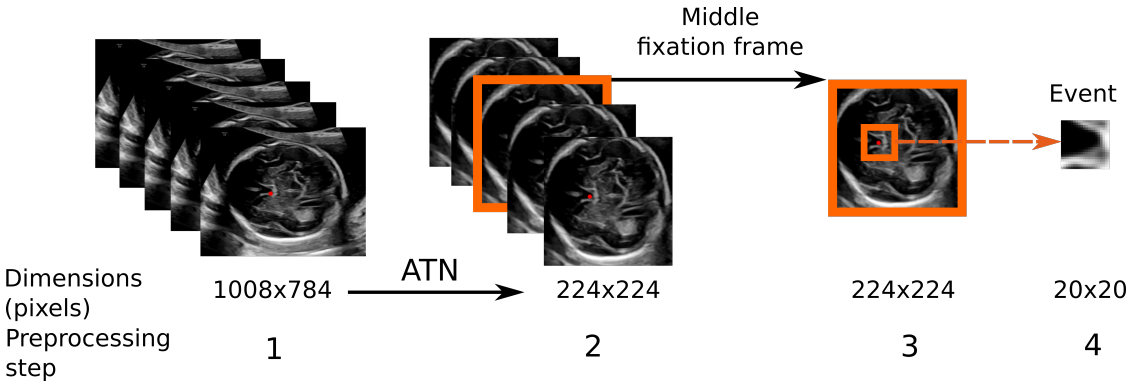


Figure 5.3: Figure demonstrating the process of generating an event to represent a fixation from the original ultrasound frames.
Step 1: Original ultrasound frames.
Step 2: Frames transformed using an affine transformer network.
Step 3: Selecting a representation of the event as the middle frame of the fixation.
Step 4: Landmark focused on, defined as an *event*.

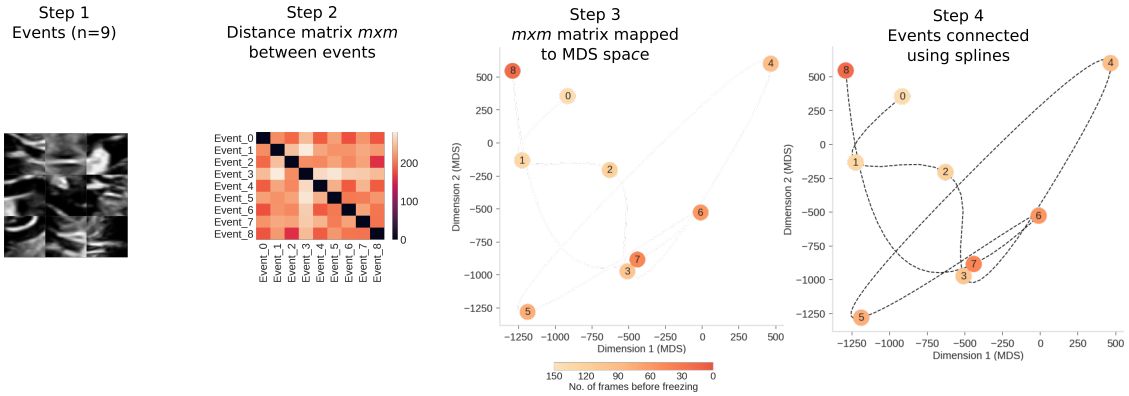


Figure 5.4: Figure summarising the process of developing the time curve (Illustrative purposes only, not to scale).
Step 1: Creating events, here number of events n=9.
Step 2: Calculating distance matrix between events, here numbered from 0 to 8.
Step 3: Distance matrix reduced to 2D using multi-dimensional scaling (MDS).
Step 4: Events connected using splines.

considered to be sufficient to capture the landmark that the sonographer focused on. The size of the crop was adjusted based on the visibility of the structures in the frames. For frames where gaze points were outside of the anatomy circumference, the landmark was substituted as a 20x20 grey square. Events which partially fell outside the anatomy circumference were zero padded with grey pixels.

Each task specific event image was flattened (Figure 5.4, step 1) to a 1×n dimensional vector. The Euclidean distance between each vector was calculated,

returning a square form distance matrix (Figure 5.4, step 2). The Euclidean distance was used as [59] has shown that a naive pixel difference for normalised images still returns informative results. The distance matrix is transformed to a lower 2D space using multi-dimensional scaling (MDS). From a m×m matrix, where m is the number of total snapshots in the dataset, a m×2 matrix is returned. These transformed data points are shown in Figure 5.4, step 3. Finally, to connect the events in temporal order (Figure 5.4, step 4), a 2nd order B-spline with a smoothing factor of 1 to interpolate between the transformed data points was used. These parameters were chosen by determining whether the interpolated curve was able to pass through or near the events. A colour mapping is used to indicate the temporal order (Figure 5.4, step 3). Light orange indicates the start of the time-series, and dark orange the end.

I provide a brief description of how to interpret the time curves based on Figure 5.4. In Step 4, there are the different colours of each event. These are used to quickly identify which events were observed toward the end of the scanning period just before freezing, where event 8 in step 4 was the last event that occured just before freezing. Event 8 is also numbered, and the darkest shade of orange. Event 0 and 8 are within proximity of each other, which suggests, given the distance matrix calculated in Step 2, that these events are similar in terms of image content. Events 3 and 7 are nearly overlapping, which suggests that these two events are also similar, if not nearly identical in image content. In fetal ultrasound, this would be a case of the sonographer revisiting a particular anatomical landmark whilst refining the final plane that needs to be captured.

The gaze scanning patterns are described (Table 5.1) using definitions of visual patterns given by [59] in Figure 17 of their paper. The examples are **cluster**, **transition**, **cycle**, **u-turn**, **outlier**, **oscillation** and **alternation**.

Pattern	Description of events
Cluster	Close in space.
Transition	Cluster of events migrated from one point in space to another.
Cycle	Start and end events are near each other and creates a closed circle loop.
U-turn	Start and end events are near each other.
Outliers	Several events far away in position compared to majority of other events.
Oscillation	Resembles a sine/cosine wave.
Alternation	Similar/identical events are repeatedly visited.

Table 5.1: Qualitative description of scanning patterns given by [59]. The patterns are cluster, transition, cycle, U-turn, outliers, oscillation and alternation.

5.3 Results

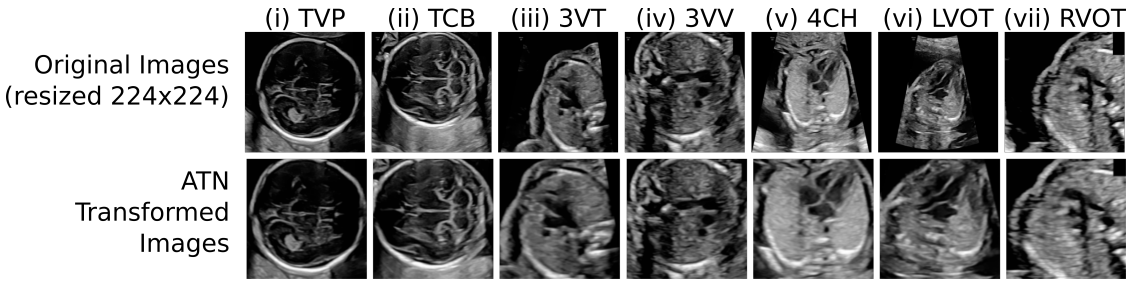


Figure 5.5: An example showing the performance of the ATN on standard plane images.
Left to right: TVP, TCB, 3VT, 3VV, 4CH, LVOT, RVOT.
Top: Original images resized to 224x224 pixels.
Bottom: Images after ATN transformation, resized to 224x224 pixels.

First visually, the result of applying the ATN to localise the anatomy plane is assessed. Examples are shown in Figure 5.5. Figure 5.5 shows the ATN model is able to localise the anatomy circumference for different standard brain and heart planes. The localisation effect is particularly effective for TCB, 3VT, 4CH and LVOT where the original images were off center and did not completely fill. The final transformed image is scaled and centered.

Several types of scanning patterns for each standard plane are presented in Figures 5.6, 5.7 and 5.8 which were observed since the primary interest is in characteristic patterns. For reference, the fully labelled time curve can be seen

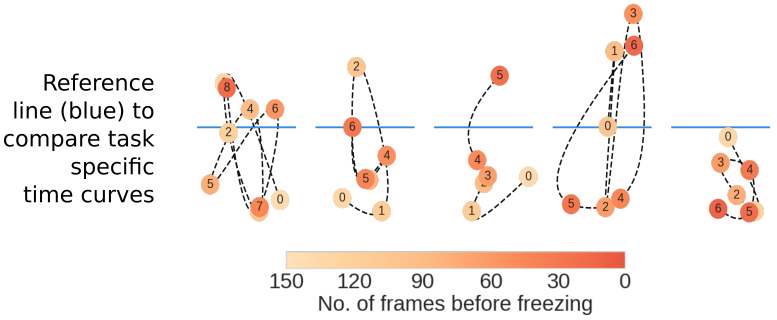


Figure 5.6: An example of several scanning patterns for brain plane TVP.

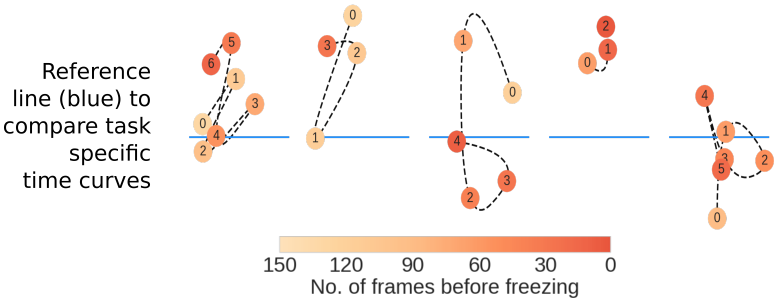


Figure 5.7: An example of several scanning patterns for brain plane TCB.

in Figure 5.4, step 4. To allow for ease of comparison between time curves, a task-specific reference line is provided in blue.

The average number of events for heart planes range from 1-2, while it ranges from 4-5 for brain planes. Figures 5.6 and 5.7 show that for brain planes there are several types of scanning patterns. Many of the scans show a combination of a **u-turn**, **cycle**, **clusters** and **transitions** where sonographers have revisited similar anatomical landmarks several times over the course of scanning.

For TVP, the majority of events in Figure 5.6 (left) are clustered below the reference line with differing temporal order (of events). For TCB, the majority of events in Figure 5.7 (right) are clustered above the reference line. There are also patterns which show that sonographers follow 3 - 5 distinct landmarks without revisiting them (Figure 5.7, right). There does not appear to be a pattern for when the events occurred, as shown by the scattered distribution of colours.

(a) An example of several scanning patterns for heart plane 3VT.

(b) An example of several scanning patterns for heart plane 3VV.

(c) An example of several scanning patterns for heart plane 4CH.

(d) An example of several scanning patterns for heart plane LVOT.

(e) An example of several scanning patterns for heart plane RVOT.

Figure 5.8: Time curves for the heart planes: 3VT, 3VV, 4CH, LVOT and RVOT.

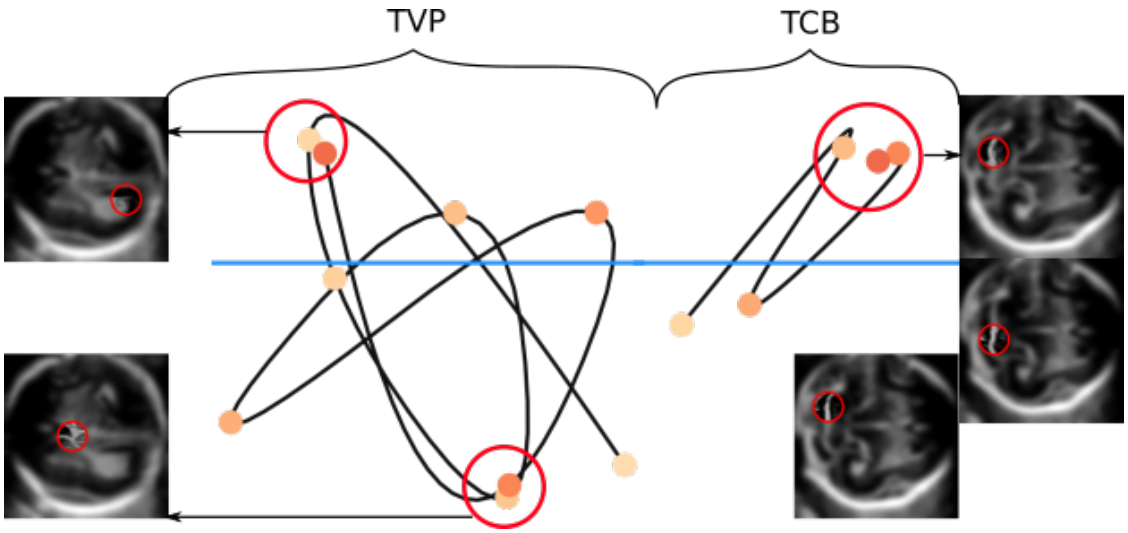


Figure 5.9: Example of landmarks which were viewed while scanning for brain planes TVP (left) and TCB (right). To avoid visual clutter, event numbers are not shown in this figure.

Figure 5.8 shows that for heart planes most of the scans only have a single event occurring, corresponding to a unique landmark that the sonographer focused on during the scan. Since the event is coloured dark orange, this indicates that the landmark was followed for the full length of 5 seconds. A small minority show patterns of loops and straight lines, where only 2-3 events have occurred.

A comparison between brain and heart plane scanning patterns show that typically it was observed that there are more unique landmarks that sonographers observe and revisit for brain planes. The majority of heart planes return a single event. Although both the brain and heart are 3-dimensional objects, the size of the heart is smaller and sonographer gaze does not need to travel as far. To obtain the correct heart plane view, the sonographer makes smaller adjustments with the probe compared to the brain. These differences are reflected in the visual patterns of brain plane time curves in Figures 5.6 and 5.7 and heart plane time curves in Figure 5.8.

A qualitative analysis of event trajectories was performed. For brain planes, these are shown in Figure 5.9. For TVP, the sonographers generally look at the choroid plexus and cavum septum pellucidem. For TCB, they look at the cerebellum where they measure the transcerebellar diameter.

Figure 5.10: Example of landmarks which were viewed while scanning for heart planes 3VT, 3VV, 4CH, LVOT, RVOT. To avoid visual clutter, event numbers are not shown in this figure.

For heart planes, some examples are presented in Figure 5.10 where the sonographer has looked at the trachea, and the intersection of the pulmonary artery and aorta (3VT), pulmonary artery (3VV), crux of the ventricular septum (4CH), aorta walls (LVOT) and bifurcation of the pulmonary artery (RVOT).

The chosen definition of an event is less able to provide defining characteristics for heart planes compared to brain planes. Many time curves returned a single event (Figures 5.8d and 5.8e) over 150 frames as their gaze points were classified as a single fixation.

5.4 Summary

In this chapter, an affine transformer network was used to localise the anatomy circumference, which provides a reference to normalise sonographer eye-tracking data recorded while performing second trimester fetal ultrasound scans. The normalised eye-tracking data was visualised using time curves. Then, task-specific sonographer scanning patterns were distinguished when searching for the heart and brain plane. This work showed that representing fixations as events to build time curve visualisations is a useful method to demonstrate observed landmarks and scanning patterns for brain planes, but less so for heart planes.

The final results demonstrate that the gaze characteristics of a sonographer is dependent on the anatomical plane being searched for and also that gaze patterns are not identical between scans. The presented method highlighted differences in the ways in which sonographers search for anatomical planes with different levels of difficulty, the brain and the heart plane. Most notably, the sonographers revisit the

brain anatomical landmarks more frequently than the heart landmarks. Broadly, these results are useful for understanding how the different anatomical structure of different fetal organs affect the gaze patterns of sonographers.

6

Eye-Tracking Based Task Classification

Contents

6.1 Introduction 81
6.1.1 Contribution 82
6.1.2 Data 82
6.1.3 Definitions 83
6.2 Methods 83
6.2.1 Normalisation of Gaze Data 84
6.2.2 Time-series Classification Models 85
6.2.3 Qualitative Visualisation of Scanpath 92
6.3 Results 93
6.3.1 Task Classification Results 93
6.3.2 Class Imbalance Models 95
6.3.3 Qualitative Results 96
6.4 Discussion 101
6.5 Summary 102

6.1 Introduction

Chapter 4 showed that sonographers have distinct gaze patterns when searching for the abdomen, brain and heart planes. Visualising these differences at the global level was useful to gain a general understanding of the differences between the landmarks located on the anatomical plane. The methods in Chapter 4 generate a

qualitative representations of sonographer gaze. Subsequently, is the gaze behaviour of sonographers sufficiently distinct for classifying separate scanning tasks?

Current challenges (discussed in Section 2.2.3) associated with task classification of eye-tracking data for fetal ultrasound is that there are various ways that a plane can be searched for. Where a sonographer starts their search differs from other scans and sonographers because the fetus is constantly moving and changing its position. It is non-trivial to use eye-tracking data in its raw form for classification. Other methods which use eye movement characteristics to classify tasks rely on the availability of a suitable eye movement classification algorithm to calculate suitable fixations and saccades metrics for comparison.

The method presented in this chapter aims to overcome these limitations by using normalised eye-tracking data to classify fetal ultrasound tasks which does not require any separation of eye-tracking data into fixations and saccades. Time-series eye-tracking data is also less computationally expensive than using images for task classification, and is an added benefit of considering using only eye-tracking data.

6.1.1 Contribution

The method presented in this chapter proposes first normalising the eye-tracking data to account for the change in scale and position of anatomy during the scan. Then, the normalised eye-tracking data is classified using a deep learning model. The best-performing model was a Gated Recurrent Unit (GRU) classification model which was able to classify the visual scanpaths of sonographers performing fetal ultrasound tasks. These two proposed steps resulted in the fetal ultrasound task being identified using only sonographer eye-tracking data.

6.1.2 Data

The dataset used was the manually labelled second trimester scans, which is described in Section 3.3.1. The anatomy labels *Ab*, *Br* and *Ht*, the abdomen, brain and heart respectively are used. These planes were chosen because sonographers spent the most time on them whilst scanning [135]. These labels do not separate

anatomy specific views. For example, the 'Br' label consists of both TVP and TCB (Table 3.1). The 'Ht' label consists of all heart views excluding the situs view. The total number of segments available is shown below.

- Total number of abdomen plane segments: 84
- Total number of brain plane segments: 160
- Total number of heart plane segments: 122

There were 10 fully qualified sonographers and 76 unique pregnant women. Eye-tracking data corresponding to the live-B mode video frames was used. The sonographer is actively searching for the anatomical plane during this time and differentiating gaze behaviour could be present. Any gaze data which was missing was interpolated, and any clips which were less than 100 frames in length were zero padded according to the procedures described in Section 3.4.2.

6.1.3 Definitions

In this chapter, a *task* is defined as the action of a sonographer looking for a specific anatomical plane. For example, the head is a separate task from the abdomen and heart.

In eye-tracking literature, a scanpath can be used to refer to eye-tracking data represented in its raw form and also as aggregated eye movements such as fixations [89]. In this chapter, a *scanpath* is defined as *the visual attention of an individual sonographer captured by eye-tracking data over time.*

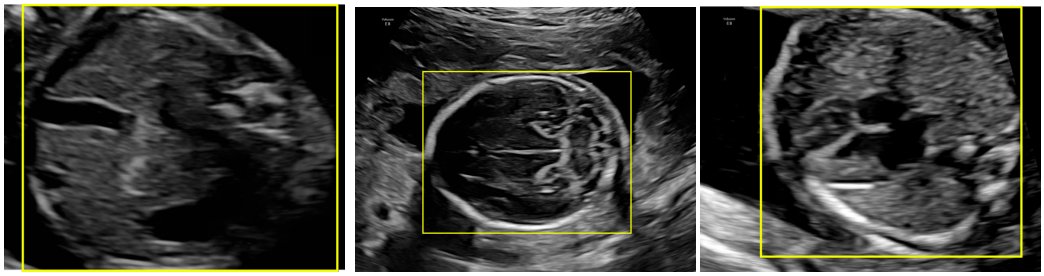
6.2 Methods

In this work, I first considered several different representations of the eye-tracking data. The purpose is to investigate whether using the eye-tracking data in its raw form or feature engineered form is more informative for the task being performed. Then, I used time-series classification models to identify the fetal ultrasound task being performed.

6.2.1 Normalisation of Gaze Data

To give the gaze data context with respect to the anatomical plane being searched for, I normalised the eye-tracking data with respect to the anatomy's circumference. The procedure is described in Section 3.4.2.

Raw gaze points recorded by the eye-tracker along the x and y axis with respect to the screen dimensions of 1920×1080 pixels are defined as G_x, G_y. Raw gaze points normalised with respect to the anatomy's circumference are given as G_{xBB}, G_{yBB} (Equation 3.2); BB is an abbreviation for bounding box, as the anatomy circumferences were drawn manually using bounding boxes [1]. An example of the drawn bounding box can be seen in Figure 6.1.



(a) Abdomen plane. **(b)** Brain plane. **(c)** Heart plane.

Figure 6.1: A bounding box drawn (in yellow) around the (a) abdomen, (b) brain and (c) heart plane's circumference. To calculate the area occupied by the plane on the frame, the area of the yellow box is divided by the area of the frame. The frame has dimensions 1008x784 pixels.

To test the effectiveness of providing context to the eye-tracking data, I consider the raw gaze points G_x, G_y as my first baseline, and G_x, G_y normalised with respect to the screen dimensions of 1920×1080 pixels as my second baseline. The gaze points normalised with respect to the screen dimensions are referred to as G_{xs}, G_{ys}, where s is an abbreviation for 'screen'.

In addition to the feature engineered gaze points, I also calculated how much space on the ultrasound machine screen the anatomical plane occupies. Typically the anatomical planes differ in scale, depending on the 'zoom' factor that the sonographer used (Figure 6.1) and this could be a factor that is important for task classification.

[1] The anatomy circumferences of the ultrasound video frames in Chapter 5 were identified using a deep learning model.

I use the bounding box drawn (Figure 6.1) to capture the difference in scale between planes. I do so by calculating the ratio of the screen that the anatomy occupies with respect to the frame. A ratio of 1 is where the anatomy image occupies the entire screen. The area of screen occupied by the bounding box (yellow box) divided by the area of the cropped image (1008x784 pixels) is given as A (Figure 6.1).

