Skill, or Style? Classification of Fetal Sonography Eye-Tracking Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a method for classifying human skill at fetal ultrasound scanning from eye-tracking and pupillary data of sonographers. Human skill characterization for this clinical task typically creates groupings of clinician skills such as expert and beginner based on the number of years of professional experience; experts typically have more than 10 years and beginners between 0-5 years. In some cases, they also include trainees who are not yet fully-qualified professionals. Prior work has considered eye movements that necessitates separating eye-tracking data into eye movements, such as fixations and saccades. Our method does not use prior assumptions about the relationship between years of experience and does not require the separation of eye-tracking data. Our best performing skill classification model achieves an F1 score of 98% and 70% for expert and trainee classes respectively. We also show that years of experience as a direct measure of skill, is significantly correlated to the expertise of a sonographer.

1 Introduction

The definition of human skill in the medical literature is most often quantified by the number of years of experience a trained medical professional has been practicing for. In fetal sonography (pregnancy ultrasound screening), this corresponds to the number of years after qualification. In [30, 27], a sonographer who has been scanning for 2 years or less is defined as newly qualified. In [17], a trained professional who has been scanning for 10 years or more is considered an expert. In other clinical sub-specialties such as surgery, skill is referenced to the number of instances the specific surgery has been performed [22, 11]. Similarly, in dentistry, the number of semesters completed by a trainee is used as a measure of skill [4, 5]. These time-based definitions are over simplified and omit other important factors that contribute to skill level. Some examples are the frequency of scanning over time, quality [31] and interpretation of the recorded image and real-time response to visual feedback [10]. Since maternal and fetal anatomy differs, no two patients will present in the same manner at any given time [10]. These measures of skill are not easily quantifiable, and current definitions used for skill groupings are domain specific.

In medical studies where eye-trackers have been used for skill assessment, researchers typically use metrics such as the number of fixations and saccades and the time taken to complete the task to differentiate groups of clinicians [29, 5, 12, 16]. For example, [18] showed that experts spent significantly less time fixating on a relevant area-of-interest and had a higher fixation count compared to trainees when viewing video clips of an ultrasound examination. These studies depend on suitable experts or eye movement classification algorithms to separate eye-tracking data into fixations, saccades, smooth pursuits, and areas-of-interest.

Separating eye-tracking data into different eye movements is challenging. Research has shown that the selected eye movement classification algorithm is heavily dependent on the chosen parameters,

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

and can return vastly different results within in the same domain-specific application [24]. In fetal sonography, separating task-specific eye tracking into eye movements is made more challenging because of the number of diagnostic planes that need to be captured and assessed; capturing and reading each anatomical plane is considered a separate task. In second-trimester scanning specifically, there are 23 planes to be captured in a 30-40 minutes appointment window [10].

A second question is how to define human skill for this task. In surgery, for instance, it has been proposed to measure skill by the time taken to complete the task and whether each suture was closed correctly [23]. In radiology, how far the radiographer deviated from the problematic areas could be an indicator that they did not identify the lesion as quickly as another expert [20, 15]. These definitions do not account for the nuances in fetal sonography, where skill metrics based on eye or probe/hand motion is non-trivial, due to the fast probe movement, fetal movement, unstructured transitions between numerous anatomical planes, sonographer experience, and maternal and fetal anatomy [10]. We test the hypothesis of whether grouping sonographer expertise based on years of experience is a suitable measure of skill.

1.1 Related Work

There are several studies that use eye-tracking data for task-specific skill classification. However, it is more typical to use tool motion data either on its own or in combination with other data modalities in applications such as surgery and fetal sonography [19, 21, 30], as opposed to only eye-tracking data. A Hidden Markov model is used in [1] to classify skill between experts and novices using both their eye-tracking and tool motion data. A statistical model was fitted to eye-tracking and tool motion data for experts and novices in endoscopic sinus surgery [2]. In the field of fetal sonography, [26] uses a combination of eye-tracking, pupillary data, and image data to classify newly qualified and expert sonographers. To define skill, [1] uses anatomical knowledge and operational knowledge of endoscopes as skill indicators. However, in fetal ultrasound, prior skill characterisation studies that use eye tracking or probe motion have used years of experience as a skill indicator [27, 26, 30, 17]. These studies combine eye-tracking with other data modalities, did not consider task-agnostic gaze behaviour, and are still largely limited by a year threshold cut-off to define skill [1, 2, 30, 31, 27].

1.2 Contribution

Our main contributions are as follows. We build a task-agnostic skill classification model using only eye-tracking and pupillary data of sonographers performing fetal ultrasound scans. We calculate the correlation between years of scanning experience and the proportion of predicted expert labels of fully qualified sonographers and its significance (at the 5% level). To determine how well a task-agnostic skill classification model performs on specific tasks, we calculate the significance for anatomical planes with different levels of difficulty.

