advances in Technology

Adv. Technol. 2024, 4(2), 146-164

Full Paper

Cost-Efficient Light-Weight YOLO V5_s for Whole Fetus Detection in Early-Stage Ultrasound Scans

U.B. Balagalla*[,a], O.V. Jayawardane[a], J.V.D. Jayasooriya[b], C. De Alwis [c], and A. Subasinghe[a]

[a]Department of Electrical and Electronic Engineering, University of Sri Jayewardenepura, Sri Lanka
[b]Essential Energy, Sydney, New South Wales, Australia
[c]School of Computer Science and Technology, University of Bedfordshire, United Kingdom

Corresponding Author: umayabalagalla@sjp.ac.lk

Received: 22 March 2024; Revised: 24 August 2024; Accepted: 25 August 2024; Published: 26 February 2025

Abstract

Routine medical Ultrasound (US) scans are recommended for expectant mothers to monitor the health and growth of the fetus. However, expectant mothers in rural areas of developing and underdeveloped countries face difficulties in receiving timely scans due to a lack of expertise and facilities. Consequently, maternal and fetal deaths occur at higher rates in these countries, especially in the first trimester. Novel concepts such as Virtual Doctors, Hospital to Home, self-scans and the Internet of Medical Things (IoMT) may address the aforementioned problem effectively. Therefore, computational-efficient algorithms which support low-end smart devices should be introduced to assist expectant mothers in rural areas to provide comfortable and timely fetal scans. In light of the above, this paper discusses computationally efficient YOLO V5_s for fetal detection in first-trimester ultrasound images using a highly diverse dataset including abnormal and multiple pregnancies. The implemented model was compared with five benchmark detection models, namely, ResNet-50 and MobileNet-based faster R-CNN, YOLO-n, YOLO-m and YOLO-l. YOLO was comparatively better than faster R-CNN. Even though YOLO-n is the most computationally inexpensive model, its mAP is 0.709, which is comparatively low, hence cannot be applied to the clinical set-up. YOLO-l model has the best performance with F-1 score and mAP of 0.978 and 0.751, respectively. However, YOLO-s has also achieved a F-1 score of 0.979 with a mAP 0.734. Therefore, a subjective test was conducted to verify using YOLO-s in the clinical set-up with five experts in the field with more than five years of experience. The subjective analysis test, assessed through Fleiss Kappa, suggests substantial agreement beyond chance ($\kappa = 0.69$), while the Intraclass Correlation Coefficient (ICC) indicates modest reliability (ICC = 0.7). The findings endorse the application of YOLO-s for real-time detection of whole fetuses in the first trimester with reduced computational complexity with further validation.

Keywords: Faster R-CNN, fetal ultrasound, first trimester, remote monitoring, whole fetus, YOLO

Introduction

Routine fetal scans of the first trimester can be performed in the first three months of the pregnancy, starting from post-conception week 6 to week 14. Receiving scans in the first trimester is important to assess the viability of the pregnancy, diagnose potential risks of fetal anomalies and abnormalities, and monitor the development of the fetus [1]. Medical imaging modalities such as Ultrasound (US), Magnetic Resonance Imaging (MRI), X-ray, Computed Tomography (CT), Positron Emission Tomography (PET), and Endoscopy are one of the key diagnostic tools used in a large clinical scope, including oncology, gynaecology and osteology which are used to visualize the interior structures of the human body [2].

Among these modalities, medical US is highly recommended in obstetrics to provide comfortable and risk-free scans to expectant mothers as it is one of the safest imaging modalities due to its non-invasive and non-ionizing nature. Medical US images are generated using the time of flight of the echo pulses. Hence, it is the safest medical imaging modality to date. Further, it provides benefits such as a real-time monitoring facility, lower cost, and the possibility to conduct without a pre-preparation [3-5]. However, the majority of the expectant mothers in underdeveloped and developing countries do not have facilities and expertise to timely receive routine fetal US scans. Unfortunately, 99% of maternal and fetal deaths are reported in these countries as expectant mothers are unable to reach expertise or good resource settings promptly [6].

The difficulties faced by expectant mothers due to geometric barriers can be addressed through the transmission of medical Ultrasound (US) images to provide remote monitoring [7, 8]. Further, remote monitoring facilitates comfortable, time and cost-efficient scanning for mothers. Automated fetal US scans enable novel biomedical engineering concepts such as "Virtual Doctor" [9, 10], "Hospital-to-Home Care (H2H care)", and "Internet of Medical Things" (IoMT) [7, 8]. Further, self-scanning is encouraged recently in which automated medical evaluations of the expectant mother and the fetus are required [11]. The novel biomedical engineering approaches facilitate comfortable, time and cost-efficient scans for the mother with minimum risk of travelling and exposure to crowded medical centres. In addition, the limitations of conventional manual fetal US scans also can be minimized by using an automated computerized system.

Comprehensively, during a conventional routine manual fetal US scan, the transducer probe is moved on the abdomen to manually capture imaging planes containing the fetus from different angles and depths [12-14]. The standard fetal scanning planes are known to be imaging planes containing clinically important fetal anatomical structures, including the whole fetus, head, femur, abdomen, and heart [15, 16]. A few of these anatomical structures are shown in Figure 1. Approximately twenty standard planes can be monitored during a fetal US scan, and each plane contains distinct anatomical structures [13-15, 17][13,14,15,17]. After capturing a standard plane, it is frozen, and the caliper is placed on specific points of the fetus manually to calculate biometric parameters such as Head Circumference (HC), Femur Length (FL), Crown Rump Length (CRL), and Abdominal Circumference (AC) using standard charts [18-21]. The biometric parameters provide important clinical information, including Gestational Age (GA), size, and weight of the fetus [22, 23]. In addition, the anatomy of the fetal heart, liver, kidney, spine, urinary bladder, and genitalia can be observed during a fetal US scan [15, 16]. Even the functionality of the fetal organs can be visualized [12]. Apart from the aforementioned information, fetal biometric parameters provide information regarding the abnormalities of the fetus, and unhealthy anatomical and physiological maternal changes [24-26]. Therefore, the assessment of the fetus and maternal health directly depend on the operator's expertise.

However, conventional manual scans may be subject to fatigue errors, systematic errors, and observer variabilities. These scans may be subject to fatigue errors due to the tedious process [17]. In addition, strain injuries due to the repetitive movement of the hand in handling the transducer and caliper placement are common among obstetricians [27, 28]. Manual scanning is a subjective process, hence tends to suffer from observer variabilities. The accuracy of the standard plane detection [13] and the caliper placement [19] is

highly dependent on expertise. Capturing the standard planes and placing the caliper requires high-level skills, domain knowledge, and experience [13, 19, 25, 29]. As a result, significant attention has been directed towards the advancement and implementation of automated routine fetal ultrasound scans [17]. The automated process is a repeatable computerised system that avoids fatigue errors. Moreover, the risk of strain injuries can be avoided by reducing the workload of obstetricians [2].