The final list of features which were used to train the model are:

1. G_x, G_y: raw gaze points.
2. G_{xs}, G_{ys}: raw gaze points normalised by screen dimensions.
3. G_{xs}, G_{ys}, A: raw gaze points normalised by screen dimensions, and area occupied by the screen.
4. G_{xBB}, G_{yBB}, A: raw gaze points normalised by the anatomy circumference, and area occupied by the bounding box relative to the screen.

6.2.2 Time-series Classification Models

In this work I consider three different classification models. I first choose a baseline model based on prior works. The first baseline model requires that the eye-tracking data is in discrete form. Hence, for the next baseline model I consider a model that can use the eye-tracking data in its continuous form. Finally, due to its success with time-series classification, I choose a suitable off-the-shelf deep learning model.

Hidden Markov Model (HMM).

The first model which was considered was a hidden Markov model (HMM), based on prior works [26]. HMM [154] is a time-series model which assumes that the process being modelled adheres to the Markov property. The Markov property is the assumption that the conditional probability distribution of future states only depends on the present state, and not past states. A hidden Markov model assumes that only the data, and therefore model parameters, is observed, while the states are hidden.

In their work, [26] first transformed the scanpath into a discrete sequence of numbers using k-means clustering. The eye-tracking data is transformed into a sequence of discrete cluster labels. The sequence is used to train the HMM. To

estimate the HMM's parameters, the Baum-Welch [12, 162] algorithm is used. The Baum-Welch algorithm estimates the model parameters of a HMM using only the observed data by updating its belief of the parameter values based on maximising a specified loss function. The loss function used to train the HMM in this chapter was the Maximum A Posteriori as shown in Equation 6.1. The maximum a posteriori estimation maximizes the posterior probability distribution with respect to variable Y given variable X is observed (Equation 6.1).

$$a^*_{\text{MAP}} = \underset{Y}{\text{argmax}} P(Y \mid X = x) \tag{6.1}$$

The final model parameters which are estimated are the probability of being in a specific state, known as the *emission probability*, and the probability of transiting between states known as the *transition probability*. In [26], each separate task was trained using a HMM, and each test sequence is scored against task specific HMMs. The predicted class is selected as the model which returns the maximum loss function value.

k-nearest neighbours Model (k-NN).

Since the eye-tracking data was represented as a sequence of discrete cluster labels, a whole time-series comparison is also considered. The purpose of doing so is to investigate whether raw gaze points are better for task classification compared to the coarse representation used in an HMM.

To classify the eye-tracking data in its continuous form, a k-nearest neighbours model (k-NN) was used. k-NN [2] is a non-parametric time-series classification model and classifies a data point based on the most frequent label amoungst its nearest k neighbours. The distance metric used to calculate its nearest neighbours is chosen by the user, for example, Euclidean distance. To select the optimal number of clusters, k, the elbow method is used. The elbow method was explained in Section 4.2.1, and an example demonstrated in Figure 4.6.

As mentioned in Section 6.1.2, zero padding was used to make the eye-tracking data equal lengths. In this instance, using the Euclidean distance is not suitable

for comparing the whole time-series[2]. The dynamic time warping (DTW) distance metric is used instead. DTW is a distance metric used to compare two time-series x and y of different lengths [24]. The algorithm calculates an optimal warped path W by matching each element in x to the closest (in Euclidean distance) element in y. The cost of the final path, given as the DTW value, is the sum of the minimum distances calculated from matching x to the nearest element in y. An example of using the dynamic time warping distance to find the closest spatial match between the x co-ordinates of a heart and brain segment can be seen in Figure 6.2.

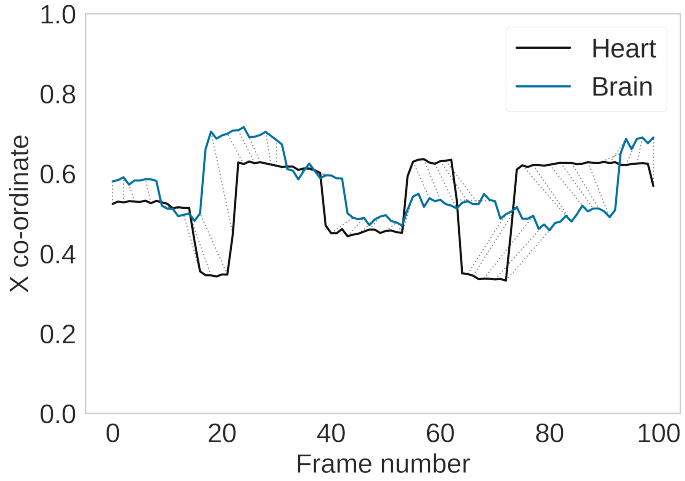


Figure 6.2: An example of calculating the distance between 2 time-series (in this instance the x-co-ordinate of the gaze data) using the dynamic time warping distance metric. The dotted lines represent the nearest match between two data points.

Gated Recurrent Unit (GRU) Model.

Given the success of deep learning models for time-series classification, I also chose an off-the-shelf deep learning model for task classification. The gated recurrent unit (GRU) model [48, 56] is a subset of recurrent neural networks (RNN) that retains long-term time dependencies between sequences. The GRU is a variation of the long short-term memory (LSTM) model [11] which is also very popular for time-series classification. However, the GRU has fewer parameters than a LSTM and have been shown to return comparable performance to the LSTM while requiring

[2]Note that this is different how the distance metric was used in Chapter 4 since there we were comparing the distance between individual data points. In this chapter, we are concerned with comparing whole time-series.

less specified parameters [107]. A LSTM model uses gates [3] to regulate which information the model retains and discards while training. The GRU has fewer parameters because the model combines 2 of the gates into a single gate. The input to the GRU model are equal-length time-series gaze data, and the output of the model returns a class label for each time-series.

Model Parameter Selection. To select the model's optimal parameters, the asynchronous successive halving algorithm (ASHA) was used. ASHA [87] is a hyperparameter tuning method that combines random search and principled early stopping asynchronously [65]. Principled early stopping uses the successive halving algorithm (SHA), also known as the multi-armed bandit algorithm. Briefly, the SHA evaluates the performance of the given set of configurations for a small number of epochs. At the next iteration, the top 50% performing configurations are kept and evaluated again with a slightly larger number of epochs. This process continues until one configuration remains.

To remove the computational bottleneck caused by using a top-down approach (i.e. starting with the entire set of configurations), the ASHA algorithm favours a bottom-up approach to [101] by selecting suitable configurations to keep. Initially, ASHA performs a random search of configurations to evaluate. For the next iteration, the algorithm finds suitable configurations to keep and discards those that did not meet a minimum required performance [101].

Loss Function. The GRU requires a specified loss function to train the model. In this chapter, 2 different loss functions are explored: cross entropy and focal loss. Cross entropy loss is a standardised function used to train classification models, and focal loss is a function similar to cross entropy but takes class imbalance of the dataset into account. For the reasons outlined in Section 3.3.1, the dataset used in this chapter is imbalanced. These two loss functions are explored to investigate whether the class imbalance of the dataset affects the final model's performance.

[3] A gate is a combination of a sigmoid neural network layer and a pointwise multiplication operation which returns an value of 0 or 1 to discard or retain information respectively [56].

The cross entropy loss function is given in Equation 6.2. The cross entropy loss function (Equation 6.2) gives equal weights to all classes in the dataset. Equation 6.2 shows the cross entropy loss function for a binary classification problem.

$$CE(p, y) = \begin{cases} -log(p) & \text{if } y = 1 \\ -log(1-p) & \text{otherwise} \end{cases} \tag{6.2}$$

- $p \in [0, 1]$ is the model's estimated probability for the class $y = 1$

The focal loss function is given in Equation 6.3 [76].

$$\mathrm{FL}\left(p_t\right) = -\alpha_t \left(1 - p_t\right)^{\gamma} \log\left(p_t\right) \tag{6.3}$$

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \tag{6.4}$$

$$\alpha_t = \begin{cases} \alpha & \text{if } y = 1 \\ 1 - \alpha & \text{otherwise} \end{cases} \tag{6.5}$$

- $\alpha_t \in [0, 1]$ is a weighting factor to address class imbalance

The focal loss function (Equation 6.3, assuming a binary class classification) is a variation of the cross entropy function that accounts for class imbalance. Unlike the cross entropy function which weights all classes equally, the focal loss forces the model to penalise the contribution of the majority class to the final loss value. The modulating factor γ in Equation 6.3 regulates the loss value of the majority class. As γ increases, the loss value of the majority class decreases and increases for the minority class. α (Equation 6.5) is a hyperparameter that acts as weighting factor on the majority and minority classes.

Implementation.

For training and testing, a 3-fold stratified cross validation was performed. 80% of the data was used for cross validation and 20% for tuning model parameters.

Gated Recurrent Unit. `RayTune`'s [88] ASHA [87] was used to tune the hyperparameters. The ASHA `config` parameters used to search for the optimal parameters were:

- n_layers: 2, 3, 4
- n_hidden: 4, 8, 16, 32, 64, 128, 256
- n_epoch: 100, 250, 500, 750, 1000
- batch_size: 1, 2, 4, 8, 12, 16
- dropout: 0.55, 0.65, 0.75, 0.85, 0.95
- lr: `loguniform`(1e-4, 1e-1)

The `ASHAScheduler` parameters were:

- time_attr: training_iteration
- metric: loss
- mode: min
- max_t: 100
- grace_period: 10
- reduction_factor: 3
- brackets: 1

The final optimised GRU hyperparameters implemented in `PyTorch` were:

- Number of hidden layers `n_hidden`: 32
- Number of recurrent layers `n_layers`: 2
- Number of epochs `n_epoch`: 250
- `batch_size`: 4
- `dropout`: 0.55
- optimiser: Adam
- learning rate `lr`: 0.003
- loss function: cross entropy

Baseline Implementations. For the hidden Markov model, the elbow method [1] was used to determine the optimal number of clusters for k-means, and a Gaussian hidden Markov model with a full covariance matrix. The optimal number of clusters was determined to be k=5. The results of using the elbow method can be seen in the Appendix, Figure 8.1.

The list of `hmmlearn GaussianHMM` parameters are:

- covariance_type: full
- n_iter: 1000
- tol: 0.001
- n_components (k): 5

The corresponding normalised gaze points G_{xBB}, G_{yBB} and their cluster labels after being classified by k-means is shown in Figure 6.3. The gaze points are clustered around the center, returning 5 clusters which are spatially close by. These cluster labels form the states which are used to train the hidden Markov model.

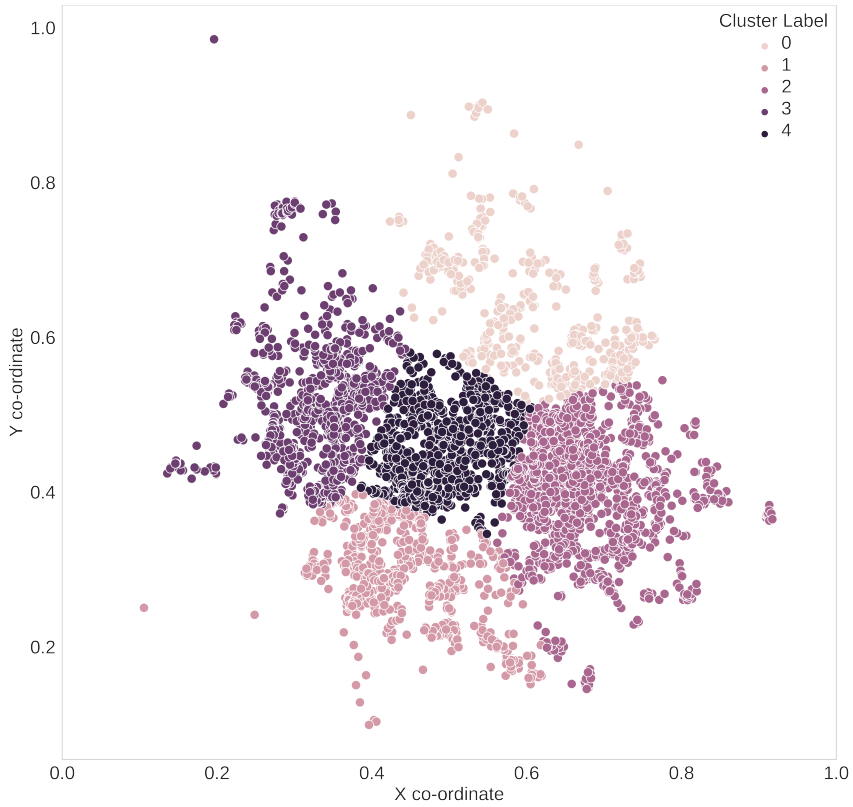


Figure 6.3: The cluster labels of gaze points after being classified by a k-means algorithm. There are 5 clusters, k, generated. These cluster labels are used as the states to train the hidden Markov models.

For the k-nearest neighbours time-series classifier, implemented using the `tslearn` [122] Python package, the optimal number of neighbours was determined to be k=2. The results of using the elbow method can be seen in the Appendix, Figure 8.2. The value of k means that the class label assigned to the gaze time-series data is dependent on its nearest 2 neighbours.

6.2.3 Qualitative Visualisation of Scanpath

A qualitative analysis was performed to understand the differences between abdomen, brain and heart plane scanpaths. The visualisation of these differences at the population level was described in Chapter 4. However, the aim of the visualisation in this section is to determine the differences between scanpaths which were classified correctly and incorrectly.

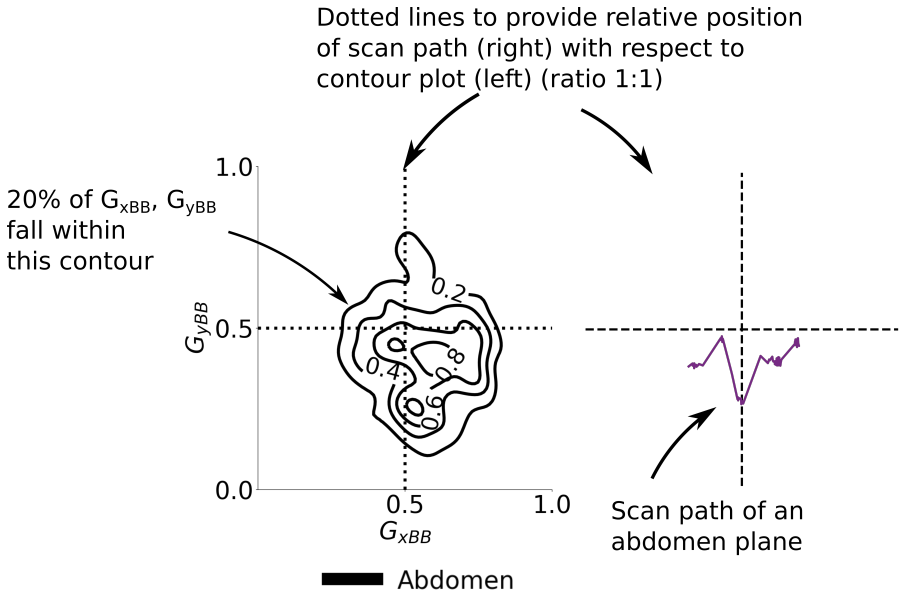


Figure 6.4: Contour density plot of abdomen gaze points normalised using the bounding box method, G_{xBB}, G_{yBB} (left), and an example abdomen scanpath (right). The dotted lines provide a 1:1 reference between the contour plot axes (left) and plotted scanpath (right). In this example, the abdomen scanpath on the left falls within the total distribution of abdomen gaze points on the right - below 0.5 on the y-axis.

This is done by visualising the distribution of the normalised gaze points, G_{xBB}, G_{xBB}, shown in Figure 6.4 on the left. Figure 6.4 (left) shows a contour density plot of training and correctly classified abdomen plane gaze points G_{xBB}, G_{xBB} with cumulative density masses at 4 equally spaced levels 0.2, 0.4, 0.6, 0.8 where 0.2 is the outer most contour and 0.8 is the inner most contour. The bi-variate distribution was calculated by superimposing a Gaussian kernel on each gaze point and returning a normalised cumulative sum. The right side of Figure 6.4 shows an example abdomen scanpath. The dotted lines on the left and right side of Figure 6.4 were drawn at a 1:1 ratio. These dotted lines provide the relative position of the example scanpath with respect to the distribution of G_{xBB}, G_{xBB}.

6.3 Results

6.3.1 Task Classification Results

Due to class imbalance, the data was randomly down sampled with respect to the minority class (abdomen plane segments), to prevent bias towards the majority

Model	Affix	Features	Weighted-F1	Average Accuracy
HMM [26]	raw	G_x, G_y	0.38±0.20	0.49±0.19
	scr	G_{xs}, G_{ys}	0.38±0.07	0.45±0.09
k-NN TSC	raw	G_x, G_y	0.57±0.05	0.57±0.06
	scr	G_{xs}, G_{ys}	0.55±0.04	0.55±0.04
	scr+A	G_{xs}, G_{ys}, A	0.52±0.05	0.54±0.04
	bb+A	G_{xBB}, G_{yBB}, A	0.63±0.03	0.64±0.02
GRU	raw	G_x, G_y	0.56±0.05	0.57±0.05
	scr	G_{xs}, G_{ys}	0.68±0.04	0.67±0.04
	scr+A	G_{xs}, G_{ys}, A	0.72±0.05	0.72±0.05
	bb+A	G_{xBB}, G_{yBB}, A (ours)	**0.84**±0.01	**0.83**±0.01

Table 6.1: Comparison of weighted F1 scores and accuracies calculated using hidden Markov model (HMM) [26], k-nearest neighbours time-series classifier (k-NN TSC), gated recurrent unit (GRU) to classify abdomen, heart and brain plane eye-tracking segments. The table shows the different eye-tracking data forms that was used to train the HMM, k-NN TSC and GRU. The four different forms of eye-tracking data were: raw eye-tracking (raw), eye-tracking normalised using the screen dimensions (scr), eye-tracking normalised using screen dimensions and area of screen occupied by the anatomical plane (scr+A) and eye-tracking normalised using the bounding box method in Section 3.4.2 and area of screen occupied by the anatomical plane (bb+A). For brevity, the column *Affix* contains the abbreviation to refer to the type of eye-tracking data used.

classes (brain plane segments). Different sets of features (shown in Table 6.1) was used to demonstrate that the proposed feature engineering method of using bounding boxes for normalisation and proposed GRU model performs better than the current baseline models. The classification results are shown in Table 6.1. The average transition and emission probabilities generated from the HMM is shown in the Appendix as Tables 8.1, 8.2, 8.3 and 8.4.

Table 6.1 shows that the best performing model normalised raw eye-tracking data with respect to the anatomy circumference (as defined in Section 3.4.2) used the GRU model architecture and performed better than previous works [26] and several baselines, returning a weighted F1 score of 0.84.

Table 6.1 shows [26] is unable to classify fetal ultrasound tasks well, where HMM(raw) and HMM(scr) returns score metrics between 38% and 49%. Instead, using k-NN TSC and GRU models improves the task classifier performance by at least 20% - HMM(raw) and HMM(scr) versus k-NN TSC(raw) and k-NN TSC(scr),

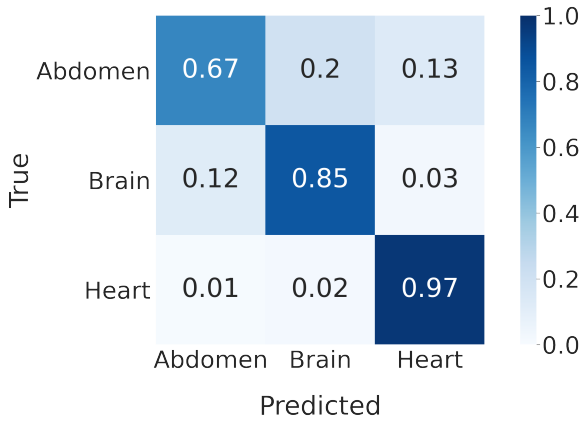


Figure 6.5: Confusion matrix for the GRU(bb+A) model normalised with respect to total number of segments per anatomy plane in the test set (106 segments in total).

GRU(raw) and GRU(scr) respectively.

Normalisation using the bounding box shows an improvement of at least 10% (Table 6.1 GRU(scr+A) versus GRU(bb+A), k-NN TSC(scr+A) versus k-NN TSC(bb+A)), returning a final F1 score of 84%. There is a slight decrease (1%-3%) in performance when including the size of the anatomy relative to the screen for models k-NN TSC, but a slight increase (4%) for GRU using scr and scr+A. GRU is better able to use the anatomy size information compared to k-NN TSC. Overall, normalising gaze points with respect to the anatomy circumference is more indicative of task type (bb), compared to how much space the anatomy occupies on the screen (A).

The confusion matrix for the GRU(bb+A) model (the best performing model) is shown in Figure 6.5. Figure 6.5 shows that heart scanpaths are the most distinct, where only 3% are misclassified. In contrast, 13% and 20% of abdomen scanpaths were misclassified as heart and brain scanpaths respectively, and 12% of brain scanpaths were predicted as abdomen scanpaths.

6.3.2 Class Imbalance Models

The initial dataset was unbalanced, with the most number of segments available for brain scanpaths and the least for abdomen scanpaths with a class imbalance ratio of 1.45. The focal loss [76] function which accounts for class imbalance when training

the model was used for the GRU(bb+A) model. The results were compared against using the cross entropy loss which did not. The effect of augmenting the dataset was also compared. The images were augmented by flipping as described in Section 3.4.1.

