Article

Ensemble Transfer Learning for Fetal Head Analysis: From Segmentation to Gestational Age and Weight Prediction

Mahmood Alzubaidi [1,*,†], Marco Agus [1,†], Uzair Shah [1], Michel Makhlouf [2], Khalid Alyafei [2] and Mowafa Househ [1,*]

1   College of Science and Engineering, Hamad Bin Khalifa University, Doha P.O. Box 34110 , Qatar
2   Sidra Medical and Research Center, Sidra Medicine, Doha P.O. Box 26999, Qatar
*   Correspondence: maal28902@hbku.edu.qa (M.A.); mhouseh@hbku.edu.qa (M.H.)
†   These authors contributed equally to this work.

Abstract: Ultrasound is one of the most commonly used imaging methodologies in obstetrics to monitor the growth of a fetus during the gestation period. Specifically, ultrasound images are routinely utilized to gather fetal information, including body measurements, anatomy structure, fetal movements, and pregnancy complications. Recent developments in artificial intelligence and computer vision provide new methods for the automated analysis of medical images in many domains, including ultrasound images. We present a full end-to-end framework for segmenting, measuring, and estimating fetal gestational age and weight based on two-dimensional ultrasound images of the fetal head. Our segmentation framework is based on the following components: (i) eight segmentation architectures (UNet, UNet Plus, Attention UNet, UNet 3+, TransUNet, FPN, LinkNet, and Deeplabv3) were fine-tuned using lightweight network EffientNetB0, and (ii) a weighted voting method for building an optimized ensemble transfer learning model (ETLM). On top of that, ETLM was used to segment the fetal head and to perform analytic and accurate measurements of circumference and seven other values of the fetal head, which we incorporated into a multiple regression model for predicting the week of gestational age and the estimated fetal weight (EFW). We finally validated the regression model by comparing our result with expert physician and longitudinal references. We evaluated the performance of our framework on the public domain dataset HC18: we obtained 98.53% mean intersection over union (mIoU) as the segmentation accuracy, overcoming the state-of-the-art methods; as measurement accuracy, we obtained a 1.87 mm mean absolute difference (MAD). Finally we obtained a 0.03% mean square error (MSE) in predicting the week of gestational age and 0.05% MSE in predicting EFW.

Keywords: image segmentation; ensemble transfer learning; fetal head; gestational age; estimated fetal weight; ultrasound

1. Introduction

Ultrasonic imaging, also known as ultrasound, is frequently utilized in clinical assessment since it does not include ionizing radiation, and it is less expensive than computed tomography (CT) and magnetic resonance imaging (MRI) [1]. Women usually have one to three ultrasounds during pregnancy. If the lady is pregnant with twins or is at high risk, ultrasounds may be required more frequently [2]. Ultrasound may be utilized in various prenatal diagnostic situations, including: confirming the pregnancy and the position of the fetus, calculating the gestational age of the fetal baby, verifying the number of fetal bodies, examining fetal development, examining the amounts of the placenta and amniotic fluid, identifying congenital disabilities, looking into complications, and other prenatal tests [3]. When ultrasound is routinely used in early pregnancy, it will result in an earlier detection of problems and an improved management of pregnancy complications, which is better than relying on clinical indicators such as bleeding in early pregnancy [4]. Halle et al. [5]

reported that 1111 women received prenatal treatment at primary care health centers in their health cohort. Ninety-five percent of women reported having some fetal ultrasound scan prior to the 19th week scan, and 64% reported having two or more scans during this period. Seventy-eight percent of women decided to participate in week 11–14 screening for fetal abnormalities. Therefore, ultrasound is the preferable option for prenatal care compared to other imaging modalities, because it allows for the recognition and measurement of anatomical structures that can be used as guidelines for physician assessment of the fetal health status [3].

