Computers in Biology and Medicine 135 (2021) 104589

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Machine learning-based analysis of operator pupillary response to assess cognitive workload in clinical ultrasound imaging

Harshita Sharma [a,*], Lior Drukker [b], Aris T. Papageorghiou [b], J. Alison Noble [a]

[a] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom
[b] Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, United Kingdom

ARTICLE INFO

Keywords:
Eye-tracking
Pupillometry
Cognitive workload
Ultrasound imaging
Time-series analysis
Sonography data science
Fetal ultrasound
Machine learning
Deep learning
Convolutional neural network
Multi-modal data

ABSTRACT

Introduction: Pupillometry, the measurement of eye pupil diameter, is a well-established and objective modality correlated with cognitive workload. In this paper, we analyse the pupillary response of ultrasound imaging operators to assess their cognitive workload, captured while they undertake routine fetal ultrasound examinations. Our experiments and analysis are performed on real-world datasets obtained using remote eye-tracking under natural clinical environmental conditions.

Methods: Our analysis pipeline involves careful temporal sequence (time-series) extraction by retrospectively matching the pupil diameter data with tasks captured in the corresponding ultrasound scan video in a multi-modal data acquisition setup. This is followed by the pupil diameter pre-processing and the calculation of pupillary response sequences. Exploratory statistical analysis of the operator pupillary responses and comparisons of the distributions between ultrasonographic tasks (fetal heart versus fetal brain) and operator expertise (newly-qualified versus experienced operators) are performed. Machine learning is explored to automatically classify the temporal sequences into the corresponding ultrasonographic tasks and operator experience using temporal, spectral, and time-frequency features with classical (shallow) models, and convolutional neural networks as deep learning models.

Results: Preliminary statistical analysis of the extracted pupillary response shows a significant variation for different ultrasonographic tasks and operator expertise, suggesting different extents of cognitive workload in each case, as measured by pupillometry. The best-performing machine learning models achieve receiver operating characteristic (ROC) area under curve (AUC) values of 0.98 and 0.80, for ultrasonographic task classification and operator experience classification, respectively.

Conclusion: We conclude that we can successfully assess cognitive workload from pupil diameter changes measured while ultrasound operators perform routine scans. The machine learning allows the discrimination of the undertaken ultrasonographic tasks and scanning expertise using the pupillary response sequences as an index of the operators' cognitive workload. A high cognitive workload can reduce operator efficiency and constrain their decision-making, hence, the ability to objectively assess cognitive workload is a first step towards understanding these effects on operator performance in biomedical applications such as medical imaging.

1. Introduction

Workload is a multi-dimensional and complex concept. It can be divided into physical load and mental load [1]. The physical component comprises the corporeal strain placed on the person during the task. The mental load involves the psychological effort to perform a specific task. In the literature, the terms cognitive workload (CW) and mental workload are synonymously used [2]. An increase in CW in medical practice can lead to undesirable consequences such as information overload, mental fatigue, decrease in situational awareness, and errors in decision-making. The starting point to better understand the implications of CW on the clinical workforce, is to develop methods to quantitatively measure and analyse CW in routine clinical practice.

A wide number of strategies have been used to measure CW. These can be broadly divided into three groups [2]. The first group includes *psychometric instruments*, with examples such as surveys, self-reports and questionnaires, which subjectively measure the perceived CW of a subject. The second group comprises of *quantitative performance*

* Corresponding author.
E-mail address: harshita.sharma@eng.ox.ac.uk (H. Sharma).

https://doi.org/10.1016/j.compbiomed.2021.104589
Received 26 March 2021; Received in revised form 12 June 2021; Accepted 15 June 2021
Available online 20 June 2021

measurements, including total duration and efficiency in completing a given task. The third group involves the use of *quantitative physiological measurements*, such as eye-tracking to measure pupil diameter changes, electrocardiogram (ECG) monitoring, and electroencephalogram (EEG) tracing. In this paper, we specifically emphasize on the subset of methods from the third group, which incorporate state-of-the-art eye--tracking technology to analyse and automatically predict CW of operators performing fetal ultrasound scans in a clinical setting.

Pupillometry, the study of eye pupil diameter changes, is a well-established modality that enables the measurement of CW [3]. Pupillary dilation is related to increase in CW *via* an increased sympathetic nervous system activity. Pupillometry has been suggested as an objective and robust method to measure CW, and otherwise unobservable insights into CW can be gleaned using pupillometric analysis [4].

Ultrasonography is one of the most widely used medical imaging technologies worldwide and the preferred choice for screening and monitoring fetal well-being due to its non-invasiveness, absence of ionising radiation, high accessibility, high reliability, and relatively low costs. In this paper, we present the first study investigating ultrasound operator pupillary response to assess their cognitive workload during routine diagnostic ultrasound imaging. We consider this problem in the context of second-trimester fetal ultrasound screening. In many countries, a second-trimester (gestational age of 18–22 weeks) ultrasound scan is offered to all pregnant women for a detailed assessment of the fetal anatomy and growth. For instance, in the United Kingdom (UK), the second-trimester scan guidelines are regulated by the National Health Service (NHS) under the Fetal Anomaly Screening Programme (FASP) [5]. During a full-length routine second-trimester ultrasound scanning session, an operator (a sonographer or fetal medicine doctor) views defined fetal anatomical structures called standard planes, including the head and brain, the heart, the abdomen, the limbs, the spine, and additional structures such as the hands and the feet, umbilical cord insertion, and maternal structures such as the uterine arteries. The second-trimester fetal ultrasound scan provides an interesting case study of cognitive workload assessment of ultrasound operators, whereby operator pupillary response is measured using non-contact eye-tracking technology, while they perform real-world scans.

1.1. Motivation

In routine obstetric ultrasound scanning, the second-trimester scan is considered to be one of the most mentally challenging scans because of the large number of mandatory standard planes, the prerequisite to comprehensively assess small structures (such as the heart which is typically 20 mm in diameter at the time of scan), and other associated challenges; specifically, ultrasound operators have to dynamically adjust the probe position in real-time to acquire the best diagnostic imaging plane, while overcoming artifacts induced by maternal habitus, fetal bone ossification, and fetal movement, thus requiring sharp hand-eye coordination skills [6]. Concurrently, the operator is required to instantaneously decide whether the imaging plane meets quality criteria and displays normal anatomical and functional features [7]. Analysis of operator clinical workflow of second-trimester ultrasound scans reveals that scanning is operator-dependent and difficult to perform, with arbitrary types, order, and time-distributions of the scanning tasks [8]. As these scans demand significant cognitive effort from the ultrasound operators, it is beneficial to quantitatively analyse operator CW to assess operators' mental overload and fatigue, which would help to avoid medical errors and improve overall patient outcomes. Moreover, assessment of CW is potentially useful to automatically characterise operator expertise and provide actionable feedback during their training, for example, on ultrasound simulators. We are motivated to explore pupillometry as a tool to assess operator CW due to its prevalence in analysing cognitive tasks in multiple other areas [3]. This paper contributes towards establishing a meaningful relationship between operators' pupillary changes and CW and developing objective computer-aided machine learning-based methods for task and skill characterization in clinical fetal ultrasound scanning.

Eye-tracking devices have advanced in recent years and provide adequate temporal resolution and precision to detect relatively small changes in the pupil diameter. Eye-tracking for continuous assessment of objective CW by measurement of pupil diameter changes is therefore possible [2]. Specifically, remote non-contact eye trackers have been increasingly used for pupillometric measurement of CW [9]. Such a capability is potentially well-suited for a clinical medical imaging setting, where remote eye trackers can be placed under the display screen facing the eyes of the operator (e.g., radiologists, pathologists, ultrasound operators including sonographers and sonologists) while they view medical images on a display monitor. In this way, operator pupillometric measurements can be made without disturbing or hindering the operator, leading to unobtrusive and continuous experiments for the analysis of CW. Thus, in routine settings, measuring CW by eye-tracking can be considered favourable compared to other categories of physiological measurements such as ECG and EEG, which are usually obtrusive and controlled. Other advantages of eye-tracking for measurement of CW include reliability (accuracy), objectivity, and continuous measurement throughout the operator activity. Hence, we are encouraged to investigate changes in pupil diameter of ultrasound imaging operators using remote eye-tracking technology.

Most literature on pupillometric measurements of CW consider strict and controlled conditions, with a very limited number of examples addressing real-world and uncontrolled scenarios, such as [10]. In contrast to the existing carefully controlled studies, our experiments have been performed using multi-modal data acquired from routine ultrasound scanning, which makes this study unique and challenging. There are two key challenges. Firstly, for real-world acquisition settings, the acquired data will contain noise, outliers, and artifacts compared to data obtained in controlled experiments. We have addressed this problem by using bespoke pupillometric data pre-processing (details in Section 5.1). Secondly, pupil diameter changes have been shown to correlate with environmental conditions, such as changes in ambient lighting, distance from the target, head positioning, and underlying eye disease, in addition to higher-level cognitive processes such as thinking or memorising [2]. In controlled pupillometric experiments, these are carefully monitored. However, in real-world experiments, such as ours, it is impossible to control these environmental factors, and this may lead to errors in observations [10]. To address this problem, we have made the following assumptions in our experiments. As a single routine ultrasound scan is usually completed in a single session of 20–40 min in the ultrasound scan room, the indoor lighting conditions are assumed to be constant during one scan. Furthermore, here the distance from the target is the distance between the operator and the display screen of the ultrasound machine (where the eye tracker is mounted); this is usually standard for a cart-based ultrasound machine and assumed constant during the scan. This distance is also recorded in the eye-tracking data (details in Section 3.2). For ultrasound operators, the head position is upright and remote eye-tracking automatically compensates for small head movements, hence, this factor has negligible effect. Lastly, all operators in our experiments have normal-to-corrected vision without any known eye condition or disease.

We explore two aspects of the ultrasound scanning process, namely, task load and operator experience, with reference to the cognitive effort exertion of ultrasound operators. Firstly, the characteristics of the undertaken ultrasonographic task can lead to varying pupillary responses depending on the perceived cognitive difficulty of the particular task. Hence, in theory, the operator pupil diameter changes can allow differentiation between easy and difficult scanning tasks. Another aspect is the scanning expertise depending on the operator's experience, which may lead to different extents of CW, hence, different pupillary responses. For instance, one may expect that an experienced operator will have a lower CW for a task when compared to a newly-qualified operator. Therefore, we can hypothesize that suitable modelling and learning of

the temporal pupil diameter changes may allow differentiation between scanning tasks and operator experience that is measured in years of scanning since qualification.