2 Method

We present an original skill classification model to differentiate trainee and fully-qualified sonographers based on their eye gaze characteristics, and use this to evaluate if years of scanning experience is an indicative measure of skill.

2.1 Skill Classification Model

An *expert* refers to any fully-qualified (FQ) sonographer, independent of their years of scanning experience. A *trainee* refers to a not yet fully-qualified sonographer, who is still learning how to scan. We define *style* as in [30]; the gaze of a sonographer is the outcome of both human skill, and personal scanning style. The purposes of the skill classification model are 2-fold. The first is to differentiate a trainee and an expert's gaze behaviour using eye-tracking and pupillary data. The second is to determine if a sonographer's *style* of scanning affects the performance of the model.

To achieve these aims, we train several groups of models using different experts in the training dataset.

Group 1: Teacher VS trainee We aim to differentiate gaze of a teacher (FQ sonographer) and a trainee. We compare the differences between using a sonographer, as 1) the Teacher, who carries out scans with a trainee present, and 2) the same sonographer carrying out scans individually. 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.

Group 2: FQ sonographers VS trainee This group of models aim to differentiate a population of FQ sonographers (FQSS) from a trainee. Here, FQ sonographers are performing scans individually and in some instances, with a trainee.

Group 3: FQ sonographer VS trainee The final group of models aim to differentiate a single FQ sonographer from a trainee. This is a reversed leave-one-out approach, analogous to 'leave-one-in'. The null hypothesis is that each individual sonographer is able to provide a representation of the gaze behaviour of all expert sonographers.

We use eye-tracking data collected when sonographers were viewing live B-mode ultrasound video streams. Live B-mode video streams are recorded when the sonographer is actively searching for the required anatomical plane. In contrast, frozen frames result when the sonographer has frozen the video and is no longer moving the probe. A fetal ultrasound video typically follows an alternating sequence of live B-mode streams and frozen streams that are referred to as live B-mode and frozen *segments*, respectively. Only live B-mode segments are used for this study.

Instead of labeling the video in terms of the anatomy being assessed at this time as in [25, 30, 31], we consider a task-agnostic approach to classify skill differences. This reduces the need for manual labelling of segments and builds a gaze-based skill classification model that is agnostic to the type of anatomical plane being searched for. To overcome the problem that live B-mode segments are of different lengths, we extract summarized gaze characteristics for each segment using the scalable feature extraction approach `tsfresh` [8].

Gaze Features Following [26] where pupillary data was used to compare differences between sonographers with > 2 years and ≤ 2 years of experience, we calculated the task-evoked pupillary response (TEPR) as a skill classification feature. Briefly, 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 Eq. 1. Following [26], we use the minimum pupil diameter d_r to represent 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.

$$\delta d_t = \frac{d_t - d_r}{d_r} \times 100\% \tag{1}$$

We also include gaze data (x and y coordinates) as features. Each live B-mode segment is represented by a $3 \times n$ feature vector, where n is its segment length. Note that n varies from segment to segment. The feature is then reduced to a $1 \times m$ feature vector, where m is the number of characteristics extracted using [8].

The feature extraction setting used in `tsfresh` was `EfficientParameters` [8] which consist of 74 unique time-series features. This setting was chosen because they provide an overview of time-series properties that are not computationally expensive to calculate and is scalable for large datasets. These features cover a range of time-series properties, such as distribution of data points, correlation properties, stationarity, entropy, and nonlinear time series analysis [8]. 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 we remove features related to length: `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 the length of the segment n, respectively.

The final dataset consists of a matrix of size $d \times m$, where d is the total number of segments available.

Implementation We consider off-the-shelf gradient boosting decision trees which have been shown to have the best performance on tabular data [3], including Categorical Boosting (CatBoost) [9], Light Gradient Boosting Machine (LightGBM) [14], and Extreme Gradient Booasting (XGBoost) [7] classifiers. Gradient boosting decision trees use an ensemble of weak decision trees to build strong predictors [9, 13]. Briefly, XGBoost is a highly scalable and efficient gradient tree boosting algorithm that can handle sparse tabular data because of its algorithmic optimisations detailed in [7]. LightGBM included two extra optimisation steps to handle large amounts of data instances and features, decreasing the computational speed and memory required compared to XGBoost [14].

CatBoost is similar to XGBoost and LightGBM but is specifically designed to handle categorical features [9], of which there are several in the extracted features from `tsfresh`. These models have been shown to be computationally efficient and suited for the classification of tabular data.

