AUTOMATED DESCRIPTION AND WORKFLOW ANALYSIS OF FETAL ECHOCARDIOGRAPHY IN FIRST-TRIMESTER ULTRASOUND VIDEO SCANS

Robail Yasrab[1] *Mohammad Alsharid*[1,3] *Md. Mostafa Kamal Sarker*[1] *He Zhao* [1]
Aris T. Papageorghiou[2] *J. Alison Noble*[1]

[1]Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[2]Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK
[3]Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, UAE

ABSTRACT

This paper presents a novel, fully-automatic framework for fetal echocardiography analysis of full-length routine first-trimester fetal ultrasound scan video. In this study, a new deep learning architecture, which considers spatio-temporal information and spatial attention, is designed to temporally partition ultrasound video into semantically meaningful segments. The resulting automated semantic annotation is used to analyse cardiac examination workflow. The proposed 2D+t convolution neural network architecture achieves an A1 accuracy of 96.37%, $F1$ of 95.61%, and precision of 96.18% with 21.49% fewer parameters than the smallest ResNet-based architecture. Automated deep-learning based semantic annotation of unlabelled video scans (n=250) shows a high correlation with expert cardiac annotations ($\rho = 0.96, p = 0.0004$), thereby demonstrating the applicability of the proposed annotation model for echocardiography workflow analysis.

Index Terms— first trimester, spatio-temporal analysis, fetal heart, ultrasound, echocardiography.

1. INTRODUCTION

Detection of cardiac abnormalities during prenatal screening is a challenging task. In accordance with NHS UK guidelines, cardiac anatomy should be evaluated in the second trimester of pregnancy. However, technological advancements have sparked a recent interest in earlier detection of congenital cardiac abnormalities [1]. The first-trimester fetal US scan (also known as the dating or nuchal scan) is carried out between 11^{+0} to 13^{+6} $weeks^{days}$ of gestation to assure pregnancy viability, accurately date the pregnancy, and assess chromosomal anomalies risk [2]. The NHS Fetal Anomaly Screening Programme (FASP) [3] guidelines dictate the acquisition of two standard planes: Nuchal Translucency (NT) and Crown Rump Length (CRL). In addition, sonographers may scan additional fetal structures, depending on their training or personal preferences. The degree of routine first-trimester screening for cardiac anomalies varies between operators and centres and may involve any of the following: assessment without cardiac examination; routine visualization of the four chambers (4C); detailed examination involving outflow-tract visualization and Doppler evaluation; or demonstrating a heart beat (spectral Doppler (SD)). The early detection of cardiac abnormalities in the first trimester complements the overarching objective of earlier diagnosis of chromosomal abnormalities.

This study aimed to investigate fetal echocardiography workflow from automated analysis of full-length first-trimester US scan videos by identifying the duration and sequence of standard imaging plane acquisition. The prerequisite for such analysis is the semantic temporal partitioning and anatomical level labelling of full-length scans. It is infeasible to carry out manual labelling of very large video datasets. Therefore, we have developed a time-efficient and accurate deep learning (DL) based model to automatically label first-trimester full-length video scans with consideration of its spatio-temporal context.

Contribution. The contribution of the paper is two-fold: 1) An original spatio-temporal DL architecture is trained to provide semantic labels for first trimester US video. We experimented with various deep neural networks to determine the best spatio-temporal (2D+t) model. We also investigated the effect of introducing spatial attention during training video frames. The proposed model with the highest performance, was assessed for similarity with expert labelled video scans. 2) To demonstrate clinical applicability, the trained model is used to semantically partition unlabelled full-length first-trimester US video scans and assess the proportion of time spent on performing fetal echocardiography tasks. For the first time, the proposed framework will describe the preferred approach/scanning-mode of sonographers to assess abnormalities. Furthermore, we will analyze the trends in echocardiography between newly qualified (NQ) and experienced sonographers (EX).

Related Work. There have been a number of approaches to automate second and third trimester US scan tasks such

Clinical data acquisition was approved by the Research Ethics Committee (Reference 18/WS/0051).

Table 1: Dataset distribution for deep learning experiments.