1.2. Contribution

We hypothesize that there are measurable variations in the operator's cognitive effort during acquisition of a routine second-trimester fetal ultrasound scan that can be measured *via* pupil diameter changes. Specifically, the study considers three questions:

1. Can pupil diameter changes, measured using eye-tracking during routine fetal ultrasound scans and acquired in sustained indoor scanning conditions, be used to assess operator cognitive workload?
2. Do pupil diameter changes of operators depend on the sonography task at hand and their scanning expertise, which, in turn determines their resulting cognitive workload?
3. Can temporal changes in pupil diameter be utilised in advanced machine learning-based models to automatically predict the sonography task and the scanning experience of the operators?

The specific contributions of the study are the following.

1. We propose a systematic multi-modal data acquisition, pre-processing, and analysis pipeline to estimate the pupil diameter changes of ultrasound operators *via* eye-tracking technology. The multi-modal data acquisition involves the simultaneous recording of scan video and eye-tracking datasets, followed by retrospective matching of pupil diameter data with events captured in the scan videos. We address the pertinent challenges of uncontrolled real-world datasets using bespoke pre-processing and pupillary response sequence calculation for our experiments.
2. We investigate operator pupil diameter change as an index of cognitive workload. We statistically study two aspects of ultrasound image acquisition that can lead to CW variations and the corresponding pupil diameter changes, namely, the ultrasonographic task and the operator experience. A quantitative measure of operator CW during scanning can be derived from the distribution of pupil diameter changes, which is found to vary more significantly among tasks with reference to the anatomical structures being scanned, and less significantly among experience groups with reference to operator expertise.
3. We develop machine learning algorithms to automatically infer the undertaken ultrasonographic task and operator experience given a measured temporal pupillary response. We comparatively evaluate the different inference models as well as the effect of windowing around event triggers.

1.3. Related work

Today, pupillometry, the study of eye pupil diameter, is a recognised modality that enables the measurement of cognitive controlled tasks [3], and several existing works suggest a meaningful relationship between pupil diameter changes and cognitive workload. We review the relevant literature in this section as the following.

It was early experimental research on pupillometry in the early 1960s that first suggested that pupillometry can be used to measure CW. Original work by Hess and Polt [11] found that mean changes in the pupil size of the eye observed during solving simple multiplication problems can be used as a direct measure of mental activity. Subsequent early studies showing that pupil diameter increase with higher cognitive demands include the pupillometric analysis for listening effort in a pitch-discrimination task [12], and pupil diameter response study during short-term memory tasks [13]. Recent studies in this direction include the assessment of pupil diameter changes in response to easy and difficult arithmetic problems using more advanced and portable eye-tracking technology [14], and the use of task-evoked pupillometry to differentiate the short-term memory capacity of children and adults [15]. A study that emphasizes real-world conditions in pupillometry experiments for four general visual search tasks (e.g., spot the difference between two images and find words hidden in a matrix) is described in Ref. [10], where the authors use task-evoked pupil dilation response to differentiate between the tasks and the users. They also use machine learning with statistical features and support vector machines (SVMs) to learn the pupillary response of event and non-event tasks. However, it was found that the binary classification performance metrics are not particularly higher than chance, owing to real-world experimental conditions and inter-observer variability. In contrast, the learning methods studied in this paper explore sophisticated hand-crafted features with classical machine learning, as well as deep learning, and as we will show, achieve reasonable classification performance metrics. A set of studies involve multi-modal fusion of physiological data such as signals (e.g., ECG, EEG) and eye-tracking information (e.g., pupil diameter response) to infer cognitive states [16], such as identifying human emotions [17]. A preliminary multi-modal learning method for operator skill characterization in clinical fetal ultrasound using spatial video, gaze, and pupillometric data is presented in Ref. [18].

In medical sciences and biomedical research, the use of advances in pupillometry has been relatively recent. A study using physician pupil diameter to distinguish between novices and experts, and perceived easy and difficult tasks, is described in Ref. [4], and within the specific domain of resuscitation medicine in Ref. [19]. A survey of mental workload assessment of clinicians using electronic health record systems, including eye-tracking based pupillometric experiments, is presented in Ref. [2]. For instance, pupil diameter changes are explored as one of the measures in usability assessment for electronic health record systems [20]. Another study [21], investigates gaze behaviour including pupillary response to differentiate between junior and expert surgeons in abdominal open surgery. Pupil response to changes in surgical difficulty for laparoscopic surgery is analysed in Ref. [22]. A study to assess mental operations by exploring differences between pupil dilation responses in alerting, orienting, and executive conflict monitoring tasks, is reported in Ref. [23].

The above studies confirm that pupil diameter changes are correlated with both the cognitive task undertaken, and the expertise of the user who performs the task. In context of routine ultrasound imaging, in a pilot study, we statistically analyse the global pupillary response as an index of CW that shows significant difference between easy and difficult task groups and operator experience [24]. In this paper, we leverage the ideas from pupillometry and biomedical data science to thoroughly analyse pupil diameter changes of ultrasound operators, in order to assess CW with respect to specific complex ultrasonographic tasks and operator experience, and to develop machine learning models that automatically infer the undertaken ultrasonographic task and the operator's expertise using only their pupillary response. To the best of our knowledge, there are no other existing pupillometric studies in routine ultrasound imaging.

2. Method overview

The overall method for analysing pupillary response of ultrasound imaging operators is presented in the block diagram in Fig. 1. Correspondingly, the paper is organised as follows. Section 3 describes the experimental setup with multi-modal data acquisition, and Section 4 explains clip extraction, annotation and resulting pupillometric datasets. Section 5 presents the pupillary data processing method, including data pre-processing and temporal sequence extraction. Section 6 reports the statistical distributions in the resulting datasets. Section 7 explains feature extraction, classical and deep machine learning models, observations, and windowing effects for ultrasonographic task classification, and Section 8 for operator experience classification. Section 9 discusses the study findings in detail, and Section 10 concludes the paper.

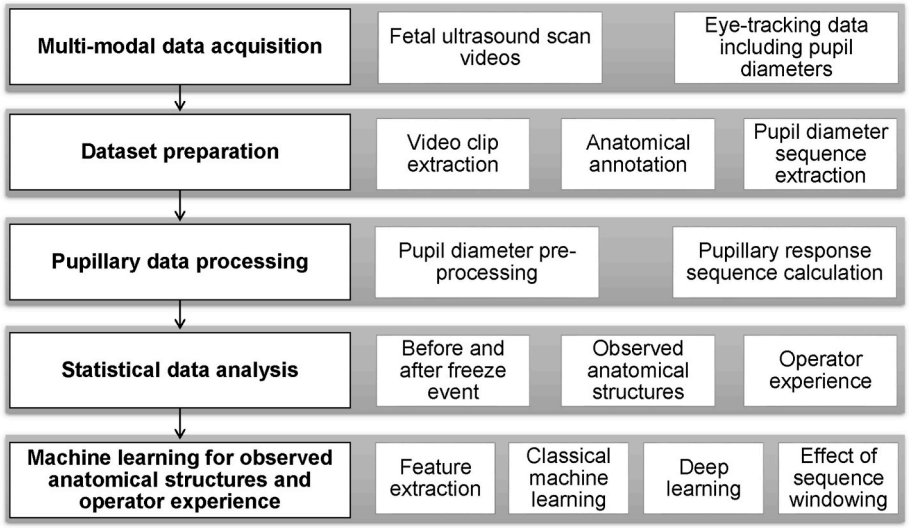

Fig. 1. Method overview.

3. Multi-modal data acquisition and experimental setup

The multi-modal data used in this paper was acquired as part of the Perception Ultrasound by Learning Sonographic Experience (PULSE) study.[1] This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051). For the purposes of this study, pregnant women with a singleton pregnancy attending pregnancy care at the Oxford University Hospitals NHS Foundation Trust, UK, were prospectively enrolled. Written informed consent was given by all participating pregnant women, as well as operators who participated in the study. Data were stored according to approved data governance rules.

3.1. Fetal ultrasound scan videos

Women who agreed were consented to have their full-length routine second-trimester ultrasound scan video recorded. All ultrasound scans included in this study were performed using a commercial Voluson E8 version BT18 (General Electric Healthcare, Zipf, Austria) ultrasound machine equipped with standard curvilinear (C2-9-D, C1-5-D), and 3D/4D (RAB6-D) probes. The LCD monitor has a resolution of 1920 × 1080 pixels and refreshes at a frequency of 60 Hz. The video signal was recorded from the scanner using lossless compression and sampled at the rate of 30 frames per second [25]. The average duration of a full-length second-trimester routine examination was 36.2 ± 11.6 min, with an average of 65,089 frames per scan video. Video data were stored as. mp4 video files, and each video file was converted to individual video frames stored as. png image files.

3.2. Eye-tracking data

Synchronised eye tracking was undertaken during the ultrasound scan acquisition using a remote eye-tracker (Tobii Eye-tracking Eye Tracker 4C, Danderyd, Sweden). The eye tracker was rigidly attached just below the display screen of the standard ultrasound system, with a magnetic mounting bracket as per the instruction of the product. The acquisition setup is shown in Fig. 2.

Pupil diameters of operators were measured by the eye tracker during scan acquisition. In the eye model of the eye tracker used, the pupil size is defined as the actual, internal physical size of the pupil [26]. The eye tracker outputs pupil size for each eye (in mm), together with each spatial gaze point (relative x and y coordinates) and 3D eye position of each eye at 90 Hz with corresponding timestamps, effectively recording three data points per video frame. The eye-tracking data were stored for later processing. For the scans studied in the datasets, the mean distance between operator pupil and display screen is 59.11 cm with a standard deviation of 5.46 cm.

Fig. 2. Acquisition setup for eye-tracking.

Operators did not have any visual or other signal to know that the eye-tracking device was functioning. Operators were free to adjust the height of the chair, and the inclination of the monitor. They operated the ultrasound probe in order to perform ultrasound examinations without being affected by the presence of an eye tracker, hence, real-world and clinically relevant eye-tracking data were recorded. The calibration of the eye tracker was previously studied in Ref. [25], where the eye tracker was calibrated for each operator following a 9-point calibration protocol.