We performed a 5-fold stratified cross-validation with 75% of our dataset, and tested on the remaining 20%. We tuned our model parameters using a grid search with 5% of the dataset. Due to class imbalance where experts form the majority class, we use Synthetic Minority Oversampling Technique (SMOTE) [6, 26] to balance our training dataset. Such imbalances in data are not uncommon, where other fetal sonography studies have also had an imbalanced expert/beginner dataset [27, 30]. This imbalance is further amplified when considering seperating sonographers on a per-year (of scanning experience) basis.

2.2 Predicting Skill Level of Fully Qualified Sonographers

We predict labels for live B-mode segments of FQ sonographers with a range of years of scanning experience. Then we calculate the proportion of segments that were labelled as expert and trainee, grouped by years of scanning experience, to test the hypothesis that years of scanning is analogous to skill [27, 30, 17].

We test the significance (at the 5% level) of years of experience and percentage of expert segments using Pearson's correlation coefficient (PC) (Eq. 2). The variables X and Y in Eq. 2 are years of experience and percentage of expert segments (between 0 and 100%) respectively. σ refers to the standard deviation, and cov refers to the covariance. The null hypothesis being tested is that there is no significant correlation (PC=0, Fig. 1) between years of experience and percentage of expert segments. Bar charts are used for visual inspection of the proportions of expert and trainee labels. An example is shown in Fig. 1.

$$PC_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

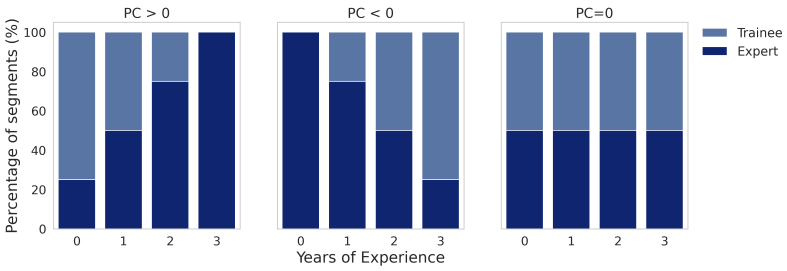

Figure 1: Example bar chart. Percentage of segments predicted as expert (dark blue) or trainee (light blue) grouped by the number of years of experience. PC refers to Pearson's correlation coefficient. $PC > 0$ suggests that the years of experience and percentage of expert segments are positively correlated. $PC < 0$ suggests that the years of experience and percentage of expert segments are negatively correlated. $PC = 0$ suggests that the years of experience and percentage of expert segments are neither positively or negatively correlated.

We also investigate how the gaze skill classification model performs at the diagnostic plane task level. This is done by predicting an expert or trainee label on labelled diagnostic plane live B-mode segments. In our work, we use the head circumference (HC), abdominal circumference (AC), and heart plane finding tasks, which have been used in [31, 30, 26]. Briefly, heart plane finding or detection is considered to be more difficult to search for because the heart is smaller in size compared to the head or abdomen and requires subtle hand movements to find the heart planes. Therefore on average, we expect that the more experienced a sonographer is (in years), the more likely their live B-mode segments would be predicted as expert when considering the heart plane.

3 Data

The sonographer's eye gaze data was acquired as part of the p**** project which received ethics committee approval. We focus specifically on second-trimester scans which were the most commonly conducted. There are two dataset partitions used. One partition had a teacher (FQ sonographer with 5 years of experience) training 4 different trainees (under independent scan sessions) (Tab. 1). The second partition had 13 FQ sonographers within 0-13 years of scanning experience (Tab. 2).

	Trainee 1	Trainee 2	Trainee 3	Trainee 4	Teacher
Number of sessions	6	6	1	1	14

Table 1: Number of unique teacher-trainee scan sessions.

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

Table 2: Number of unique scan sessions performed by fully qualified sonographers.

Gaze Preprocessing Eye-tracking and pupillary data were collected using a Tobii Eye Tracker 4C which was sampled at 90 Hz. We follow the pupillary data preprocessing method outlined in [26], and the eye-tracking data preprocessing method outlined by [28]. Briefly, we discard any pupil diameters <1.5mm and >9.0mm, and linearly interpolate any missing values. For gaze data, we discard any segments with >210ms of gaze data missing and linearly interpolate any other gaps.

3.1 Training Data

Model Grouping	Expert's Data Abbreviation	Expertise (years)	≈ Class Imbalance Ratio
Teacher VS trainee	Teacher	5	3
	Teacher+$FQ_{2,5}$	2-5	12
FQ sonographers VS trainee	Teacher+$FQ_{0,13}$	0-13	18
	$FQ_{0,13}$	0-13	14
FQ sonographer VS trainee	$FQ_{1,2}$	1-2	23
	$FQ_{2,3}$	2-3	8
	$FQ_{0,3}$	0-3	17
	$FQ_{10,11}$	10-11	3
	$FQ_{14,15}$	14-15	15

Table 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 $FQ_{a,b}$, where FQ stands for fully qualified, and a, b represents the lower and upper bound of number of years of scanning experience.