Dataset	Frames	Distribution(%)
Training	20,520	77.9
Test	4,970	22.1

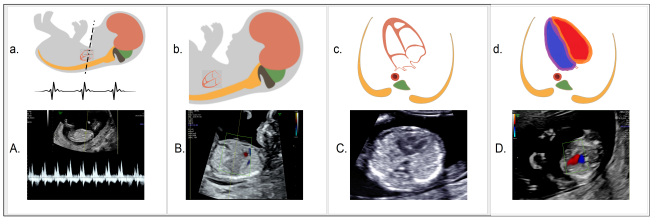

Fig. 1: First-trimester fetal echocardiography: A, a: Spectral Doppler. B, b: Sagittal cardiac view. C, c: Four-chamber view and D, d: Doppler evaluation. Lowercase letter represents a graphical abstraction, and the uppercase letter represents the equivalent ultrasound image.

as standard plane detection [4], segmentation [5], and fetal biometry [6]. Regarding ultrasound workflow analysis, Sharma et al. [7] proposed a spatio-temporal VGG variant CNN model for second trimester US partitioning and description. We include that architecture in our comparison. However, few works [8, 9] have studied the first-trimester fetal heart and none have considered echocardiography workflow analysis. According to Karim et *al.* [1], a detailed sonographic examination of the fetal heart in the first trimester can optimize the detection of fetal cardiac anomalies. According to the study by Bardi et *al.* [10], 91% of cardiac defects can be diagnosed prenatally, and 9% are detected in the first trimester. This finding is not surprising given that first-trimester anatomical screening programs do not include cardiac screening. It is observed that cardiac defects are more subtle and only identifiable by the trained eye when scanning the heart in a systematic way and by use of Doppler flow [1]. Thus, the proposed study aims to conduct a machine learning-based analysis of routine ultrasound scans performed during the first trimester in order to determine the time and method employed for echocardiography in practice.

2. DATA ACQUISITION AND PRE-PROCESSING

Routine clinical first-trimester fetal US scans were available from a large-scale study PULSE [11]. US video was acquired through screen-grab signals at 30 frames per second of a GE Voluson E8 version BT18 (GE Healthcare, Zipf, Austria) US machine. The echocardiography US frames were extracted from full-length first trimester scans of 250 subjects and manually annotated by experienced clinical and engineering researchers. Four anatomical categories ["classes"] were defined: 'sagittal cardiac view evaluation' [SV] (17.9%), 'Four-Chamber view' [4C] (26.4%), 'Doppler evaluation' [DP] (19.6%), 'Spectral Doppler (heartbeat)' [SD] (19.8%) and Other anatomy (CRL, NT, Brain) [OT] (16.3 %). For

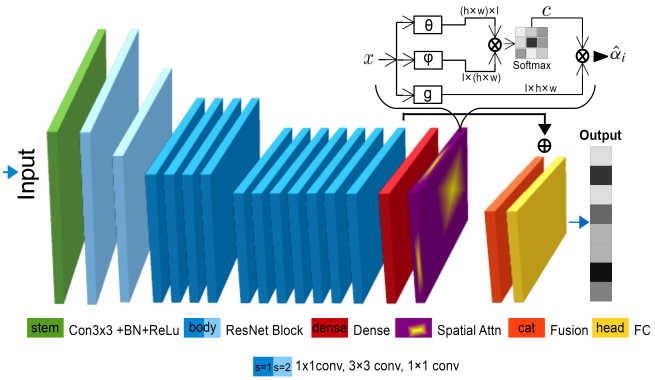

Fig. 2: Proposed spatial attention-based CNN architecture.

the spatial CNN training, images were sampled every eighth frame of a video to incorporate a wide variety of anatomical views and spatial diversity to concurrent frames. Table. 1 summarizes the dataset.

3. METHODS

We experimented with different CNN models for US video partitioning. We investigated;

1. Spatial modelling (2D CNN),
2. Spatio-temporal modelling (Recurrent Neural Networks (RNN/LSTM)) (2D+t).

For 2D spatial modelling, we trained and tested VGG [12], ResNet [13] and RegNet[14] architectures due to their established high benchmark classification performance on public datasets. The quantitative results in Table 2 show RegNet2D consistently outperforms the other CNN benchmarks (precision score=0.91) with a significant reduction in trainable parameter overhead (2.32 M). Hence, it was selected as the 2D backbone for subsequent spatio-temporal analysis. RegNet2D is designed through an optimized neural architecture search, resulting in a low-dimensional design space that leads to a simple and efficient network architecture. The detailed network architecture is shown in Figure 2. To incorporate temporal information, long-short term memory (LSTM) architecture was considered. We also considered both 2D (RegNet2Dt) and 3D (RegNet3D) convolution kernels for training a 2D+t architecture. We also considered adding an attention module for one model variant (RegNet2Dt-At) to encourage that model to focus on salient anatomical features. Specifically, we adopted a spatial attention block (SAB) [15] that generates a spatial attention map by utilizing the inter-spatial relationship of features. The detailed structure of SAB is shown in Figure 2, where $x \in R^{C\times H\times W}$ in an input feature map sent to a convolutional layer to produce $\{\theta, \varphi\} \in R^{1\times h\times w}$. The input feature maps are reshaped to $R^{1\times N}$, where $N = h \times w$ is the number of pixels. The transposes of θ and φ are multiplied at the Softmax layer to

Table 2: Quantitative Analysis of Proposed Network.