4. Dataset preparation

4.1. Video clip extraction

The pupil diameter datasets were created corresponding to video clips of pre-defined scanning tasks extracted from full-length scan videos, as explained next. For video clip extraction, the scanning parameters were automatically extracted for each video frame in the full-length US scan video using optical character recognition [27] on the display screen. Under these, the 'Freeze' state for each frame was automatically detected and recorded as a technical annotation. A video clip was defined with respect to a freeze frame. In the scan, video is

[1] Project PULSE, funded by the European Research Council (grant ERC ADG-2015 694581) https://www.eng.ox.ac.uk/pulse.

typically frozen when a standard plane, according to the UK FASP protocol [5], is found. After freezing, the operator performs, for example, diagnostic inspections, biometric measurements, measurements using Doppler or Pulse Doppler and obtaining the optimal surface rendering. Hence, from the pupillometric point of view, freezing of the video can be detected as an 'event' or action when the operator CW is expected to change. For instance, before freezing, operators are often refining the view selection (fine-tuning), and after freezing, they are interpreting the content on the screen. Using the video clip definition and extracted 'Freeze' states, a full-length ultrasound scan video was automatically segmented in time to extract video clips. Specifically, 100 frames (3.3 s) were selected before the first detected frozen frame and a variable number after this frame, depending on the location of the last sequential frozen frame. A schematic of video clip extraction is shown in Fig. 3.

4.2. Anatomical annotation

A semi-automatic annotation method [8] was applied to obtain non-overlapping labelled video clips from full-length ultrasound scan videos depicting individual scanning tasks based on the viewed anatomy. Here, the extracted video clips were either visually inspected and manually annotated, or automatically annotated using a machine learning-based annotation model. In the latter case, fixed-length manually labelled clips were used for training deep spatio-temporal neural networks, and annotations of unlabelled clips were inferred using the trained networks. For manual annotation, a total of twenty-three labels were identified by a clinical expert. It was found that the most commonly occurring ultrasonographic tasks were 'Heart' and 'Brain', indicating that operators spent most time on one of these two tasks. These two tasks also showed a high accuracy in automatic annotation algorithm (>80%) [8,28]. These two tasks were selected for pupillometry analysis. The automatically labelled clips of these tasks were further manually inspected to filter out misclassified clips. An example video clip for each of these ultrasonographic tasks is depicted in Fig. 4.

4.3. Datasets for pupillometric experiments

For our experiments, we select a large subset of the PULSE dataset containing 380 routine full-length manually and automatically labelled second-trimester fetal ultrasound scans from 380 different women. These scans were undertaken by a total of 12 operators. The operators were separated into two groups based on their experience in sonography, namely newly-qualified (NQ) operators and experienced (XP) operators. The NQ group consists of three operators with less than or equal to two years of scanning experience (O_1, O_2 and O_3), and XP group has nine operators with more than two years of scanning experience (O_4–O_{12}). The total number of scans performed by the NQ group were 231, and 149 scans were performed by the XP group. The distribution of the ultrasound scans among the different operators is given in Table 1.

Using the extracted and labelled video clips from full-length scans, we extract the corresponding pupil diameter sequences of the two ultrasonographic tasks, namely, Brain and Heart. Each pupil diameter sequence consists of pupil diameters for the left and the right eye, as recorded by the eye tracker. A total of 769 Brain sequences (505 for NQ operators and 264 for XP operators) were extracted with a mean duration (standard deviation) of 31.5 (24.4) seconds. A total of 1694 Heart sequences (1062 for NQ operators and 632 for XP operators) were extracted with a mean duration (standard deviation) of 26.1 (33.8) seconds. For the operator experience groups, a total of 1567 sequences were extracted for NQ operators with a mean duration (standard deviation) of 19.4 (18.7) seconds, and 896 sequences for XP operators with a mean duration (standard deviation) of 24.3 (28.3) seconds. The sequence durations show a high variability, as in a second-trimester fetal ultrasound scan, the type, duration and order of the ultrasonographic tasks are arbitrary, and depend on several factors such as complexity of the task, fetal position, and operator expertise [8]. We have explained the difference in sequence durations in detail in Section 6. The distributions of the Brain and Heart sequences among the different operators (and corresponding experience groups) are shown in Table 1.

5. Pupillary data processing

We outline the pupillary data processing methods in this section. Firstly, the recorded raw pupil diameter data is processed using bespoke pre-processing. Then, a task-evoked pupillary response is calculated from the processed pupil diameter.

5.1. Pupil diameter pre-processing

Raw eye-tracking data collected directly from the eye tracker in uncontrolled experiments may contain noise or artifact samples, and discontinuities due to acquisition conditions (e.g., direction of gaze may not always be towards the screen when the operator interacts with the subject leading to missing data, brightness of the room may vary during the scan). These factors can lead to unwanted changes in pupillary response in addition to those associated with CW, thus, adversely affecting pupil size analysis. Therefore, it is necessary to perform pre-processing of the pupil diameter data to increase the reliability of CW inference from the corresponding pupil diameter variations.

To pre-process the pupil diameter data, we follow the guidelines prescribed by Ref. [29] for biomedical applications. This involves three steps, namely, removing noise and outliers, interpolating missing data and smoothing the raw signal. The sequential steps are briefly described as follows. Firstly, invalid samples (noise and outliers) are identified as samples outside the median absolute deviation of pupil dilation speeds, edge artifacts, trend-line deviation outliers, temporally isolated samples, and samples that are simply outside of a predefined feasible range, typically 1.5 mm–9 mm. This step identifies and rejects outliers due to system errors, blinks and look-away moments. After obtaining valid raw signal samples, gaps between these samples are interpolated by upsampling at a high sampling rate (1000 Hz). The resulting signal is smoothed using a zero-phase low-pass filter, with a cutoff frequency of 4 Hz. All the parameters in the pre-processing step are kept identical to those recommended in Ref. [29].

Examples of pupil diameter sequences before and after the pre-processing stage are shown in Fig. 5 for the Brain task and Fig. 6 for the Heart task. Observe that noisy samples are removed, gaps in data are filled, and the signal is smoothed. Moreover, in the processed Brain sequence, the dilation after freeze frame may be indicative of increase in

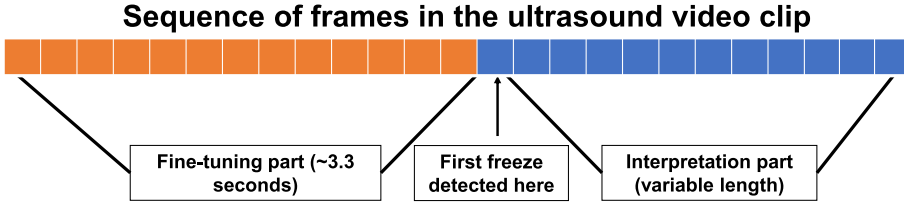

Fig. 3. Schematic of video clip extraction.

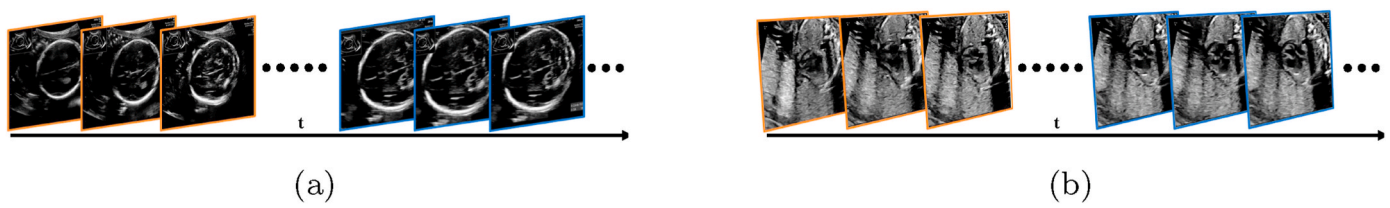

Fig. 4. Example video clips for the two ultrasonographic tasks (a) Brain (b) Heart. Orange outline represents non-frozen frames and blue outline represents frozen frames.

Table 1
Distribution of scans and pupil diameter sequences with operators and their scanning experience.

Experience Group	Newly-qualified (NQ)			Experienced (XP)								Total
Operator (Years)	O_1(0)	O_2(1)	O_3(2)	O_4(3)	O_5(5)	O_6(6)	O_7(7)	O_8(8)	O_9(10)	O_{10}(10)	O_{11}(14)	O_{12}(15)	
Ultrasound Scans	122	88	21	6	4	16	1	3	2	28	83	6	380
Brain Sequences	229	232	44	9	10	39	1	7	3	51	132	12	769
Heart Sequences	665	245	152	24	27	60	6	32	11	113	321	38	1694

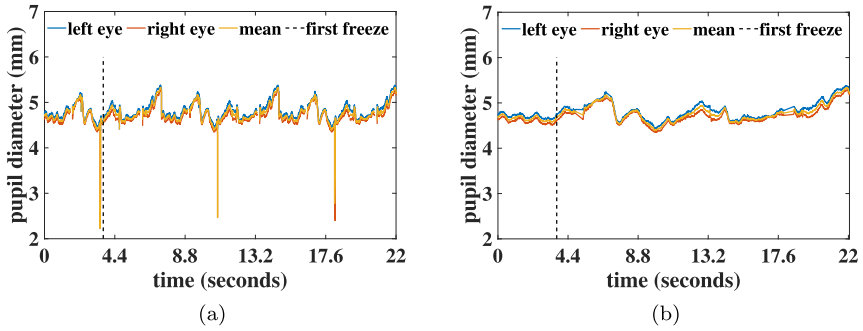

Fig. 5. Example pupil diameter Brain sequence (a) before and (b) after pre-processing.

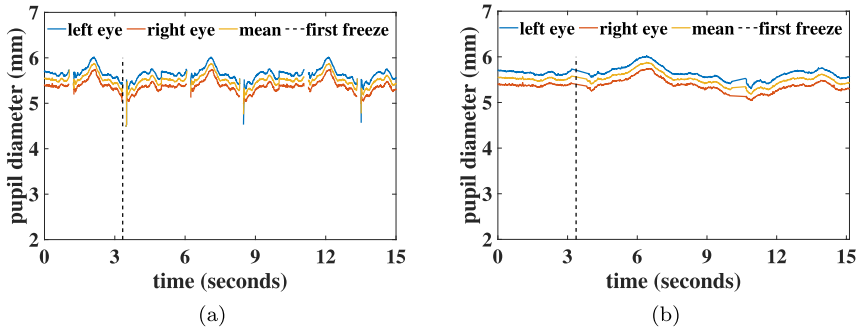

Fig. 6. Example pupil diameter Heart sequence (a) before and (b) after pre-processing.

operator CW just after freezing the screen to perform biometric measurements; in contrast, constriction after first freeze in the processed Heart sequence represents decrease in workload when a desirable plane is found (as a sign of relief or when the operator views the beating heart), but then the workload increases with time when different heart views need to be inspected and the probe needs to be carefully manoeuvred again.