Downsampled	Augmented	Loss function	Weighted-F1	Average Accuracy
False	False	Focal loss	0.81±0.01	0.81±0.02
True	False	Cross entropy	0.79±0.04	0.78±0.05
False	True	Focal loss	0.83±0.01	0.81±0.02
True	**True**	**Cross entropy**	**0.84±0.01**	**0.83±0.01**

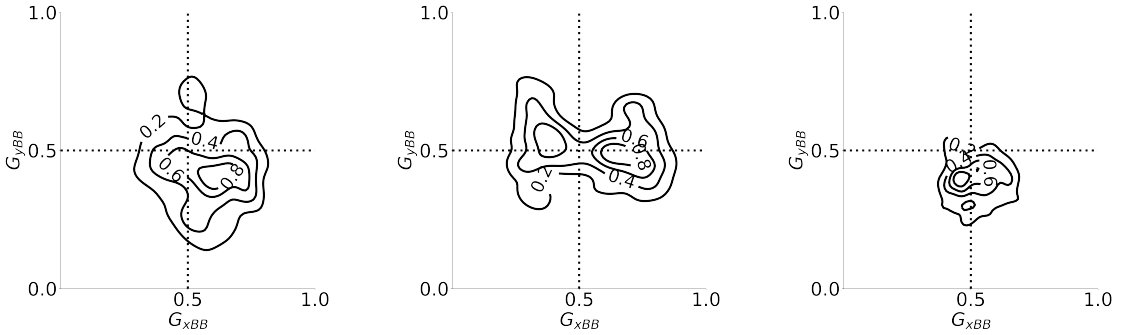
Table 6.2: Weighted F1 scores and accuracies using our proposed GRU model comparing the original, downsampled and augmented datasets. The best performing GRU model's weighted F1 scores (taken from Table 6.1) is shown in the last row of the table. The first row corresponds to the original dataset; the second row the original dataset downsampled with respect to the minority class; the third row was the original dataset that was augmented using the flipping method described in Section 3.4.1.

The results in Table 6.2 show that using an augmented dataset does not improve the performance when comparing the balanced (row 4) and imbalanced (row 3) datasets, where the weighted F1 score showed a difference of 1%. The effect of using the downsampled dataset is seen when considering a smaller dataset (row 1 vs row 2) (drop of 2-3%). Using focal loss returns more consistent results than that of using cross entropy, where the original dataset (row 1) returns a lower standard deviation across folds compared to the downsampled dataset (row 2). Overall, using an augmented downsampled dataset did not affect the performance of our model negatively (row 3 vs row 4), but increasing the size of the dataset through augmentation (row 2 vs row 4) improved performance by 4-5%.

6.3.3 Qualitative Results

A qualitative investigation was performed to understand why brain and heart scanpaths are more likely to be confused with abdomen scanpaths, and why abdomen and brain scanpaths are more often misclassified with each other compared to the heart. First, the distribution of all G_{xBB}, G_{yBB} for the abdomen, brain and heart scanpaths are plotted using contour plots (explained in Section 6.2.3) in Figure 6.6.

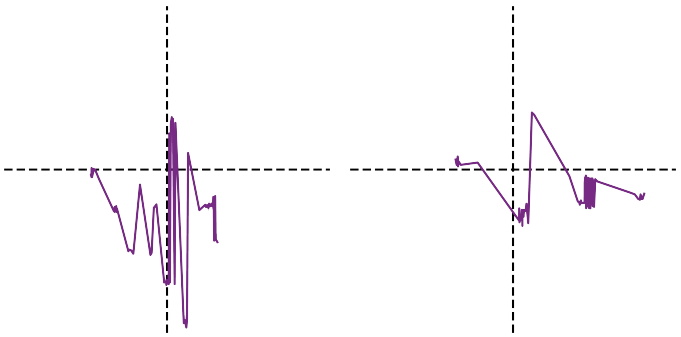
Figure 6.6a shows that for abdomen planes, sonographer scanpaths are concentrated within the central area of the anatomy. Figure 6.6b shows that for



(a) Abdomen. **(b)** Brain. **(c)** Heart.

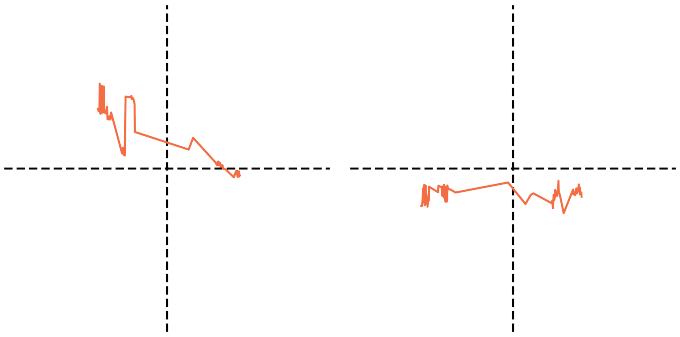
Figure 6.6: Contour plots of normalised eye-tracking data G_{xBB}, G_{yBB} of abdomen, brain and heart plane scanpaths.

brain planes, sonographer scanpaths are more elongated compared to abdomen and heart scanpaths, where the sonographer has looked across the midline of the anatomical plane. Figure 6.6c shows that for heart scanpaths, the sonographers gaze is concentrated at the center of the image where the sonographer has looked focused on and followed a single landmark. An example of individual scanpaths are shown in Figures 6.7, 6.8 and 6.9.



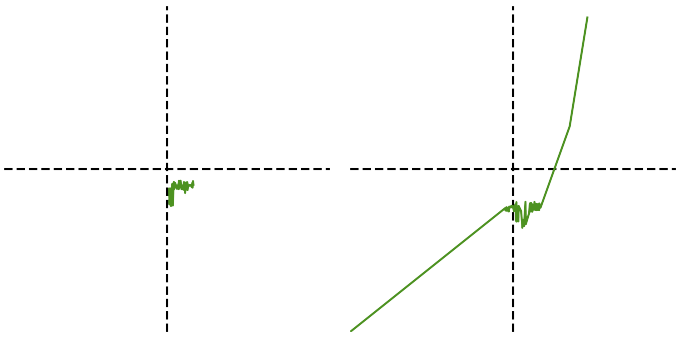
(a) Abdomen. **(b)** Abdomen.

Figure 6.7: Examples of individual abdomen scanpaths. The gaze points G_{xBB}, G_{yBB} are plotted as scatterpoints and joined with connecting lines in temporal order. In these examples, the sonographers have focused on the middle of the abdomen plane where the landmarks are located. In Figure 6.7a, the sonographer's gaze alternated between the different landmarks (up and down).



(a) Brain. **(b)** Brain.

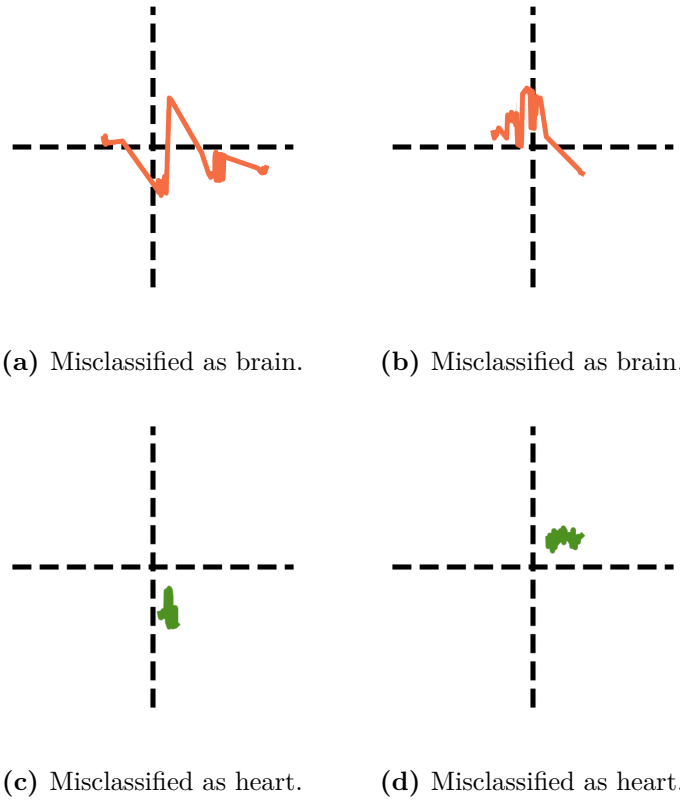
Figure 6.8: Examples of individual brain plane scanpaths. The gaze points G_{xBB}, G_{yBB} are plotted as scatterpoints and joined with connecting lines in temporal order. In these examples, the sonographers' gaze have traversed from the left to right representing their visual scanpath across the midline of the plane.



(a) Heart. **(b)** Heart.

Figure 6.9: Examples of individual heart plane scanpaths. The gaze points G_{xBB}, G_{yBB} are plotted as scatterpoints and joined with connecting lines in temporal order.In these examples, the sonographers' gaze have focused near the center of the heart plane. In Figure 6.9b, the sonographer found the landmark they were interested in and quickly honed in. When they had finished scanning, their gaze quickly veered off towards a different point.

The scanpaths which were mislabelled were also investigated. Figure 6.10 shows the abdomen scanpaths which were mislabelled as brain (Figures 6.10a and 6.10b) and heart (Figures 6.10c and 6.10d). For abdomen scanpaths predicted as heart, the sonographer focused on a single area (Figure 6.10c and 6.10d) similar to how sonographers visually search for the heart. For scanpaths predicted as brain, the sonographer moved the probe, causing their gaze to shift accordingly with the image (ii), or had moved their gaze across the screen (i) similar to how sonographers search for the brain.



(a) Misclassified as brain. **(b)** Misclassified as brain.

(c) Misclassified as heart. **(d)** Misclassified as heart.

Figure 6.10: Abodomen scanpaths which were misclassified as brain and heart scanpaths. For scanpaths which were mislabelled as brain (Figures 6.10a and 6.10b), the sonographers' visual scanpath mimiced a brain scanpath, traversing 'across' the plane. For scanpaths which were mislabelled as heart (Figures 6.10c and 6.10d), the sonographer had focused in the middle of the plane.

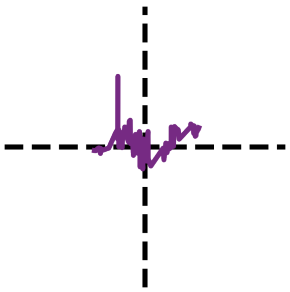
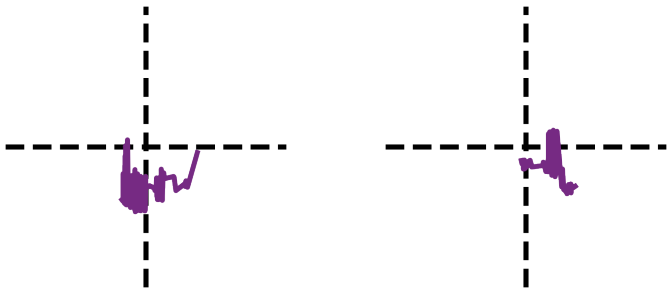


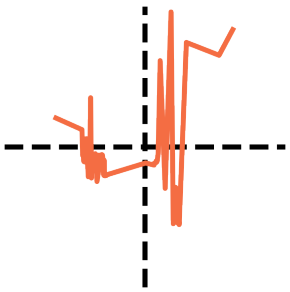
Figure 6.11: Brain scanpaths which were misclassified as an abdomen scanpath. The sonographer had not scaled the brain plane to fit the screen fully, and did not focus along the midline horizontally but diagonally across the plane.

For misclassified brain scanpaths (Figure 6.11), the image was small and occupied <50% of the screen, and the sonographer did not focus along the midline horizontally

but diagonally across the plane.



(a) Misclassified as abdomen. **(b)** Misclassified as abdomen.



(c) Misclassified as brain.

Figure 6.12: Heart scanpaths which were misclassified as an abdomen and brain scanpaths. For those mislabelled as abdomen scanpaths (Figures 6.12a and 6.12b), the sonographer had looked around the area of focus, as opposed to just focusing on the landmark and 'following' the landmark as they adjusted the probe. For the heart scanpath which was mislabelled as a brain scanpath, the sonographer was moving their gaze and the image simultaneously, resulting in a scanpath which was 'sweeping', rather than focused.

Misclassified heart scanpaths (Figure 6.12c) showed that the image itself was moving, indicating that the probe was moving, causing the sonographer to shift their gaze accordingly or the sonographer was looking around the walls of the heart cavity.

6.4 Discussion

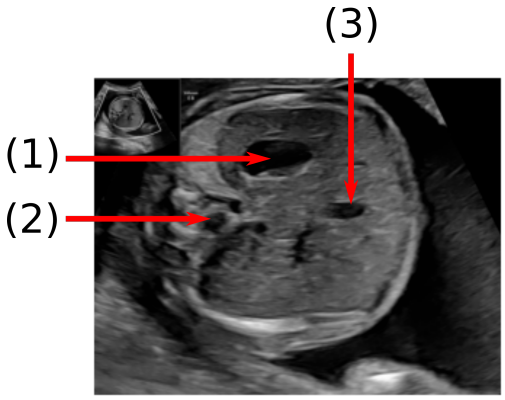


Figure 6.13: Labelled anatomical landmarks of the abdomen plane: stomach, aorta and umbilical vein.

Given eye-tracking data of sonographers, differences between type of task being performed were classified with a weighted F1 score of 84%. Abdomen and brain plane scanpaths were misclassified because the sonographer did not necessarily look at the distinguishing landmarks, and showed a slightly different visual search strategy. On average, the population of sonographers looked at the distinguishing landmarks. However, in some instances, such as the abdomen plane, the sonographer focused at the center point of all three landmarks: stomach, aorta and umbilical vein (Figure 6.13). This specific search strategy is more similar to heart scanpaths, where the sonographer focused on and followed a single landmark.

This chapter also showed the normalising eye-tracking data with respect to the area-of-interest, in this case the fetal anatomical plane being searched for, is important for contextualising the gaze data. Without this prior knowledge, the classification model performs quite poorly, with a weighted F1 score of 56% (Table 6.1). The area of the screen occupied is not that important a factor to consider. The classification results only improve by 4% (Table 6.1, using the GRU model) compared to the feature engineering sets scr, and scr+A. This suggests the sonographers on average scale the fetal plane to approximately the same size. This

is not surprising since the guidelines in [90] suggest that the fetal plane should occupy at least $\frac{2}{3}$ of the screen.

6.5 Summary

In this chapter, the method presented was able to classify the fetal ultrasound task being performed with a weighted F1 score of 84%. Gaze was normalised to provide context of the sonographer's behaviour with respect to the plane being searched for. Then, a time-series classification model was used to classify the normalised eye-tracking data into separate planes: abdomen, brain and heart. The final model which returned the best performance was the gated recurrent unit model.

The final classification results demonstrate that the location of the anatomical landmarks and method of sonographer scanning affects how well the gaze data can be classified. These findings are useful for understanding that, even if the scan presents the plane in a different position and location on the screen, sonographers largely observe the same landmarks and or look within the vicinity of the final expected landmark position.

7

Skill and Style Classification

Contents

7.1 Introduction 103
7.1.1 Contribution 105
7.1.2 Data 105
7.1.3 Definitions 105
7.1.4 Software Packages 106
7.2 Method 106
7.2.1 Generating Features from Eye-Tracking Data 107
7.2.2 Classification Models 108
7.2.3 Skill and Style of Sonographer Gaze 109
7.2.4 Class Imbalance Sampling 111
7.2.5 Years of Scanning and Level of Expertise 112
7.2.6 Implementation 115
7.3 Results 116
7.3.1 Skill Classification Models 116
7.3.2 Qualitative Analysis of Trees 117
7.3.3 Relationship between Years of Scanning and Level of Expertise 119
7.4 Discussion 121
7.5 Summary 123

7.1 Introduction

Classification of human skill in the context of medical applications is still largely limited by the definition of skill. The motivation behind quantifying or classifying hu-

man skill at fetal sonography is to find out whether there *are* quantitative differences between an experienced and less experienced sonographer. If so, these differences can be used to improve the scanning skill of the less experienced sonographer.

In fetal ultrasound, the number of years a sonographer has spent scanning is typically used to group sonographers into different expert categories. Current studies do not necessarily test whether this skill definition is suitable, i.e., whether years of experience is sufficient as an indicator of skill (Section 2.3.4). They also do not consider other factors such as a sonographer's personal scanning *style* can affect the definition of skill. Style is defined as in [123]; the gaze of a sonographer is the outcome of both human skill, and personal scanning style. Subsequently, there is a need to validate whether using years of experience to build skill classification models is suitable, and whether a sonographer's style of scanning is a factor that needs to be taken into account.

In this chapter, I aim to determine whether a sonographer's scanning style is a factor that affects the chosen definition of expertise. Then, I aim to present a method that will test the hypothesis that skill and years of scanning experience are correlated. The main goal is to understand whether years of experience is a suitable measure of skill, since this has been widely used in prior studies.

To remove the 'years-of-experience' constraint when defining skill, a skill classification model is trained to distinguish between experts and non-experts, where an expert is defined as a fully qualified sonographer, independent of number of years spent scanning, and a non-expert as a trainee sonographer. Then, to consider whether a sonographer's style affects the model's results, the population of experts used to train the model is stratified based on individual sonographers. The effect of changing the model's training data on the test dataset is investigated. This method was based on the paper which was published and presented at *Gaze Meets ML workshop in conjunction with NeurIPS 2022*, but its contents have been modified to suit the format of this thesis.

7.1.1 Contribution

The presented method builds a task-agnostic skill classification model using only eye-tracking and pupillary data of sonographers performing fetal ultrasound scans. The correlation coefficient is calculated between years of scanning experience and skill of expert sonographers and its significance (at the 5% level). The presented method shows that task-agnostic eye-tracking data can be classified into expert and trainee with a weighted average F1 score of 98% (expert) and 70% (trainee) using a light gradient boosting machine model, and that, depending on the skill model used, generally there can be a positive relationship between years of experience and expertise.

7.1.2 Data

The dataset used to build the task-agnostic skill classification model was eye-tracking trainee data described in Section 3.3.4 and the eye-tracking data from the population of expert sonographers performing second trimester scans as detailed in Section 3.3.5. Any missing data was interpolated using the method described in Section 3.4.2.

7.1.3 Definitions

Before describing the method, I define some of the terms in this chapter.

- Expert: refers to a fully-qualified (FQ) sonographer who has been qualified to work with the National Health Service (NHS), independent of their years of scanning experience. An expert could be a sonographer with 2 years of scanning experience, or 10 years of scanning experience for instance.
- Trainee: refers to an individual who is learning how to scan and training to become a fully qualified sonographer. A trainee is an individual who has not been accredited with the qualifications to scan with the NHS. An expert sonographer is performing the scan and training them simultaneously.
- Style: the gaze of a sonographer is the outcome of both human skill, and personal scanning style [123].

7.1.4 Software Packages

I used the software package `tsfresh` to calculate time-series features based on the eye-tracking data to build the skill classification model. In the package `tsfresh`, one of the parameters that needs to be defined is the type of features to be calculated. The features are categorised based on the computational speed required to calculate them. These features are pre-defined in `tsfresh` and the full list of features can be found in [83]. In this chapter, I used the category `EfficientParameters`. Both `tsfresh` and `EfficientParameters` are described briefly below.

- `tsfresh`: A `Python` package which calculates a range of time-series properties, such as distribution of data points, correlation properties, stationarity, entropy, and nonlinear time-series analysis [83].
- `EfficientParameters`: A parameter in `tsfresh` which specifies which time-series properties to calculate – those that are not computationally expensive and is scalable for large datasets [83]. The features are a combination of continuous and categorical variables. In total there are 74 unique time-series features. The full list of features are found in the tables in Appendix 8.4.

7.2 Method

In this work, I first consider whether a sonographer's gaze behaviour is a function of both skill and style. Then, I consider whether years of experience is a suitable indicator of human skill.

In this subsection, I first describe the gaze features which were used to build the skill models. Then I describe the classification models to classify skill and sampling methods used to account for class imbalance. After which, I present the methods which were used to answer the two research questions posed: whether skill and style are entangled in sonographer gaze behaviour, and if years of experience is a suitable indicator of skill.

7.2.1 Generating Features from Eye-Tracking Data

In prior fetal sonography skill studies [138, 123, 150], specific tasks such as the brain and heart anatomical plane were analysed. This requires a labelling method to identify the planes of interest. Instead, in this chapter, I consider a task-agnostic approach. A task-agnostic approach means that the eye-tracking data does not need to be separated into different segments corresponding to specific anatomical plane reducing labelling efforts. To analyse skill, only eye-tracking data related to the live B-mode segments where sonographers are actively searching for an anatomical plane are considered.

Following [136] where pupillary data was used to compare skill differences between sonographers with > 2 years and ≤ 2 years of experience, task-evoked pupillary response (TEPR) is used as a skill classification feature. The TEPR is used to calculate the cognitive load of the participant whose gaze is being recorded. TEPR measures the change in pupil dilation with respect to a baseline pupil diameter. A larger change in TEPR is indicative of a higher cognitive load, and vice versa. The equation for calculating TEPR is given as δd_t in Equation 7.1. The minimum pupil diameter d_r represents the sonographer's pupil diameter while resting. d_t represents the pupil diameter at time t and δd_t represents the TEPR at time t [136].

$$\delta d_t = \frac{d_t - d_r}{d_r} \times 100\% \tag{7.1}$$

Gaze data (x and y co-ordinates) is also used as features. Each live B-mode segment is represented by a $3 \times n$ feature vector, where n is its segment length and 3 is the number of final features that were used to train the model - gaze x and y co-ordinates and δd_t. Note that n varies from segment to segment.

To overcome the problem that the live B-mode segments are of different lengths, summarized gaze characteristics are extracted for each segment using the scalable feature extraction approach `tsfresh` [83] (described in Section 7.1.4) returning the eye-tracking data in tabular data form. Tabular data is a data form where each row represents a single unique instance of the time-series data, and each column

represents a specific type of feature. The feature is then reduced to a $1 \times m$ feature vector, where m is the number of characteristics extracted using `tsfresh` [83].

The feature extraction setting used in `tsfresh` was `EfficientParameters` [83]. In fetal ultrasound, due to the unstructured nature of searching for anatomical planes, the time taken per segment is not necessarily a fair indicator of skill. Hence features related to length are removed: `length`, and `ratio_value_number_to_time_series_length`, which calculate the length of the segment n and the number of unique values in the segment divided by n, respectively. The `impute` method in `tsfresh` is used to select relevant features for prediction.

7.2.2 Classification Models

To build the task-agnostic model, gradient boosting decision tree models were chosen because it has been shown that they work best with tabular data, rather than time-series data (like that in Chapter 6). Gradient boosting decision trees use an ensemble of weak decision trees to build strong predictors [84, 14]. Random forests utilise a bagging approach, where the final classification is taken as the majority vote. However, gradient boosting trees use a sequential approach - the model's parameters are updated sequentially based on the residual of each individual tree. The gradient is *boosted* over each tree. I briefly describe the 3 models, Extreme Gradient Boosting, Light Gradient Boosting Machine and Categorical Boosting, which have been shown to work best with tabular data. These models are used in this chapter to classify skill.