	Network	P	R	F1	A1	Pr	GFlops
Spatial	VGG-16 [12]	0.78	0.73	0.69	68.63±0.01	134.7m	15.55
	ResNet-18 [13]	0.82	0.80	0.81	90.10±0.03	11.18m	1.83
	ResNet-50 [13]	0.89	0.86	0.87	93.09±0.11	23.90	4.14
	RegNet2D [14]	0.91	0.90	0.89	93.76±0.08	2.32m	0.20
	RegNet2D-At	0.94	0.92	0.92	94.59±0.18	2.45m	0.37
Spt. T.	RegNet-LSTM	0.81	0.74	0.72	81.22±0.02	4.78m	0.58
	RegNet-3D	0.95	0.94	0.94	95.01±0.03	3.56m	0.54
	Sharma et al. [7]	0.96	0.94	0.95	96.11±0.01	23.0m	15.36
	RegNet2Dt	0.95	0.93	0.92	95.41±0.08	4.78m	0.65
	RegNet2Dt-At	0.96	0.95	0.95	96.37±0.03	4.91m	0.76

compute the intermediate spatial-attention-map $c \in R^{N \times N}$;

$$c_{ij} = \frac{\exp(\theta_i \cdot \varphi_j)}{\sum_{i=1}^{N} \exp(\theta_i \cdot \varphi_j)}, \tag{1}$$

where c_{ij} calculates the impact of the ith position on the jth position. Correlation is substantial if both positions have a similar surrounding texture. Matrix multiplication between c and the transpose of g results in the final spatial attention map:

$$\hat{\alpha} = \sum_{j=1}^{N} (c_{ij} g_j). \tag{2}$$

Implementation Details: The CNN architectures were implemented using PyTorch v1.8.0. US video frames were scaled to 224×224 pixels. Standard data augmentation was used (rotation $[-30°, 30°]$, horizontal flip, Gaussian noise, and shear (≤ 0.2)). Images were normalised to zero-mean and unit variance. Batch size was adjusted according to model size and GPU memory restrictions. All CNN models were trained using a cross-entropy loss function for 200 epochs, constantly reducing the learning rate ($\times 0.1$ every 20 epochs).

4. RESULTS AND DISCUSSION

4.1. Quantitative Evaluation of Trained Models

Recall (R), Precision (P), F1-score (F1), and Top-1 accuracy (A1) were used to assess the performance of the classification models. Table 2 reports the quantitative performance and the number of trainable parameters for each model. We observe that RegNet2D outperforms the other 2D vanilla CNNs. Adding an attention model to RegNet gives further improvement and the best 2D result (F1-score (3.0%) and A1(0.83%)). Spatio-temporal models that use video frame sequences, such as 2D+t, perform better than the LSTMs and 3D-Conv (2D+t) representations. The best performing model, for all evaluation metrics, is RegNet2Dt which describes the spatio-temporal properties of video clips trained using random initialization weights. The Sharma et *al.* [7] 2D+t model offers competitive results with the pre-trained weight model but has 23.0 million trainable parameters, making it a slow inference model for large scale US video datasets.

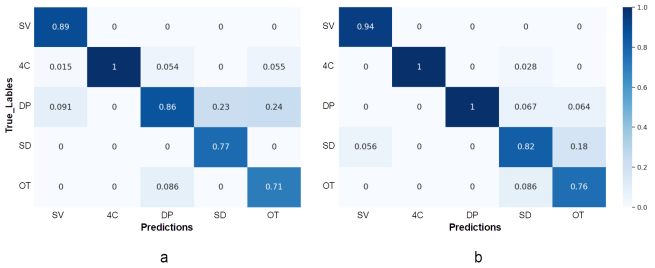

Fig. 3: Confusion matrix for automated semantic annotations vs manual annotations (a) RegNet2Dt, (b) RegNet2Dt-At.

Table 3: Dataset distribution for deep learning experiments.

Cardiac Views	Total Time Spent(%)
Four-Chamber View [4C]	52.75
Spectral Doppler (heartbeat) [SD]	33.34
Doppler Evaluation [DP]	13.76
Sagittal Cardiac View Evaluation [SV]	0.15

Automated annotation of the test set revealed a high correlation ($\rho = 0.96, p = 0.0004$) with manual expert annotations. The confusion matrix in Figure 3 shows an excellent agreement with manual and automatic semantic labelling. Note that even classes with few samples (e.g. SV) are detected with high accuracy.