In our work, we considered different FQ sonographers to represent experts in our training dataset for skill classification. These models were outlined in Section 2.1 as *Teacher VS trainee*, *FQ sonographers VS trainee* and *FQ sonographer VS trainee*. *Teacher VS trainee* models used the same sonographer where they taught (Teacher) and performed scans on their own ($FQ_{2,5}$). Due to data imbalance (Tab. 2), we use sonographers with the most (top 5) gaze data in Tab. 2 to represent our expert population for our *FQ sonographer VS trainee* models.

We set aside 20% of the FQ data in Tab. 2, abbreviated as $FQ_{0,13}$, for training and testing our skill classification model. The remaining 80% is used to predict skill level of FQ sonographers. Any anatomy-specific segments were labelled using optical character recognition.

4 Results

4.1 Skill Classification

Model	LightGBM		XGBoost		CatBoost	
Data	Expert	Trainee	Expert	Trainee	Expert	Trainee
Teacher	**0.94±0.01**	**0.79±0.02**	0.94±0.01	0.78±0.03	0.91±0.00	0.71±0.01
Teacher+$FQ_{2,5}$	0.97±0.00	0.72±0.04	0.97±0.00	0.74±0.02	0.95±0.01	0.62±0.04
Teacher+$FQ_{0,13}$	0.98±0.00	0.60±0.02	0.98±0.00	0.60±0.02	0.95±0.01	0.38±0.03
$FQ_{0,13}$	**0.98±0.00**	**0.70±0.03**	0.98±0.00	0.66±0.04	0.96±0.00	0.50±0.01
$FQ_{1,2}$	0.99±0.00	0.71±0.04	0.99±0.00	0.74±0.01	0.97±0.00	0.58±0.02
$FQ_{2,3}$	0.98±0.00	0.71±0.03	0.97±0.00	0.65±0.03	0.96±0.00	0.54±0.01
$FQ_{0,3}$	0.99±0.00	0.84±0.02	0.99±0.00	0.80±0.02	0.98±0.00	0.68±0.04
$FQ_{10,11}$	0.95±0.00	0.86±0.01	**0.95±0.01**	**0.88±0.02**	0.94±0.01	0.84±0.02
$FQ_{14,15}$	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 4: Average F1 scores using the different training datasets described in Tab. 3. In bold, the best performing model *Teacher VS trainee* model using LightGBM, the best performing model *FQ sonographers VS trainee* model using LightGBM, the best performing *FQ sonographer VS trainee* model using XGBoost.

Table 4 shows the 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 uses an XGBoost architecture and $FQ_{10,11}$ as the expert. It achieves an F1 score of 95% for the expert class and 88% for the trainee class.

Teacher VS trainee: A comparison of Teacher and Teacher+$FQ_{2,5}$ shows a 7% decrease in performance, suggesting that an expert's gaze is affected by the presence of a trainee. It could be that the expert teaches trainees in a specific textbook manner, but uses their own style when scanning individually.

FQ sonographers VS trainee: The best performing model which considers a range of years of experience achieves a 98% and 70% F1 score for expert and trainee classes respectively. This model only included sonographers performing scans individually. By including gaze data where the teacher was actively training a trainee, Teacher+$FQ_{0,13}$, the performance drops by 10%.

FQ sonographer VS trainee: The performance of the trainee class depends on which experts were used in training, with F1 scores between 60% and 84%. When comparing similar years of experience, $FQ_{14,15}$ and $FQ_{10,11}$, $FQ_{0,3}$ and $FQ_{1,2}$, there is a difference of at least 13% (Tab. 4). 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.

4.2 Skill Prediction of Fully Qualified Sonographers

We calculate the proportion of expert segments predicted by the models on 80% of $FQ_{0,13}$ and use bar charts to display the average proportion of segments predicted as trainee and expert. We use the LightGBM model as overall it had the best results (Tab. 4). For brevity, we show the bar charts of the top 3 datasets which returned the best performing models: Teacher, $FQ_{10,11}$, $FQ_{0,3}$.

Figure 2 shows the proportion of expert-trainee labels, where the expert used in the training dataset was the Teacher with 5 years of experience. Note that sonographers with 2 and 5 years of experience have the highest proportion of their segments labelled as expert. This is because the Teacher also performed some scan sessions individually, which was presented in the $FQ_{0,13}$ dataset. In the $FQ_{0,13}$ dataset, the sonographer had 2 years of experience. At the time of teaching, they had 5 years of experience.