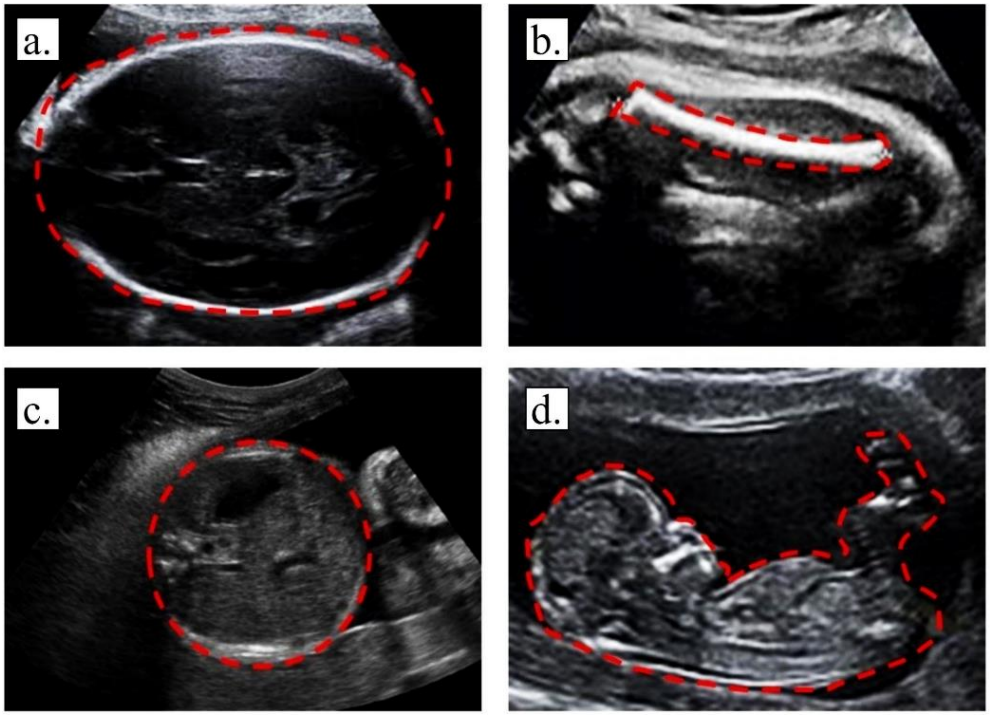

Figure 1. Commonly captured fetal anatomical structures during US scans; a. head, b. femur, c. abdomen, d. whole fetus

Automated fetal US scan consists of four main stages: pre-processing, fetus detection and segmentation, compression, and biometric parameter-based health evaluation. This process is illustrated in Figure 2. In the first stage, images are pre-processed using techniques such as de-noising and contrast-enhancing to ensure the smooth flow of the subsequent stages. In the second stage, fetal US image segmentation follows fetus detection. Depending on the requirement, either compression or biometric-parameter measurements evaluation is conducted. Compression is performed to achieve higher efficiency in image transmission for remote monitoring. The step, "fetus detection and segmentation" is a key stage that directly contributes to the efficiency and reliability of the automated process.

Region-of-Interest (ROI) is the clinically important area in a medical image. In fetal US images, ROI contains fetus/es. During a routine fetal US scan, ROI is observed and analyzed to monitor the growth of the fetus. In addition, health problems such as aneuploidies and abnormalities related to the fetus, or the mother can be diagnosed through routine scans [30]. Hence, detection and segmentation of the ROI are required to improve the accuracy and reliability of automated scans and improve the efficiency of the process.

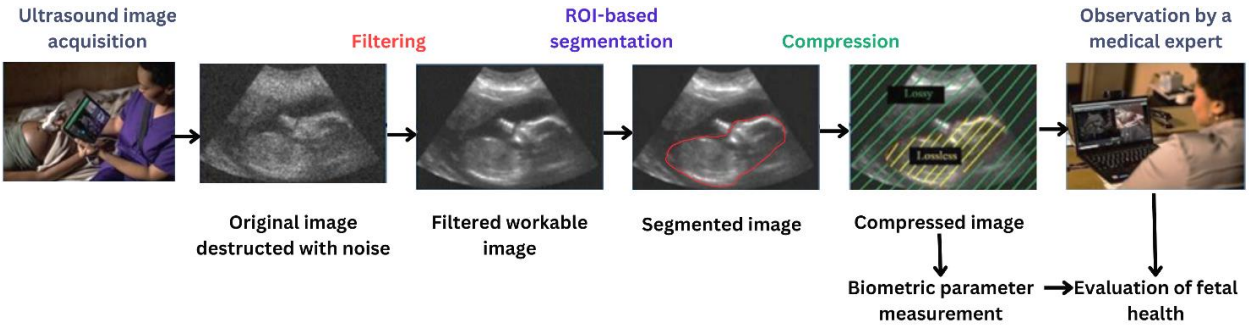

Figure 2. The process of automated fetal ultrasound scanning

Fetal US image segmentation facilitates efficient transmission via compression. Figure 1, d) illustrates an example of a whole fetus in a fetal US image in the mid-sagittal plane. Crown Rump Length (CRL) and Nuchal Translucency (NT), which can be measured using the mid-sagittal plane of the whole fetus, are vital biometric parameters measured in the first trimester. The CRL, which is shown in Figure 3, is the distance between the fetal head (crown) and the buttock. CRL is used to estimate the gestational age of the fetus. Moreover, the cardiac activity of the fetus should be present when CRL is greater than 7 mm according to current guidelines of the Society of Radiologists in Ultrasound (SRU) [31]. NT, which is illustrated in Figure 4, is the subcutaneous fluid accumulated behind the fetal neck. NT is an indication of fetal aneuploidy, cardiac abnormalities, and structural defects [32].

Nevertheless, the implementation of automated segmentation of fetal US images faces several challenges. Fetal US images are susceptible to suffering from various types of distortions [33]. Distortions pose a significant challenge in fetal US image segmentation and fetus detection [34-36]. Moreover, blurry edges and low contrast between the fetal and maternal tissues challenge automated fetus detection and segmentation [37]. Further, arbitrary shapes of the fetus and the similarity of the fetus to the nearby tissues are challenging [22, 35]. Further, the small size of the fetus is challenging, especially in the first trimester [38]. In addition, differences may appear in the images due to maternal structures, even when the images are captured from the fetuses in the same GA in the same standard plane. For example, US signals are highly attenuated when the maternal mass increases [39]. Moreover, maternal respiration may cause image differences [40]. Special cases such as twins and triplets pose a challenge in automated fetal US image segmentation. In addition, unhealthy fetuses may appear with differences in the anatomical structures compared to healthy fetuses, consequently causing differences in the fetal US images. Figure 5 illustrates examples for blurry edges and speckle noise in fetal US images.

Recent literature extensively explores deep learning for medical image processing, including segmentation and fetus detection applications, as it provides comparatively higher accuracy and reliable outputs in clinical diagnoses and therapy [30, 41, 42]. The visual perception-based training and advanced pattern recognition capabilities of deep learning algorithms cater to working with the distorted nature of fetal US images and arbitrary complex shapes that appear in the images. Fetal US images tend to suffer from speckle noise, acoustic shadows, different intensity distributions, and blurry edges [43]. Hence, the conventional image segmentation and fetus detection algorithms do not precisely work on fetal US images. Moreover, depending on the position and the location of the fetus, maternal tissues, pressure applied on the mother's

abdomen by the transducer probe, and motion artifacts due to maternal breathing and fetal movements may produce arbitrary shapes in the fetal US images [44]. Other advantages of deep learning algorithms are that these architectures do not require pre-defined parameters to initiate the process and have minimal human intervention requirements.

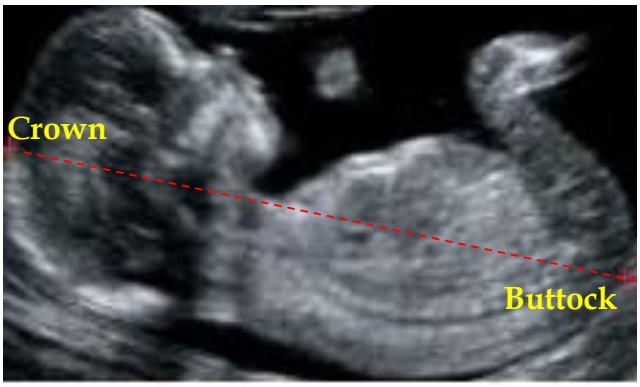

Figure 3. Crown-Rump Length (CRL) of the fetus

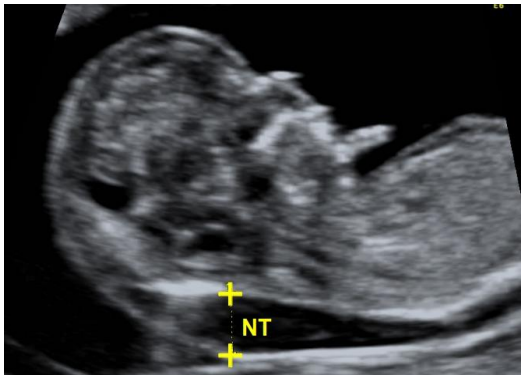

Figure 4. Nuchal Translucency (NT) of the Fetus [43]