5.2. Pupillary response sequence calculation

Absolute pupil diameters have a heterogeneous numerical range and exhibit inter-observer variability depending on a subject's biological traits. Hence, absolute pupil diameter measurements need to be normalised to make them biologically invariant, such that measurements derived from multiple subjects can be combined.

Task-Evoked Pupillary Response (TEPR) is defined as the stimulus-induced increase in pupil diameter relative to a pre-stimulus baseline period [3,9]. We calculate the TEPR from the processed pupil diameters to make the pupillary response invariant to the biological characteristics of individual users. However, since the data is acquired 'in the wild' during real-world ultrasound scans, thus, there is no defined pre-stimulus baseline task. Selecting the beginning of the routine scan as the pre-stimulus baseline is not a good choice because operators may deeply concentrate on determining the fetal position in this time. Therefore, this parameter is replaced by a *rest pupil diameter* that is computed as the minimum pupil diameter of the operator while performing a scan, obtained from the processed pupil diameters for a given scan. The rest pupil diameter is analogous to a pre-stimulus baseline, as

we assume that an increase in the CW would lead to pupil dilation, which represents an increase in the pupil diameter from rest. Hence, the TEPR Δd for a task-evoked pupil diameter d_{te}, given rest pupil diameter d_r is calculated by Equation (1).

$$\Delta d(\%) = \frac{d_{te} - d_r}{d_r} \times 100\% \quad (1)$$

The task-evoked pupillary response is illustrated in Fig. 7. A TEPR sequence is computed from each corresponding mean pupil diameter sequence. Temporal changes in operator pupillary response are represented by the TEPR sequences.

6. Statistical data analysis

We performed an exploratory analysis on the raw pupil diameter measurements for the ultrasound scan datasets. For this, we plotted the histogram and box-whisker plot of the raw pupil diameter data. We found that the mean, median, and quantiles of the distribution are 4.61 mm, 4.49 mm, 4.01 mm (0.25 quantile), and 5.20 mm (0.75 quantile). From the histogram in Fig. 8(a), we observed that the distribution has a small positive (right) skew compared to a normal distribution, and the skewness was computed as 0.31. The depicted normal distribution has the parameters as mean 4.61 mm and standard deviation 0.82 mm. In Fig. 8(b), the box-whisker plot also shows a small difference between the median and the mean (green dot); notches were used to represent the confidence interval of the median, but these are very small and not noticeable in the figure.

After establishing the distribution characteristics, we compare the statistical distributions for the pupil diameter responses recorded before and after the freeze event, for observed anatomical structures based on the ultrasonographic task at hand, and operator experience groups. We present box-whisker plots for each group in the comparison, and report Δ as the difference in the mean TEPR percentage of the two compared distributions $meanTEPR_{group1} - meanTEPR_{group2}$. To test the statistical significance of the difference between distributions of the two compared groups, we perform a two-sample Kolmogorov-Smirnov test [30], a non-parametric test that returns a decision for the null hypothesis that the data in the groups are from the same continuous distribution, and the alternative hypothesis is that the data in groups are from different continuous distributions. The result of the test is 1 if the test rejects the null hypothesis at the 5% significance level ($p < 0.05$), and 0 otherwise.

6.1. Before and after freeze event

Pupillary responses are compared before and after the first freeze event in the TEPR sequences. Fig. 9(a) shows the box-whisker plot of the pupillary response distributions. This shows a significantly higher overall TEPR values after the freeze event compared to before the freeze event in the sequences ($\Delta = 1.99\%, p < 0.05$). This result is indicative of higher overall cognitive load of operators after they freeze the scan video on the ultrasound machine including interpretation of the frozen frames, such as diagnostic inspection and biometric measurements, in comparison to the fine-tuning stage before freezing the live scan video.

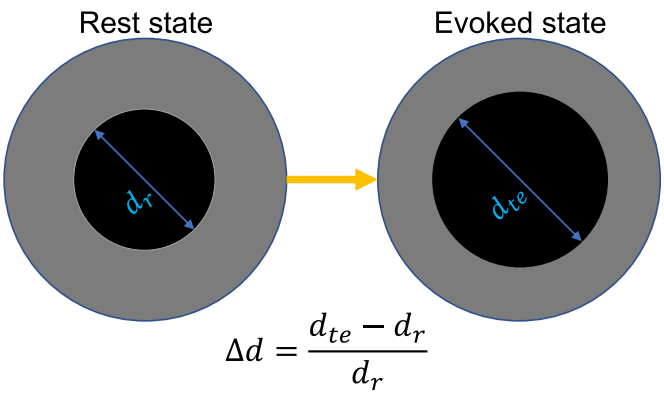

Fig. 7. Task-evoked pupillary response calculation for our experiments.

In Fig. 9(b) and (c), we plot the mean TEPR sequence with 95% CI, before and after the freeze event, respectively. We find that mean pupillary response begins to increase near freeze in the TEPR sequence before the freeze action is performed, representing an increase in CW of operators at this time. A high value just after freeze action and a sharp decrease afterwards shows that the operators' CW decreases immediately after they find a suitable view and freeze it. We also observe that the pupillary response gradually increases after the first freeze in the TEPR sequence, indicative of CW build-up during the interpretation of frozen frames.

6.2. Observed anatomical structures

Pupillary responses of the Brain and Heart ultrasonographic tasks are compared. Fig. 10(a) shows the box-whisker plot of the pupillary response distributions. It shows a significantly higher overall TEPR for Brain compared to Heart ($\Delta = 2.20\%, p < 0.05$). This result is indicative of higher cognitive load of operators while observing the fetal head and brain anatomical structures compared to the fetal heart and thorax.

In Fig. 10(b), we plot the mean TEPR sequence with the 95% CI for each of the two ultrasonographic tasks. We observe that the mean TEPR sequence decreases with time for Brain and increases with time for Heart sequences. This is interesting to observe, and suggests that CW for observing Brain and neighbouring structures is higher in the fine-tuning stage and the beginning of the interpretation task and then decreases gradually, whereas, an opposite phenomenon is observed for Heart and nearby anatomies. For the Brain sequences, a higher CW in the fine-tuning stage suggests that the operators exert higher amount of cognitive efforts to perform the precise localisation of the brain for accurate biometric measurements. For the Heart sequences, the increase in CW during interpretation may be explained by the complexity of interpreting the frozen standard planes of the fetal heart in comparison to the fetal brain at 20 weeks of gestation. There are two standard fetal brain planes, namely, Transventricular (TV) and Transcerebellar (TC), and five standard fetal heart planes, namely, Three Vessel Trachea (3 V T), Right Ventricular Outflow Tract (RVOT), Left Ventricular Outflow Tract (LVOT), Four Chamber View (4CH) and Situs, along with Doppler views. Each of the heart standard planes have more features to be assessed in comparison to any other fetal organ. Additionally, heart is a smaller structure and requires more operator concentration during interpretation of frozen frames.

Furthermore, an important observation is that Heart sequences have a shorter mean duration than Brain sequences as depicted in the figure. The difference in the mean durations of the Heart and the Brain sequences can be explained as the following. In a previous related study, operator clinical workflow was analysed in the second-trimester clinical fetal ultrasound scans [8], where it was found that a freehand scan is difficult to perform and operator-dependent, as it consists of arbitrary task types, order, and time-distributions, with multiple attempts and repetitions if the tasks were not successfully completed earlier. The reason for difference in the mean durations is two-fold. Firstly, the acquisition of the Brain sequences involves precise biometric measurements such as the head circumference (HC) and the biparietal diameter (BPD), in contrast there are no biometric measurements involved for the acquisition of the Heart sequences, leading to longer duration for the Brain sequences. Secondly, heart is a smaller structure and a higher number of heart standard planes are needed to be acquired compared to the brain, hence, operators may require multiple attempts to scan the heart, leading to a higher number of repetitions if the planes are not satisfactorily captured earlier. In contrast, during the brain scanning, operators can analyse the visual information with fewer attempts and repetitions in the scan. This has led to a lower mean duration for the

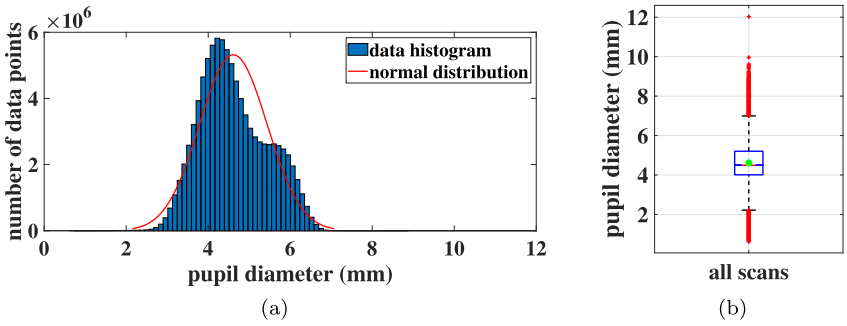

Fig. 8. Distribution of raw pupil diameters (a) Histogram (b) Box-whisker plot.