4.2. Clinical Workflow Analysis

Despite the advancements in fetal echocardiography, the effectiveness of first-trimester fetal heart screening has not been adequately investigated. This study investigated the operator's echocardiography workflow analysis for routine first-trimester US scans. We have investigated different visualization methods used by sonographers to visualize fetal cardiac structures. The RegNet2Dt-At model was applied to label 250 unseen full-length US video scans. This shows which cardiac views the operator was viewing as a time-line and thus provides a description of clinical work flow. The subjects analyzed in study had an average maternal age ± SD = 31.6 ± 5.4 years and average body mass index (BMI) 25.3 ± 5.8.

On average, a recorded first-trimester US video scan takes 15.7 ± 4.2 minutes, with an average of $28,237 \pm 7,534$ frames per video scan. In order to assess only cardiac workflow, we have excluded other anatomical views from the analysis. The total amount of time spent on echocardiography during the first trimester is 18.91 percent (2.97 minutes) of total scan time. Table 3 illustrates the distribution of time between different views during echocardiography. According to our findings, 4-chamber heart views are the most commonly observed views (52.75 %) for fetal heart assessment. In addition, we observed that spectral Doppler is also an important method for observing the heartbeat of the fetus. The reason is that sonographers typically look for a 4C view of the fetal heart to observe a beating heart to ensure the fetus is alive. Spectral Doppler helps sonographers to check fetal heartbeat (regular

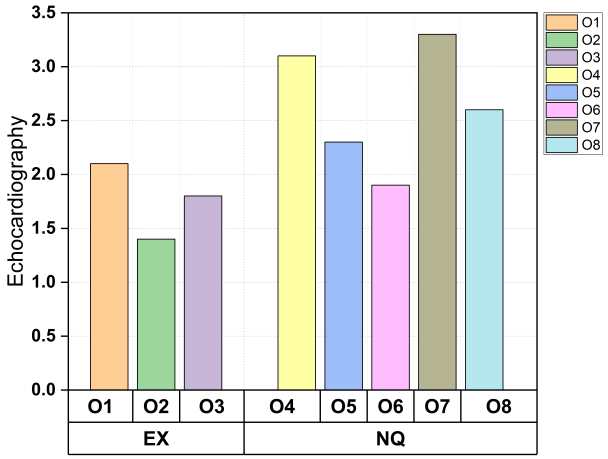

Fig. 4: Average time duration spent on the key cardiac views for EX and NQ groups. This shows that the NQ group spends more time locating cardiac planes. Conversely, the EX group spends less time.

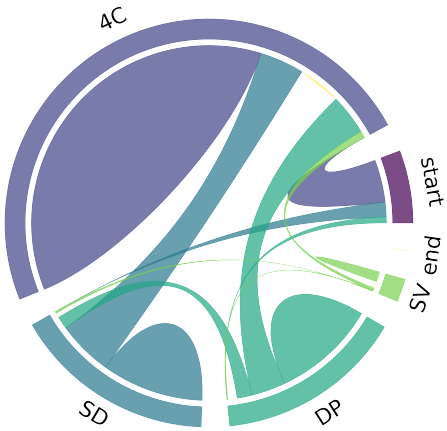

Fig. 5: A visualisation derived from a task transition matrix for operators US workflow.

heartbeat between 110 and 160 beats per minute). The use of Doppler evaluation (DP) is the least significant (counts for only 13.76 %) because in the early weeks of first-trimester pregnancy sometimes the heart is not fully developed, therefore Doppler based evaluation is not possible. There were 9 sonographers involved in the study, however eight (n= O1, O2...O8) of those looked at the fetal heart. These operators were divided into two groups based on experience (2 years); expert (EX) and newly qualified (NQ). Figure 4 shows that NQ operators take more time to scan the fetal heart compared to EX operators.

Table 4: The Apriori algorithm used to extract the most frequent anatomical workflows adopted by operators.