Extreme Gradient Boosting. (XGBoost) XGBoost is a highly scalable and efficient gradient tree boosting algorithm that can handle sparse tabular data because of its algorithmic optimisations detailed in [62]. Briefly, their optimisations include a sparsity aware algorithm, a weighted quantile sketch and a cache-aware implementation. The sparsity aware algorithm learns the patterns of missing values in the dataset (≈ 50 speed up compared to without), where missing values could be due to frequent zeros in statistics or one-hot encoding. Their weighted quantile

sketch algorithm finds optimal split points amoung the datasets. Lastly, their cache-aware implementation ($\approx$ 2 speed up) speeds up computation by pre-fetching gradient statistics stored in an internal buffer [62].

Light Gradient Boosting Machine. (LightGBM) LightGBM includes two extra optimisation steps to handle large amounts of data instances and features, decreasing the computational speed and memory required compared to XGBoost. Their two optimisation methods are *gradient-based one-side sampling* and *exclusive feature bundling*. Gradient-based one-side sampling amplifies the under-trained data points (data points with small gradients) when calculating information gain. Exclusive feature bundling groups features which are mutually exclusive into a single feature [74].

Categorical Boosting. CatBoost is similar to XGBoost and LightGBM but is specifically designed to handle categorical features. Instead of pre-processing categorical features (e.g. one-hot encoding), they are handled during training by using ordered target statistics, i.e. each category has an estimated target value. There is an $\approx$ 25-60 times speed up when compared to XGBoost and LightGBM [84].

7.2.3 Skill and Style of Sonographer Gaze

To understand whether gaze patterns of a sonographer scanning is influenced by their own personal style, I built several skill models and compare their classification results. First, a skill classification model is built using a population of expert and trainee sonographers eye-tracking data. This model is abbreviated as $EX_{0,16}$ in Table 7.1; the abbreviation refers to the years of scanning experience of sonographers in the training dataset, 0 to 16 years of scanning after qualifying. The full list of abbreviations and corresponding years of experience is seen in Table 7.1. To build the model, all the trainees data was used, and 20% of the experts' data was used [1].

[1]The reason why 80% of the experts' data was set aside is explained in detailed Section 7.2.5.

Expert	Expertise (years)
20% of $EX_{0,16}$	0-16
$EX_{1,2}$	1-2
$EX_{2,3}$	2-3
$EX_{0,3}$	0-3
$EX_{10,11}$	10-11
$EX_{14,15}$	14-15

Table 7.1: Table of groups of experts represented in the training dataset for skill classification, with their corresponding number of years of scanning experience. The abbreviation for these experts are $EX_{a,b}$, where EX stands for expert, and a, b represents the lower and upper bound of number of years of scanning experience.

Data used to build the 'leave-one-in' skill models		Percentages of Data Used
Trainee Class	Expert Class	
Trainees 1-4 (as in Section 3.3.4)	20% of $EX_{0,16}$ $EX_{1,2}$ $EX_{2,3}$ $EX_{0,3}$ $EX_{10,11}$ $EX_{14,15}$	75% Train+Validate 20% Test 5% Tune

Table 7.2: Table showing the percentage of data used to build the skill classification models. For the 'leave-one-out' models, the individual experts' data was used to build the model. For the general skill model that used all the experts' data to represent the expert class, only 20% of the total experts' data was used. 80% was set aside to test the correlation between years of experience and expertise. For the trainee class, all the trainee data was used.

Then, I consider a reversed leave-one-out approach, analogous to 'leave-one-in'. A leave-one-out approach removes an individual sonographer from the training dataset to investigate how the individual influences the classification results. Conversely, a leave-one-in approach uses only a single sonographer in the training dataset. It seems counter intuitive, but the aim of the experiment is to determine whether one expert's gaze patterns are similar enough to another, that the classification results would also likewise be similar.

Due to data imbalance, experts with the most (top 5, Table 7.2) eye-tracking data is shown. These models are abbreviated as $EX_{1,2}$, $EX_{2,3}$, $EX_{0,3}$, $EX_{10,11}$, $EX_{14,15}$ in Table 7.1 and 7.2. If a sonographer's unique scanning style does not

affect their skill, then the results of the different models $EX_{1,2}$, $EX_{2,3}$, $EX_{0,3}$, $EX_{10,11}$, $EX_{14,15}$ would not vary by much. A description of the data used to train the trainee and expert class is shown in Table 7.2. The data breakdown for training, testing, validating and tuning the model's parameters is shown in Table 7.2. A 5-fold stratified cross-validation is carried out with 75% of the dataset, and tested on the remaining 20%. Model parameters are tuned using a `GridSearch` with 5% of the dataset. In all 5 models, the trainee class consists of all 4 trainees' data. The expert class consists of only the specific expert sonographer's data.

7.2.4 Class Imbalance Sampling

Expert	Expertise (years)	≈ Class Imbalance Ratio
20% of $EX_{0,16}$	0-16	14
$EX_{1,2}$	1-2	23
$EX_{2,3}$	2-3	8
$EX_{0,3}$	0-3	17
$EX_{10,11}$	10-11	3
$EX_{14,15}$	14-15	15

Table 7.3: Table of groups of experts represented in the training dataset for skill classification, with their corresponding number of years of scanning experience. The table also includes a class imbalance ratio of the expert class and trainee segments available for training; the expert class is the majority class. The abbreviation for these experts are $EX_{a,b}$, where EX stands for expert, and a, b represents the lower and upper bound of number of years of scanning experience.

The data is imbalanced towards the expert class and a breakdown of the class imbalance ratio is shown in Table 7.3. Due to class imbalance where experts form the majority class, Synthetic Minority Oversampling Technique (SMOTE) [15, 136] is used to balance the training dataset. SMOTE is a sampling method to address class imbalance - the minority class is over-sampled by creating synthetic data points based on the original data points. The synthetic points are generated along any line segments which connect any of the minority class data points. Such imbalances in data are not uncommon, where other fetal sonography studies have also had an imbalanced expert/beginner dataset [137, 123]. This imbalance is

further amplified when considering separating sonographers on a per-year (of scanning experience) basis.

7.2.5 Years of Scanning and Level of Expertise

After determining whether style is a factor that influences gaze patterns, I want to investigate whether years of experience is a suitable measure of expertise. In prior studies, expertise of sonographers is defined as years of scanning experience. Since the data was collected over a period of several years, the years of experience is taken as the number of years of scanning the sonographer had when they performed the scan.

In this subsection, *level of expertise* is defined as the percentage of eye-tracking data segments of an expert that is classified as expert. That is, using a skill classification model, predict whether an expert's data is an expert or trainee, and calculate the total percentage of their segments which were labelled as expert. The trained skill classification model can identify expert segments which are more similar to trainee segments (i.e., expert segments which are misclassified as trainee segments).

The skill classification models built in Section 7.2.3 are used to predict whether an experts' data is classified as a trainee or an expert. For each model, the data used to predict the level of expertise of experts' is shown in Table 7.4. These data were not used to train, test, validate and tune the model in Section 7.2.3.

Data used to build the 'leave-one-in' and population skill models		Dataset aside for prediction of expertise level
Trainee Class	Expert Class	
Trainees 1-4	20% of $EX_{0,16}$ $EX_{1,2}$ $EX_{2,3}$ $EX_{0,3}$ $EX_{10,11}$ $EX_{14,15}$	80% of $EX_{0,16} \notin$ 20% of $EX_{0,16}$ $EX_{0,16} \notin EX_{1,2}$ $EX_{0,16} \notin EX_{2,3}$ $EX_{0,16} \notin EX_{0,3}$ $EX_{0,16} \notin EX_{10,11}$ $EX_{0,16} \notin EX_{14,15}$

Table 7.4: Table showing the data which was used to build the skill classification models in Section 7.2.3, and the data used for predicting the level of expertise in Section 7.2.5. To read the notation used in the column *'dataset aside for prediction'*, $EX_{0,16} \notin EX_{14,15}$ means that all the experts' data in $EX_{0,16}$ was used to predict trainee/expert except $EX_{14,15}$.

To calculate whether years of experience and level of expertise exhibit a positive monotonic[2] relationship, the Spearman's rank correlation coefficient (SC) between the two variables, years of experience and levels of experience, is computed using (Equation 7.2), where 'R' denotes the rank of the variable. The rank of the variable can be calculated as follows. If X={22%, 37%, 10%, 60%}, the rank of X, R_X, is returned as R_X={3, 2, 4, 1}, where the highest rank 1 is the highest score.

$$SC_{XY} = \frac{cov(R_X, R_Y)}{\sigma_{R_X}\sigma_{R_Y}} \tag{7.2}$$

The SC measures the monotonic relationship between two variables, X and Y (Equation 7.2), by calculating the covariance between the ranks of X and Y, $cov(R_X, R_Y)$, divided by the standard deviation of the rank of X and Y multiplied. The equation for calculating the Spearman's rank correlation coefficient is given as SC_{XY} in Equation 7.2. σ refers to the standard deviation, and *cov* refers to the covariance. The range of the SC is between -1 and 1.

[2]A function which is either entirely nonincreasing or nondecreasing. `https://mathworld.wolfram.com/MonotonicFunction.html`

Spearman's Coefficient	Analysis
SC > 0	Expertise increases with years of experience.
SC < 0	Expertise decreases with years of experience.
SC ≈ 0	Expertise and years of experience are not monotonically related.

Table 7.5: Table showing the interpretation of the Spearman's rank correlation coefficient (SC) in relation to years of experience and expertise.

In general, SC > 0 signifies X and Y are increasing. SC < 0 signifies as X increases Y decreases. SC $= 0$ signifies that X and Y do not share an increasing or decreasing relationship. The variables X and Y in Equation 7.2 are years of scanning experience (X variable) and percentage of expert segments (between 0 and 100%) (Y variable) respectively. The rank R_X, R_Y of the X and Y variables are calculated as shown below. A summarised version of this analysis is shown in Table 7.6.

Years of Experience (X)	1	4	5	7	8	9	10
R_X	1	2	3	4	5	6	7
Level of Expertise	10%	20%	25%	30%	40%	55%	55%
R_Y	1	2	3	4	5	7	7

Table 7.6: Calculating the rank of the variables: years of experience and level of expertise. The higher the level of expertise, the higher the rank. The higher the number of years of experience, the higher the rank. Values which are identical share the same rank. For example, where Y=$\{55\%\}$, $R_Y = 7$.

The behaviour that I am investigating is qualitatively shown in Figure 7.1. Figure 7.1 shows the percentage of segments predicted as expert increasing with years of experience; SC > 0.

As the data was collected over a period of several years (2018 to 2023), there were 2-3 sonographers whose years of experienced increased. For example, a sonographer who started scanning for the PULSE project in 2018 might have also scanned for the project in 2019. Their recorded years of experience would then increase by one (Table 7.7). However, to perform the Spearman's coefficient correlation, there is an assumed independence between the 2 variables being compared. To overcome this challenge, the analysis is performed as follows.

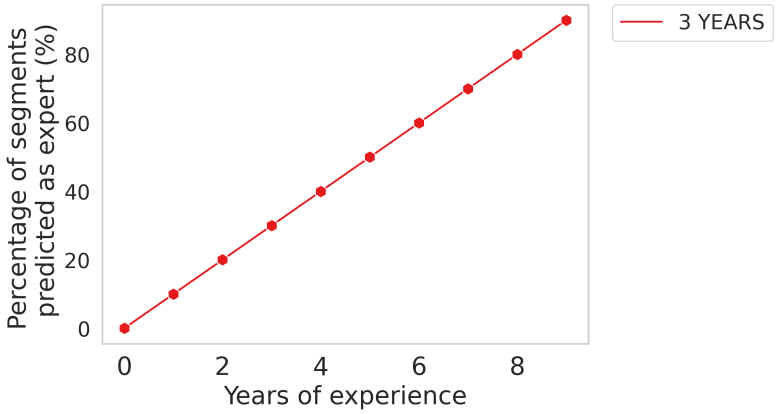


Figure 7.1: An example lineplot of a positive monotonic relationship between years of scanning experience and percentage of segments predicted as expert. In this subsection, I have defined percentage of segments predicted as expert as *levels of expertise*. The legend '3 YEARS' refers to the years of experience the expert used to train the skill classification model.

Years of Experience	1	2	3	4	5	6	7
Level of Expertise	10%	20%	25%	30%	40%	50%	55%
Sonographer ID	A	B	Richard	Richard	C	David	David

Table 7.7: Example of data where a sonographer's data appears in 2 different bins of years of experience. Years of experience is the independent variable (X), and level of expertise is the dependent variable (Y). David and Richard's data appears in 2 different years of experience.

Where a sonographer's ID appears multiple times, the Spearman's coefficient is calculated as follows. A coefficient is calculated using X={10%, 20%, 25%, 40%, 50%} and Y={1, 2, **3**, 5, **6**}. Another coefficient is calculated using X={10%, 20%, 30%, 50%, 55%} and Y={1, 2, **4**, 5, **7**}. The analysis was performed this way to ensure that that assumption independence between X and Y is not violated.

7.2.6 Implementation

The following parameters were tuned using `GridSearch` to find the optimal parameters for the classification decision tree models. To prevent overfitting, the max depth of the decision trees are limited to 10.

The tuning parameters used for Light GBM were:

- learning_rate: `numpy.arange(0.1, 1, 0.1)`
- num_leaves: `range(2, 20, 3)`

- max_depth: `range(3, 10, 2)`

The tuning parameters used for xGBoost were:

- learning_rate: `numpy.arange(0.1, 1, 0.2)`
- max_depth: `range(3, 10, 2)`
- min_child_weight': `range(1, 6, 2)`
- gamma: `numpy.arange(0.1, 1, 0.2)`
- eval_metric: mlogloss

The tuning parameters used for CatBoost were:

- learning_rate: `numpy.arange(0.1, 1, 0.2)`
- max_depth: `range(2, 10, 2)`

7.3 Results

7.3.1 Skill Classification Models

Table 7.8 shows the average results of the model's performance on the test set across the 5 folds. On average, both LightGBM and XGBoost outperform CatBoost. This is not unexpected since the number of continuous features in the dataset is more than the number of categorical features. Given that class imbalance favours the majority class (expert), it is not surprising that the performance of the expert class is much better than that of the trainee class, with average F1 scores of at least 94%. The best performing model based on the trainee class performance uses an XGBoost architecture and $EX_{10,11}$ as the expert. It achieves an F1 score of 95% for the expert class and 88% for the trainee class.

Model	LightGBM		XGBoost		CatBoost	
Data	Expert	Trainee	Expert	Trainee	Expert	Trainee
$EX_{0,16}$	0.98±0.00	**0.70±0.03**	0.98±0.00	0.66±0.04	0.96±0.00	0.50±0.01
$EX_{1,2}$	0.99±0.00	0.71±0.04	0.99±0.00	0.74±0.01	0.97±0.00	0.58±0.02
$EX_{2,3}$	0.98±0.00	0.71±0.03	0.97±0.00	0.65±0.03	0.96±0.00	0.54±0.01
$EX_{0,3}$	0.99±0.00	0.84±0.02	0.99±0.00	0.80±0.02	0.98±0.00	0.68±0.04
$EX_{10,11}$	0.95±0.00	0.86±0.01	**0.95±0.01**	**0.88±0.02**	0.94±0.01	0.84±0.02
$EX_{14,15}$	0.98±0.00	0.72±0.02	0.98±0.00	0.71±0.02	0.95±0.01	0.52±0.02

Table 7.8: Average F1 scores on the test dataset using the models built from the different training datasets described in Table 7.4.

The performance of the trainee class depends on which experts were used in training, with F1 scores between 71% and 86% (Table 7.8, LightGBM). When comparing similar years of experience, $EX_{14,15}$ and $EX_{10,11}$, $EX_{0,3}$ and $EX_{1,2}$, there is a difference of at least 13% (Table 7.8). These results suggest that when considering a skill classification model, a sonographer's style is also a factor that is not easily disentangled from their skill. As a result, misclassification of trainee segments is dependent on the style of the expert's gaze and how similar their gaze data was to the expert.

The trainee class was highly imbalanced in some of the training data, such as $EX_{0,3}$ and $EX_{1,2}$ (factors of 17 and 23 respectively, Table 7.3). A comparison of $EX_{0,3}$ and $EX_{10,11}$, which had an imbalance ratio of 17 and 3 respectively, returned similar results for the best-performing model. When comparing $EX_{0,3}$ and $EX_{1,2}$, both had between 0-3 years of experience but a 13% difference in performance for the trainee class. Similarly, $EX_{10,11}$ and $EX_{14,15}$ had a 14% difference. These results suggest that although class imbalance could have caused the minority class (trainee) to perform worse than the expert class, it is more likely that the gaze behaviour of a sonographer is dependent on their scanning style, causing different representations of experts to return a range of model performances.

7.3.2 Qualitative Analysis of Trees

A qualitative analysis of the best performing model general skill model ($EX_{0,16}$) is discussed here. The best performing model is in bold in Table 7.8, where the sonographer used to train the 'expert' class was the population of sonographers $EX_{0,16}$ using LightGBM. As a 5-fold stratified cross validation was performed, the analysis from the fold with the least F1 score is presented here. The full list of features and their importance is shown in the tables in the Appendix 8.4. The feature importance is presented as the numbers of times the feature is used in the model, following the in-built method of the `lightgbm` package.

The fold with the lowest F1 score was the 2nd fold, with an F1 score of 0.98 (expert) and 0.65 (trainee) and its qualitative results are presented here.

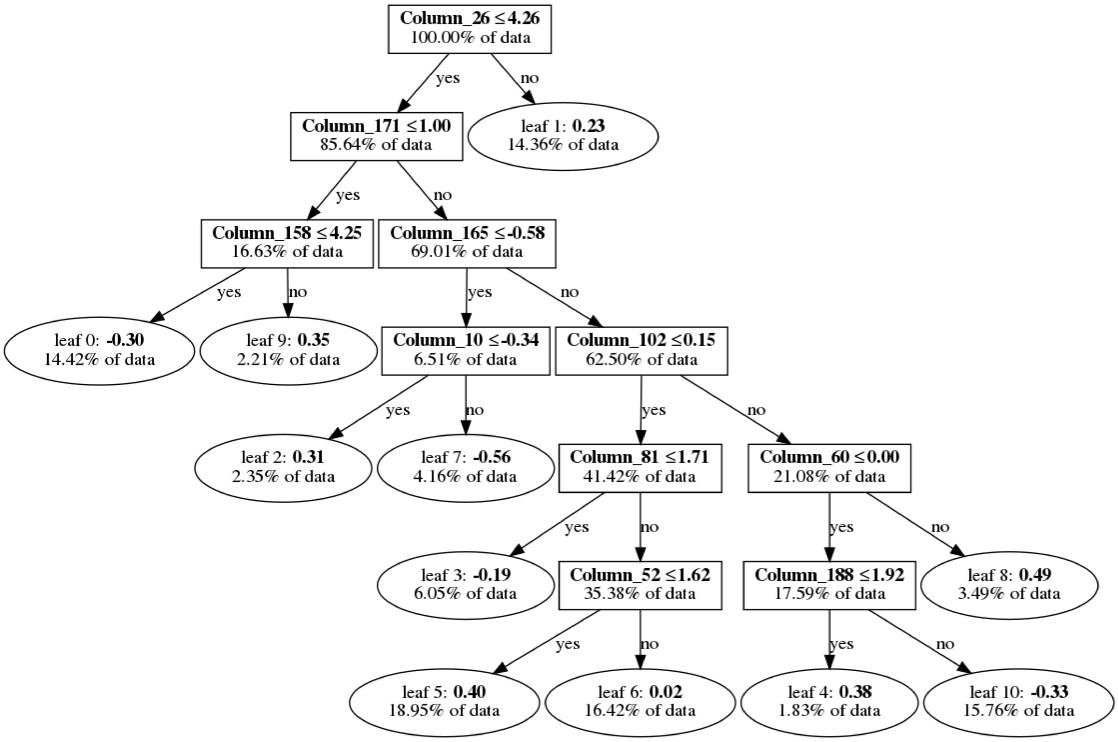


Figure 7.2: The 50th boosted tree (100 trees were trained in total) trained. The tree was generated using the in-built LightGBM `plot_tree` function. The tree shows the percentage of data which falls into each leaf. The names of the features which were used at the splits are given as columns 26, 171, 158, 165, 10, 102, 81, 60, 52, 188. The list of features can be found in the list below.

The list of features in Figure 7.2 are given below:

1. Column 26: x_permutation_entropy___dimension_5___tau_1
2. Column 171: TEPR_range_count___max_1___min_-1
3. Column 158: x_change_quantiles___f_agg_"mean"___isabs_True ___qh_0.8___ql_0.4
4. Column 165: TEPR_cwt_coefficients___coeff_14___w_5___widths _(2, 5, 10, 20)
5. Column 10: x___benford_correlation
6. Column 102: x_change_quantiles___f_agg_"mean"___isabs_False ___qh_0.4___ql_0.2
7. Column 81: x_change_quantiles___f_agg_"mean"___isabs_True ___qh_0.6___ql_0.4

8. Column 60: TEPR__ratio__beyond__r__sigma___r__6
9. Column 52: TEPR__permutation__entropy___dimension__3___tau__1
10. Column 188: x__change__quantiles___f__agg__"mean"___isabs__True ___qh__0.6___ql__0.2

One interesting observation is that the gaze feature TEPR (task-evoked pupillary response) has a similar number of features which contribute to the skill classification model, compared to the gaze co-ordinates. Specifically, there are 90, 44 and 61 x, y and TEPR features which has non-zero feature importance values (Appendix 8.4) In particular, there were 5, 4 and 7 x, y and TEPR features which have a feature importance of >10. This is in line with prior research discussed in the literature review (Section 2.3) which used pupillary response to separate groups of experts. TEPR could also be affected by the luminance of the screen whilst scanning. For example, an expert might spend less time looking at the screen compared to someone who was less confident, or spend less time overall looking at specific landmarks. Hence, the importance of TEPR could be a result of both differences in cognitive load and also of skill. The importance of the x co-ordinate also suggests that how the sonographer is traversing the plane horizontally (looking left to right), rather than vertically, is of importance.