Figure 2: Percentage of segments predicted as expert (dark blue) or trainee (light blue) on $FQ_{0,13}$, grouped by the number of years of experience. The expert used for training the model was Teacher.

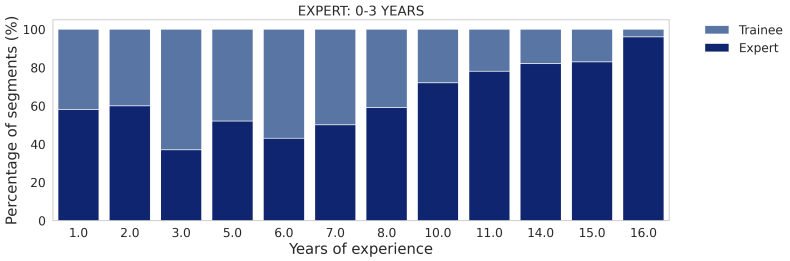

Figure 3: Percentage of segments predicted as expert (dark blue) or trainee (light blue) for $FQ_{0,13}$, grouped by number of years of experience. The expert used for training the model was $FQ_{0,3}$.

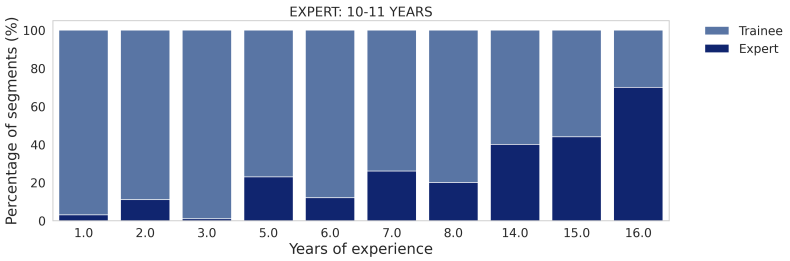

Figure 4: Percentage of segments predicted as expert (dark blue) or trainee (light blue) for $FQ_{0,13}$, grouped by number of years of experience. The expert used for training the model was $FQ_{10,11}$.

	Teacher	Teacher+$FQ_{0,13}$	Teacher+$FQ_{2,5}$	$FQ_{0,13}$
p-value	0.82	0.41	0.95	0.44
PC	-0.07	-0.26	-0.02	0.25

	$FQ_{1,2}$	$FQ_{2,3}$	$FQ_{0,3}$	$FQ_{10,11}$	$FQ_{14,15}$
p-value	0.15	0.46	0.00	0.00	0.0
PC	-0.47	-0.24	0.84	0.92	0.80

Table 5: Table of Pearson's coefficient (PC) and p-values between years of experience and percentage of expert segments predicted.

Both Fig. 3 and 4 show that the number of years of scanning is positively correlated with the proportion of segments being labelled as experts. The PC is significant (p-value < 0.05). This is unlike Fig. 2 where the coefficient is -0.07 and is not significant.

4.3 Anatomy Specific Skill

We also investigate how well our task-agnostic gaze skill classification model performs when considering specific anatomical planes. The anatomical planes that were considered are head circumference

	Teacher+$FQ_{2,3}$	$FQ_{1,2}$	$FQ_{0,3}$	$FQ_{10,11}$	$FQ_{14,15}$
AC	0.58			0.82	
HC	0.62			0.78	
Heart		-0.61	0.74	0.86	0.58

Table 6: Table of Pearson's coefficient (PC) between years of experience and percentage of expert segments predicted for head circumference (HC), abdominal circumference (AC), and heart. We only show the correlation coefficients which were found to be significant at the 5% level. Teacher, Teacher+$FQ_{0,13}$, $FQ_{0,13}$ and $FQ_{2,3}$ did not have any significant coefficients. Empty entries correspond to p-value > 0.05. Entries with a coefficient value correspond to p-values < 0.05.

(HC), abdominal circumference (AC), and heart. The anatomy-specific PC results are shown in Tab. 6 which suggests that there is little significant correlation for anatomical planes AC and HC (2 out of 9 models show a significant correlation). Conversely, there is some significant correlation for heart planes (4 out of 9). These results suggest that there is significant difference in gaze behaviour between FQ sonographers when searching for the heart, but not for the HC and AC. The results of [27, 26] show that the heart is more difficult to search for because of its relatively smaller size in comparison to the abdomen and brain. Therefore, it is not surprising that the number of years of experience in scanning is positively correlated with the proportion of expert segments for the heart (Figs. 5 and 6).