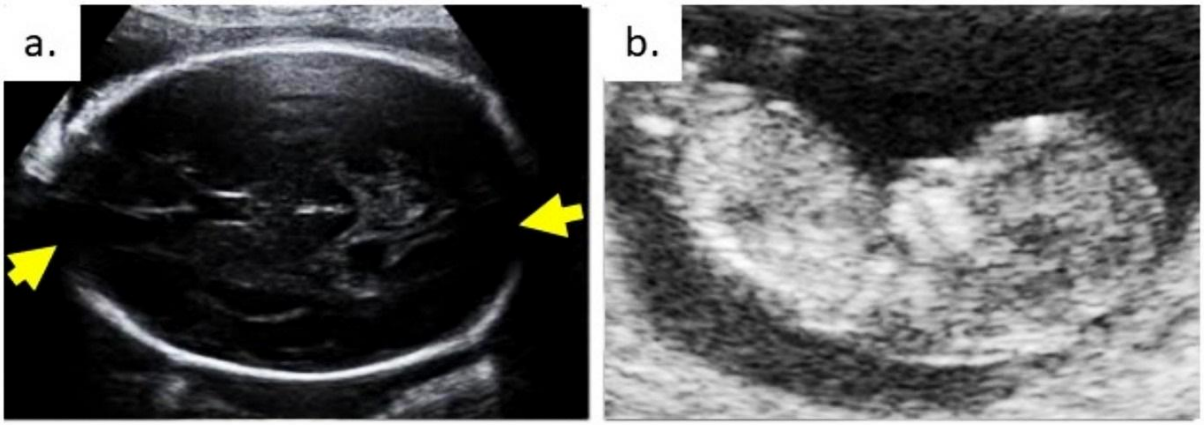

Figure 5. Challenges in fetal detection in US images: a: blurry edges, b: highly distorted nature

Fiorentino et al., (2023) presented a review on fetal US image analysis using deep learning [45]. Komatsu et al. (2021) proposed an algorithm to detect cardiac structures in three-vessel trachea view and four-chamber heart view of fetal US videos using Supervised Object detection with Normal data Only (SONO) [46]. This algorithm was developed based on the Convolution Neural Network (CNN). Zhang et al., (2021) proposed a Feature Pyramid Network-based faster-CNN algorithm to detect the fetal heart, head, and abdomen [47]. Al-Battal et al., (2021)] proposed an optical flow CNN to track the objects in medical US videos in real-time [48]. Nurmaini et al., (2021) also analyzed the performance of RPN-based CNN in the detection of structures of the fetal heart [49]. Teodor et al., (2022) analyzed the performance of YOLOV3 in detecting gestational sac, yolk sac, and embryo in week 6 - week 10 US images [40]. Sinclair et al., (2018) introduced a robust real-time algorithm to measure HC and BPD from a standard trans-ventricular (TV)

brain view plane using FCN-16 using 1948, 539, 216 images for training, testing, and validating respectively. Moreover, an observer variability study was conducted using 100 random test images with two experts: an engineer and a sonographer. A similar architecture named Cascaded FCN (CasFCN) was presented by Wu et al., (2017). This algorithm is an improved version of FCN-8 and achieved better results than basic Convolution Neural Networks, FCN-8, and U-Net [50]. Sundarasan et al., (2017) evaluated the performance of FCN-8, FCN-16, and FCN-32 in detecting three common fetal heart standard planes [51]. FCN-16 was comparatively accurate among the three models, with a low localization error and a minimum classification error. Skeika et al., (2020) evaluated the performance of a modified algorithm developed upon FCN and V-Net (VNet-C) using 297 grayscale 2D US images. This algorithm produced promising results compared to three state-of-the-art approaches presented by van den Heuvel et al., (2018) and Sobhaninia et al., in 2020 and 2019 [11,30]. Sobhaninia et al., (2019) presented an end-to-end Multi-Task Network based on Link-net architecture (MLTN). Sobhaninia et al., (2020) proposed a multi-scale Mini-LinkNet encoder-decoder architecture to address the limitations of conventional LinkNet namely, low accuracy in distorted images, a large number of training parameters, time in-efficiency, and loss of spatial information. Ryou et al., (2019) introduced the FCN-based algorithm for standard plane detection and semantic segmentation of the 3D fetal US volumes in the first trimester [52]. Szentimrey et al., (2022) proposed an algorithm using shape prior-based U-net to segment 3D neonatal cerebral ventricles US image segmentation [53]. This algorithm was comparatively accurate and time-efficient compared to 2D multiplane U-Net and 2D SegNet architecture.

This paper delves into the utilization of YOLO-s in a clinical setting to detect the whole fetus during the first trimester. The remainder of this paper is structured as follows: The second section, 'Materials and Methods,' provides a comprehensive description of the materials and methods followed. The third section presents the results obtained and a critical analysis of these findings. Lastly, the 'Conclusion' section offers recommendations based on the outcomes.

Materials and Methods

In this section, materials and methods employed in the study are extensively described. The experiments were conducted using publicly available 2D fetal US images, specifically, the whole fetus within the mid-sagittal plane during the first trimester. Data augmentation techniques and transfer learning were utilized to enhance the model's robustness. Our primary approach, YOLO-s, was implemented and benchmarked against additional models including YOLO-n, YOLO-m, YOLO-l, MobileNet, and ResNet. Additionally, we conducted a Likert-based subjective analysis to assess the performance and usability of the proposed model to provide valuable insights into its practical clinical applicability. The rest of the section extensively describes the materials and methods used for the experiment under the subsections, Dataset, Data Augmentation, Data Annotation, Transfer Learning, Implementation of Deep Learning Models, Evaluation Parameters and Statistical Analysis of Subjective Data.

Dataset

The majority of the investigations conducted to detect fetuses in US images focus on anatomical structures such as the head, femur, heart, and abdomen in the second and third trimesters. However, Crown Rump Length (CRL), which is the distance between the crown and the buttock of the fetus, is the main biometric parameter measured in the first trimester. This paper explores the detection of the whole fetus in 2D ultrasound images captured in the mid-sagittal plane during the first trimester. The dataset used for the experiment is small yet highly diverse. Publicly available 171 images which carry the fetus within week 6 to week 12 were included into the data set. These images were captured using various US machines by different experts. The original images were resized using zero padding, and the image size was set to 512 x 512. Image size 512 x 512 supports the implementation of deep learning algorithms, and it is a compatible resolution to view images in smart devices. The image dataset was initially categorized into three normal fetuses, special cases and null, as per the examples shown in Figure 6. The category "normal fetuses" contained healthy fetuses. "Special cases" contained the images of twins and unhealthy fetuses. "Null" image dataset category carried the images of non-pregnant US scans in the mid-sagittal plane. Data distribution among three categories is shown in Figure 7.

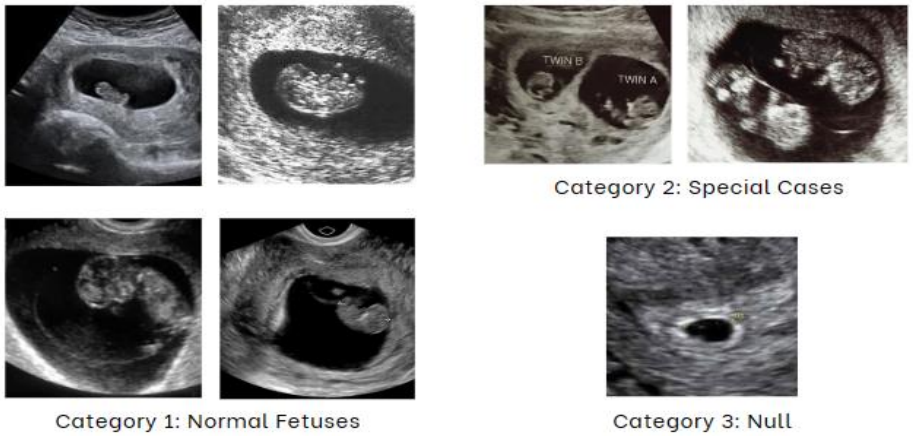

Figure 6. Initial categories of the dataset: Category 1: Normal fetuses, Category 2: Special cases, Category 3: Null

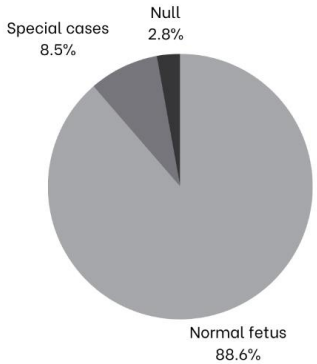

Figure 7. Data distribution among the three categories

The dataset was randomly divided into 70%, 20% and 10% for training, validating, and testing respectively using random uniform distribution as given in Table 1.