7.3.3 Relationship between Years of Scanning and Level of Expertise

To investigate the correlation between years of scanning experience, the Spearman's correlation coefficient is calculated between years of scanning experience and percentage of expert segments predicted as expert. I use the best performing $EX_{0,16}$ (representing a population of expert sonographers) model architecture from Section 7.3.1, which was the light gradient boosting machine.

The results are shown in Table 7.9. Table 7.9 shows that the models using experts $EX_{0,3}$, $EX_{10,11}$, $EX_{14,15}$ to train the model returned a positive monotonic relationship between years of experience and level of expertise. This is unlike the

models which were trained using 20% of all experts' data $EX_{0,16}$, $EX_{1,2}$ and $EX_{2,3}$. The classification results in Section 7.3.1 showed that the style and skill are not easily disentangled and Table 7.9 corroborates these results.

Model	(Spearman's Coefficient, p-value)
$EX_{0,16}$	(0.13, 0.75) (-0.04, 0.93)
$EX_{1,2}$	(-0.47, 0.21) (-0.52, 0.15)
$EX_{2,3}$	(-0.09, 0.81) (-0.08, 0.85)
$EX_{0,3}$	**(0.78, 0.01)** **(0.70, 0.04)**
$EX_{10,11}$	**(0.86, 0.01)** **(0.86, 0.01)**
$EX_{14,15}$	**(0.81, 0.02)** **(0.85, 0.01)**

Table 7.9: Table of Spearman's coefficient and p-values between years of experience and percentage of expert segments predicted. The coefficient was calculated twice where a sonographer's ID appeared twice using the method outlined in Section 7.2.5.

The results are investigated qualitatively, as shown in Figures 7.3 and 7.4. Generally, Figure 7.3 shows a positive monotonic relationship plot of the level of expertise against years of experience for models $EX_{0,3}$, $EX_{10,11}$, $EX_{14,15}$. Figure 7.3 shows the lineplots of the level of expertise against years of experience for models $EX_{0,16}$, $EX_{1,2}$ and $EX_{2,3}$ where there is generally neither a strong positive or negative relationship between years of experience and expertise. These lineplots show a different behaviour to that of Figure 7.3. When every expert is represented in the class ($EX_{0,16}$), the Spearman's coefficient is 0.13 and -0.04, suggesting that expertise neither increases or decreases with experience.

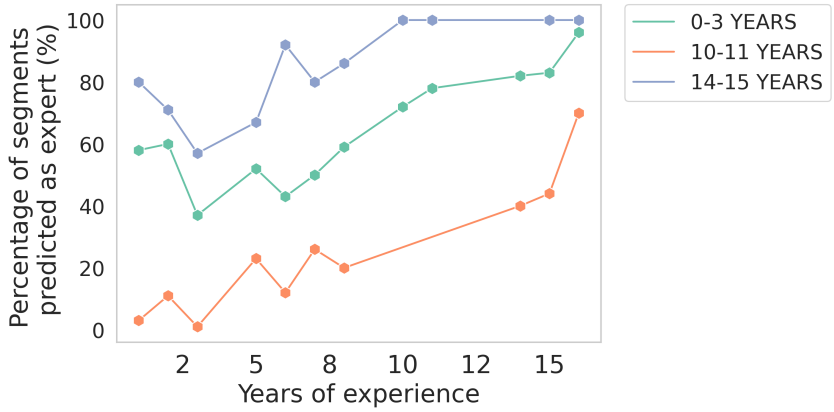


Figure 7.3: Lineplots of models: $EX_{0,3}$, $EX_{10,11}$ and $EX_{14,15}$ which demonstrates a positive monotonic relationship.

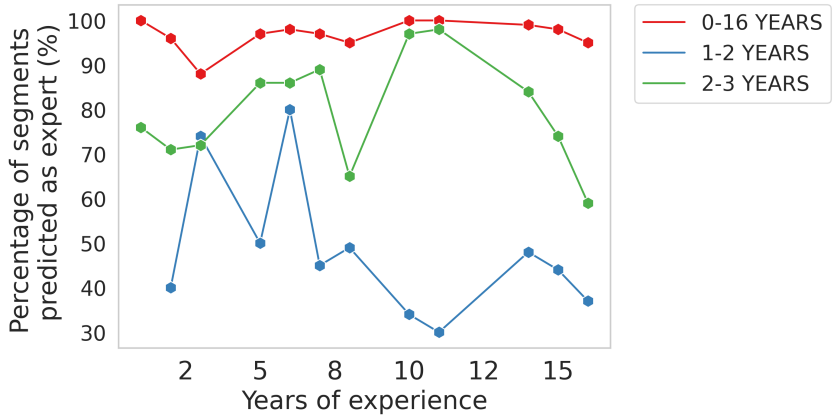


Figure 7.4: Lineplots of models: $EX_{0,16}$, $EX_{1,2}$ and $EX_{2,3}$ which did not show a strong positive or negative relationship between years of experience and expertise.

7.4 Discussion

Based on the results, the performance of a gaze skill classification model is dependent on the expert(s) which was used to the train the skill model. The model was used to predict whether an expert's years of experience demonstrates a positive monotonic relationship to an experts' level of expertise, where the level of expertise was defined as the percentage of segments predicted as expert. The Spearman's correlation coefficient test showed that when using $EX_{0,3}$, $EX_{10,11}$, and $EX_{14,15}$ as the expert to train the skill model, there is a positive relationship between the number of scanning years and the percentage of expert segments. In contrast, when using $EX_{0,16}$ and $EX_{2,3}$, there did not appear to be any positive or negative relationship,

with coefficients ranging from -0.09 to 0.13 (Table 7.9). Finally, when using $EX_{1,2}$ there was a negative relationship between expertise and years of experience, with a coefficient of -0.47 and -0.52 (Table 7.9).

These results suggest that, while style is a factor that cannot be disentangled easily from skill, without making any prior assumptions about the relationship between scanning years and expertise, there can be a positive relationship between the 2 variables – 3 out of the 6 models in Table 7.9 had a positive coefficient. However this cannot be generalised across the population of sonographers, and a deeper examination would be required to understand why the other models did not return a strongly positive monotonic relationship between expertise and experience – in particular, model $EX_{1,2}$ the relationship was strongly negative with coefficients of -0.47 and -0.52.

One interesting thing to note is that as as the individual expert's years of experience increases, the Spearman's coefficient also increases. From Table 7.9, models' $EX_{1,2}$, $EX_{2,3}$, $EX_{0,3}$, $EX_{10,11}$, and $EX_{14,15}$ returns coefficients of -0.47, -0.09, 0.78, 0.86 and 0.81 (and 0.85). This interesting result also warrants further examination – it suggests that a sonographer with less experience is less likely to be predicted as expert, when compared to someone with many more years of experience. Model $EX_{0,16}$ returned a coefficient of 0.13 and -0.04 which suggests neither a strong positive or negative relationship. I hypothesize that this could be because the different years of expertise was represented in the training set which created a generalisable model of skill that was less able to account for style.

The results also showed that even though the data was imbalanced, with the experts' as the majority class, in models where the experts' had similar years of experience, the classification results still showed a >10% difference in performance for the trainee class. This further strengthens the notion that an experts' style of scanning is difficult to entangle from their skill.

One limitation that should be considered is that the dataset did not include 'intermediate' years of expertise, where there is an experience gap of 6 years. Hence

it would be interesting to see if these methods would return the similar trends when applied on different eye-tracking medical datasets (for example dentistry).

7.5 Summary

A skill classification model was presented, where experts were defined as fully qualified sonographers independent of their years of scanning experience, and trainees were defined as sonographers learning how to scan. The best performing model considering a range of years of experience used a LightGBM and returned F1 scores of 98% and 70% for expert and trainee classes respectively. It was also shown that sonographer gaze behaviour is indicative of both skill and style, with performance differences of up to 16% depending on which experts' data was used to train the model. Finally, without making any prior assumptions of the relationship between years of experience as a direct measure of skill, generally there can be a positive relationship between years of scanning and expertise when considering task-agnostic gaze characteristics.

The final results presented demonstrate that skill and years of experience is positively correlated, but that this result is also dependent on whose data was used to define the expert population of sonographers. Consequently, sonographer scanning style is also a factor that is not easily disentangled from years of experience. Whether a sonographer is considered an expert is dependent both on how the expert population was represented, and also whether the sonographer's style of scanning closely mimicked that of the expert population. These findings are useful because they validate that years of experience can be a useful indicator of skill, but more importantly, that there are other factors such as personal scanning style which are not easily quantified that contribute to whether a sonographer is considered skillful or not.

8 Conclusion

Contents

8.1 Summary 124
8.1.1 Distinct Task-Specific Sonographer Gaze Patterns . . . 125
8.1.2 Human Skill and Style of Sonographer Scanning 126
8.2 Limitations 127
8.2.1 Hardware 127
8.2.2 Trimester Differences 130
8.2.3 Implementation 130
8.3 Future Avenues 131
8.4 Conclusion 132

8.1 Summary

I summarise the challenges that were discussed in the literature review in Chapter 2 and how the contributions in Chapters 4 to 7 have addressed these questions. I finish with some considerations that are unique to the PULSE study that could be explored in future work.

Briefly, the methods in Chapters 4 to 6 have shown us similarities and differences in *how* and *where* sonographers gaze at different areas-of-interests in abdomen, brain and heart planes. The skill classification method in Chapter 7 also showed that style and skill of a sonographer are not easily disentangled.

8.1.1 Distinct Task-Specific Sonographer Gaze Patterns

In Section 2.1.3, I mentioned that current eye-tracking visualisation methods designed for videos consider coloured images but in fetal ultrasound, the videos are recorded in grayscale. They also usually require pre-labelled areas-of-interest. In fetal ultrasound, there are many diagnostic planes which need to be assessed [127]. Studies which investigated non-parametric methods to reduce labelling efforts use high sampling frequency eye-trackers ($>$ 250Hz) which was not available in the PULSE study.

Given these considerations, Chapters 4 and 5 presented 2-dimensional visualisation methods for analysing fetal ultrasound videos which requires minimal manual labelling of eye-tracking data, and was able to capture different gaze characteristics of anatomical planes and capture specific areas-of-interests that the sonographer looks at during their scan. Specifically, I used unsupervised methods that allowed the analysis of eye-tracking data without the use of any manual labels of different types of eye movements. I used 2 types of visuals. The first were contour plots which visualised spatio-temporal gaze characteristics of sonographers when searching for abdomen, brain and heart planes. The second was an event-based visual of gaze scanning patterns that characterised skill based on which anatomical landmarks the sonographer had looked at during the scan. Finally, I built an affine transformer network which normalised gaze with respect to the anatomy circumference, providing context to the gaze data with respect to the anatomical circumference. The network reduced the burden of manually labelling bounding boxes (Section 3.4.2). The final results was able to show both global and local gaze characteristics of sonographers when searching for the abdomen, brain and heart plane.

In these chapters, the main contributions I have made are the following. First, I have shown that there are distinct gaze patterns when sonographers are scanning for the 3 anatomies (of which they spend the majority of their time on during the scan) abdomen, brain and heart. I have also shown that the manner of scanning differs between the brain (typically easier) and heart (typically harder due to its relative smaller size).

With the knowledge learned from Chapters 4 and 5, this naturally followed on with the question of whether sonographer eye-tracking data when searching for these planes were distinct enough to be classified into abdomen, brain and heart plane tasks. In Section 2.2.3, I mentioned that current task classification methods which used eye-tracking data require suitable eye movement classification algorithms to separate eye-tracking data into fixations and saccades. Prior work also considered static image tasks such as reading a page as opposed to watching a video. Finally, prior studies used simulated environments that did not account for external factors which can influence the participant's behaviour.

Given these considerations, Chapter 6 presented a method which classified 3 different anatomical planes (abdomen, brain and heart) using real-world sonographer scanpaths without requiring the separation of eye-tracking data into different eye movements. The final results showed that the success of the classification model was dependent on the location of the anatomical landmarks for each of the planes, where the normalised gaze position improved the results as the normalisation method gave meaningful information about where the landmark was located with respect to the boundaries of the anatomical plane.

In this chapter, the main contributions I have made are the following. First I have shown the importance of using gaze data in context of the AOI, in this chapter, the anatomy's circumference, to provide the classification model further information of where the sonographer is looking at. I have also shown that the gaze patterns are not just qualitatively distinct, but quantitatively via the classification model. Where the model misclassified gaze data, I have also shown that there are individual sonographers whose gaze patterns differ from the population of sonographers. This could have implications for future models which use gaze in other fetal ultrasound studies.

8.1.2 Human Skill and Style of Sonographer Scanning

Finally, I explored the widely used definition of sonographer skill in Chapter 7. In Section 2.3.4, I mentioned that current skill classification methods which used eye-

tracking data consider suitable eye movement classification algorithms to calculate quantitative metrics and differences between groups of experts. An example metric is the total number of fixations. They also usually use years of experience to separate groups of participants into experts and non experts. Other complexities such as differing anatomical presentations of mother and fetus, image quality and interpretation are not easily quantified using years of experience.

Given these considerations, Chapter 7 presented a method which trained a skill classification model using eye-tracking data of trainees and experts, where an expert was defined as a fully qualified sonographer independent of years of experience, and trainees were sonographers who were still learning to scan. The final results were important in validating the use of years of experience as an indicator of skill, but also showed that skill classification models need to consider sonographer style as a factor that affects the model's performance.

In this chapter, the main contributions I have made are the following. First I have shown that both pupillary and gaze data contain valuable information regarding a sonographer's skill of scanning, where the pupillary features contributes heavily towards the classification of skill, in line with prior work done by [136]. I have also shown that a sonographer's skill and style of scanning are not trivial to separate; specifically, the performance of the classification model depends on which sonographers are represented in the training data. Finally, I also showed that generally there can be a positive relationship between years of experience and human scanning expertise.

8.2 Limitations

Although the presented methods have revealed meaningful insights of sonographer gaze behaviour, there was some limitations which I now highlight.

8.2.1 Hardware

The PULSE data was collected at a single site (John Radcliffe Hospital, Oxford), using a single ultrasound machine and eye-tracker. Consequently, using a different

sampling frequency eye-tracker or an ultrasound machine which does produce as high quality images may not present the visualisations in the same way. In Chapter 5, events are defined using crops of the image centered around the recorded gaze point. In Chapter 5, the clusters may form differently if the eye-tracking sampling frequency is higher and is less susceptible to noise. Since the results were also inspected qualitatively, the overall takeaways of where and how the sonographer reads the ultrasound video would remain the same.

Eye-Tracking Accuracy and Precision Implications.

The main assumption of this thesis's work is that the eye-tracking data recorded is sufficiently accurate to what the sonographer looked at during the scan. In [110], their study showed that the precision of the eye-tracker was not affected between calibration and later use, so I focus on how the accuracy could affect the results.

The eye-tracking median accuracy and precision were reported in Section 1.5[1]. Eye-tracking accuracy and precision can affect the outcomes and conclusions of experiments, where errors can be propagated from its raw form and into analyses. I discuss those specific to my thesis, namely, the use of eye-tracking to:

- Determine AOIs (Chapter 4).
- Determine fixations (Chapter 5) using the I-VT algorithm.
- Determine image-based features (Chapters 4, 5).
- Differentiate between tasks (Chapter 6) and experience (Chapter 7).

Determine AOIs. In Chapter 4, unsupervised clustering is used to determine AOIs for 3 different tasks. Unsupervised clustering relies on spatial proximity of the gaze points. Taking the accuracy of the eye-tracker into account, it is unlikely that the conclusion of Chapter 4 would be different because the landmarks are sufficiently far apart for the abdomen and brain plane. For the heart plane, the conclusion in Chapter 4 also remains true because the sonographers' gaze did not change much over time, returning a closely packed cluster of gaze points. The implications of

[1]The median accuracy was 0.65 degrees, with a precision of 0.09 degrees (Section 1.5).

the precision and accuracy could affect planes where separation between important landmarks fall within the interval of 0.65 degrees ± 0.16 degrees[2].

Determine fixations and image-based features. In Chapter 5, I used the I-VT algorithm to first identify fixations and saccades, and then used fixations to generate meaningful events. This is followed by an image-based calculation, using the gaze point as the centroid. A bounding box of 1.5 degrees (in the x and y direction) around the gaze point is used as a snapshot. Similarly, in Chapter 5, a crop size of 20x20 pixels (of a resized 224x224 ultrasound frame) is used to capture the image of the event which occurred. For the I-VT calculation, the identified fixations and saccades could have been affected by the eye-tracker's characteristics, where the associated risk is missing short saccades. However fixations are used here. That, and the short time-series being considered (3-5 seconds) and prior work [81] which used I-VT in their analysis, it is likely that the overall conclusion of brain versus heart analysis would not differ. To obtain events and snapshots, the image provides (pixel-level) information that would also compensate for eye-tracking accuracy and precision errors. The snapshot and cropped images provide the general area of the sonographer's gaze. Rather than a 'point'-based analysis, the crop of the image centered around the gaze gives a suitable and wide enough margin that differences between tasks can still be calculated.

Differentiate tasks and experience. In Chapters 6, the eye-tracking data was used in its raw and feature engineered form. The eye-tracker accuracy is less likely to affect the conclusions because the entire length of time-series gaze data was being used to classify task rather than calculate specific eye characteristics. Any loss of accuracy during that short period of time (100-150 frames, or during the length of the scan of 30-40 minutes) is likely to be negligible. There might have been periods where the sonographer did not scan for the project, and the loss of accuracy could be applicable. In this case, the general gaze pattern and subsequent

[2]0.16 degrees is the recorded loss of accuracy over time.

observed landmarks remains largely unchanged, as seen in Chapters 4 and 5, where the gaze patterns were qualitatively distinct enough.

In Chapter 7, the skill of the sonographer was being classified using both gaze and pupil data. Given that the lighting conditions of the room remain reasonably constant, and the participants were adults (and not young children who might look away from the screen more often), the baseline pupil size is unlikely to change significantly. Since entire lengths of time-series were used for building the skill model, there were instances of eye-tracking data which were not able to be used due to tracking errors or missing data, or interpolation gaps which were too large to be considered viable. That coupled with the differences in available data for some years, whilst the conclusion of Chapter 7 is most likely to remain unchanged (i.e. style and skill are hard to disentangle), I discuss potential remedies in Section 8.3.

8.2.2 Trimester Differences

The work presented here also used only second trimester data. The purposes of the first and third trimester have different aims and different sized fetus (smaller in the first, larger in the third). For example, in the first trimester, the sonographer measures the nuchal translucency and crown rump length, usually with a side view of the fetus [151]. In the second trimester, the fetus anatomical planes are 2D cross-sections of the different anatomies (e.g. abdomen, brain and heart). Sonographer gaze behaviour in Chapters 4 to 7 would return different insights unique to the goals of the first and third trimester.

8.2.3 Implementation

Visualisation of data and how best to use these insights in clinical practice is difficult. This thesis has focused on using data analysis of sonographer gaze behaviour to reveal similarities and differences depending on the type of anatomical plane searched for and years of experience. Translating these results into clinical practice would require careful design of graphical user interfaces to ensure that the eye-tracker benefits the scanning experience.

8.3 Future Avenues

In this thesis, I have presented the analysis of spatio-temporal gaze characteristics of different tasks based on segments of the sonographer's searching behaviour, i.e., unfrozen frames. These segments are calculated from just before freezing (for example, 100 frames before freezing in Chapter 6) which is only possible having access to the full ultrasound video. There is still room to explore *real-time gaze behaviour* based on the identified gaze characteristics in Chapters 4 and 6. Real-time analysis of whether their gaze behaviour follows a certain pattern, or more likely that they have not, could provide sonographers-in-training real-time feedback so that they can adjust the probe accordingly to get the view that they are looking for. This would involve a multi-modal analysis, where the ultrasound image also has to be analysed using appropriate deep learning classification models to determine which landmarks are present on the screen.

There is also room to explore the concept of *skill assessment*, using the definitions described in Chapter 7, where an expert is compared to a trainee. The work presented in Chapter 7 provides a stepping stone to consider comparison of expert and non-expert sonographers, without using a time-based definition of skill. Gaze can also be combined with other sources of data, for example the quality of the image [150]. Then, the sonographer's skill would be a function of both gaze and image quality. Recording the probe's movement using IMU sensors proved to be more challenging because of the thickness of the cable (of the probe) where the sensor was mounted on. In this instances, optical flow of the ultrasound image [152] could be used instead as a proxy of the probe's movement.

A brief note on future eye-tracking experiments that can be carried out to mitigate some of the known hardware limitations of eye-trackers [143]. For anatomical planes where important landmarks fall within the error bounds of the accuracy and precision values, analysis of any gaze characteristics should also include an error bound. Where skill is being assessed, if possible, a comprehensive study of data which is not used (due to missingness) and how that can affect the results should be carried out (for example using an ablation study).

8.4 Conclusion

This thesis has presented several applications combining visualisation, supervised and unsupervised learning to sonographer eye-tracking data. I have presented 4 research contributions structured in 3 chapters, on the topics of data visualisation, task and skill classification. These methods have contributed to the goals of the PULSE project – a deeper understanding of how sonographers perform second trimester fetal ultrasound scans.

Appendix

Rules used to filter the read sonographer text via the OCR algorithm

Algorithm 2 Pseudocode showing several rule-based filters which were used to label the heart views after being processed by `pytesseract`.

```
label ←string picked up by OCR
number_of_labels ←number of unique strings picked up by OCR in a frozen segment
if SL in label then
    label ←Situs
end if
if (H in label) or (CH in label) then
    label ←4CH
end if
if (VT in label) then
    label ←3VT
end if
if (WT in label) or (LVOT in label) then
    label ←LVOT
end if
if WW in label and number_of_labels == 1 then
    label ←None                                  ▷ Erroneous 3VV detected.
end if
if (WW and SL in label) or (WW and SITUS in label) then
    label ←Situs
end if
```

Chapter 6 Task Classification Model Training Parameters

Hidden Markov Model

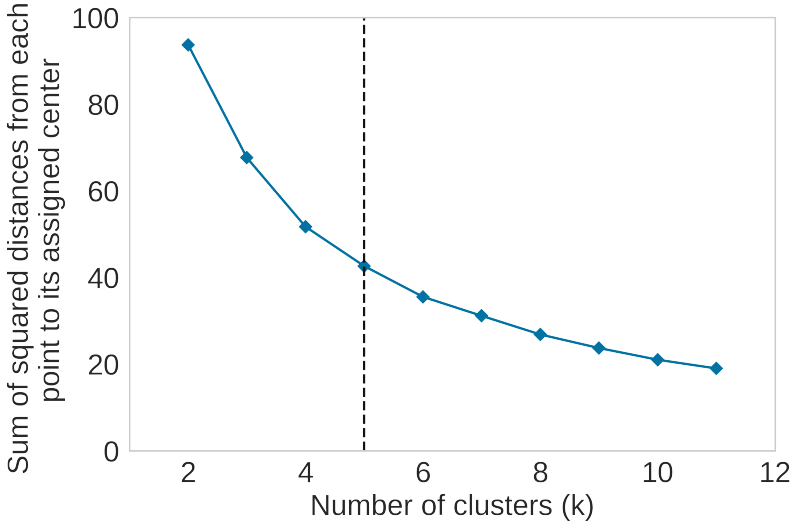


Figure 8.1: Elbow method used to determine optimal number of states to use for training the Hidden Markov Models. Here the figure shows that the optimal number of clusters is 5. The scoring metric used here was `distortion` in the `YellowBrick` package [80], which is the sum of the squared distances between the gaze point and its assigned cluster center.