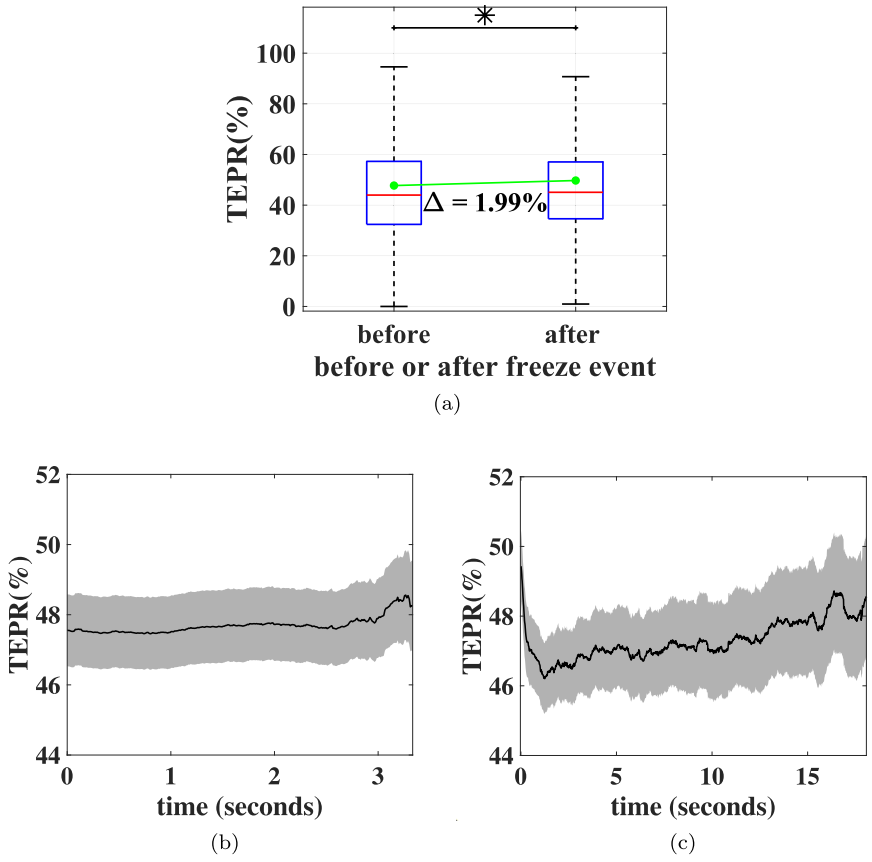

Fig. 9. Pupillary response before and after freeze event. (a) Distributions of TEPR values and Mean TEPR sequences with 95% CI (b) before and (c) after freeze event.

individual Heart sequences compared to the Brain sequences.

6.3. Operator experience

Pupillary responses of the NQ and XP experience groups are compared. Fig. 11(a) shows the box-whisker plot of the pupillary response distributions for NQ and XP experience groups. It shows slightly higher overall TEPR for XP compared to NQ ($\Delta = 0.84\%$, $p < 0.05$), indicating a higher CW exertion by experienced operators compared to newly-qualified operators. This observation is contrary to intuition. A pilot study was previously performed to compare the pupillary responses of newly-qualified and experienced operators for the predominant anatomical tasks in ultrasound scans [24]. In that study, we observed that the global pupillary response of the NQ operators was slightly higher compared to the XP operators. However, in this paper, the Heart and the Brain ultrasonographic tasks are selected due to the reasons outlined in Section 4.2, and the distribution is found only for the task-specific TEPR sequences. This observation indicates that experienced operators demonstrate a slightly higher CW compared to newly-qualified operators for these two complex tasks. Moreover, the overall pupillary response difference between the two skill groups is lower compared to the previous comparisons, which indicates that the pupillary response values are closer for the two experience groups, and it can be more difficult to discriminate between experience levels based on these values.

In Fig. 11(b), we plot the mean TEPR sequence with the 95% CI for each of the two skill groups. We observe that the mean TEPR sequence gradually increases with time for NQ operators, whereas it sharply decreases with time for XP operators. The observation suggests that experienced operators concentrate more during the fine-tuning of the

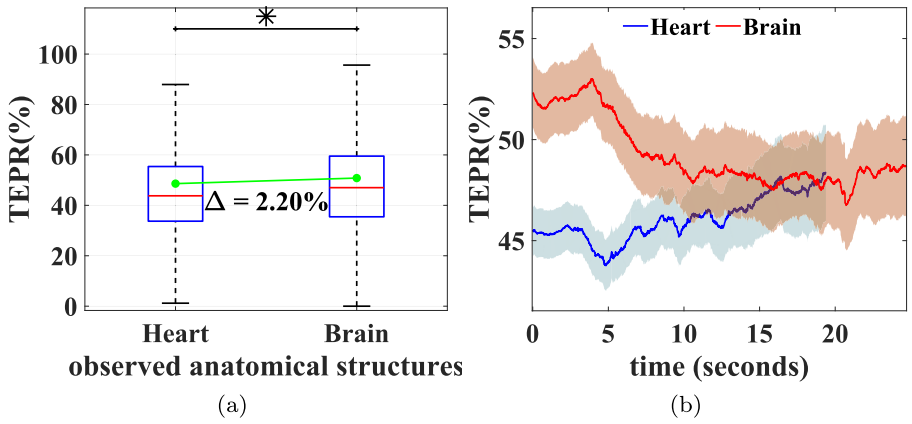

Fig. 10. Pupillary response for Heart and Brain ultrasonographic tasks. (a) Distributions of TEPR values (b) Mean TEPR sequence with 95% CI.

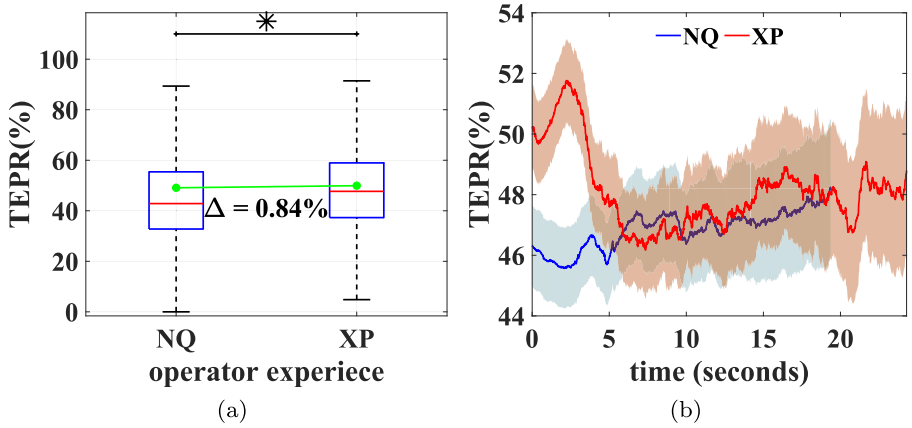

Fig. 11. Pupillary response for NQ and XP experience groups. (a) Distributions of TEPR values (b) Mean TEPR sequence with 95% CI.

anatomical views and rapidly feel a lower CW as they progress into the interpretation of the given anatomical structures, whereas the newly-qualified operators experience a higher CW into the interpretation stage including inspection and biometric measurements.

Moreover, we observe that NQ sequences have a shorter mean duration than XP sequences. The difference in the mean duration of the sequences captured by the NQ and the XP operators can be explained as the following. From clinical workflow analysis [8], it was observed that the NQ operators show higher disorder (entropy), with a greater number of attempts and repetitions of the scan tasks. For complex anatomies such as the Heart and the Brain, it is intuitive that individual sequence durations of the NQ operators would be lower compared to the XP operators, due to a lower clinical experience, thereby, less capability to complete these tasks in succession and repeating these later if they were not satisfactorily performed earlier in the scan.

The above comparative analyses of the TEPR values suggest that there exist perceivable variations in the pupillary responses, thereby indicating measurable differences in the CW of ultrasound operators, depending on the observed anatomical structures and their scanning experience. We further explore these two factors by training machine learning models with features extracted from temporal TEPR sequences to automatically distinguish between ultrasonographic tasks and operator experience groups.

7. Machine learning for observed anatomical structures

After extracting the corresponding TEPR sequences from the processed pupil diameters, each sequence is treated as a separate time-series with associated binary class labels for the ultrasonographic task, namely, Brain or Heart. Machine learning models are trained to automatically classify the TEPR sequences into one of the two tasks. Uniform windowing of the sequences on either side of the triggering event is comparatively evaluated using the best-performing learned models.

7.1. Feature extraction and classification

Hand-crafted features including temporal, spectral and time-frequency features, are extracted from the pupillary response sequences and used to train classical machine learning models inferring the ultrasonographic tasks. Deep learning models are also explored in this context which do not require hand-crafted feature extraction. The feature extraction and classification methods are explained next.

7.1.1. Temporal features

In this category, the temporal characteristics of the TEPR sequences are emphasized, and statistical features are directly computed for each time-series. The computed eight temporal features are mean, median, minima, maxima, standard deviation, inter-quartile range, skewness, and kurtosis.

7.1.2. Spectral features

Here, the discrete Fourier transform of each TEPR time-series is computed. This is followed by finding the statistics of the spectrum, specifically, frequencies of the $K = 4$ amplitude peaks, mean frequency,

median frequency, maximum amplitude, total power, and mean power, resulting in a total of nine spectral features.

7.1.3. Time-frequency features

This feature group involves wavelet analysis of the TEPR time-series to obtain a time-frequency representation of the signal [31]. Here, each TEPR sequence is first partitioned into windows, followed by the application of a multi-scale wavelet transform at five levels of decomposition. For the wavelet coefficients, eight statistical features are computed, namely, the mean, variance, standard deviation, energy, kurtosis, skewness, waveform length and entropy. A total of 240 time-frequency features are obtained for this feature group.

7.1.4. Classical machine learning

The hand-crafted features are used for binary classification using support vector machine (SVM) learners. SVM models are trained using radial basis function (RBF) kernels to distinguish between the two ultrasonographic tasks. Default hyperparameter settings [32] are used as we perform experiments for different classification problems (ultrasonographic tasks and operator experience), and we want to demonstrate multi-task characteristics of the machine learning method using the pupillary response data. Since the datasets are unbalanced for both the observed anatomical structures and the operator expertise, the Synthetic Minority Oversampling Technique (SMOTE) [33] is applied to balance the binary class labels for the extracted features.

7.1.5. Deep learning using convolutional neural networks

We explore Convolutional Neural Networks (CNN) to learn the salient features of temporal pupillary response sequences, and to classify these into one of the two ultrasonographic tasks. The CNN models are designed in two ways. First, a one-dimensional (1D) CNN model is investigated that directly learns features from one-dimensional TEPR time-series. Second, a two-dimensional (2D) CNN based on the ResNet-18 CNN architecture is learnt using two-dimensional images obtained from wavelet scalograms from the corresponding TEPR time-series.

7.1.5.1. 1D CNN. A simple and lightweight 1D CNN architecture is employed, which extracts features from the TEPR sequence to perform classification. The CNN architecture comprises of a cascade of three identical layers of 1×5 1D kernel convolutions with 64 filters followed by ReLU nonlinearity, batch normalisation, dropout ($p = 0.2$) and maxpooling, followed by a global average pooling, fully-connected, and softmax layer for classification. A low complexity CNN architecture is designed to allow efficient parametrisation with respect to the limited size of the datasets. The total number of trainable parameters in the proposed 1D CNN architecture is 1,857,714. The weights of the 1D CNN architecture are randomly initialised for training the TEPR sequences, as there is no existing pre-trained 1D CNN with a closely associated domain to the temporal pupil diameter datasets that could be used for transfer learning. Initially, randomly initialised long-short term memory (LSTM) units, popular with time-series data, were also investigated but were not found suitable for the pupillary response signals in the studied datasets. The reasons for such behaviour could be a smaller-sized data compared to large-scale data required for complex training of the LSTM units, and the nature of the pupillary response signal itself.