Table 1. Random distribution of data into training, validating, and testing

Category	Number of Images	Percentage (%)
Training	126	70
Validating	41	20
Testing	22	10

Data Augmentation

Data augmentation was performed as a solution to the difficulty in collecting sufficient data to train a learning architecture. 8 images were generated from 1 image using data augmentation. Basic data augmentation methods, including rotation, RGB shift, downscale, compression, horizontal flip, intensity jitter and blur, were used. In addition, random speckle noise was added as it is a common type of distortion in US images. Data augmentation improves network efficiency and prevents overfitting. Figure 8 presents a few examples of the appearance of the image after performing augmentation.

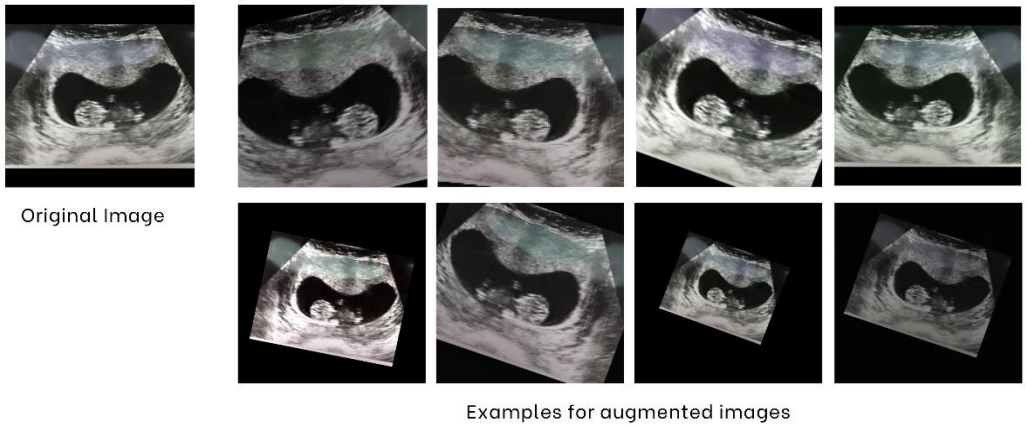

Figure 8. Examples for the appearance of augmentations in the images

Data Annotation

Data annotation was conducted manually by two experts: a senior radiologist with more than five years of experience and a Biomedical engineer. The "Roboflow" toolbox was used to label the images by marking the bounding box [54].

Transfer Learning

Due to the difficulty in accessing a higher number of fetal US images in the first trimester, transfer learning was used to train the model. The pre-trained weights on COCO were used for this step. In addition, using transfer learning improved the time efficiency of the process.

Implementation of Detection Models

Faster Region-based Convolutional Neural Network (Faster R-CNN) is an object detection algorithm rooted in deep learning. It focuses on enhancing the speed and accuracy of conventional R-CNN models. Faster R-CNN optimizes object detection by comprising two main components: the Region Proposal Network (RPN) and the detection network. The RPN employs a small network to generate region proposals by traversing the convolutional feature map. Subsequently, the detection network classifies these proposals and refines their bounding box coordinates. By sharing convolutional features and incorporating the RPN, Faster R-CNN enables end-to-end training and avoids the computationally expensive selective search used in previous methods, making it faster and more efficient for object detection tasks.

ResNet-50 is a convolutional neural network architecture introduced by Microsoft Research. It consists of 50 layers, including residual blocks that enable deep network training without encountering vanishing gradients. ResNet-50 employs skip connections, allowing the network to bypass certain layers and learn residual mappings. ResNet-50 has achieved notable performance in various computer vision tasks, showcasing its effectiveness in image classification, object detection, and semantic segmentation.

MobileNet is the comparatively fastest architecture which has been developed to date. It is a lightweight deep neural network which originated from the VGG-16 base network. MobileNet is specifically designed for mobiles and embedded vision applications. Therefore, it works in limited resource setups, even onset-ups, low-end smart devices. MobileNet performs depth-wise separable convolution to create a lightweight neural network. Consequently, it reduces the computational complexity compared to the networks based on regular convolution. MobileNet V3 and MobileNet V3-320 are variations of the same neural network architecture. However, they have differences in terms of input resolution. MobileNet V3 is a general architecture design that can work with various input resolutions. In contrast, MobileNet V3-320 is a specific variant of MobileNet V3 with an input resolution fixed at 320x320 pixels. The selection of input resolution affects the model's inference speed, accuracy, and memory requirements. Generally, lower resolutions like 320x320 result in faster inference but may compromise fine-grained detail in the output.

YOLO v5 is an object detection algorithm that builds upon the previous versions of YOLO (You Only Look Once). It offers versions; YOLO v5-n, YOLO v5-s, YOLO v5-m and YOLO v5-l. YOLO v5-n (small) has a smaller model size and fewer parameters, making it faster but with slightly lower accuracy. It is well-suited for real-time applications on devices with constrained computational resources.

YOLO v5-n is smaller and faster than YOLO v5-s, yet it may compromise the accuracy. YOLO v5-m strikes a balance between speed and accuracy, providing a good trade-off between the two. It offers improved performance over YOLO v5-s while maintaining a manageable model size. YOLO v5-l is a larger model that achieves higher accuracy but at the cost of slower inference speed and increased memory usage. It is well-suited for scenarios where accuracy is of utmost importance. These variations allow users to choose the model that aligns with their needs, striking a balance between inference speed, model size, and detection accuracy.

Evaluation Parameters

Several evaluation parameters are used for object detection. These parameters are closely related and include Intersection over Union (IoU), Precision, Recall, F-1 Score, and Mean Average Precision (mAP). The evaluation parameters used in this experiment are described below [55].

IoU is a metric that evaluates how much the ground truth overlaps the prediction given in Equation 1. IoU lies between 0 to 1 where 0 implies no overlap, and 1 implies complete overlap. To meaningfully integrate the metric IoU into detection, a threshold is set. Here, the threshold is set to 0.5, which is the true positive. IoU is vital in fetal detection in US images for evaluating the accuracy of object detection models by measuring how well the predicted bounding boxes align with the actual fetal structures. High IoU ensures precise localization, which is critical for accurate diagnosis and monitoring of fetal development.

$$IoU=\frac{\text{ground truth}\cap\text{prediction}}{\text{ground truth}\cup\text{prediction}} \qquad \textit{Equation 1}$$

Precision defines the correctness of the model in detecting only the relevant objects in each class. It is the ratio between True Positives (TP) over the total number of detections made by the model as given in Equation 2. Precision is important as it indicates the proportion of correctly identified fetal structures among all detected instances. High precision minimizes false positives, ensuring that only accurate and relevant detections are used for diagnosis and monitoring.

$$\text{Precision}=\frac{\text{True positives (TP)}}{\text{Total number of detections}} \qquad \textit{Equation 2}$$

The ability of detecting the ground truths by the model is defined by Recall. It is the ratio between True Positives (TP) over all the ground truths as given in Equation 3. Recall measures the ability of the model to identify all relevant fetal structures. High recall ensures that no significant fetal features are missed, which is crucial for comprehensive monitoring and accurate diagnosis of fetal health.

$$\text{Recall}=\frac{\text{True positives (TP)}}{\text{Total number of ground truths}} \qquad \textit{Equation 3}$$