Many clinical ultrasonography diagnostics necessitate the use of anatomical structure measurements that are clear and reliable. These measurements are used to estimate fetal gestational age and weight, which is essential for monitoring growth patterns during pregnancy [6]. Abdominal circumference (AC), femur length (FL), crown–rump length (CRL), occipitofrontal diameter (OFD), biparietal diameter (BPD), and head circumference (HC) are some of the biological characteristics that may be measured during a prenatal checkup [7]. In the 13th to 25th week of pregnancy, obstetricians and gynecologists may calculate the fetus's gestational age and weight, evaluate the fetus's growth, and decide if aberrant head development is suspected, by measuring the fetus's HC [8]. When measuring HC in clinical practice, the procedure is performed manually by either overlaying an ellipse on the fetal skull or by recognizing landmarks that delimit the central head axis. Despite this practice, the manual delineation raises concerns about measurement repeatability and time consumption, since ultrasound imaging is prone to various errors, including motion blurring, missing borders, acoustic shadows, speckle noise, and a low signal-to-noise ratio [9]. As a result, interpreting ultrasound images becomes extremely difficult, necessitating the use of skilled operators. Figure 1 shows ultrasound image samples that are noisy and indistinct, with an incomplete head contour; additionally, the fetal skull is not evident enough to be detected in the first trimester, as indicated in the samples obtained from the public dataset [10].

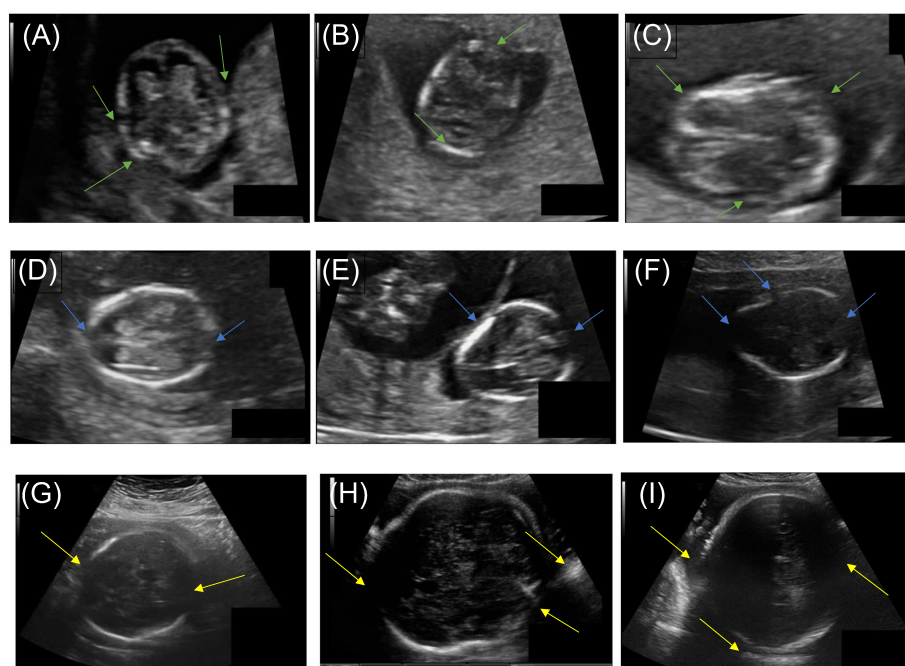**Figure 1.** Typical prenatal ultrasound images from each trimester. (**A–C**) First trimester, green arrows indicate blurred fetal head and artifacts. (**D–F**) Second trimester, blue arrows indicate poor signal-to-noise ratio and reflection from the fetal membranes and amniotic fluid interface. (**G–I**) Third trimester, yellow arrows indicate speckle noise and standard sutures or ultrasonography artifacts.

Traditional approaches for fetal biometric segmentation and measures have been under investigation for the past decade. As a result of the development of these approaches, workflow efficiency has been increased by lowering the number of steps required for routine fetal measures and examination time [6]. The randomized Hough transform [11], semi-supervised patch-based graphs [12], multilevel thresholding circular shortest paths [13], boundary fragment models [14], haar-like features [7], active contouring [15], morphological operators [16], the difference of Gaussians [17], and deformable models [18] have all been used in previous HC measurement studies.

With the advancement of deep learning technology in recent years, integrating medical images and artificial intelligence has emerged as a popular study area in medicine [19]. Convolutional neural networks (CNNs) have rapidly gained popularity as a powerful tool for many image processing applications, including classification, object identification, segmentation, and registration, among others [20]. As a result, the field of medical image segmentation is exploding with new applications. A few representative designs of CNNs are fully convolutional networks (FCNs) [21], UNet [22], and three-dimensional VNet [23].