Furthermore, we show the bar charts for $FQ_{0,3}$ and $FQ_{10,11}$. Visually, they confirm the results of the significance test, where there is a noticeable increase in predicted expert segments with years of experience for $FQ_{10,11}$ compared to $FQ_{0,3}$ for AC and HC.

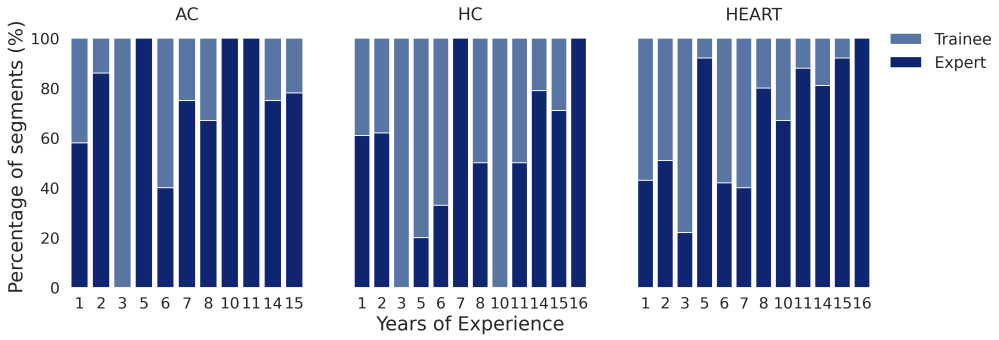

Figure 5: Percentage of segments predicted as expert (dark blue) or trainee (light blue) for anatomical planes HC, AC and Heart in $FQ_{0,13}$, grouped by number of years of experience. The expert used for training the model was $FQ_{0,3}$.

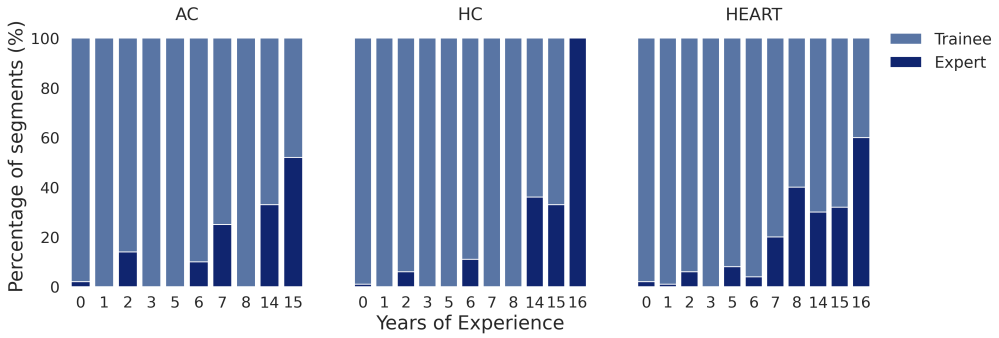

Figure 6: Percentage of segments predicted as expert (dark blue) or trainee (light blue) for anatomical planes HC, AC and Heart in $FQ_{0,13}$, grouped by number of years of experience. The expert used for training the model was $FQ_{10,11}$.

A comparison between Figs. 5 and 6 also show that there is a larger proportion of predicted expert segments when being compared against FQ sonographer with less scanning experience, $FQ_{0,3}$ (0-3 years), than that of $FQ_{10,11}$ (10-11 years). The same behaviour can also be observed in Fig. 3 and 4. These results suggest that it is 'more difficult' to have gaze patterns of a sonographer who has been scanning for several years. It is likely that sonographers developed their own style over time, possibly moving away from their scanning style from earlier years when they first qualified.

5 Discussion

In our work, we have shown that the performance of a gaze skill classification model is dependent on the sonographer representing the expert population. We then used the model to predict whether a FQ sonographer's years of experience is positively correlated to the proportion of expert segments predicted. The Pearson's correlation coefficient test showed that when using $FQ_{0,3}$, $FQ_{10,11}$, and $FQ_{14,15}$ as the expert benchmark, there is a significant positive correlation between the number of scanning years and the percentage of expert segments. These results suggest that, without making any prior assumptions about the relationship between scanning years and expertise, there is a positive correlation between the 2 variables. With more years of scanning experience, a FQ sonographer is likely to have a higher proportion of predicted expert segments. This relationship is also seen when sonographers are searching for heart planes.

The trainee class was highly imbalanced in some of the training data, such as $FQ_{0,3}$ and $FQ_{1,2}$. A comparison of $FQ_{0,3}$ and $FQ_{10,11}$, which had an imbalance ratio of 17 and 3 respectively, returned similar results for the best-performing model. When comparing $FQ_{0,3}$ and $FQ_{1,2}$, both had between 0-3 years of experience but a 13% difference in performance for the trainee class. Similarly, $FQ_{10,11}$ and $FQ_{14,15}$ 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.