F-1 score is a single metric which combines precision and recall as shown in Equation 4. F-1 score is important because it balances precision and recall, providing a single metric that reflects both the accuracy and completeness of the detections. A high F1 score indicates that the model effectively identifies fetal structures with both minimal false positives and minimal missed detections, ensuring reliable assessment of fetal health.

$$\text{F-1 score}=\frac{2\times\text{recall}\times\text{precision}}{\text{recall}+\text{precision}} \qquad \textit{Equation 4}$$

mAP provides a single measure of overall accuracy. It quantifies the average precision across multiple classes or categories in object detection. Equation to calculate mAP is given in *Equation 5*. *N* is the total number of classes and AP_i is the average precision for class *i.*

$$\text{mAP}=\frac{1}{N}\sum_{i=1}^{N} \text{AP}_i \qquad \textit{Equation 5}$$

Statistical Analysis of Subjective Data

Considering the comparable objective analysis results of YOLO-s to the YOLO-l and its low computer complexity, a subjective test was conducted to validate the usability of YOLO-s in a clinical set-up. 5 expert professionals including medical practitioners in the field of gynecology and biomedical engineers working with medical US imaging were engaged in conducting the subjective analysis. Each of the participants in the group was provided with a set of 17 sample images randomly extracted from the output images of the YOLO-s object detection model and had them rate the usability of images clinically using a Likert scale, ranging from 1 (poor) to 5 (excellent). To avoid the observer variabilities, 3 repeated images were included in the questionnaire. In addition, the order of presentation of images was randomized. Moreover, the procedure was repeated at least two days apart from the initial attempt.

Different statistical techniques were used to assess the inter-rater agreement and intra-rater agreements for data and rating reliability respectively. Cohen's Kappa given in Equation 6 was used to assess pairwise inter-rater agreement beyond chance, as well as intra-rater agreement between the same rater on the same image but repeated attempts.

$$k = \frac{1-P_e}{P_o-P_e} \qquad \textit{Equation 6}$$

In Equation 6, Kappa Score (k), calculated using relative observed agreement among raters (Po), hypothetical probability of chance agreement (Pe). Table 2 presents the Kappa score interpretations for both methods.

Table 2. Kappa score interpretation

Kappa Score (k)	Agreement beyond chance
$k < 0$	Poor agreement
$0 \leq k \leq 0.2$	Slight agreement
$0.2 \leq k \leq 0.4$	Fair agreement
$0.4 \leq k \leq 0.6$	Moderate agreement
$0.6 \leq k \leq 0.8$	Substantial agreement
$0.8 \leq k \leq 1.0$	Perfect agreement

To calculate the collective agreement between all raters, Fleiss Kappa (k_{fleiss}) given in Equation 7 was used.

$$k_{fleiss} = \frac{\bar{P}-\bar{P_e}}{1-\bar{P_e}} \qquad \textit{Equation 7}$$

The Intraclass Correlation Coefficient (ICC) was used to evaluate the reliability of ratings made by different raters. The method was adapted in such a way that ordinal data was treated as interval data, implying equal spacing between rating categories, and assuming the data are continuous and normally distributed. Equation 8 presents the equation used to calculate ICC using a 2-way mixed-effects model, which considers both the mean square of differences between targets (BMS) and within targets (WMS), adjusted for the number of raters (k). Table 3 presents the ICC score interpretations.

$$\text{Equation 8}$$

Table 3. Intraclass Correlation Coefficient (ICC) score interpretation

ICC Value	Interpretation
$0 \leq \text{ICC} \leq 0.4$	Poor reliability
$0.4 \leq \text{ICC} \leq 0.75$	Modest reliability
$0.75 \leq \text{ICC} \leq 1.0$	Excellent reliability

Results and Discussion

The performance of the implemented 7 models were compared with each other for object detection in fetal US images. Table 4 presents the comparison results.

Table 4. Comparison of the performance evaluation metrics

Method	Precision @ 0.5	Recall @ 0.5	F-1 score	mAP @ 0.5:095
ResNet-50 faster RCNN	0.950	0.966	0.958	0.583
MobileNet SSD faster R-CNN	0.955	0.960	0.958	0.583
MobileNet SSD 320 faster RCNN	0.912	0.924	0.918	0.535
YOLO V5-n	0.980	0.963	0.972	0.709
YOLO V5-s	0.991	0.968	0.979	0.734
YOLO V5-m	0.969	0.963	0.966	0.747
YOLO V5-l	0.982	0.975	0.978	0.751

According to Table 4, faster R-CNN based detection models provide mAP in-between 0.535 to 0.583 while the mAP of YOLO-v5 is ranging from 0.709 to 0.751. Accordingly, YOLO-v5 works better for fetal detection in US images compared to faster R-CNN. YOLO v5-l has shown the best performance for fetal detection in US images by achieving best F-1 score and mAP of 0.978 and 0.751, respectively. Figure 9 visualizes the results given in Table 1 in a graph.

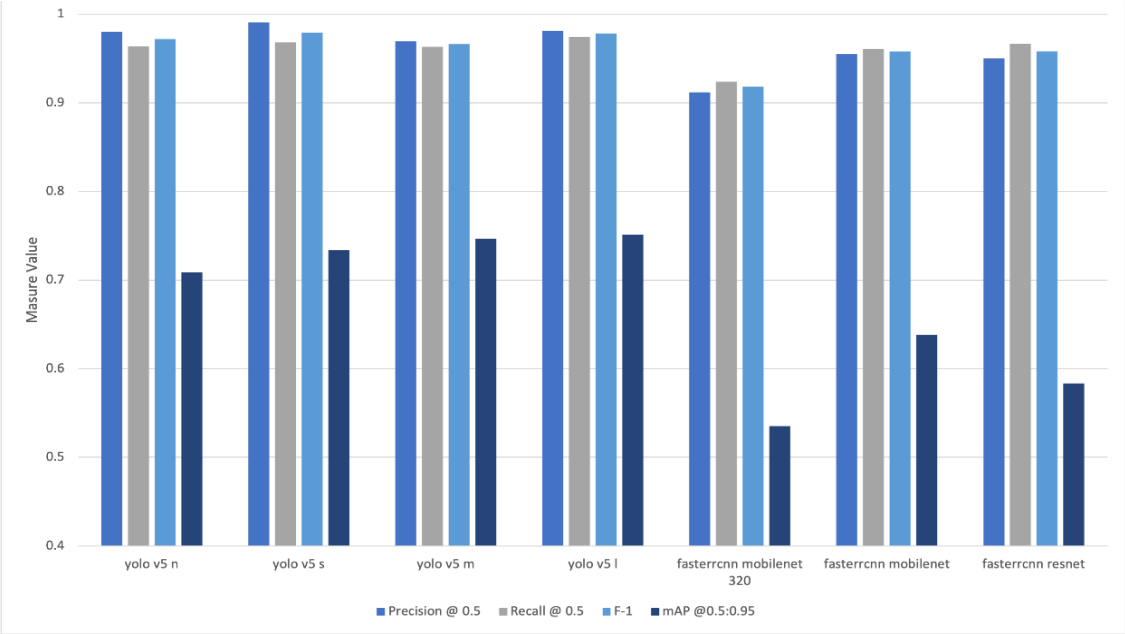

Figure 9. Graphical comparison of performance parameters of the implemented algorithms

However, as the ultimate objective of this research work is to provide the scanning facilities in low-end devices, computer complexity analysis for YOLO-v5 versions was conducted. The comparison of the computer complexity of YOLO-v5 versions is presented in Table 5.

Table 5. Computer Complexity Analysis of YOLO V5 Versions

YOLO-V5 Model	No. of Parameters	GFlops
Nano (n)	1,761,871	4.2
Small (s)	7,025,023	15.9
Medium (m)	20,856,975	48.0
Large (l)	46,113,663	107.8

According to Table 5, among YOLO-V5 models, YOLO-n and YOLO-s have demonstrated significantly low computer complexity. Figure 10 presents a comparative analysis of the outputs from YOLO-n and YOLO-s. The first row displays the outputs from YOLO-n, while the second row shows the outputs from YOLO-s. In each image, the red box highlights the detected object and provides the corresponding accuracy in numerical values.