Abdomen	0.22±0.07	0.10±0.02	0.32±0.16	0.19±0.11	0.17±0.07
Brain	0.34±0.09	0.19±0.04	0.28±0.12	0.19±0.07	0.0±0.0
Heart	0.22±0.17	0.32±0.03	0.12±0.08	0.22±0.12	0.12±0.09

Table 8.1: Emission probabilities of being in states 1-5 for the abdomen, brain and heart model. The mean probability and its standard deviation across 3 folds is reported in this table.

	State 1	State 2	State 3	State 4	State 5
State 1	0.93±0.0	0.02±0.0	0.03±0.02	0.02±0.01	0.01±0.01
State 2	0.03±0.0	0.92±0.01	0.01±0.0	0.02±0.02	0.02±0.01
State 3	0.02±0.01	0.00±0.0	0.93±0.01	0.03±0.01	0.02±0.01
State 4	0.02±0.01	0.01±0.01	0.04±0.01	0.91±0.01	0.02±0.0
State 5	0.02±0.02	0.02±0.0	0.04±0.01	0.02±0.0	0.89±0.02

Table 8.2: Transition probabilities between states 1-5 for the abdomen HMM model. The mean probability and its standard deviation across 3 folds is reported in this table.

	State 1	State 2	State 3	State 4	State 5
State 1	0.90±0.06	0.04±0.01	0.03±0.02	0.03±0.03	0.0±0.0
State 2	0.04±0.01	0.90±0.01	0.03±0.02	0.02±0.0	0.0±0.0
State 3	0.03±0.02	0.03±0.01	0.91±0.06	0.04±0.03	0.0±0.0
State 4	0.03±0.02	0.02±0.01	0.03±0.02	0.92±0.05	0.01±0.0
State 5	0.02±0.02	0.0±0.0	0.06±0.05	0.08±0.04	0.85±0.07

Table 8.3: Transition probabilities between states 1-5 for the brain HMM model. The mean probability and its standard deviation across 3 folds is reported in this table.

	State 1	State 2	State 3	State 4	State 5
State 1	0.94±0.01	0.02±0.01	0.01±0.0	0.03±0.02	0.01±0.01
State 2	0.02±0.01	0.94±0.01	0.01±0.01	0.01±0.01	0.01±0.0
State 3	0.04±0.04	0.04±0.02	0.86±0.09	0.02±0.01	0.05±0.0
State 4	0.03±0.03	0.02±0.03	0.01±0.01	0.94±0.03	0.01±0.0
State 5	0.03±0.01	0.04±0.02	0.0±0.0	0.02±0.04	0.91±0.03

Table 8.4: Transition probabilities between states 1-5 for the heart HMM model. The mean probability and its standard deviation across 3 folds is reported in this table.

k-Nearest Neighbors Time-Series Classifier

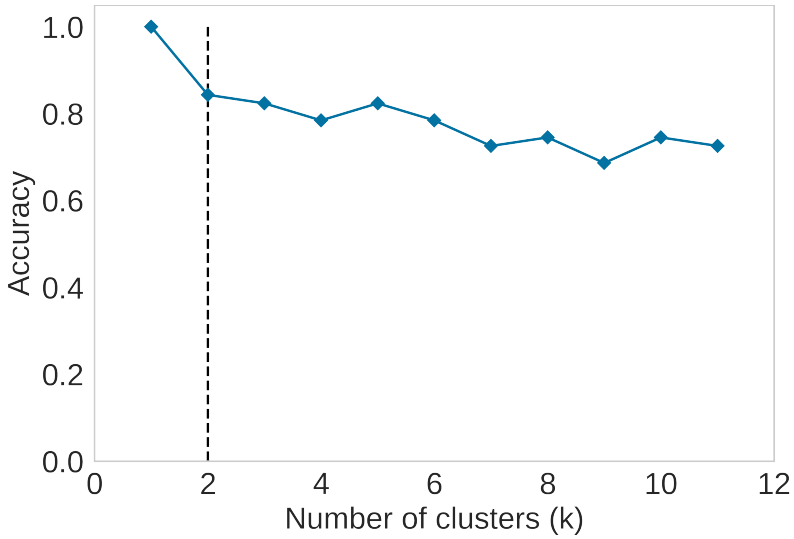


Figure 8.2: Elbow method used to determine optimal number of neighbours to train the k-nearest neighbour time-series classifier. Here the figure shows that the optimal number of clusters (k) is 2. The scoring metric used here was the accuracy of the tuned dataset using the `tslearn` package [122].

Chapter 7 Skill Classification

List of `tsfresh` features in `EfficientFeatures`

Feature name	Description of features	Settings
abs_energy	Returns the absolute energy of the time-series which is the sum over the squared values	
absolute_maximum	Calculates the highest absolute value of the time-series x.	
absolute_sum_of_changes	Returns the sum over the absolute value of consecutive changes in the series x	
agg_autocorrelation	Descriptive statistics on the autocorrelation of the time-series.	{f_agg: [“mean", “median", “var"], maxlag: 40}
agg_linear_trend	Calculates a linear least-squares regression for values of the time-series that were aggregated over chunks versus the sequence from 0 up to the number of chunks minus one.	{“attr": [“rvalue”, “intercept”, “slope”, “stderr”], chunk_len: [5, 10, 50], “f_agg": [“max”, “min”, “var”, “mean”] }
ar_coefficient	This feature calculator fits the unconditional maximum likelihood of an auto-regressive AR(k) process.	{“coeff”: (0, 10), k: 10}
augmented_dickey_fuller	Does the time-series have a unit root?	“attr”: [“teststat”, “pvalue”, “usedlag”]
autocorrelation	Calculates the autocorrelation of the specified lag	“lag”: (0, 9)
benford_correlation	Useful for anomaly detection applications	
binned_entropy	First bins the values of x into max_bins equidistant bins.	“max_bins”: 10
c_3	Uses c3 statistics to measure non linearity in the time-series	“lag”: (1, 3)

Continued on next page

Table 8.5 – continued from previous page

Feature name	Description of features	Settings
change_quantiles	First fixes a corridor given by the quantiles ql and qh of the distribution of x.	“f_agg”: [“mean”, “var”], “isabs”: [false, true], “qh”: [0.2, 0.4, 0.6, 0.8, 1.0], “ql”: [0, 0.2, 0.4, 0.6, 0.8, 1.0]
“cid_ce”	This function calculator is an estimate for a time-series complexity (A more complex time-series has more peaks, valleys etc.).	“normalise”: [true, false]
count_above	Returns the percentage of values in x that are higher than t.	t: 0
count_above_mean	Returns the number of values in x that are higher than the mean of x.	null
count_below	Returns the percentage of values in x that are lower than t.	t: 0
count_below_mean	Returns the number of values in x that are lower than the mean of x.	null
cwt_coefficients	Calculates a Continuous wavelet transform for the Ricker wavelet, also known as the “Mexican hat wavelet”.	“coeff”: (0, 14), “w”: [2, 5, 10, 20], “widths”: [2, 5, 10, 20]
energy_ratio_by_chunks	Calculates the sum of squares of chunk i out of N chunks expressed as a ratio with the sum of squares over the whole series.	num_segments: [10], segment_focus: (0, 9)
fft_aggregated	Returns the spectral centroid (mean), variance, skew, and kurtosis of the absolute fourier transform spectrum.	“aggtype”: [“centroid”, “variance”, “skew”, “kurtosis”]

Continued on next page

Table 8.5 – continued from previous page

Feature name	Description of features	Settings
fft_coefficient	Calculates the fourier coefficients of the one-dimensional discrete Fourier Transform for real input by fast.	“attr”: [“real”, “imag”, “abs”, “angle”], “coeff”: (0, 99)
first_location_of_maximum	Returns the first location of the maximum value of x.	null
first_location_of_minimum	Returns the first location of the minimum value of x.	null
fourier_entropy	Calculate the binned entropy of the power spectral density of the time-series (using the welch method).	“bins”: [2, 3, 5, 10, 100]
friedrich_coefficients	Coefficients of polynomial h(x), which has been fitted to.	“coeff”: (0, 3), “m”: 3, “r”: 30
has_duplicate	Checks if any value in x occurs more than once.	null
has_duplicate_max	Checks if the maximum value of x is observed more than once.	null
has_duplicate_min	Checks if the minimal value of x is observed more than once.	null
index_mass_quantile	Calculates the relative index i of time-series x where q% of the mass of x lies left of i.	“q”: (0.1, 0.9)
kurtosis	Returns the kurtosis of x (calculated with the adjusted Fisher-Pearson standardized moment coefficient G2).	null
large_standard_deviation	Does time-series have large standard deviation?	“r”: (0.05, 0.95)
last_location_of_maximum	Returns the relative last location of the maximum value of x.	

Continued on next page

Table 8.5 – continued from previous page

Feature name	Description of features	Settings
last_location_of_minimum	Returns the last location of the minimal value of x.	
lempel_ziv_complexity	Calculate a complexity estimate based on the Lempel-Ziv compression algorithm.	“bins”: [2, 3, 5, 10, 100]
length	Returns the length of x.	null
linear_trend		“attr”: [“pvalue”, “rvalue”, “intercept”, “slope”, “stderr”]
linear_trend_timewise	Calculate a linear least-squares regression for the values of the time-series versus the sequence from 0 to length of the time-series minus one.	“attr”: [“pvalue”, “rvalue”, “intercept”, “slope”, “stderr”]
longest_strike_above_mean	Returns the length of the longest consecutive subsequence in x that is bigger than the mean of x.	null
longest_strike_below_mean	Returns the length of the longest consecutive subsequence in x that is smaller than the mean of x.	null
matrix_profile	Calculates the 1-D Matrix Profile and returns Tukeys Five Number Set plus the mean of that Matrix Profile.	“feature”: [“mean”, “median”, “min”, “max”, 25, 75], “threshold”: 0.98
max_langevin_fixed_point	Largest fixed point of dynamics :math:argmax_x {h(x)=0} estimated from polynomial h(x).	“m”: 3, “r”: 30
maximum	Calculates the highest value of the time-series x.	null
mean	Returns the mean of x.	null
median	Returns the median of x.	
minimum	Returns the minimum of x.	
standard_deviation	Returns the standard deviation of x.	

Continued on next page

Table 8.5 – continued from previous page

Feature name	Description of features	Settings
mean_abs_change	Average over first differences.	
mean_change	Average over time-series differences.	
mean_n_absolute_max	Calculates the arithmetic mean of the n absolute maximum values of the time-series.	“number_of_maxima”: 7
mean_second_derivative_central	Returns the mean value of a central approximation of the second derivative.	null
number_crossing_m	Calculates the number of crossings of x on m.	“m”: [-1, 0, 1]
number_cwt_peaks	Number of different peaks in x.	“n”: [1, 5]
number_peaks	Calculates the number of peaks of at least support n in the time-series x.	“n”: [1, 3, 5, 10, 50]
partial_autocorrelation	Calculates the value of the partial autocorrelation function at the given lag.	“lag”: (0, 9)
percentage_of_reoccurring_datapoints_to_all_datapoints	Returns the percentage of non-unique data points.	null
percentage_of_reoccurring_values_to_all_values	Returns the percentage of values that are present in the time-series more than once.	null
permutation_entropy	Calculate the permutation entropy.	“dimension”: (3, 7), “tau”: 1
quantile	Calculates the q quantile of x.	“q”: (0.1, 0.9)

Continued on next page

Table 8.5 – continued from previous page

Feature name	Description of features	Settings
query_similarity_count	This feature calculator accepts an input query subsequence parameter, compares the query (under z-normalized Euclidean distance) to all subsequences within the time-series, and returns a count of the number of times the query was found in the time-series (within some predefined maximum distance threshold).	"query": null, "threshold": 0
range_count	Count observed values within the interval [min, max).	"max": 1, "min": -1
range_count	Count observed values within the interval [min, max).	"max": 0, "min": -1000000000000
range_count	Count observed values within the interval [min, max).	"max": 1000000000000, "min": 0
ratio_beyond_r_sigma	Ratio of values that are more than r * std(x) (so r times sigma) away from the mean of x.	"r": [0.5, 1, 1.5, 2, 2.5, 3, 5, 6, 7, 10]
ratio_value_number_to_time_series_length	Returns a factor which is 1 if all values in the time-series occur only once, and below one if this is not the case.	null
root_mean_square	Returns the root mean square (rms) of the time-series.	null
skewness	Returns the sample skewness of x (calculated with the adjusted Fisher-Pearson standardized moment coefficient G1).	null

Continued on next page

Table 8.5 – continued from previous page

Feature name	Description of features	Settings
spkt_welch_density	This feature calculator estimates the cross power spectral density of the time-series x at different frequencies.	"coeff": [2, 5, 8]
sum_of_reoccurring_data_points	Returns the sum of all data points, that are present in the time-series more than once.	null
sum_of_reoccurring_values	Returns the sum of all values, that are present in the time-series more than once.	null
sum_values		Calculates the sum over the time-series values.
symmetry_looking	Boolean variable denoting if the distribution of x looks symmetric.	"r": (0, 0.95, 0.05)
time_reversal_asymmetry_statistic	Returns the time reversal asymmetry statistic.	"lag": [1, 2, 3]
value_count	Count occurrences of value in time-series x.	"value": (-1, 1)
variance	Returns the variance of x.	null
variance_larger_than_standard_deviation	Is variance higher than the standard deviation?	null
variation_coefficient	Returns the variation coefficient (standard error / mean, give relative value of variation around mean) of x.	null

Table 8.5: List of features in `EfficientParameters` from the Python package `tsfresh`. Their feature name, description and settings are as presented [83].

LightGBM: Feature Importance

Name	Feature Importance
change_quantiles_f_agg_"mean"_isabs_False_qh_0.6_ql_0.4	15

permutation_entropy_dimension_4_tau_1	15
change_quantiles_f_agg_"mean"_is-abs_False_qh_0.4_ql_0.2	14
permutation_entropy_dimension_3_tau_1	13
fft_coefficient_attr_"real"_coeff_75	12
fft_coefficient_attr_"imag"_coeff_81	10
benford_correlation	10
agg_linear_trend_attr_"inter-cept"_chunk_len_50_f_agg_"max"	9
change_quantiles_f_agg_"mean"_is-abs_True_qh_0.6_ql_0.2	8
permutation_entropy_dimension_5_tau_1	8
change_quantiles_f_agg_"mean"_is-abs_True_qh_0.8_ql_0.4	8
change_quantiles_f_agg_"mean"_is-abs_True_qh_0.4_ql_0.2	8
mean	8
change_quantiles_f_agg_"mean"_is-abs_False_qh_0.6_ql_0.2	8
change_quantiles_f_agg_"mean"_is-abs_True_qh_0.6_ql_0.4	7
ar_coefficient_coeff_0_k_10	7
change_quantiles_f_agg_"var"_is-abs_True_qh_0.8_ql_0.6	7
minimum	6
change_quantiles_f_agg_"mean"_is-abs_False_qh_1.0_ql_0.6	6
cwt_coefficients_coeff_1_w_5_widths_(2, 5, 10, 20)	6
change_quantiles_f_agg_"var"_is-abs_False_qh_0.8_ql_0.6	6
max_langevin_fixed_point_m_3_r_30	5
cwt_coefficients_coeff_5_w_2_widths_(2, 5, 10, 20)	5
energy_ratio_by_chunks_num_segments_10_seg-ment_focus_8	5
change_quantiles_f_agg_"var"_is-abs_True_qh_0.6_ql_0.4	5
permutation_entropy_dimension_6_tau_1	5
change_quantiles_f_agg_"mean"_is-abs_True_qh_0.8_ql_0.6	5
quantile_q_0.1	5
cwt_coefficients_coeff_2_w_2_widths_(2, 5, 10, 20)	4
change_quantiles_f_agg_"var"_is-abs_False_qh_0.8_ql_0.4	4
change_quantiles_f_agg_"var"_is-abs_False_qh_0.6_ql_0.4	4

agg_linear_trend_attr_"intercept"_chunk_len_50_f_agg_"min"	4
cwt_coefficients_coeff_4_w_20_widths_(2, 5, 10, 20)	4
change_quantiles_f_agg_"var"_isabs_False_qh_0.4_ql_0.2	4
quantile_q_0.8	4
cwt_coefficients_coeff_2_w_10_widths_(2, 5, 10, 20)	3
longest_strike_below_mean	3
mean_n_absolute_max_number_of_maxima_7	3
change_quantiles_f_agg_"mean"_isabs_False_qh_0.8_ql_0.4	3
fft_coefficient_attr_"abs"_coeff_1	3
change_quantiles_f_agg_"mean"_isabs_False_qh_0.8_ql_0.2	3
agg_linear_trend_attr_"intercept"_chunk_len_50_f_agg_"mean"	3
cwt_coefficients_coeff_3_w_2_widths_(2, 5, 10, 20)	3
cwt_coefficients_coeff_1_w_2_widths_(2, 5, 10, 20)	3
cwt_coefficients_coeff_14_w_10_widths_(2, 5, 10, 20)	3
quantile_q_0.4	2
permutation_entropy_dimension_7_tau_1	2
cwt_coefficients_coeff_14_w_20_widths_(2, 5, 10, 20)	2
quantile_q_0.7	2
median	2
cwt_coefficients_coeff_12_w_10_widths_(2, 5, 10, 20)	2
quantile_q_0.2	2
autocorrelation_lag_3	2
cwt_coefficients_coeff_5_w_5_widths_(2, 5, 10, 20)	2
agg_linear_trend_attr_"intercept"_chunk_len_5_f_agg_"mean"	2
agg_autocorrelation_f_agg_"median"_maxlag_40	2
cwt_coefficients_coeff_10_w_20_widths_(2, 5, 10, 20)	2
agg_linear_trend_attr_"intercept"_chunk_len_10_f_agg_"mean"	2
quantile_q_0.9	2
number_peaks_n_10	2
agg_linear_trend_attr_"stderr"_chunk_len_5_f_agg_"var"	2
cwt_coefficients_coeff_9_w_20_widths_(2, 5, 10, 20)	1
quantile_q_0.6	1
absolute_maximum	1
change_quantiles_f_agg_"mean"_isabs_False_qh_0.8_ql_0.6	1
agg_linear_trend_attr_"intercept"_chunk_len_5_f_agg_"min"	1
number_peaks_n_3	1

agg_linear_trend_attr_"intercept"_chunk_len_5_f_agg_"max"	1
cwt_coefficients_coeff_13_w_5_widths_(2, 5, 10, 20)	1
number_cwt_peaks_n_1	1
number_peaks_n_5	1
cwt_coefficients_coeff_7_w_5_widths_(2, 5, 10, 20)	1
cwt_coefficients_coeff_5_w_20_widths_(2, 5, 10, 20)	1
cwt_coefficients_coeff_12_w_5_widths_(2, 5, 10, 20)	1
cwt_coefficients_coeff_8_w_20_widths_(2, 5, 10, 20)	1
cwt_coefficients_coeff_5_w_10_widths_(2, 5, 10, 20)	1
agg_linear_trend_attr_"intercept"_chunk_len_10_f_agg_"max"	1
cwt_coefficients_coeff_11_w_5_widths_(2, 5, 10, 20)	1
agg_autocorrelation_f_agg_"mean"_maxlag_40	1
cwt_coefficients_coeff_7_w_20_widths_(2, 5, 10, 20)	1
cwt_coefficients_coeff_8_w_10_widths_(2, 5, 10, 20)	1
cwt_coefficients_coeff_12_w_20_widths_(2, 5, 10, 20)	1
cwt_coefficients_coeff_11_w_10_widths_(2, 5, 10, 20)	1
agg_autocorrelation_f_agg_"var"_maxlag_40	1
cwt_coefficients_coeff_3_w_10_widths_(2, 5, 10, 20)	1
cwt_coefficients_coeff_6_w_20_widths_(2, 5, 10, 20)	1
autocorrelation_lag_9	1
autocorrelation_lag_7	1
cwt_coefficients_coeff_13_w_10_widths_(2, 5, 10, 20)	1
cwt_coefficients_coeff_4_w_10_widths_(2, 5, 10, 20)	1

Table 8.6: Gaze x co-ordinate non-zero features, where feature importance is calculated as numbers of times the feature is used in the model.