7.1.5.2. Wavelet scalograms with 2D CNN. Wavelet scalograms are time-frequency plots of the absolute value of the wavelet transform of the signal representing the proportion of energy for each wavelet coefficient. These are useful tools in time-series analysis allowing the detection of the most representative scales in the signals [34]. We compute a wavelet scalogram for each TEPR sequence and the resulting two-dimensional data is used to train a 2D CNN. A ResNet-18 CNN architecture [35] is selected for the image-based training, as it provides a good balance of network size and accuracy on general image classification. Deeper and heavier networks may provide a higher accuracy for general images, but are unsuitable for smaller datasets leading to overfitting. A ResNet-18 CNN pre-trained on ImageNet [36] is used, and weights are fine-tuned using the wavelet scalograms of the TEPR sequences. The last fully-connected and softmax layers are replaced for the given class labels. The total number of trainable parameters in the 2D CNN architecture is 11,181,314.

7.1.5.3. Training process. The training of both deep learning models was performed with the following hyperparameter settings. Stochastic gradient descent with momentum ($\mu = 0.9$) was used as the optimiser to update network parameters. The initial learning rate was set to 0.01 with a drop of 0.1 after every 10 epochs. The networks were trained for a total of 50 epochs. A batch size of 32 was used during training.

Since deep learning models, like classical machine learning, can be adversely affected by imbalanced datasets, leading to undesirable biases in the learnt models, we trained the CNNs using balanced data. This was achieved by random undersampling, as the SMOTE method was not applicable here due to a different nature of the data consisting of TEPR sequences of uneven length and 2D scalogram images, in contrast to equal-length hand-crafted features. In each training round, the number of samples of each class was set equal to the number of samples of the least represented class of the training dataset. For instance, training data in one round consisted of 592 samples each for the Brain and the Heart ultrasonographic tasks.

In order to process the TEPR sequences by the CNNs, these also needed to be adjusted for length due to their uneven durations. Zero-padding, a standard operation used for time-series length adjustment, was applied to the TEPR sequences for the 1D CNN. Truncation was not performed to prevent information loss. In contrast, scalograms for 2D CNN were computed after making sequences equal in length to the average sequence length, which involved either zero-padding for shorter sequences, or truncation for larger sequences. It is important to note that zero-padding to maximum length was not employed for 2D CNNs, as this led to unreasonable aspect ratios in the resulting scalogram images.

7.2. Performance evaluation

7.2.1. Cross-validation

The discrimination ability of the extracted features learnt with classical machine learning method, and deep learning methods was evaluated through five-fold cross-validation to differentiate between the pupillary response of Heart and Brain sequences. A scan-wise split was implemented in each round of cross-validation. The same cross-validation splits were used for evaluating the different machine learning methods, to ensure a fair comparison between the methods. The reported standard metrics for both binary classification experiments include the mean and standard deviation of the accuracy of each class label and the overall accuracies, computed over the cross-validation rounds.

Table 2 depicts the evaluation results for learning the pupillary

Table 2
Performance evaluation for discrimination of ultrasonographic tasks. Values are mean and standard deviation computed over the five-fold cross-validation rounds.

Extracted Features	Learning Method	Accuracy Brain	Accuracy Heart	Overall Accuracy
Temporal	SVM	0.44 ± 0.05	0.67 ± 0.02	0.60 ± 0.03
Spectral	SVM	0.45 ± 0.07	0.70 ± 0.06	0.62 ± 0.03
Time-Frequency	SVM	0.59 ± 0.04	0.83 ± 0.02	0.75 ± 0.02
TEPR Sequence	1D CNN	**0.84 ± 0.17**	**0.96 ± 0.06**	**0.87 ± 0.10**
Scalogram	ResNet18 2D CNN	0.64 ± 0.15	0.58 ± 0.10	0.60 ± 0.03

response sequences using the explored machine learning methods to discriminate between the Heart or Brain tasks.

From the above results, we observe that the best-performing method for this problem learns the temporal TEPR sequence directly using a 1D CNN model, with an accuracy of 0.87, and outperforms other models in both Brain and Heart accuracy. Wavelet-based time-frequency hand-crafted features learnt by classical SVM models also show reasonable performance, with an accuracy of 0.75. We observe that the scalogram with image-based CNN model is not as successful as the former mentioned models. Specifically, it is unable to achieve good classification for Heart sequences; the main reason being the sequence length adjustment that has led to either zero-padding or truncation of unevenly long sequences. For the Heart sequences, a high standard deviation of duration (Section 4.3) indicates that these sequences have highly uneven lengths, thus processing these for uniform length, followed by scalogram computation may not be suitable and leads to lower performance. Furthermore, the simpler temporal and spectral features with classical machine learning do not lead to as accurate classification as the other proposed methods. Overall, the 1D CNN-based deep learning models outperformed the 2D CNN and hand-crafted features with classical machine learning, to achieve a more accurate classification discriminating between the ultrasonographic tasks as Heart or Brain.

7.2.2. Effect of sequence windowing

The experiments for comparative evaluation of machine learning methods preserved the original length of the TEPR sequences. However, recent research has shown promising results using symmetric windowing of the pupillary response sequences on either side of event triggers [10]. Hence, in this section, we explore symmetric windows around the first detected freeze, which is the triggering event for the pupillary response sequences.

Specifically, sequence windowing is explored for classification of ultrasonographic tasks using all the described machine learning methods. For each method, we comparatively analyse receiver operating characteristics (ROCs) and area under the curve values for the four windowing cases: original sequence length (asymmetric), a 1 s window on each side of first freeze (symmetric-1 sec), a 2 s window on each side of first freeze (symmetric-2 sec), and a 3 s window on either side of first freeze (symmetric-3 sec).

We observe that the quantitative results of sequence windowing are consistent with our observations in Table 2 for ultrasonographic task classification. The best ROC metrics in all the four windowing cases are achieved using the 1D CNN method, which is also the best-performing method in the cross-validation experiment. Moreover, performance of the other methods after windowing effects shows the same order as obtained earlier, *i.e.*, time-frequency features, spectral features, scalogram-based 2D CNN, and temporal features (in decreasing order of classification accuracy). Hence, we subsequently report in detail, the effect of sequence windowing for the 1D CNN method.

The comparative ROC curves for 1D CNN method are depicted in Fig. 12 for ultrasonographic task classification. From the ROC curves, we observe that for the discrimination of ultrasonographic tasks, symmetric windowing of the pupillary response sequences on each side of the freeze event shows superior performance compared to the original asymmetric windows. This observation is consistent with the results reported in Ref. [10]. A reason for a lower area under curve (AUC) for asymmetric length sequences is that the zero-padding applied to the TEPR sequence for length adjustment to train the 1D CNN model, can perturb the composition of the original pupillary response sequence. Therefore, using the experimentally best-performing machine learning method and the best sequence windowing, we achieve an average area under ROC curve as 0.98 for ultrasonographic task classification.

7.2.3. Generalisability

Generalisability of machine learning models trained on physiological data is a known challenge. We previously explored the generalisability of machine learning models trained on the acquired datasets in the PULSE study for other problems such as operator workflow analysis using videos [8] and skill assessment using probe motion data [37]. We observed that the performance of the models generally decreases when prediction is performed on unseen operators. This behaviour can be attributed to a limited amount of data for individual operators with high inter-operator variability and the scan imbalance among operators and experience groups (Table 1). In Ref. [37], we used domain adaptation to make the models operator invariant and overcome the generalisability issue.

To investigate the generalisation power of the proposed machine learning models, we performed operator hold-out validations which could give a better idea about how well the models predict ultrasonographic tasks from the pupillary responses of unseen operators. From Table 1, we observe that there are more XP operators (9) compared to the NQ operators (3), however, the number of scans is usually lower for individual XP operators than NQ operators. Hence, we selected hold-out test sets consisting of all the pupillary response sequences from certain NQ (O_3) and XP (O_4, O_{10}, O_{12}) operators, and the training set consisted of the sequences from the remaining operators. The selection was performed such that the hold-out sets do not reduce the size of the training set substantially and have sufficient sequences for testing from the different categories. We report the best-generalising model as time-frequency features with SVM-based classical machine learning method, with accuracy for Brain as 0.54 ± 0.03, accuracy for Heart as 0.78 ± 0.04 and overall accuracy as 0.72 ± 0.02. As expected, the metrics are lower compared to cross-validation due to high inter-operator variability and limited sized datasets with data imbalance among operators and experience groups. Overall, the classical machine learning (shallow) models outperform the deep learning models, as the latter show bias towards one of the two classes, which reiterates the requirement of more large-scale data to train the deep networks.

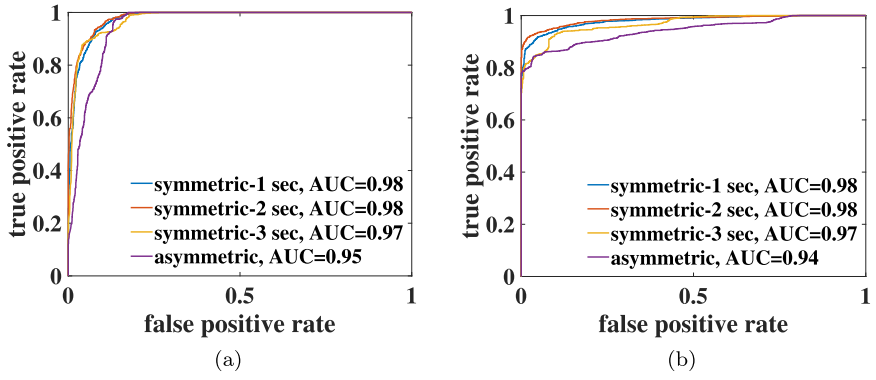

Fig. 12. ROC curves for sequence windowing for ultrasonographic task classification for (a) Brain and (b) Heart.

8. Machine learning for operator experience

Each TEPR time-series has an associated binary class label as the operator experience group based on the scanning experience the ultrasound operator, namely, NQ or XP group. Machine learning models are trained to automatically classify the TEPR sequences into one of the two experience groups. Uniform windowing of the sequences on either side of the triggering event is comparatively evaluated using the best-performing machine learning models as in Section 7.