No.	Antecedents	Consequents	Antecedent support (%)
1	(4C, SD)	(DP)	93.91
2	(4C, DP)	(SD)	87.14
3	(4C, SD)	(SV)	41.29

Task-occurrence probability is calculated using the Apriori algorithm [16] to extract the most frequent anatomical task-occurrence probability for each scan. The Apriori is a fitting choice to extract the most frequent anatomical task-occurrence probability of each operator's (O) task $I = \{i_1, i_2, \ldots, i_n\}$ and task transition matrix $T = \{t_1, t_2, \ldots, t_m\}$ named as database of anatomical transactions. Here, each transaction t_x in T has a unique transaction-ID with the subset of item sets in I. The Apriori rule for any two anatomical activities (X, Y) stated as $X \Rightarrow Y$, where, X is 'Antecedent' and Y is 'Consequent' such as $X, Y \subseteq I$. Table 4 shows a different combination of operators' preferences for starting the first-trimester scan. The task-occurrence probability from 4C-SD or 4C-DP has the highest confidence (Table 4). It appears that in most scans, 4C and SD appear sequentially; this can be explained by the fact that SD is usually observed in the same plane as 4C. The chord diagram in Figure 5 illustrates the resultant workflow association of most frequent fetal cardiac analysis activities. Each view is represented by a fragment on the outer part of the circular layout. The arcs drawn between these anatomical structures show the workflow patterns of operators. The thickness of these arcs is proportional to the frequency of transition between a certain view or task to another that the operator follows during US scanning. The thicker the arc, the more common and 'traditional' this workflow pattern is. The purpose of this study was to observe routine first-trimester ultrasound examination from a cardiac perspective. The results demonstrate that despite the fact that observing the fetal heart is not part of the FASP [3] protocol for trimester scans, sonographers nonetheless do so in practice. This observation also implies that NHS-trained sonographers tend to look into the fetal heart due to their training history or personal preferences.

5. CONCLUSION

This paper shows that spatio-temporal modelling with attention-based feature refinement gave the best performing model for automatically labelling full-length routine first-trimester fetal US scan videos. The best performing model was applied to a large-scale first-trimester clinical video dataset to provide quantitative insight on echocardiography work flow. First-trimester fetal echocardiography is a very useful method for the detection of fetal cardiac abnormalities. Although second-trimester cardiac scanning remains the gold standard for anomaly detection, we have observed that most of operators scanned cardiac structures during the first trimester as well. Additionally, this study indicates that 4C and SD are the most preferred fetal heart assessment tasks performed by sonographers during the first trimester scan. In the future, we intend to use insights from this study to develop a standard echocardiography assessment protocol for the early detection

of fetal heart anomalies.

6. ACKNOWLEDGMENTS

This work is supported by the ERC (ERC-ADG-2015694581, project PULSE), EPSRC (EP/R013853/1 and EP/T028572/1) and the NIHR Oxford Biomedical Research Centre.

7. REFERENCES

[1] JN Karim et al., "First-trimester ultrasound detection of fetal heart anomalies: systematic review and meta-analysis," *UoG*, vol. 59, no. 1, pp. 11–25, 2022.

[2] LJ Salomon et al., "Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan," *UOG*, vol. 37, no. 1, pp. 116–126, 2011.

[3] "Fetal anomaly screen programme handbook," Report, NHS Screening Programmes, London, UK, 2015.

[4] H Chen et al., "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE JBHI*, vol. 19, no. 5, pp. 1627–1636, 2015.

[5] S Rueda et al., "Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge," *IEEE TMI*, vol. 33, no. 4, pp. 797–813, 2013.

[6] LH Lee et al., "Calibrated bayesian neural networks to estimate gestational age and its uncertainty on fetal brain ultrasound images," in *Proc. ASMUS*, pp. 13–22. Springer, 2020.

[7] H Sharma et al., "Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans," in *Proc. ISBI*. IEEE, 2019, pp. 987–990.

[8] R Stoean et al., "Deep learning for the detection of frames of interest in fetal heart assessment from first trimester ultrasound," in *Proc. IWANN*. Springer, 2021, pp. 3–14.

[9] MN Rachmatullah et al., "Convolutional neural network for semantic segmentation of fetal echocardiography based on four-chamber view," *BEEI*, vol. 10, no. 4, pp. 1987–1996, 2021.

[10] F Bardi et al., "Prenatal diagnosis and pregnancy outcome of major structural anomalies detectable in the first trimester: A population-based cohort study in the netherlands," *Paediatric and Perinatal Epidemiology*, 2022.

[11] L Drukker et al., "Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.

[12] K Simonyan and A Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] K He et al., "Deep residual learning for image recognition," in *Proc. CCVPR*, 2016, pp. 770–778.

[14] I Radosavovic et al., "Designing network design spaces," in *Proc. CVF*, 2020, pp. 10428–10436.

[15] W Shi et al., "(sarn) spatial-wise attention residual network for image super-resolution," *The Visual Computer*, vol. 37, no. 6, pp. 1569–1580, 2021.

[16] R Agrawal et al., "Fast algorithms for mining association rules," in *VLDB*. Citeseer, 1994, vol. 1215, pp. 487–499.