Name	**Feature Importance**
permutation_entropy_dimension_3_tau_1	25
friedrich_coefficients_coeff_1_m_3_r_30	13
kurtosis	12
spkt_welch_density_coeff_5	12
range_count_max_1_min_-1	11
large_standard_deviation_r_0.2	11
spkt_welch_density_coeff_8	11
energy_ratio_by_chunks_num_segments_10_segment_focus_0	10
fft_coefficient_attr_"angle"_coeff_33	10
change_quantiles_f_agg_"mean"_isabs_True_qh_0.4_ql_0.0	10
agg_autocorrelation_f_agg_"mean"_maxlag_40	10
permutation_entropy_dimension_4_tau_1	10
agg_linear_trend_attr_"intercept"_chunk_len_50_f_agg_"var"	9

large_standard_deviation_r_0.1	8
agg_autocorrelation_f_agg_"median"_maxlag_40	8
index_mass_quantile_q_0.1	7
permutation_entropy_dimension_7_tau_1	7
permutation_entropy_dimension_5_tau_1	6
index_mass_quantile_q_0.2	6
ratio_beyond_r_sigma_r_0.5	6
augmented_dickey_fuller_attr_"pvalue"_autolag_"AIC"	6
ratio_beyond_r_sigma_r_10	5
variation_coefficient	5
large_standard_deviation_r_0.15000000000000002	5
fourier_entropy_bins_100	5
ratio_beyond_r_sigma_r_6	5
ar_coefficient_coeff_0_k_10	4
cwt_coefficients_coeff_14_w_5_widths_(2, 5, 10, 20)	4
ratio_beyond_r_sigma_r_5	4
autocorrelation_lag_2	4
permutation_entropy_dimension_6_tau_1	4
quantile_q_0.1	4
partial_autocorrelation_lag_1	3
friedrich_coefficients_coeff_2_m_3_r_30	3
friedrich_coefficients_coeff_3_m_3_r_30	3
has_duplicate_max	3
change_quantiles_f_agg_"var"_isabs_True_qh_0.2_ql_0.0	3
autocorrelation_lag_9	3
max_langevin_fixed_point_m_3_r_30	3
fft_coefficient_attr_"abs"_coeff_16	3
cwt_coefficients_coeff_1_w_2_widths_(2, 5, 10, 20)	3
lempel_ziv_complexity_bins_100	2
cid_ce_normalize_True	2
change_quantiles_f_agg_"mean"_isabs_True_qh_0.2_ql_0.0	2
augmented_dickey_fuller_attr_"teststat"_autolag_"AIC"	2
cwt_coefficients_coeff_2_w_2_widths_(2, 5, 10, 20)	2
agg_linear_trend_attr_"intercept"_chunk_len_50_f_agg_"min"	1
quantile_q_0.2	1
lempel_ziv_complexity_bins_2	1
change_quantiles_f_agg_"var"_isabs_False_qh_0.2_ql_0.0	1
cwt_coefficients_coeff_14_w_10_widths_(2, 5, 10, 20)	1
maximum	1

mean_n_absolute_max_number_of_maxima_7	1
autocorrelation_lag_4	1
lempel_ziv_complexity_bins_5	1
autocorrelation_lag_5	1
lempel_ziv_complexity_bins_10	1
autocorrelation_lag_8	1
binned_entropy_max_bins_10	1
ratio_beyond_r_sigma_r_7	1
change_quantiles_f_agg_"var"_is-abs_False_qh_0.4_ql_0.0	1

Table 8.7: TEPR non-zero features, where feature importance is calculated as numbers of times the feature is used in the model.

Name	**Feature Importance**
change_quantiles_f_agg_"mean"_is-abs_False_qh_0.8_ql_0.4	18
permutation_entropy_dimension_3_tau_1	13
has_duplicate_max	13
agg_linear_trend_attr_"stderr"_chunk_len_50_f_agg_"var"	12
variation_coefficient	9
friedrich_coefficients_coeff_3_m_3_r_30	9
augmented_dickey_fuller_attr_"usedlag"_auto-lag_"AIC"	8
partial_autocorrelation_lag_4	7
friedrich_coefficients_coeff_0_m_3_r_30	7
permutation_entropy_dimension_4_tau_1	7
cwt_coefficients_coeff_14_w_20_widths_(2, 5, 10, 20)	6
ar_coefficient_coeff_4_k_10	6
abs_energy	5
change_quantiles_f_agg_"mean"_is-abs_False_qh_0.8_ql_0.6	5
linear_trend_attr_"pvalue"	5
cwt_coefficients_coeff_1_w_5_widths_(2, 5, 10, 20)	5
quantile_q_0.1	5
agg_linear_trend_attr_"stderr"_chunk_len_5_f_agg_"var"	5
agg_linear_trend_attr_"inter-cept"_chunk_len_50_f_agg_"mean"	3
agg_linear_trend_attr_"stderr"_chunk_len_10_f_agg_"var"	3
cwt_coefficients_coeff_8_w_5_widths_(2, 5, 10, 20)	3
agg_linear_trend_attr_"inter-cept"_chunk_len_10_f_agg_"min"	3
agg_linear_trend_attr_"inter-cept"_chunk_len_50_f_agg_"min"	3
cwt_coefficients_coeff_2_w_2_widths_(2, 5, 10, 20)	3
number_peaks_n_10	3

benford_correlation	2
cwt_coefficients_coeff_7_w_5_widths_(2, 5, 10, 20)	2
partial_autocorrelation_lag_3	2
cwt_coefficients_coeff_4_w_5_widths_(2, 5, 10, 20)	2
agg_linear_trend_attr_"stderr"_chunk_len_50_f_agg_"min"	2
friedrich_coefficients_coeff_2_m_3_r_30	2
agg_linear_trend_attr_"stderr"_chunk_len_50_f_agg_"mean"	2
agg_linear_trend_attr_"inter-cept"_chunk_len_5_f_agg_"min"	2
cwt_coefficients_coeff_3_w_5_widths_(2, 5, 10, 20)	2
cwt_coefficients_coeff_2_w_5_widths_(2, 5, 10, 20)	2
cwt_coefficients_coeff_4_w_2_widths_(2, 5, 10, 20)	2
friedrich_coefficients_coeff_1_m_3_r_30	1
quantile_q_0.2	1
cwt_coefficients_coeff_5_w_5_widths_(2, 5, 10, 20)	1
cwt_coefficients_coeff_1_w_2_widths_(2, 5, 10, 20)	1
agg_linear_trend_attr_"stderr"_chunk_len_10_f_agg_"max"	1
agg_linear_trend_attr_"inter-cept"_chunk_len_5_f_agg_"max"	1
linear_trend_attr_"intercept"	1
agg_linear_trend_attr_"inter-cept"_chunk_len_10_f_agg_"max"	1

Table 8.8: Gaze y co-ordinate non-zero features, where feature importance is calculated as numbers of times the feature is used in the model.

Bibliography

[1] Robert L Thorndike. "Who belongs in the family?" In: *Psychometrika* 18.4 (Dec. 1953), pp. 267–276. DOI: 10.1007/bf02289263.

[2] Thomas M. Cover and Peter E. Hart. "Nearest neighbor pattern classification". In: *IEEE Trans. Inf. Theory* 13 (1967), pp. 21–27.

[3] T. Caliński and J. Harabasz. "A Dendrite Method For Cluster Analysis". In: *Communications in Statistics* 3.1 (1974), pp. 1–27. ISSN: 00903272. DOI: 10.1080/03610927408827101.

[4] David L Davies and Donald W Bouldin. "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), pp. 224–227. DOI: 10.1109/TPAMI.1979.4766909.

[5] Jackson Beatty. "Task-evoked pupillary responses, processing load, and the structure of processing resources." In: *Psychological Bulletin* 91.2 (1982), pp. 276–292. ISSN: 1939-1455. DOI: 10.1037/0033-2909.91.2.276.

[6] M. Eizenman, R.C. Frecker, and P.E. Hallett. "Precise non-contacting measurement of eye movements using the corneal reflex". In: *Vision Research* 24.2 (Jan. 1984), pp. 167–174. ISSN: 00426989. DOI: 10.1016/0042-6989(84)90103-2.

[7] Peter J Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7.

[8] M. J. D. Powell. "A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation". In: *Advances in Optimization and Numerical Analysis*. Dordrecht: Springer Netherlands, 1994, pp. 51–67. DOI: 10.1007/978-94-015-8330-5_4.

[9] M. J. D. Powell. "A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation". In: *Advances in Optimization and Numerical Analysis*. Dordrecht: Springer Netherlands, 1994, pp. 51–67. DOI: 10.1007/978-94-015-8330-5_4.

[10] Calvin F Nodine et al. "Nature of expertise in searching mammograms for breast masses". In: *Academic Radiology* 3.12 (Dec. 1996), pp. 1000–1006. ISSN: 1076-6332. DOI: 10.1016/S1076-6332(96)80032-8.

[11] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9 (1997), pp. 1735–1780.

[12] Jeff A. Bilmes. "A gentle tutorial of the em algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models". In: *CTIT technical reports series* (1998).

[13] Dario D. Salvucci and Joseph H. Goldberg. "Identifying fixations and saccades in eye-tracking protocols". In: *Proceedings of the symposium on Eye tracking research & applications - ETRA '00*. New York, New York, USA: ACM Press, 2000, pp. 71–78. ISBN: 1581132808. DOI: 10.1145/355017.355028.

[14] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5 (Oct. 2001). ISSN: 0090-5364. DOI: 10.1214/aos/1013203451.

[15] N. V. Chawla et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (June 2002), pp. 321–357. ISSN: 1076-9757. DOI: 10.1613/jair.953.

[16] Benjamin Law et al. "Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment". In: *Proceedings of the Eye tracking research & applications symposium on Eye tracking research & applications - ETRA'2004*. New York, New York, USA: ACM Press, 2004, pp. 41–48. ISBN: 1581138253. DOI: 10.1145/968363.968370.

[17] Henry C. Lin et al. "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions". In: *Computer Aided Surgery* 11.5 (Jan. 2006), pp. 220–230. ISSN: 1092-9088. DOI: 10.3109/10929080600989189.

[18] David Manning et al. "How do radiologists do it? The influence of experience and training on searching for chest nodules". In: *Radiography* 12.2 (May 2006), pp. 134–142. ISSN: 1078-8174. DOI: 10.1016/j.radi.2005.02.003.

[19] G. Megali et al. "Modelling and Evaluation of Surgical Performance Using Hidden Markov Models". In: *IEEE Transactions on Biomedical Engineering* 53.10 (Oct. 2006), pp. 1911–1919. ISSN: 0018-9294. DOI: 10.1109/TBME.2006.881784.

[20] A. James et al. "Eye-Gaze Driven Surgical Workflow Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007*. Ed. by Nicholas Ayache, Sébastien Ourselin, and Anthony Maeder. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 110–117. ISBN: 978-3-540-75759-7.

[21] Damien Litchfield et al. "Learning from others: effects of viewing another person's eye movements while searching for chest nodules". In: *Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment*. Vol. 6917. International Society for Optics and Photonics, Mar. 2008, p. 691715. DOI: 10.1117/12.768812.

[22] Pieter Blignaut. "Fixation identification: The optimum threshold for a dispersion algorithm". In: 71.7 (2009), pp. 1439–1459. DOI: 10.3758/APP.

[23] Lindsey Cooper et al. "Radiology image perception and observer performance: how does expertise and clinical information alter interpretation? Stroke detection explored through eye-tracking". In: (Jan. 2009). URL: https://repository.lboro.ac.uk/articles/Radiology_image_perception_and_observer_performance_how_does_expertise_and_clinical_information_alter_interpretation_Stroke_detection_explored_through_eye-tracking/9404030.

[24] Toni Giorgino. "Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package". In: *Journal of Statistical Software* 31.7 (2009). ISSN: 1548-7660. DOI: 10.18637/jss.v031.i07.

[25] Fernando Vilariño et al. "Eye Tracking Search Patterns in Expert and Trainee Colonoscopists: A Novel Method of Assessing Endoscopic Competency?" In: *Gastrointestinal Endoscopy* 69.5 (Apr. 2009), AB370. ISSN: 0016-5107. DOI: 10.1016/j.gie.2009.03.1110.

[26] Narges Ahmidi et al. "Surgical Task and Skill Classification from Eye Tracking and Tool Motion in Minimally Invasive Surgery". In: *Lecture Notes in Computer Science*. Vol. 6363 LNCS. PART 3. 2010, pp. 295–302. DOI: 10.1007/978-3-642-15711-0_37.

[27] M. Dorr et al. "Variability of eye movements when viewing dynamic natural scenes". In: *Journal of Vision* 10.10 (Aug. 2010), pp. 28–28. ISSN: 1534-7362. DOI: 10.1167/10.10.28.

[28] Oleg V. Komogortsev et al. "Standardization of Automated Analyses of Oculomotor Fixation and Saccadic Behaviors". In: *IEEE Transactions on Biomedical Engineering* 57.11 (Nov. 2010), pp. 2635–2645. ISSN: 0018-9294. DOI: 10.1109/TBME.2010.2057429.

[29] Lee Richstone et al. "Eye Metrics as an Objective Assessment of Surgical Skill". In: *Annals of Surgery* 252.1 (2010). ISSN: 0003-4932. URL: https://journals.lww.com/annalsofsurgery/Fulltext/2010/07000/Eye_Metrics_as_an_Objective_Assessment_of_Surgical.28.aspx.

[30] Narges Ahmidi et al. "An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data". In: *International Forum of Allergy & Rhinology* 2.6 (Nov. 2012), pp. 507–515. ISSN: 2042-6984. DOI: 10.1002/alr.21053.

[31] Gennady Andrienko et al. "Visual analytics methodology for eye movement studies". In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2889–2898. ISSN: 10772626. DOI: 10.1109/TVCG.2012.276.

[32] Natalia Andrienko et al. "Visual analytics for understanding spatial situations from episodic movement data". In: *KI-Künstliche Intelligenz* 26.3 (2012), pp. 241–251.

[33] Philip J. Benson et al. "Simple Viewing Tests Can Detect Eye Movement Abnormalities That Distinguish Schizophrenia Cases from Controls with Exceptional Accuracy". In: *Biological Psychiatry* 72.9 (Nov. 2012), pp. 716–724. ISSN: 00063223. DOI: 10.1016/j.biopsych.2012.04.019.

[34] R. Borgo et al. "State of the art report on video-based graphics and video visualization". In: *Computer Graphics Forum* 31.8 (2012), pp. 2450–2477. ISSN: 14678659. DOI: 10.1111/j.1467-8659.2012.03158.x.

[35] Shahram Eivazi et al. "Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings". In: *Proceedings of the Symposium on Eye Tracking Research and Applications*. New York, NY, USA: ACM, Mar. 2012, pp. 377–380. ISBN: 9781450312219. DOI: 10.1145/2168556.2168641.

[36] Stefan Mathe and Cristian Sminchisescu. "Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition". In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 842–856. ISBN: 978-3-642-33709-3.

[37] Matthew S. Mould et al. "A simple nonparametric method for classifying eye fixations". In: *Vision Research* 57 (2012), pp. 18–25. ISSN: 00426989. DOI: 10.1016/j.visres.2011.12.006.

[38] Anneli Olsen. "The Tobii I-VT Fixation Filter: Algorithm description". In: *Tobii Technology* (2012), p. 21. URL: https://connect.tobii.com/s/article/What-is-Fixation-Filter?language=en_US.

[39] I. Sarris et al. "Intra- and interobserver variability in fetal ultrasound measurements". In: *Ultrasound in Obstetrics and Gynecology* 39.3 (2012), pp. 266–273. ISSN: 14690705. DOI: 10.1002/uog.10082.

[40] Rita Borgo et al. "Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications". In: *Eurographics 2013 - State of the Art Reports*. Ed. by M Sbert and L Szirmay-Kalos. The Eurographics Association, 2013. DOI: 10.2312/conf/EG2013/stars/039-063.

[41] Michael Burch, Andreas Kull, and Daniel Weiskopf. "AOI Rivers for Visualizing Dynamic Eye Gaze Frequencies". In: *Computer Graphics Forum* 32 (2013). DOI: 10.1111/cgf.12115.

[42] Ricardo J G B Campello, Davoud Moulavi, and Joerg Sander. "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Jian Pei et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–172. ISBN: 978-3-642-37456-2.

[43] Oleg V. Komogortsev and Alex Karpov. "Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades". In: *Behavior Research Methods* 45.1 (2013), pp. 203–215. ISSN: 1554351X. DOI: 10.3758/s13428-012-0234-9.

[44] Kuno Kurzhals and Daniel Weiskopf. "Space-Time Visual Analytics of Eye-Tracking Data for Dynamic Stimuli". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (Dec. 2013), pp. 2129–2138. ISSN: 1077-2626. DOI: 10.1109/TVCG.2013.194.

[45] Marcus Nyström, Ignace Hooge, and Kenneth Holmqvist. "Post-saccadic oscillations in eye movement data recorded with pupil-based eye trackers reflect motion of the pupil inside the iris". In: *Vision Research* 92 (Nov. 2013), pp. 59–66. ISSN: 00426989. DOI: 10.1016/j.visres.2013.09.009.

[46] Yu Akaishi et al. "Validity of direct ophthalmoscopy skill evaluation with ocular fundus examination simulators". In: *Canadian Journal of Ophthalmology* 49.4 (Aug. 2014), pp. 377–381. ISSN: 00084182. DOI: 10.1016/j.jcjo.2014.06.001.

[47] T Blascheck et al. "State-of-the-Art of Visualization for Eye Tracking Data". In: *Eurographics Conference on Visualization (EuroVis)* (2014), pp. 1–20. DOI: 10.2312/eurovisstar.20141173.

[48] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (2014), pp. 1724–1734. DOI: 10.3115/v1/d14-1179.

[49] Seth D. König and Elizabeth A. Buffalo. "A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds". In: *Journal of Neuroscience Methods* 227 (2014), pp. 121–131. ISSN: 1872678X. DOI: 10.1016/j.jneumeth.2014.01.032.

[50] Kuno Kurzhals, Florian Heimerl, and Daniel Weiskopf. "ISeeCube: Visual analysis of gaze data for video". In: *Eye Tracking Research and Applications Symposium (ETRA)* (2014), pp. 43–50. DOI: 10.1145/2578153.2578158.

[51] Kuno Kurzhals et al. "Benchmark data for evaluating visualization and analysis techniques for eye tracking for video stimuli". In: *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. New York, NY, USA: ACM, Nov. 2014, pp. 54–60. ISBN: 9781450332095. DOI: 10.1145/2669557.2669558.

[52] Delwyn Nicholls, Linda Sweet, and Jon Hyett. "Psychomotor skills in medical ultrasound imaging: An analysis of the core skill set". In: *Journal of Ultrasound in Medicine* 33.8 (2014), pp. 1349–1352. ISSN: 15509613. DOI: 10.7863/ultra.33.8.1349.

[53] Christian Szegedy et al. "Going Deeper with Convolutions". In: *CoRR* abs/1409.4 (2014). URL: http://arxiv.org/abs/1409.4842.

[54] Max Jaderberg et al. "Spatial Transformer Networks". In: *CoRR* abs/1506.0 (2015). URL: http://arxiv.org/abs/1506.02025.

[55] Linnéa Larsson et al. "Detection of fixations and smooth pursuit movements in high-speed eye-tracking data". In: *Biomedical Signal Processing and Control* 18 (Apr. 2015), pp. 145–152. ISSN: 17468094. DOI: 10.1016/j.bspc.2014.12.008.

[56] Christopher Olah. *Understanding LSTM Networks – colah's blog*. 2015. URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[57] M. Ahmed and J. A. Noble. "An eye-tracking inspired method for standardised plane extraction from fetal abdominal ultrasound volumes". In: *Proceedings - International Symposium on Biomedical Imaging* 2016-June (2016), pp. 1084–1087. ISSN: 19458452. DOI: 10.1109/ISBI.2016.7493454.

[58] M. Ahmed and J.A. Noble. "Fetal Ultrasound Image Classification Using a Bag-of-words Model Trained on Sonographers' Eye Movements". In: *Procedia Computer Science* 90 (2016), pp. 157–162. DOI: 10.1016/j.procs.2016.07.021.

[59] Benjamin Bach et al. "Time Curves: Folding Time to Visualize Patterns of Temporal Evolution in Data". In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (Jan. 2016), pp. 559–568. ISSN: 10772626. DOI: 10.1109/TVCG.2015.2467851.

[60] David P. Bahner et al. "Language of Transducer Manipulation: Codifying Terms for Effective Teaching". In: *Journal of Ultrasound in Medicine* 35.1 (2016), pp. 183–188. ISSN: 15509613. DOI: 10.7863/ultra.15.02036.

[61] Michael Burch et al. "Color bands: visualizing dynamic eye movement patterns". In: *2016 IEEE Second Workshop on Eye Tracking and Visualization (ETVIS)*. IEEE, Oct. 2016, pp. 40–44. ISBN: 978-1-5090-4731-4. DOI: 10.1109/ETVIS.2016.7851164.

[62] Tianqi Chen and Carlos Guestrin. "XGBoost". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, Aug. 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785.

[63] Kuno Kurzhals et al. "Fixation-image charts". In: *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. New York, NY, USA: ACM, Mar. 2016, pp. 11–18. ISBN: 9781450341257. DOI: 10.1145/2857491.2857507.

[64] Kuno Kurzhals et al. "Gaze Stripes: Image-Based Visualization of Eye Tracking Data". In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (Jan. 2016), pp. 1005–1014. ISSN: 1077-2626. DOI: 10.1109/TVCG.2015.2468091.

[65] Lisha Li et al. "Efficient Hyperparameter Optimization and Infinitely Many Armed Bandits". In: *CoRR* abs/1603.06560 (2016). arXiv: 1603.06560. URL: http://arxiv.org/abs/1603.06560.

[66] Prithiviraj K. Muthumanickam et al. "Supporting Exploration of Eye Tracking Data". In: *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization - BELIV '16*. New York, New York, USA: ACM Press, 2016, pp. 70–77. ISBN: 9781450348188. DOI: 10.1145/2993901.2993905.

[67] Thiago Santini et al. "Bayesian identification of fixations, saccades, and smooth pursuits". In: *Eye Tracking Research and Applications Symposium (ETRA)* 14 (2016), pp. 163–170. DOI: 10.1145/2857491.2857512.

[68] S. Swaroop Vedula et al. "Task-Level vs. Segment-Level Quantitative Metrics for Surgical Skill Assessment". In: *Journal of Surgical Education* 73.3 (2016), pp. 482–489. ISSN: 18787452. DOI: 10.1016/j.jsurg.2015.11.009.

[69] Gezheng Wen et al. "Computational assessment of visual search strategies in volumetric medical images". In: *Journal of Medical Imaging* 3.1 (Jan. 2016), p. 015501. DOI: 10.1117/1.jmi.3.1.015501.

[70] Richard Andersson et al. "One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms". In: *Behavior Research Methods* 49.2 (2017), pp. 616–637. ISSN: 15543528. DOI: 10.3758/s13428-016-0738-9.

[71] J. Timothy Balint, Dustin Arendt, and Leslie M. Blaha. "Storyline visualizations of eye tracking of movie viewing". In: *Proceedings of the 2nd Workshop on Eye Tracking and Visualization, ETVIS 2016* (2017), pp. 35–39. DOI: 10.1109/ETVIS.2016.7851163.