According to the subjective analysis test results of the YOLO model, inter-rater Cohen's Kappa values ranged from 0.34 (fair agreement) to 0.74 (substantial agreement), where the majority fell under 0.4 and 0.6, denoting a moderate level of agreement. On the other hand, the majority of experts demonstrated perfect intra-rater reliability (Kappa score of 1), except for a few deviations. Meanwhile, Fleiss' Kappa resulted in a value of 0.69 multi-rater agreement, indicating substantial agreement on the between raters beyond chance. An ICC value of 0.7 indicated modest reliability. The results suggest that despite some differences in individual ratings, the expert ratings were fairly consistent with each other.

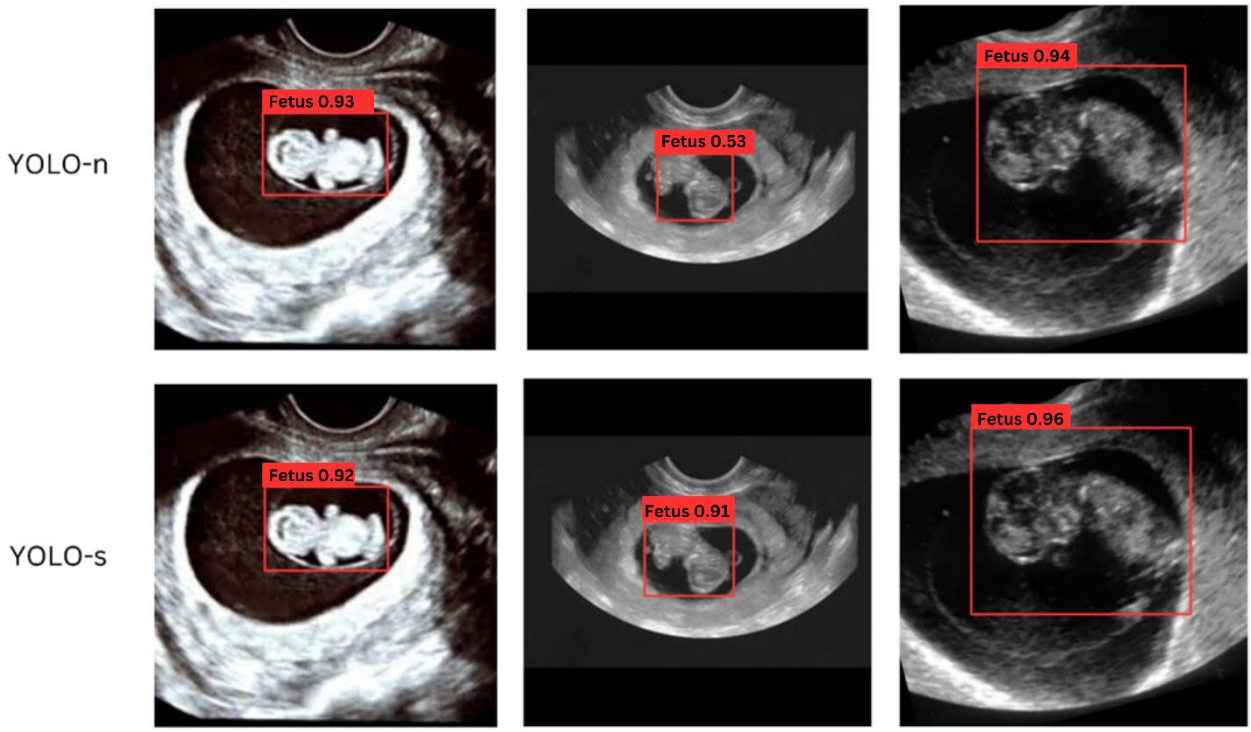

Figure 10. Examples for a comparison of outputs of YOLO-n and YOLO-s

Limitations of the Presented Work

Benchmark common detection models, faster RCNN, and YOLO-V5 may fail in some cases to detect small objects as they might not appear across all feature maps. Hence, detection of the fetus in the first few weeks is challenging [46]. However, improving the image resolution resolves the issue to some extent. Increasing the diversity of the dataset by including images of the different health conditions may assure the reliability and accuracy of the algorithm in a clinical set-up [37] . Therefore, the data distribution can be improved by including more unhealthy fetuses and special cases such as twins and triplets [60]. In addition, data annotation can be strengthened by using annotated data by five or more experts, and the same image can be annotated more than once by one expert. This process reduces the inter-observer variability and intra-observer variability in the algorithm [5]. Typical routine US scans capture the videos of the fetuses from different angles [41, 45]. Therefore, the detection of the fetus in video frames following accurate fetal US standard plane extraction from the videos is suggested. That is, mid-sagittal plane extraction from the fetal US videos is required prior to fetal detection. The work presented in this paper is limited to first-trimester fetuses captured in the mid-sagittal plane. Working with other fetal US planes in all trimesters is required to adapt YOLO V5-s to the clinical set-up.

Conclusion

Ultrasound is recommended for expectant mothers due to its non-invasive, non-ionizing nature, offering a safe means to monitor fetal health and growth. However, in rural areas of underdeveloped and developing countries, access to timely routine scans is hindered by a lack of expertise and facilities. This delay can lead

to severe consequences, increasing the risk of fetal and maternal deaths, particularly during the critical first trimester. To address this issue, innovative solutions such as remote monitoring and self-scans utilizing low-end smart devices are proposed. These technologies aim to bridge the gap, providing accessible and timely care to expectant mothers in resource-constrained settings. The study was conducted using a collection of 171 2D fetal ultrasound images depicting the entire fetus captured in the mid-sagittal plane during the first trimester.

The identification of fetuses in ultrasound images is a crucial component of remote monitoring and self-scanning initiatives. However, this task is inherently challenging for various reasons. These reasons include distortions such as blurry edges and low contrast, arbitrary shapes and the small sizes of the fetuses, large variability in images, special cases such as twins, and health disparities. In this study, the focus is on evaluating the efficacy of YOLO-s in fetal detection. To assess its performance, the model is compared against benchmark algorithms, including ResNet-50 and MobileNet-based faster R-CNN, as well as YOLO-n, YOLO-m, and YOLO-l. This comparative analysis aims to provide insights into the strengths and weaknesses of YOLO-s in relation to established detection algorithms, contributing valuable information to the advancement of fetal detection in ultrasound imaging for remote monitoring and self-scanning applications.

This study explores the use of computationally efficient YOLO V5_s for fetal detection in ultrasound images. The model is compared with benchmark detection models, including faster R-CNN with base networks ResNet-50 and MobileNet, YOLO-n, YOLO-m, and YOLO-l. YOLO-based models outperform faster R-CNN models, with YOLO-l exhibiting the best performance with an F-1 score of 0.978 and an mAP of 0.751. Despite the low computer complexity, YOLO-n's lower mAP of 0.709 renders it unsuitable for clinical use. A subjective test involving experts with over five years of experience validates the utility of YOLO-s in the clinical setting. The collective mean of independent image ratings indicates above-average clinical usability. Cohen's Kappa values reveal a fair level of agreement among raters, while Fleiss' Kappa of 0.69 indicates substantial agreement beyond chance, and an ICC value of 0.7 suggests modest reliability consistency between raters. Future Overall results endorse the application of YOLO-s for real-time fetal detection in the first trimester, offering a promising solution for remote monitoring in resource-constrained settings. However, further validation through strong subjective analysis is recommended.

Commonly used detection models such as faster RCNN and YOLO-V5 may result in inaccurate predictions due to the small size of fetuses in the first trimester. This can be overcome by enhancing image resolution and diversifying datasets. Strengthening data annotation through multiple expert annotations reduces variability improving the validity. While this study focuses on mid-sagittal plane extraction for first-trimester fetuses, future work should extend to other fetal ultrasound planes in all trimesters for comprehensive clinical applicability.