8.1. Feature extraction and classification

Hand-crafted features, such as temporal, spectral and time-frequency features, are extracted from the pupillary response sequences for training classical machine learning models. Under deep learning, convolutional neural networks are explored. The feature extraction and classification methods for operator experience classification are identical to those explained in Section 7.1 for ultrasonographic task classification.

8.2. Performance evaluation

8.2.1. Cross-validation

Cross-validation is performed for operator experience classification in the same manner as explained in Section 7.2.1 for ultrasonographic task classification. Table 3 depicts the evaluation results for learning pupillary response sequences to discriminate between the operator experience groups as NQ and XP.

We observe that, in general, the performance of the explored methods at discriminating experience groups is lower compared to that for discriminating ultrasonographic tasks, which can be attributed to the former being a more complex classification problem. We note this because, during the statistical analysis (Section 6), we observed lower difference in the distributions of TEPR values between the two groups NQ and XP, suggesting a harder classification problem. Moreover, an interesting observation is that hand-crafted methods with classical machine learning have outperformed deep learning methods. A possible explanation for this observation is the use of random under sampling to balance datasets in deep learning (as opposed to SMOTE of features in classical machine learning), which may have resulted in lower-sized datasets in certain cross-validation rounds due to a high class imbalance between the NQ and XP samples, whereas it is well-established that deep learning works favourably with large-scale data. Wavelet-based features with SVM models outperform the other methods in all metrics, with a mean accuracy of 0.75. Other proposed methods have a comparable performance. 1D CNN does not perform favourably for the XP operators.

8.2.2. Effect of sequence windowing

Sequence windowing is explored for the classification of operator experience group using all the machine learning methods. For each method, we compare the receiver operating characteristics (ROCs) and area under the curve values for the four sequence windowing cases, as mentioned in Section 7.2.2.

From detailed analysis we observe that, consistent with our observation in Table 3, the highest ROC metrics in all the four windowing cases are achieved using time-frequency features with SVM-based classical machine learning method. Moreover, performance of the other classification methods after windowing effects shows the same order as obtained earlier, *i.e.*, temporal features, spectral features, scalogram-based 2D CNN, and 1D CNN (in decreasing order of classification accuracy). The best-performing method, namely, time-frequency features with SVM is considered for investigating the effect of sequence windowing for operator experience classification.

The ROC curves of the four windowing cases using time-frequency features with SVMs for operator experience classification are shown in Fig. 13. A comparative analysis of the ROC curves shows that the original asymmetric TEPR sequences are consistently more effective than the symmetrically windowed sequences for experience classification. The observation suggests that it is important to preserve the original length of the pupillary response sequences in order to infer operator experience using machine learning, as this may not only depend on the pupil diameter changes in the vicinity of the freeze event trigger, but on the overall pupillary response of operators during the entire scan. We can conclude that the best-performing classification model has an average area under ROC curve of 0.80 for operator experience classification.

Table 3
Performance evaluation for discrimination of operator experience. Values are mean and standard deviation computed over the five-fold cross-validation rounds.

Extracted Features	Learning Method	Accuracy NQ	Accuracy XP	Overall Accuracy
Temporal	SVM	0.69 ± 0.04	0.59 ± 0.04	0.66 ± 0.03
Spectral	SVM	0.70 ± 0.01	0.51 ± 0.05	0.63 ± 0.03
Time-Frequency	SVM	**0.81 ± 0.04**	**0.63 ± 0.05**	**0.75 ± 0.04**
TEPR Sequence	1D CNN	0.72 ± 0.15	0.38 ± 0.24	0.62 ± 0.06
Scalogram	ResNet18 2D CNN	0.57 ± 0.16	0.55 ± 0.14	0.58 ± 0.03

8.2.3. Generalisability

In order to investigate the generalisation power of the proposed machine learning methods for operator experience classification, we performed operator hold-out validations similar to those explained in Section 7.2.3. This was done to understand whether the models are able to predict the operator experience from the pupillary responses of unseen operators. Interestingly, the best-generalising model is the same as observed in cross-validation and sequence windowing, *i.e.*, time-frequency features with SVM-based classical machine learning method, with accuracy for NQ as 0.57 ± 0.02, accuracy for XP as 0.42 ± 0.10 and overall accuracy as 0.54 ± 0.04. The performance for discriminating operator experience is lower compared to discriminating ultrasonographic tasks, due to a more complex problem, as observed in the previous experiments. The metrics are lower compared to cross-validation due to similar reasons as for ultrasonographic task classification. Again, the classical machine learning (shallow) models outperform the deep learning models, as explained in Section 7.2.3.

9. Discussion

In the data processing stage, we performed a systematic preprocessing of the pupil diameters using bespoke guidelines and observed more desirable characteristics of the processed data compared to the raw data. On performing exploratory statistical analysis on the operators' pupillary response, we observed measurable differences between before and after triggering events, anatomical structures, and scanning expertise, that helped us assess the operator's mental workload during the different stages of the scan. Two classification problems were investigated using the temporal pupillary response sequences and machine learning methods, namely, discrimination of ultrasonographic tasks and operator experience. For discriminating the pupillary response for observed anatomical structures based on undertaken ultrasonographic tasks, deep learning methods proved to be more accurate (Brain 84%, Heart 96%). For discriminating the pupillary response between different operator experience groups, time-frequency features with classical (shallow) machine learning achieved better classification performance than deep learning (NQ 81%, XP 63%). From this context, we identify the factors to consider in interpreting our findings, limitations, and opportunities for future work as the following.

Firstly, the correlation of the operator CW was found stronger with the ultrasonographic task at hand, than with the operator experience,

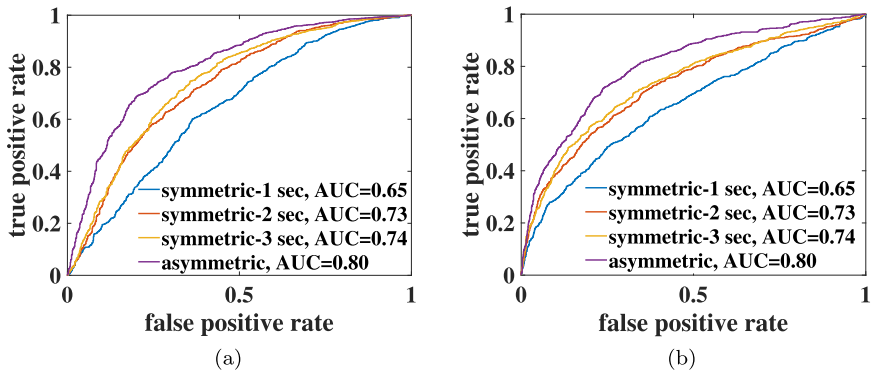

Fig. 13. ROC curves for sequence windowing for operator experience classification for (a) NQ and (b) XP groups.

both statistically and *via* machine learning. This result may be influenced by the currently assumed definitions of the NQ and XP experience groups with a threshold at 2 years. The choice of 2 years as the threshold was recommended by fetal ultrasound specialists and led to a lower data imbalance compared to other thresholds. An interesting future direction would be to vary this threshold, if more data was available. A second consideration is that the classification rates reflect the relatively limited size of datasets and class imbalance in the routinely acquired data, underlying complexity of the naturally acquired pupil diameters, and difficulty of the given classification tasks. Our generalisability analysis suggests that inter-operator variability and data imbalance among operators could lead to overfitting, and the proposed methods need to be tested on large-scale datasets with higher number of operators in each experience group and more scans per operator. Thus, acquisition of more data would allow further refinement of modelling. Lastly, if our analysis was to be applied to other biomedical imaging applications, our initial assumptions about environmental conditions during real-world ultrasound scanning may require revision.

In summary, the study proposes a systematic multi-modal data acquisition, pre-processing, and analysis pipeline to estimate the pupil diameter changes of ultrasound operators *via* non-contact eye-tracking technology. From the perspective of practical usability, the study provides insights to inform understanding of the operators' mental efforts during diagnostic medical imaging that can help increase the acquisition efficiency and reduce human errors in the clinical care pathways. Other potential practical applications include objective operator skill assessment for training and education, and enhancement of human-computer interfaces for better imaging technology design. Under this vision, in future, it would be interesting to investigate the manual activities of image acquisition and understand the physical component of operator workload (for example, repetitive strain injuries of ultrasound operators) for a comprehensive workflow analysis in the clinical scan room.

10. Conclusion

This paper explores ideas of *Sonography Data Science* and *Pupillometry* to analyse pupil diameter changes of ultrasound imaging operators for the assessment of their cognitive or mental workload in the context of clinical fetal ultrasound scanning. We presented a comprehensive pipeline to acquire and analyse multi-modal data in real-world settings, including remote eye-tracking and scan video data. We performed systematic pre-processing and sequence extraction (time-series) of the pupillary responses of multiple ultrasound imaging operators. We performed an exploratory statistical analysis on the operators' pupillary responses and observed measurable variations in the scan reflecting their cognitive workload, for example, between ultrasonographic tasks and scanning expertise. Furthermore, we explored machine learning models using classical (shallow) and deep learning for the automatic inference of the ultrasonographic task and scanning experience using the temporal pupillary response sequences. We investigated the effects of sequence windowing around triggering events and generalisability on the classification performance of the learned models. The ability to objectively assess cognitive workload can be the first step towards understanding how this may affect the observer's performance in routine diagnostic medical imaging.

11. Implementation details

MATLAB® 2020a was used for the data analysis, visualisation, and performance evaluation. Hand-crafted features were extracted and SVMs for classical machine learning were trained using MATLAB functions. Wavelet analysis for classical machine learning was based on the functions in MECLab Toolbox [38,39] with required modifications. MATLAB Deep Learning Toolbox was used for the design, training and classification using the 1D CNN and 2D CNN models. One NVIDIA GTX 1070 (8 GB) graphics card was used for deep learning. All the data analysis was performed on a workstation containing Microsoft® Windows 10 64-bit operating system with Intel®Core i7 processor at 3.60 GHz.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge the ERC (ERC-ADG-2015 694581 project PULSE), the EPSRC (EP/MO13774/1), and the Oxford Partnership Comprehensive Biomedical Research Centre funded by the NIHR Biomedical Research Centre (BRC) funding scheme. We thank Pierre Chatelain for his contribution in the data acquisition setup. We thank Richard Droste for his contribution in the extraction of scanning parameters for video data, and the raw eye-tracking data. We thank the anonymous reviewers for their invaluable comments that helped to improve the paper.