[72] T. Blascheck et al. "Visualization of Eye Tracking Data: A Taxonomy and Survey". In: *Computer Graphics Forum* 36.8 (Dec. 2017), pp. 260–284. ISSN: 01677055. DOI: 10.1111/cgf.13079.

[73] J. Hanley et al. "Visual Interpretation of Plain Radiographs in Orthopaedics Using Eye-Tracking Technology". English. In: *The Iowa orthopaedic journal* 37 (2017). Cited By :6, pp. 225–231. URL: www.scopus.com.

[74] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems*. Ed. by I Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[75] Kuno Kurzhals et al. “A Task-Based View on the Visual Analysis of Eye-Tracking Data”. In: *Eye Tracking and Visualization*. Ed. by Michael Burch et al. Cham: Springer International Publishing, 2017, pp. 3–22. ISBN: 978-3-319-47024-5.

[76] Tsung-Yi Lin et al. “Focal Loss for Dense Object Detection”. In: *CoRR* abs/1708.02002 (2017). arXiv: 1708.02002. URL: http://arxiv.org/abs/1708.02002.

[77] Ilya Loshchilov and Frank Hutter. “Fixing Weight Decay Regularization in Adam”. In: *CoRR* abs/1711.0 (2017). URL: http://arxiv.org/abs/1711.05101.

[78] Leland McInnes, John Healy, and Steve Astels. “hdbscan: Hierarchical density based clustering”. In: *The Journal of Open Source Software* 2.11 (2017), p. 205.

[79] Rudolf Netzel and Daniel Weiskopf. “Hilbert attention maps for visualizing spatiotemporal gaze data”. In: *Proceedings of the 2nd Workshop on Eye Tracking and Visualization, ETVIS 2016* (2017), pp. 21–25. DOI: 10.1109/ETVIS.2016.7851160.

[80] Benjamin Bengfort et al. *Yellowbrick*. Version 0.9.1. Nov. 14, 2018. DOI: 10.5281/zenodo.1206264.

[81] Yifan Cai et al. “Multi-task SonoEyeNet: Detection of Fetal Standardized Planes Assisted by Generated Sonographer Attention Maps”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by Alejandro F Frangi et al. Cham: Springer International Publishing, 2018, pp. 871–879. ISBN: 978-3-030-00928-1.

[82] Nora Castner et al. “Scanpath comparison in medical image reading skills of dental students”. In: *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. New York, NY, USA: ACM, June 2018, pp. 1–9. ISBN: 9781450357067. DOI: 10.1145/3204493.3204550.

[83] Maximilian Christ et al. “Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh A Python package)”. In: *Neurocomputing* 307 (Sept. 2018), pp. 72–77. ISSN: 09252312. DOI: 10.1016/j.neucom.2018.03.067.

[84] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. “CatBoost: gradient boosting with categorical features support”. In: (Oct. 2018). URL: http://arxiv.org/abs/1810.11363.

[85] Andrew T. Duchowski et al. “The Index of Pupillary Activity”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2018, pp. 1–13. ISBN: 9781450356206. DOI: 10.1145/3173574.3173856.

[86] S Erridge et al. “Comparison of gaze behaviour of trainee and experienced surgeons during laparoscopic gastric bypass”. In: *British Journal of Surgery* 105.3 (Feb. 2018), pp. 287–294. ISSN: 0007-1323. DOI: 10.1002/bjs.10672.

[87] Liam Li et al. “Massively Parallel Hyperparameter Tuning”. In: *CoRR* abs/1810.05934 (2018). arXiv: 1810.05934. URL: http://arxiv.org/abs/1810.05934.

[88] Richard Liaw et al. “Tune: A Research Platform for Distributed Model Selection and Training”. In: *CoRR* abs/1807.05118 (2018). arXiv: 1807.05118. URL: http://arxiv.org/abs/1807.05118.

[89] Vsevolod Peysakhovich and Christophe Hurter. "Scan path visualization and comparison using visual aggregation techniques". In: *Journal of Eye Movement Research* 10.5 (Jan. 2018). ISSN: 1995-8692. DOI: 10.16910/jemr.10.5.9.

[90] Public Health England (PHE). "NHS Fetal Anomaly Screening Programme Handbook". In: August (2018). URL: https://www.gov.uk/government/publications/fetal-anomaly-screening-programme-handbook/20-week-screening-scan.

[91] Damla Topalli and Nergiz Ercil Cagiltay. "Eye-Hand Coordination Patterns of Intermediate and Novice Surgeons in a Simulation-Based Endoscopic Surgery Training Environment". In: *Journal of Eye Movement Research* 11.6 (2018), pp. 1–14. ISSN: 19958692. DOI: 10.16910/JEMR.11.6.1.

[92] Chia-Kai Yang and Chat Wacharamanotham. "Alpscarf: Augmenting Scarf Plots for Exploring Temporal Gaze Patterns". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI EA 18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–6. ISBN: 9781450356213. DOI: 10.1145/3170427.3188490.

[93] Raimondas Zemblys et al. "Using machine learning to detect events in eye-tracking data". In: *Behavior Research Methods* 50.1 (Feb. 2018), pp. 160–181. ISSN: 1554-3528. DOI: 10.3758/s13428-017-0860-3.

[94] Mohammad Alsharid et al. "Captioning Ultrasound Images Automatically". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Cham: Springer International Publishing, 2019, pp. 338–346. ISBN: 978-3-030-32251-9.

[95] Valentin Bruder et al. "Space-time volume visualization of gaze and stimulus". In: *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. New York, NY, USA: ACM, June 2019, pp. 1–9. ISBN: 9781450367097. DOI: 10.1145/3314111.3319812.

[96] Romuald Carette et al. "Learning to Predict Autism Spectrum Disorder based on the Visual Patterns of Eye-tracking Scanpaths". In: *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications, 2019, pp. 103–112. ISBN: 978-989-758-353-7. DOI: 10.5220/0007402601030112.

[97] Richard Droste et al. "Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 11492 LNCS. Springer Verlag, 2019, pp. 592–604. ISBN: 9783030203504. DOI: 10.1007/978-3-030-20351-1_46.

[98] Eric Fichtel et al. "Eye tracking in surgical education: gaze-based dynamic area of interest can discriminate adverse events and expertise". In: *Surgical Endoscopy* 33.7 (2019), pp. 2249–2256. ISSN: 14322218. DOI: 10.1007/s00464-018-6513-5.

[99] Nishan Gunawardena et al. "Assessing surgeons skill level in laparoscopic cholecystectomy using eye metrics". In: *Eye Tracking Research and Applications Symposium (ETRA)* June (2019). DOI: 10.1145/3314111.3319832.

[100] Matthew C H Lee et al. "Image-and-Spatial Transformer Networks for Structure-guided Image Registration". In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 2019.

[101] Liam Li. *Massively parallel hyperparameter optimization*. Dec. 2019. URL: https://blog.ml.cmu.edu/2018/12/12/massively-parallel-hyperparameter-optimization/.

[102] Prithiviraj K. Muthumanickam et al. "Identification of temporally varying areas of interest in long-duration eye-tracking data sets". In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 87–97. ISSN: 19410506. DOI: 10.1109/TVCG.2018.2865042.

[103] Subhankar Roy et al. "Unsupervised Domain Adaptation Using Feature-Whitening and Consensus Loss". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2019-June. IEEE, June 2019, pp. 9463–9472. ISBN: 978-1-7281-3293-8. DOI: 10.1109/CVPR.2019.00970.

[104] H Sharma et al. "Spatio-Temporal Partitioning And Description Of Full-Length Routine Fetal Anomaly Ultrasound Scans". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019, pp. 987–990. DOI: 10.1109/ISBI.2019.8759149.

[105] Harshita Sharma and Alison J. Noble. *PULSE Annotation Methods Reference Document*. White Paper. 2019.

[106] Mikhail Startsev, Ioannis Agtzidis, and Michael Dorr. "1D CNN with BLSTM for automated classification of fixations, saccades, and smooth pursuits". In: *Behavior Research Methods* 51.2 (2019), pp. 556–572. ISSN: 15543528. DOI: 10.3758/s13428-018-1144-2.

[107] Peter T. Yamak, Li Yujian, and Pius K. Gadosey. "A comparison between ARIMA, LSTM, and GRU for time series forecasting". In: *ACM International Conference Proceeding Series* (2019), pp. 49–55. DOI: 10.1145/3377713.3377722.

[108] Robert G Alexander et al. "What do radiologists look for? Advances and limitations of perceptual learning in radiologic search". en. In: *J. Vis.* 20.10 (Oct. 2020), p. 17.

[109] Yifan Cai et al. "Spatio-temporal visual attention modelling of standard biometry plane-finding navigation". In: *Medical Image Analysis* 65 (2020). ISSN: 13618423. DOI: 10.1016/j.media.2020.101762.

[110] Pierre Chatelain et al. "Evaluation of Gaze Tracking Calibration for Longitudinal Biomedical Imaging Studies". In: *IEEE Transactions on Cybernetics* 50.1 (2020), pp. 153–163. DOI: 10.1109/TCYB.2018.2866274.

[111] Long-Fei Chen, Yuichi Nakamura, and Kazuaki Kondo. "User behavior analysis toward adaptive guidance for machine operation tasks: Analysis of behavior differences through skill-improving experiments". In: *Green, Pervasive, and Cloud Computing*. Cham: Springer International Publishing, 2020, pp. 288–302. ISBN: 9783030642426.

[112] R. Droste et al. "Discovering Salient Anatomical Landmarks by Predicting Human Gaze". In: *Proceedings - International Symposium on Biomedical Imaging* 2020-April (2020), pp. 1711–1714. ISSN: 19458452. DOI: 10.1109/ISBI45749.2020.9098505.

[113] Richard Droste et al. "Towards Capturing Sonographic Experience: Cognition-Inspired Ultrasound Video Saliency Prediction". In: *Medical Image Understanding and Analysis*. Ed. by Yalin Zheng, Bryan M. Williams, and Ke Chen. Cham: Springer International Publishing, 2020, pp. 174–186. ISBN: 978-3-030-39343-4.

[114] Andrew T. Duchowski et al. "The Low/High Index of Pupillary Activity". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2020, pp. 1–12. ISBN: 9781450367080. DOI: 10.1145/3313831.3376394.

[115] Gonca Gokce Menekse Dalveren and Nergiz Ercil Cagiltay. "Distinguishing Intermediate and Novice Surgeons by Eye Movements". In: *Frontiers in Psychology* 11.September (2020), pp. 1–10. ISSN: 16641078. DOI: 10.3389/fpsyg.2020.542752.

[116] Kyriaki Mengoudi et al. "Augmenting Dementia Cognitive Assessment With Instruction-Less Eye-Tracking Tests". In: *IEEE Journal of Biomedical and Health Informatics* 24.11 (Nov. 2020), pp. 3066–3075. ISSN: 2168-2194. DOI: 10.1109/JBHI.2020.3004686.

[117] Juan Francisco Ortega-Morán et al. "Using Eye Tracking to Analyze Surgeons Cognitive Workload During an Advanced Laparoscopic Procedure". In: 2020, pp. 3–12. DOI: 10.1007/978-3-030-31635-8_1.

[118] Youssef Othman et al. "Eye-To-Eye: Towards Visualizing Eye Gaze Data". In: *2020 24th International Conference Information Visualisation (IV)*. IEEE, Sept. 2020, pp. 729–733. ISBN: 978-1-7281-9134-8. DOI: 10.1109/IV51561.2020.00128.

[119] Arijit Patra and J. Alison Noble. "Incremental Learning of Fetal Heart Anatomies Using Interpretable Saliency Maps". In: 2020, pp. 129–141. DOI: 10.1007/978-3-030-39343-4_11.

[120] Subhankar Roy et al. "Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound". In: *IEEE Transactions on Medical Imaging* 39.8 (Aug. 2020), pp. 2676–2687. ISSN: 0278-0062. DOI: 10.1109/TMI.2020.2994459.

[121] Harshita Sharma et al. "Knowledge Representation and Learning of Operator Clinical Workflow from Full-length Routine Fetal Ultrasound Scan Videos". In: (2020). DOI: https://doi.org/10.1016/j.media.2021.101973.

[122] Romain Tavenard et al. "Tslearn, A Machine Learning Toolkit for Time Series Data". In: *Journal of Machine Learning Research* 21.118 (2020), pp. 1–6. URL: http://jmlr.org/papers/v21/20-091.html.

[123] Yipei Wang et al. "Differentiating Operator Skill During Routine Fetal Ultrasound Scanning Using Probe Motion Tracking". In: vol. 1. Springer International Publishing, 2020, pp. 180–188. ISBN: 9783030603342. DOI: 10.1007/978-3-030-60334-2_18.

[124] Chuhao Wu et al. "Eye-Tracking Metrics Predict Perceived Workload in Robotic Surgical Skills Training". In: *Human Factors* 62.8 (2020), pp. 1365–1386. DOI: 10.1177/0018720819874544.

[125] Mohammad Alsharid. "Generating textual captions for ultrasound visuals in an automated fashion". PhD thesis. Oxford University Research Archives, 2021.

[126] Richard Droste. "Advancing ultrasound image analysis by capturing operator gaze patterns". PhD thesis. Oxford University Research Archives, Jan. 2021.

[127] Lior Drukker et al. "Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video". In: *Scientific Reports* 11.1 (2021), p. 14109. ISSN: 2045-2322. DOI: 10.1038/s41598-021-92829-1.

[128] Benedikt Hosp et al. "Differentiating Surgeons Expertise solely by Eye Movement Features". In: *Companion Publication of the 2021 International Conference on Multimodal Interaction*. New York, NY, USA: ACM, Oct. 2021, pp. 371–375. ISBN: 9781450384711. DOI: 10.1145/3461615.3485437.

[129] Hyeju Jang et al. "Classification of Alzheimers disease leveraging multi-task machine learning analysis of speech and eye-movement data". en. In: *Frontiers in human neuroscience* 15 (2021), p. 716670. ISSN: 1662-5161. DOI: 10.3389/fnhum.2021.716670.

[130] Konstantinos-Filippos Kollias et al. "The contribution of machine learning and eye-tracking technology in autism spectrum disorder research: A systematic review". en. In: *Electronics* 10.23 (2021), p. 2982. ISSN: 2079-9292. DOI: 10.3390/electronics10232982.

[131] Maela Le Lous et al. "Impact of Physician Expertise on Probe Trajectory During Obstetric Ultrasound: A Quantitative Approach for Skill Assessment". In: *Simulation in Healthcare* 16.1 (2021). ISSN: 1559-2332. URL: https://journals.lww.com/simulationinhealthcare/Fulltext/2021/02000/Impact_of_Physician_Expertise_on_Probe_Trajectory.10.aspx.

[132] Lok Hin Lee, Yuan Gao, and J Alison Noble. "Principled Ultrasound Data Augmentation for Classification of Standard Planes". In: *Information Processing in Medical Imaging*. Ed. by Aasa Feragen et al. Cham: Springer International Publishing, 2021, pp. 729–741. ISBN: 978-3-030-78191-0.

[133] Arijit Patra et al. "Multimodal Continual Learning with Sonographer Eye-Tracking in Fetal Ultrasound". In: *Simplifying Medical Ultrasound*. Ed. by J. Alison Noble et al. Cham: Springer International Publishing, 2021, pp. 14–24. ISBN: 978-3-030-87583-1.

[134] Khaled Saab et al. "Observational Supervision for Medical Image Classification Using Gaze Data". In: 2021, pp. 603–614. DOI: 10.1007/978-3-030-87196-3_56.

[135] Harshita Sharma et al. "Knowledge Representation and Learning of Operator Clinical Workflow from Full-length Routine Fetal Ultrasound Scan Videos". In: *Medical Image Analysis* 69 (2021), p. 101973. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2021.101973.

[136] Harshita Sharma et al. "Machine learning-based analysis of operator pupillary response to assess cognitive workload in clinical ultrasound imaging". In: *Computers in Biology and Medicine* 135 (2021), p. 104589. ISSN: 18790534. DOI: 10.1016/j.compbiomed.2021.104589.

[137] Harshita Sharma et al. "Multi-Modal Learning from Video, Eye Tracking, and Pupillometry for Operator Skill Characterization in Clinical Fetal Ultrasound". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, Apr. 2021, pp. 1646–1649. ISBN: 978-1-6654-1246-9. DOI: 10.1109/ISBI48211.2021.9433863.

[138] Harshita Sharma et al. "Multi-modal learning from video, eye tracking, and pupillometry for operator skill characterization in clinical fetal ultrasound". In: *Proceedings - International Symposium on Biomedical Imaging* 2021-April (2021), pp. 1646–1649. ISSN: 19458452. DOI: 10.1109/ISBI48211.2021.9433863.

[139] Juraj Skunda et al. "Schizophrenia Detection Using Convolutional Neural Network". In: *2021 International Symposium ELMAR*. IEEE, Sept. 2021, pp. 151–154. ISBN: 978-1-6654-4437-8. DOI: 10.1109/ELMAR52657.2021.9550955.

[140] Clare Teng et al. "Towards Scale and Position Invariant Task Classification Using Normalised Visual Scanpaths in Clinical Fetal Ultrasound". In: *Simplifying Medical Ultrasound*. Ed. by J. Alison Noble et al. Cham: Springer International Publishing, 2021, pp. 129–138. ISBN: 978-3-030-87583-1.

[141] Nora Castner et al. "LSTMs can distinguish dental expert saccade behavior with high plaque-urracy". In: *2022 Symposium on Eye Tracking Research and Applications*. New York, NY, USA: ACM, June 2022, pp. 1–7. ISBN: 9781450392525. DOI: 10.1145/3517031.3529631.

[142] Zeyu Fu et al. *Anatomy-Aware Contrastive Representation Learning for Fetal Ultrasound*. 2022. DOI: 10.48550/ARXIV.2208.10642.

[143] Kenneth Holmqvist et al. *Eye tracking: empirical foundations for a minimal reporting guideline*. en. Apr. 2022. DOI: 10.3758/s13428-021-01762-8.

[144] Maksim Kholiavchenko et al. "Gaze-based attention to improve the classification of lung diseases". In: *Medical Imaging 2022: Image Processing*. Ed. by Ivana Igum and Olivier Colliot. SPIE, 2022.

[145] Maurice Koch, Daniel Weiskopf, and Kuno Kurzhals. "A Spiral into the Mind". In: *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 5.2 (May 2022), pp. 1–16. ISSN: 2577-6193. DOI: 10.1145/3530795.

[146] Xin Liu et al. "When medical trainees encountering a performance difficulty: evidence from pupillary responses". In: *BMC Medical Education* 22.1 (Dec. 2022), p. 191. ISSN: 1472-6920. DOI: 10.1186/s12909-022-03256-3.

[147] Lena Maier-Hein et al. "Surgical data science from concepts toward clinical translation". In: *Medical Image Analysis* 76 (Feb. 2022), p. 102306. ISSN: 13618415. DOI: 10.1016/j.media.2021.102306.

[148] Clare Teng et al. "Skill Characterisation of Sonographer Gaze Patterns during Second Trimester Clinical Fetal Ultrasounds using Time Curves". In: *2022 Symposium on Eye Tracking Research and Applications*. ETRA '22. Seattle, WA, USA: Association for Computing Machinery, 2022. ISBN: 9781450392525. DOI: 10.1145/3517031.3529637.

[149] Clare Teng et al. "Visualising Spatio-Temporal Gaze Characteristics for Exploratory Data Analysis in Clinical Fetal Ultrasound Scans". In: *2022 Symposium on Eye Tracking Research and Applications*. ETRA '22. Seattle, WA, USA: Association for Computing Machinery, 2022. ISBN: 9781450392525. DOI: 10.1145/3517031.3529635.

[150] Yipei Wang et al. "Task model-specific operator skill assessment in routine fetal ultrasound scanning". In: *International Journal of Computer Assisted Radiology and Surgery* 17.8 (Aug. 2022), pp. 1437–1444. ISSN: 1861-6429. DOI: 10.1007/s11548-022-02642-y.

[151] Robail Yasrab et al. "End-to-End First Trimester Fetal Ultrasound Video Automated CRL And NT Segmentation". In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. 2022, pp. 1–5. DOI: 10.1109/ISBI52829.2022.9761400.

[152] He Zhao et al. "Towards Unsupervised Ultrasound Video Clinical Quality Assessment with Multi-modality Data". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by Linwei Wang et al. Cham: Springer Nature Switzerland, 2022, pp. 228–237. ISBN: 978-3-031-16440-8.

[153] Clare Teng et al. "Skill, or Style? Classification of Fetal Sonography Eye-Tracking Data". In: *Proceedings of The 1st Gaze Meets ML workshop*. Ed. by Ismini Lourentzou et al. Vol. 210. Proceedings of Machine Learning Research. PMLR, Mar. 2023, pp. 184–198. URL: https://proceedings.mlr.press/v210/teng23a.html.

[154] Monica Franzese and Antonella Iuliano. *Hidden markov model*. Accessed in 2021. URL: https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/hidden-markov-model.

[155] Samuel Hoffstaetter. *Madmaze/pytesseract: A Python wrapper for google tesseract*. Accessed in 2022. URL: https://github.com/madmaze/pytesseract.

[156] Openvinotoolkit. *CVAT*. Accessed in 2021. URL: https://github.com/openvinotoolkit/cvat.

[157] PyTesseract. *Improving the quality of the output*. Accessed in 2022. URL: https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html.

[158] Python. *difflib Helpers for computing deltas*. Accessed in 2022. URL: https://docs.python.org/3/library/difflib.html.

[159] Python. *minimize(method=COBYLA)*. Accessed in 2022. URL: https://docs.scipy.org/doc/scipy/reference/optimize.minimize-cobyla.html.

[160] Phil Roth. *Demonstration of k-means assumptions*. Accessed in 2023. URL: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html#sphx-glr-auto-examples-cluster-plot-kmeans-assumptions-py.

[161] Tobii. *Types of Eye Movements*. Accessed in 2020. URL: https://connect.tobii.com/s/article/types-of-eye-movements.

[162] Stephen Tu. *Derivation of Baum-Welch algorithm for Hidden Markov models - GitHub Pages*. Accessed in 2022. URL: https://stephentu.github.io/writeups/hmm-baum-welch-derivation.pdf.