1.1. Contributions

Numerous challenges remain for prior traditional and deep learning methods, including segmenting regions with missing edges, the absence of textural contrast, the specification of a region of interest (ROI), and background detection. These difficulties can be overcome using ensemble learning. Nowadays, CNNs are evolving towards lightweight architectures that can be integrated in edge computing frameworks [24], but prior mentioned techniques required a lengthy training period, high network parameters, high image resolution, and costly resources to run a heavy model. However, these issues may be mitigated by fine-tuning a pre-trained lightweight network. Finally, earlier studies did not explore the feasibility of utilizing machine learning and segmented image measurements to determine fetal gestational age (GA), estimated fetal weight (EFW), and abnormality signs. In this regard, this work proposes a complete pipeline for automatic segmentation and measuring the fetal head in two-dimensional (2D) ultrasound images, followed by a prediction of the fetal gestation age and weight. Below is a summary of technical contributions:

1. We fine-tuned eight segmentation networks using a pre-trained lightweight network (EffientNetB0) and employed weighted voting ensemble learning on the trained segmentation networks to obtain the optimal segmentation result.
2. We extensively evaluated the ensemble transfer learning model (ETLM) by performing three-level evaluations: fetal head segmentation evaluation, predicted mask and post-processing quality assessment, and head measurement evaluation.
3. We generated a new fetal head measurement dataset and manually labeled it by adding fetal gestation age and weight.
4. We trained multiple regressions model to predict fetal GA and EFW to address the limitation of the current formulas (Equations (21) and (22)).
5. We evaluated the regression model result using an expert obstetrician, and a longitudinal reference using Pearson's correlation coefficient (Pearson's r).

1.2. Organization

The following is the paper's organization: Section 2 discusses relevant research on fetal head segmentation, HC measurement, and fetal GA and EFW calculation. Section 3.1 discusses the dataset and our methodology pipeline in depth. Section 4 contains details about the experiment and evaluation methods. Section 5 presents the results, discussion, and a comparison with state-of-the-art works. Section 6 highlights the strengths and limitations of the research. Finally, Section 7 covers a conclusion and future work.

2. Related Work

Our works deals with fetal head segmentation using traditional approaches and deep learning, HC measurement, and the calculation of GA and EFW. It is impossible to provide

here an extensive overview of the literature related to these topics. We refer readers to the survey and review [4,25–27]. In the following, we discuss the methods that are most closely related to our work.

2.1. Fetal Head Segmentation

2.1.1. Traditional Approaches

Many works have used a variety of machine learning algorithms for fetal head segmentation. One example is the probabilistic boosting tree (PBT), which has been utilized for AC measurement [28]. A random Hough transform approach developed by Lu et al. [29] has been used to recognize incomplete ellipses in images with severe noise. However, their method may fail to detect the fetal head in low-contrast ultrasound images. Zhang et al. [30] developed multi-scale and multi-directional filter banks to extract anatomical structures and texture characteristics from fetal anatomical structure and texture images. Li et al. [31] used a prior knowledge of fetal head circumference to obtain the region of interest with random forest and detect the fetal head edge with phase symmetry. They found that their method performed poorly on fitting the fetal skull from ultrasound images with partially missing features taken in late pregnancy. A complex approach, such as [10], retrieved the HC by using haar-like characteristics to train a random forest classifier to detect the fetal skull, and employed the Hough transform, dynamic programming, and elliptical fitting. Even though these previous approaches produced promising findings, they were only tested on small datasets of specific pregnancy trimesters, and fetal ultrasound images at different stages of pregnancy vary in their inherent characteristics. Therefore, aspects such as the efficiency and accuracy of current traditional methods for automatic fetal head segmentation and HC biometry performance need to be improved because with current limitations, they are not adequate for accurate and reliable diagnosis by physicians.