6 Conclusion

In this paper, we have presented a skill classification model, 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. Our 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. We have also showed that sonographer gaze behaviour is indicative of both skill and style, with performance differences of up to 16%. Finally, without making any prior assumptions of the correlation between years of experience as a direct measure of skill, we show that there is a significant correlation between years of scanning and expertise when considering task-agnostic gaze characteristics and task-specific planes such as the heart.

References

[1] N. Ahmidi, G. D. Hager, L. Ishii, G. Fichtinger, G. L. Gallia, and M. Ishii. Surgical Task and Skill Classification from Eye Tracking and Tool Motion in Minimally Invasive Surgery. In *Lecture Notes in Computer Science*, volume 6363 LNCS, pages 295–302. 2010. doi: 10.1007/978-3-642-15711-0{_}37. URL `http://link.springer.com/10.1007/978-3-642-15711-0_37`.

[2] N. Ahmidi, M. Ishii, G. Fichtinger, G. L. Gallia, and G. D. Hager. An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data. *International Forum of Allergy & Rhinology*, 2(6):507–515, 11 2012. ISSN 2042-6984. doi: 10.1002/alr.21053.

[3] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967, 3 2021. ISSN 0269-2821. doi: 10.1007/s10462-020-09896-5. URL `https://link.springer.com/10.1007/s10462-020-09896-5`.

[4] N. Castner, E. Kasneci, T. Kübler, K. Scheiter, J. Richter, T. Eder, F. Hüttig, and C. Keutel. Scanpath comparison in medical image reading skills of dental students. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, New York, NY, USA, 6 2018. ACM. ISBN 9781450357067. doi: 10.1145/3204493.3204550.

[5] N. Castner, J. Frankemölle, C. Keutel, F. Huettig, and E. Kasneci. LSTMs can distinguish dental expert saccade behavior with high "plaque-urracy". In *2022 Symposium on Eye Tracking Research and Applications*, pages 1–7, New York, NY, USA, 6 2022. ACM. ISBN 9781450392525. doi: 10.1145/3517031.3529631.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 6 2002. ISSN 1076-9757. doi: 10.1613/jair.953.

[7] T. Chen and C. Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA, 8 2016. ACM. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL `https://dl.acm.org/doi/10.1145/2939672.2939785`.

[8] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*, 307:72–77, 9 2018. ISSN 09252312. doi: 10.1016/j.neucom.2018.03.067.

[9] A. V. Dorogush, V. Ershov, and A. Gulin. CatBoost: gradient boosting with categorical features support. 10 2018. URL `http://arxiv.org/abs/1810.11363`.

[10] L. Drukker, H. Sharma, R. Droste, M. Alsharid, P. Chatelain, J. A. Noble, and A. T. Papageorghiou. Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. *Scientific Reports*, 11(1):14109, 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-92829-1. URL `https://doi.org/10.1038/s41598-021-92829-1`.

[11] S. Erridge, H. Ashraf, S. Purkayastha, A. Darzi, and M. H. Sodergren. Comparison of gaze behaviour of trainee and experienced surgeons during laparoscopic gastric bypass. *British Journal of Surgery*, 105(3):287–294, 2 2018. ISSN 0007-1323. doi: 10.1002/bjs.10672.

[12] E. Fichtel, N. Lau, J. Park, S. Henrickson Parker, S. Ponnala, S. Fitzgibbons, and S. D. Safford. Eye tracking in surgical education: gaze-based dynamic area of interest can discriminate adverse events and expertise. *Surgical Endoscopy*, 33(7):2249–2256, 2019. ISSN 14322218. doi: 10.1007/s00464-018-6513-5. URL `http://dx.doi.org/10.1007/s00464-018-6513-5`.

[13] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 10 2001. ISSN 0090-5364. doi: 10.1214/aos/1013203451.

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf`.

[15] E. A. Krupinski, J. Chao, R. Hofmann-Wellenhof, L. Morrison, and C. Curiel-Lewandrowski. Understanding visual search patterns of dermatologists assessing pigmented skin lesions before and after online training. *Journal of Digital Imaging*, 27(6):779–785, 12 2014. ISSN 1618-727X. doi: 10.1007/s10278-014-9712-1.

[16] B. Law, M. S. Atkins, A. E. Kirkpatrick, and A. J. Lomax. 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*, pages 41–48, New York, New York, USA, 2004. ACM Press. ISBN 1581138253. doi: 10.1145/968363.968370.