References

[1] M.S. Young, K.A. Brookhuis, C.D. Wickens, P.A. Hancock, State of science: mental workload in ergonomics, Ergonomics 58 (1) (2015) 1–17, https://doi.org/10.1080/00140139.2014.956151.

[2] B.A. Wilbanks, S.P. McMullan, A review of measuring the cognitive workload of electronic health records, CIN, Computers, Informatics, Nursing 36 (12) (2018) 579–588, https://doi.org/10.1097/CIN.0000000000000469. https://journals.lww.com/cinjournal/Abstract/2018/12000/A_Review_of_Measuring_the_Cognitive_Workload_of.3.aspx.

[3] P. van der Wel, H. van Steenbergen, Pupil dilation as an index of effort in cognitive control tasks: a review, Psychon. Bull. Rev. 25 (6) (2018) 2005–2015, https://doi.org/10.3758/s13423-018-1432-y.

[4] A. Szulewski, D. Kelton, D. Howes, Pupillometry as a tool to study expertise in medicine, number: 3, Frontline Learning Research 5 (3) (2017) 55–65, https://doi.org/10.14786/flr.v5i3.256, https://journals.sfu.ca/flr/index.php/journal/article/view/256.

[5] D. Kirwan, NHS Fetal Anomaly Screening Programme, 18+ 0 to 20+ 6 Weeks Fetal Anomaly Scan National Standards and Guidance for England, NHS, 2010. http://www.perinatal.nhs.uk/ultrasound/Final_ultrasound_standards.pdf.

[6] L. Drukker, J. Noble, A.T. Papageorghiou, Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology, Ultrasound Obstet. Gynecol. 56 (4) (2020) 498–505, https://doi.org/10.1002/uog.22122. https://obgyn.onlinelibrary.wiley.com/doi/full/10.1002/uog.22122.

[7] M. Yaqub, B. Kelly, H. Stobart, R. Napolitano, J. Noble, A. Papageorghiou, Quality-improvement program for ultrasound-based fetal anatomy screening using large-scale clinical audit, Ultrasound Obstet. Gynecol. 54 (2) (2019) 239–245, https://doi.org/10.1002/uog.20144. https://obgyn.onlinelibrary.wiley.com/doi/full/10.1002/uog.20144.

[8] H. Sharma, L. Drukker, P. Chatelain, R. Droste, A.T. Papageorghiou, J.A. Noble, Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos, Med. Image Anal. 69 (2021) 101973, https://doi.org/10.1016/j.media.2021.101973. https://www.sciencedirect.com/science/article/pii/S1361841521000190.

[9] J. Klingner, R. Kumar, P. Hanrahan, Measuring the task-evoked pupillary response with a remote eye tracker, in: Proceedings of the 2008 Symposium on Eye Tracking Research & Applications - ETRA '08, ACM Press, Savannah, Georgia, 2008, p. 69, https://doi.org/10.1145/1344471.1344489. http://portal.acm.org/citation.cfm?doid=1344471.1344489.

[10] N.V.K. Medathati, R. Desai, J. Hillis, Towards inferring cognitive state changes from pupil size variations in real world conditions, in: ACM Symposium on Eye Tracking Research and Applications, ETRA '20 Full Papers, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–10, doi:10.1145/3379155.3391319, https://doi.org/10.1145/3379155.3391319.

[11] E.H. Hess, J.M. Polt, Pupil size in relation to mental activity during simple problem-solving, Science 143 (3611) (1964) 1190–1192, https://doi.org/10.1126/science.143.3611.1190.

[12] D. Hahnemann, J. Beatty, Pupillary responses in a pitch-discrimination task, Percept. Psychophys. 2 (3) (1967) 101–105, doi:10.3758/BF03210302, https://doi.org/10.3758/BF03210302.

[13] D. Kahneman, J. Beatty, Pupil diameter and load on memory, Science 154 (3756) (1966) 1583–1585, https://doi.org/10.1126/science.154.3756.1583.

[14] A. Szulewski, S.M. Fernando, J. Baylis, D. Howes, Increasing pupil size is associated with increasing cognitive processing demands: a pilot study using a mobile eye-tracking device, 02 (01), Open J. Emerg. Med. (2014) 8–11, https://doi.org/10.4236/ojem.2014.21002. scirp.org/journal/doi.aspx?DOI=10.4236/ojem.2014.21002.

[15] E.L. Johnson, A.T. Miller Singley, A.D. Peckham, S.L. Johnson, S.A. Bunge, Task-evoked pupillometry provides a window into the development of short-term memory capacity, publisher: Frontiers, Front. Psychol. 5 (2014), https://doi.org/10.3389/fpsyg.2014.00218, https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00218/full.

[16] E. Debie, R.F. Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. Anavatti, M. Garratt, H. A. Abbass, Multimodal Fusion for Objective Assessment of Cognitive Workload: A Review, IEEE Transactions on Cybernetics, 1–14Conference Name: IEEE Transactions on Cybernetics, 2019, https://doi.org/10.1109/TCYB.2019.2939399.

[17] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, A. Cichocki, Emotion meter: a multimodal framework for recognizing human emotions, IEEE Transactions on Cybernetics 49 (3) (2019) 1110–1122, https://doi.org/10.1109/TCYB.2018.2797176, conference Name: IEEE Transactions on Cybernetics.

[18] H. Sharma, L. Drukker, A.T. Papageorghiou, J.A. Noble, 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), 2021, pp. 1646–1649, https://doi.org/10.1109/ISBI48211.2021.9433863.

[19] A. Szulewski, A. Gegenfurtner, D.W. Howes, M.L.A. Sivilotti, J.J.G. van Merriënboer, Measuring physician cognitive load: validity evidence for a physiologic and a psychometric tool, Adv. Health Sci. Educ. 22 (4) (2017) 951–968, doi:10.1007/s10459-016-9725-2, https://doi.org/10.1007/s10459-016-9725-2.

[20] P.R. Mosaly, L. Mazur, L.B. Marks, Usability evaluation of electronic health record system (EHRs) using subjective and objective measures, North Carolina, USA, in: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, CHIIR '16, Association for Computing Machinery, Carrboro, 2016, pp. 313–316. doi:10.1145/2854946.2854985, https://doi.org/10.1145/2854946.2854985.

[21] T. Tien, P.H. Pucher, M.H. Sodergren, K. Sriskandarajah, G.-Z. Yang, A. Darzi, Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair, Surg. Endosc. 29 (2) (2015) 405–413, https://doi.org/10.1007/s00464-014-3683-7.

[22] B. Zheng, X. Jiang, M.S. Atkins, Detection of changes in surgical difficulty: evidence from pupil responses, Surg. Innovat. 22 (6) (2015) 629–635, https://doi.org/10.1177/1553350615573582.

[23] R. Geva, M. Zivan, A. Warsha, D. Olchik, Alerting, orienting or executive attention networks: differential patters of pupil dilations, publisher: Frontiers, Front. Behav. Neurosci. 7 (2013), https://doi.org/10.3389/fnbeh.2013.00145, https://www.frontiersin.org/articles/10.3389/fnbeh.2013.00145/full.

[24] H. Sharma, L. Drukker, R. Droste, P. Chatelain, A. Papageorghiou, J. Noble, Oc10.02: task-evoked pupillary response as an index of cognitive workload of sonologists undertaking fetal ultrasound, 28. arXiv:https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1002/uog.22266, Ultrasound Obstet. Gynecol. 56 (S1) (2020) 28. doi:10.1002/uog.22266, https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/uog.22266.

[25] P. Chatelain, H. Sharma, L. Drukker, A.T. Papageorghiou, J.A. Noble, Evaluation of gaze tracking calibration for longitudinal biomedical imaging studies, IEEE Transactions on Cybernetics 50 (1) (2020) 153–163, https://doi.org/10.1109/TCYB.2018.2866274.

[26] Are pupil size calculations possible with Tobii Eye Trackers? (aug 2015). URL https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/is-pupil-size-calculations-possible-with-tobii-eye-trackers/.

[27] A. Kay, Tesseract: an open-source optical character recognition engine, Linux J. 159 (2007) 2, 2007, https://www.linuxjournal.com/article/9676.

[28] H. Sharma, R. Droste, P. Chatelain, L. Drukker, A.T. Papageorghiou, J.A. Noble, Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 987–990. https://pubmed.ncbi.nlm.nih.gov/31993109/.

[29] M.E. Kret, E.E. Sjak-Shie, Preprocessing pupil size data: guidelines and code, Behav. Res. Methods 51 (3) (2019) 1336–1342, doi:10.3758/s13428-018-1075-y, https://doi.org/10.3758/s13428-018-1075-y.

[30] G. Marsaglia, W.W. Tsang, J. Wang, Evaluating Kolmogorov's distribution, 1–4, number: 1, J. Stat. Software 8 (1) (2003), https://doi.org/10.18637/jss.v008.i18, https://www.jstatsoft.org/index.php/jss/article/view/v008i18.

[31] D.B. Percival, A.T. Walden. Wavelet Methods for Time Series Analysis vol. 4, Cambridge University Press, 2000.

[32] Train support vector machine (SVM) classifier for one-class and binary classification - MATLAB fitcsvm - MathWorks United Kingdom. https://uk.mathworks.com/help/stats/fitcsvm.html.

[33] R. Blagus, L. Lusa, SMOTE for high-dimensional class-imbalanced data, BMC Bioinf. 14 (1) (2013) 106, doi:10.1186/1471-2105-14-106, https://doi.org/10.1186/1471-2105-14-106.

[34] V.J. Bolós, R. Benítez, The wavelet scalogram in the study of time series, in: Advances in Differential Equations and Applications, Springer, 2014, pp. 147–154. https://link.springer.com/chapter/10.1007/978-3-319-06953-1_15.

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252, https://doi.org/10.1007/s11263-015-0816-y. https://link.springer.com/article/10.1007/s11263-015-0816-y.

[37] Y. Wang, R. Droste, J. Jiao, H. Sharma, L. Drukker, A.T. Papageorghiou, J.A. Noble, Differentiating operator skill during routine fetal ultrasound scanning using probe motion tracking, in: Y. Hu (Ed.), Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis, Springer, 2020, pp. 180–188. https://link.springer.com/chapter/10.1007/978-3-030-60334-2_18.

[38] A.D. Chan, G.C. Green, Myoelectric control development toolbox, CMBES Proceedings 30 (2007).

[39] R. Khushaba, Feature extraction using multisignal wavelet transform decom. https://github.com/RamiKhushaba/getmswtfeat, 2020.