Conflicts of Interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants;

participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

[1] Lee, W.A., Nelson, G., and Grogan, S.P., *Sonography 1st Trimester Assessment, Protocols, and Interpretation*, in *StatPearls*. **2025**, StatPearls Publishing Copyright © 2025, StatPearls Publishing LLC.: Treasure Island (FL) ineligible companies. Disclosure: Grant Nelson declares no relevant financial relationships with ineligible companies. Disclosure: Scott Grogan declares no relevant financial relationships with ineligible companies.

[2] Hermawati, F.A., Tjandrasa, H., Sugiono, Sari, G.I.P., and Azis, A., Automatic femur length measurement for fetal ultrasound image using localizing region-based active contour method. Journal of Physics: Conference Series, **2019**. 1230*(1)*, 012002. 10.1088/1742-6596/1230/1/012002.

[3] Kumar, C. and Prakash, R., Ultrasound medical image denoising using threshold-based wavelet transformation method. Journal of Medical Imaging and Health Informatics, **2020**. 10, 1825-1830.

[4] Meiburger, K.M., Acharya, U.R., and Molinari, F., Automated localization and segmentation techniques for B-mode ultrasound images: A review. Comput Biol Med, **2018**. 92, 210-235. 10.1016/j.compbiomed.2017.11.018.

[5] Rawat, V., Jain, A., and Shrimali, V., Automated Techniques for the Interpretation of Fetal Abnormalities: A Review. Appl Bionics Biomech, **2018**. 2018, 6452050. 10.1155/2018/6452050.

[6] van den Heuvel, T.L.A., de Bruijn, D., de Korte, C.L., and Ginneken, B.V., Automated measurement of fetal head circumference using 2D ultrasound images. PLoS One, **2018**. 13*(8)*, e0200412. 10.1371/journal.pone.0200412.

[7] Alqaralleh, B.A.Y., Vaiyapuri, T., Parvathy, V.S., Gupta, D., Khanna, A., and Shankar, K., Blockchain-assisted secure image transmission and diagnosis model on Internet of Medical Things Environment. Personal and Ubiquitous Computing, **2024**. 28*(1)*, 17-27. 10.1007/s00779-021-01543-2.

[8] Joyia, G.J., Liaqat, R.M., Farooq, A., and Rehman, S., Internet of medical things (IoMT): applications, benefits and future challenges in healthcare domain. Journal of Communications, **2017**. 12, 240-247.

[9] Keswani, B., Mohapatra, A.G., Mishra, T.C., Keswani, P., Mohapatra, P.C.G., Akhtar, M.M., and Vijay, P., *World of virtual reality (VR) in healthcare in Advanced Computational Intelligence Techniques for Virtual Reality in Healthcare*. **2020**: Springer.

[10] Shi, Y., *Reducing uncertainties in virtual consultation: the impact of media naturalness and mental model alignment on patient satisfaction in doctor*. **2020**.

[11] Honeyman, L., *Personal use of handheld fetal doppler monitors by women in New Zealand*. 2021, Open Access Te Herenga Waka-Victoria University of Wellington.

[12] Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., and Gratacós, E., Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Scientific Reports, **2020**. 10*(1)*, 10200. 10.1038/s41598-020-67076-5.

[13] Sinclair, M., Baumgartner, C.F., Matthew, J., Bai, W., Martinez, J.C., Li, Y., Smith, S., Knight, C.L., Kainz, B., and Hajnal, J., *Human-level performance on automatic head biometrics in fetal ultrasound using fully convolutional neural networks*,

in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. **2018**, IEEE. pp. 714-717.

[14] Sobhaninia, Z., Emami, A., Karimi, N., and Samavi, S., *Localization of fetal head in ultrasound images by multiscale view and deep neural networks*, in *25th International Computer Conference*. **2020**. pp. 1-5.

[15] Salomon, L.J., Alfirevic, Z., Berghella, V., Bilardo, C., Hernandez-Andrade, E., Johnsen, S.L., Kalache, K., Leung, K.Y., Malinger, G., Munoz, H., Prefumo, F., Toi, A., Lee, W., and Committee, I.C.S., Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound Obstet Gynecol, **2011**. 37*(1)*, 116-26. 10.1002/uog.8831.

[16] MacGregor, S.N. and Sabbagha, R.E., Assessment of Gestational Age by Ultrasound. The Global Library of Women's Medicine, **2009**. 10.3843/glowm.10206.

[17] Sarris, I., Ioannou, C., Chamberlain, P., Ohuma, E., Roseman, F., Hoch, L., and Altman, D.G., Intra- and interobserver variability in fetal ultrasound measurements. Ultrasound Obstet Gynecol, **2012**. 39*(3)*, 266-73. 10.1002/uog.10082.

[18] Loughna, P., Chitty, L., Evans, T., and Chudleigh, T., Fetal Size and Dating: Charts Recommended for Clinical Obstetric Practice. **2009**. 17*(3)*, 160-166. 10.1179/174313409x448543.

[19] Yazdi, B., Zanker, P., Wanger, P., Sonek, J., Pintoffl, K., Hoopmann, M., and Kagan, K.O., Optimal caliper placement: manual vs automated methods. Ultrasound Obstet Gynecol, **2014**. 43*(2)*, 170-5. 10.1002/uog.12509.

[20] Verburg, B.O., Steegers, E.A., De Ridder, M., Snijders, R.J., Smith, E., Hofman, A., Moll, H.A., Jaddoe, V.W., and Witteman, J.C., New charts for ultrasound dating of pregnancy and assessment of fetal growth: longitudinal data from a population-based cohort study. Ultrasound Obstet Gynecol, **2008**. 31*(4)*, 388-96. 10.1002/uog.5225.

[21] Altman, D.G. and Chitty, L.S., New charts for ultrasound dating of pregnancy. Ultrasound Obstet Gynecol, **1997**. 10*(3)*, 174-91. 10.1046/j.1469-0705.1997.10030174.x.

[22] Chen, H., Wu, L., Dou, Q., Qin, J., Li, S., Cheng, J.Z., Ni, D., and Heng, P.A., Ultrasound Standard Plane Detection Using a Composite Neural Network Framework. IEEE Trans Cybern, **2017**. 47*(6)*, 1576-1586. 10.1109/TCYB.2017.2685080.

[23] Figueras, F. and Gratacos, E., Update on the diagnosis and classification of fetal growth restriction and proposal of a stage-based management protocol. Fetal Diagn Ther, **2014**. 36*(2)*, 86-98. 10.1159/000357592.

[24] Banos, N., Perez-Moreno, A., Julia, C., Murillo-Bravo, C., Coronado, D., Gratacos, E., Deprest, J., and Palacio, M., Quantitative analysis of cervical texture by ultrasound in mid-pregnancy and association with spontaneous preterm birth. Ultrasound Obstet Gynecol, **2018**. 51*(5)*, 637-643. 10.1002/uog.17525.

[25] Kagan, K. and Sonek, J., How to measure cervical length. Ultrasound in Obstetrics & Gynecology, **2015**. 45, 358-362.

[26] Nicolaides, K.H., Syngelaki, A., Ashoor, G., Birdir, C., and Touzet, G., Noninvasive prenatal testing for fetal trisomies in a routinely screened first-trimester population. Am J Obstet Gynecol, **2012**. 207*(5)*, 374 e1-6. 10.1016/j.ajog.2012.08.033.

[27] Li, J., Wang, Y., Lei, B., Cheng, J.Z., Qin, J., Wang, T., Li, S., and Ni, D., Automatic Fetal Head Circumference Measurement in Ultrasound Using Random Forest and Fast Ellipse Fitting. IEEE J Biomed Health Inform, **2018**. 22*(1)*, 215-223. 10.1109/JBHI.2017.2703890.

[28] Janga, D. and Akinfenwa, O., Work-related repetitive strain injuries amongst practitioners of obstetric and gynaecological ultrasound worldwide. Arch Gynecol Obstet, **2012**. 286*(2)*, 353-6. 10.1007/s00404-012-2306-6.