[17] M. Le Lous, F. Despinoy, M. Klein, E. Fustec, V. Lavoue, and P. Jannin. Impact of Physician Expertise on Probe Trajectory During Obstetric Ultrasound: A Quantitative Approach for Skill Assessment. *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`.

[18] W. F. Lee and J. Chenkin. Exploring Eye-tracking Technology as an Assessment Tool for Point-of-care Ultrasound Training. *AEM Education and Training*, 5(2), 4 2021. ISSN 24725390. doi: 10.1002/aet2.10508.

[19] H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. *Computer Aided Surgery*, 11(5):220–230, 1 2006. ISSN 1092-9088. doi: 10.3109/10929080600989189.

[20] D. Manning, S. Ethell, T. Donovan, and T. Crawford. How do radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography*, 12(2):134–142, 5 2006. ISSN 1078-8174. doi: 10.1016/j.radi.2005.02.003. URL `http://www.sciencedirect.com/science/article/pii/S1078817405000131`.

[21] G. Megali, S. Sinigaglia, O. Tonet, and P. Dario. Modelling and Evaluation of Surgical Performance Using Hidden Markov Models. *IEEE Transactions on Biomedical Engineering*, 53(10):1911–1919, 10 2006. ISSN 0018-9294. doi: 10.1109/TBME.2006.881784.

[22] J. F. Ortega-Morán, J. B. Pagador, V. Luis-del Campo, J. C. Gómez-Blanco, and F. M. Sánchez-Margallo. Using Eye Tracking to Analyze Surgeons' Cognitive Workload During an Advanced Laparoscopic Procedure. pages 3–12. 2020. doi: 10.1007/978-3-030-31635-8{_}1. URL `http://link.springer.com/10.1007/978-3-030-31635-8_1`.

[23] C. E. Reiley and G. D. Hager. Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5761 LNCS(PART 1):435–442, 2009. ISSN 03029743. doi: 10.1007/978-3-642-04268-3{_}54.

[24] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the symposium on Eye tracking research & applications - ETRA '00*, pages 71–78, New York, New York, USA, 2000. ACM Press. ISBN 1581132808. doi: 10.1145/355017.355028. URL `http://portal.acm.org/citation.cfm?doid=355017.355028`.

[25] H. Sharma, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Multi-modal learning from video, eye tracking, and pupillometry for operator skill characterization in clinical fetal ultrasound. *Proceedings - International Symposium on Biomedical Imaging*, 2021-April:1646–1649, 2021. ISSN 19458452. doi: 10.1109/ISBI48211.2021.9433863.

[26] H. Sharma, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Machine learning-based analysis of operator pupillary response to assess cognitive workload in clinical ultrasound imaging. *Computers in Biology and Medicine*, 135:104589, 2021. ISSN 18790534. doi:

10.1016/j.compbiomed.2021.104589. URL https://doi.org/10.1016/j.compbiomed.2021.104589.

[27] H. Sharma, L. Drukker, A. T. Papageorghiou, and 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)*, pages 1646–1649. IEEE, 4 2021. ISBN 978-1-6654-1246-9. doi: 10.1109/ISBI48211.2021.9433863. URL https://ieeexplore.ieee.org/document/9433863/.

[28] C. Teng, H. Sharma, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Towards Scale and Position Invariant Task Classification Using Normalised Visual Scanpaths in Clinical Fetal Ultrasound. In J. A. Noble, S. Aylward, A. Grimwood, Z. Min, S.-L. Lee, and Y. Hu, editors, *Simplifying Medical Ultrasound*, pages 129–138, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87583-1.

[29] D. Topalli and N. E. Cagiltay. Eye-Hand Coordination Patterns of Intermediate and Novice Surgeons in a Simulation-Based Endoscopic Surgery Training Environment. *Journal of Eye Movement Research*, 11(6):1–14, 2018. ISSN 19958692. doi: 10.16910/JEMR.11.6.1.

[30] Y. Wang, R. Droste, J. Jiao, H. Sharma, L. Drukker, A. T. Papageorghiou, and J. A. Noble. *Differentiating Operator Skill During Routine Fetal Ultrasound Scanning Using Probe Motion Tracking*, volume 1. Springer International Publishing, 2020. ISBN 9783030603342. doi: 10.1007/978-3-030-60334-2{_}18. URL http://dx.doi.org/10.1007/978-3-030-60334-2_18.

[31] Y. Wang, Q. Yang, L. Drukker, A. Papageorghiou, Y. Hu, and J. A. Noble. Task model-specific operator skill assessment in routine fetal ultrasound scanning. *International Journal of Computer Assisted Radiology and Surgery*, 17(8):1437–1444, 8 2022. ISSN 1861-6429. doi: 10.1007/s11548-022-02642-y.