[29] Wu, L., Cheng, J.Z., Li, S., Lei, B., Wang, T., and Ni, D., FUIQA: Fetal Ultrasound Image Quality Assessment With Deep Convolutional Networks. IEEE Trans Cybern, **2017**. 47*(5)*, 1336-1349. 10.1109/tcyb.2017.2671898.

[30] Sobhaninia, Z., Rafiei, S., Emami, A., Karimi, N., Najarian, K., Samavi, S., and Reza Soroushmehr, S.M., Fetal Ultrasound Image Segmentation for Measuring Biometric Parameters Using Multi-Task Deep Learning. Annu Int Conf IEEE Eng Med Biol Soc, **2019**. 2019, 6545-6548. 10.1109/EMBC.2019.8856981.

[31] Murugan, V., Murphy, B., Dupuis, C., Goldstein, A., and Kim, Y., *Role of ultrasound in the evaluation of first-trimester pregnancies in the acute setting*. *Ultrasonography*. . 2020. p. 178-189.

[32] Pauta, M., Martinez-Portilla, R.J., and Borrell, A., Diagnostic yield of next-generation sequencing in fetuses with isolated increased nuchal translucency: systematic review and meta-analysis. Ultrasound Obstet Gynecol, **2022**. 59*(1)*, 26-32. 10.1002/uog.23746.

[33] Kalbhor, M., Shinde, S., Wajire, P., and Jude, H., CerviCell-detector: An object detection approach for identifying the cancerous cells in pap smear images of cervical cancer. Heliyon, **2023**. 9*(11)*. 10.1016/j.heliyon.2023.e22324.

[34] Skeika, E.L., Da Luz, M.R., Fernandes, B.J.T., Siqueira, H.V., and De Andrade, M.L.S.C., Convolutional neural network to detect and measure fetal skull circumference in ultrasound imaging. IEEE Access, **2020**. 8, 191519-191529.

[35] Wu, L., Xin, Y., Li, S., Wang, T., Heng, P.A., and Ni, D., *Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation*, in *IEEE 14th international symposium on biomedical imaging*. **2017**, IEEE. pp. 663-666.

[36] Lu, W., Tan, J., and Floyd, R., Automated fetal head detection and measurement in ultrasound images by iterative randomized Hough transform. Ultrasound Med Biol, **2005**. 31*(7)*, 929-36. 10.1016/j.ultrasmedbio.2005.04.002.

[37] Brattain, L.J., Telfer, B.A., Dhyani, M., Grajo, J.R., and Samir, A.E., Machine learning for medical ultrasound: status, methods, and future opportunities. Abdom Radiol (NY), **2018**. 43*(4)*, 786-799. 10.1007/s00261-018-1517-0.

[38] Sridar, P., Kumar, A., Li, C., Woo, J., Quinton, A., Benzie, R., Peek, M.J., Feng, D., Kumar, R.K., Nanan, R., and Kim, J., Automatic Measurement of Thalamic Diameter in 2-D Fetal Ultrasound Brain Images Using Shape Prior Constrained Regularized Level Sets. IEEE J Biomed Health Inform, **2017**. 21*(4)*, 1069-1078. 10.1109/JBHI.2016.2582175.

[39] Dashe, J.S., McIntire, D.D., and Twickler, D.M., Effect of maternal obesity on the ultrasound detection of anomalous fetuses. Obstet Gynecol, **2009**. 113*(5)*, 1001-1007. 10.1097/AOG.0b013e3181a1d2f5.

[40] Teodor, O.M., Cirstoiu, M.M., Ichim, L., and Popescu, D. *Detection of Early Pregnancy in Ultrasound Images Using YOLOv3*. in *2022 30th Mediterranean Conference on Control and Automation (MED)*. **2022**. pp. 188-193.

[41] Wang, Z., *Deep learning in medical ultrasound image segmentation: A review*. **2020**.

[42] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., and Sanchez, C.I., A survey on deep learning in medical image analysis. Med Image Anal, **2017**. 42, 60-88. 10.1016/j.media.2017.07.005.

[43] Karaoğlu, O., Bilge, H.Ş., and Uluer, İ., Removal of speckle noises from ultrasound images using five different deep learning networks. Engineering Science and Technology, an International Journal, **2022**. 29. 10.1016/j.jestch.2021.06.010.

[44] Torrents-Barrena, J., Monill, N., Piella, G., Gratacos, E., Eixarch, E., Ceresa, M., and Gonzalez Ballester, M.A., Assessment of Radiomics and Deep Learning for the Segmentation of Fetal and Maternal Anatomy in Magnetic Resonance Imaging and Ultrasound. Acad Radiol, **2021**. 28*(2)*, 173-188. 10.1016/j.acra.2019.11.006.

[45] Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E., and Moccia, S., A review on deep-learning algorithms

for fetal ultrasound-image analysis. Medical Image Analysis, **2023**. 83, 102629. https://doi.org/10.1016/j.media.2022.102629.

[46] Komatsu, M., Sakai, A., Komatsu, R., Matsuoka, R., Yasutomi, S., Shozu, K., Dozen, A., Machino, H., Hidaka, H., Arakaki, T., Asada, K., Kaneko, S., Sekizawa, A., and Hamamoto, R., Detection of Cardiac Structural Abnormalities in Fetal Ultrasound Videos Using Deep Learning. Applied Sciences, **2021**. 11*(1)*, 371. 10.3390/app11010371.

[47] Zhang, B., Liu, H., Luo, H., and Li, K., Automatic quality assessment for 2D fetal sonographic standard plane based on multitask learning. Medicine (Baltimore), **2021**. 100*(4)*, e24427. 10.1097/md.0000000000024427.

[48] Al-Battal, A.F., Gong, Y., Xu, L., Morton, T., Du, C., Bu, Y., Lerman, I.R., Madhavan, R., and Nguyen, T.Q., A CNN Segmentation-Based Approach to Object Detection and Tracking in Ultrasound Scans with Application to the Vagus Nerve Detection. Annu Int Conf IEEE Eng Med Biol Soc, **2021**. 2021, 3322-3327. 10.1109/embc46164.2021.9630522.

[49] Nurmaini, S., Rachmatullah, M.N., Sapitri, A.I., Darmawahyuni, A., Tutuko, B., Firdaus, F., Partan, R.U., and Bernolian, N., Deep Learning-Based Computer-Aided Fetal Echocardiography: Application to Heart Standard View Segmentation for Congenital Heart Defects Detection. Sensors (Basel), **2021**. 21*(23)*. 10.3390/s21238007.

[50] Ronneberger, O., Fischer, P., and Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. **2015**. Cham: Springer International Publishing. pp. 234-241.

[51] Sundaresan, V., Bridge, C.P., Ioannou, C., and Noble, J.A. *Automated characterization of the fetal heart in ultrasound images using fully convolutional neural networks*. in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. **2017**. pp. 671-674.

[52] Ryou, H., Yaqub, M., Cavallaro, A., Papageorghiou, A.T., and Alison Noble, J., Automated 3D ultrasound image analysis for first trimester assessment of fetal health. Phys Med Biol, **2019**. 64*(18)*, 185010. 10.1088/1361-6560/ab3ad1.

[53] Szentimrey, Z., de Ribaupierre, S., Fenster, A., and Ukwatta, E., Automated 3D U-net based segmentation of neonatal cerebral ventricles from 3D ultrasound images. Med Phys, **2022**. 49*(2)*, 1034-1046. 10.1002/mp.15432.

[54] Aljabri, M., AlAmir, M., AlGhamdi, M., Abdel-Mottaleb, M., and Collado-Mesa, F., Towards a better understanding of annotation tools for medical imaging: a survey. Multimed Tools Appl, **2022**. 81*(18)*, 25877-25911. 10.1007/s11042-022-12100-1.

[55] Xiao, Y., Tian, Z., Yu, J., Zhang, Y., Liu, S., Du, S., and Lan, X., A review of object detection based on deep learning. Multimedia Tools and Applications, **2020**. 79*(33)*, 23729-23791. 10.1007/s11042-020-08976-6.