2.1.2. Deep Learning

Deep learning techniques began to grow in popularity because of advancements in technology. This method had significantly better skills in image processing tasks due to their promising capabilities. In particular, CNN has emerged as a top choice for medical image classification, segmentation, and object detection [4]. UNet [22] is a network often used for biomedical image segmentation because of the symmetric structure observed in the images, allowing for the efficient use of skip connection layers and the reduced computing complexity. First, a feature map is extracted from an image via the encoders in the UNet architecture. Then, the decoders cascade their corresponding encoded feature maps to extract even more spatial information from the image. Several modified U-shape networks [32–34] been used to segment fetal ultrasound images, and have achieved notable results. The segmented images obtained can be utilized to detect the elliptic fetal skull and calculate the fetal HC. Sobhaninia et al. [32] proposed a multi-task deep network structure based on the LinkNet topology. They segmented fetal ultrasound images using LinkNet [35] capabilities. Their experimental results revealed that multi-task learning yields better segmentation outcomes than a single-task network. Qiao and Zulkernine [36] presented an expanded UNet model [22] with dilated convolution layers and Squeeze-and-Excitation (SE) blocks to enhance segmentation of the fetal skull border and skull in 2D ultrasound images. They used dilated convolution extracting features from a more extensive spatial range to detect edges without increasing the model complexity, and to measure fetal HC.

Desai et al. [37] proposed the DUNet architecture based on the UNet. The image and its scattering coefficients (SC) are inputs for the DUNet. Each of these inputs has an encoder. The encoders' outputs are combined and sent into a single decoder, eliminating data augmentation and reducing the training time. Aji et al. [38] utilized UNet with pixel-wise classification to increase ROI image classification performance. Each pixel is divided into four classes: maternal networks have horizontal direction patterns, higher head borders have concave arc patterns, lower head boundaries have convex arc patterns,

and the rest. The LinkNet network [35] was used as inspiration for the multi-scale and low-complexity structure of the proposed network by Sobhaninia et al. [39]. They were able to lower the number of convolutional layers in mini-LinkNet. The LinkNet network includes four encoder blocks; however, the mini-LinkNet network has just three encoder blocks, which appear to be more efficient and may retain image characteristics. These researchers demonstrate that employing a light network for the segmentation of the fetal head can lead to the intended result. Brahma et al. [40] proposed accurate binary DSCNNs for medical image segmentation. The networks' encoder and decoder structures use parameter-free skip connections to binarize them. Asymmetric encoder–decoder DSCNNs feature pyramid networks with asymmetric decoders and spatial pyramid pooling with atrous convolutions are evaluated on the fetal head image. An intensely supervised attention-gated (DAG) VNet method was introduced by Zeng et al. [41] for automated two-dimensional ultrasound image segmentation of the fetal head. Attention gates (AGs) and deep supervision were added to the original VNet architecture. Multi-scale loss functions for deep supervision were also introduced. The suggested DAG VNet technique increased segmentation accuracy while increasing the convergence speed by including the attention mechanism and deep supervision strategy. Xu et al. [42] proposed a vector self-attention layer (VSAL) and a context aggregation loss (CAL) in CNN. Geometric priors and multi-scale calibration were developed for long-range spatial reasoning. Unlike nonlocal neural networks, VSAL could concurrently attend to spatial and channel information, and VSAL consider multi-scale information by applying geometric priors and multi-scale calibration. They also introduced context aggregation loss (CAL) as an additional benefit to VSAL. CAL analyzes global contextual information and intra- and inter-class dependencies. Then, they use VSAL as the backbone to replace the convolutional layers. The suggested VSAL outperforms various mainstream methods on prenatal images. It also shows the method's adaptability to various segmentation networks. Skeika et al. [43] presented an innovative approach for automatically segmenting a fetal head in 2D ultrasound images. The suggested approach, called VNet-c, uses the original VNET [23] but includes several modifications. The modifications include pre-processing, batch normalization, dropout use, data augmentation, loss function, and network depth adjustments. The authors in [23] evaluated the suggested method's performance quantitatively using negative and positive rates. The fetal head and abdomen segmentation in an ultrasound image was performed by Wu et al. [44] using a cascaded FCN in combination with context information. Sinclair et al. [45] used an VGG-16 FCN to segment the fetal head in ultrasound images taken during the second trimester. Object detection is also used with fetal ultrasound images, using fast regions convolutional neural networks (R-CNN) and FCN. Al Bander et al. [46] developed a method to identify the fetal head boundary using a combination of fast R-CNN and FCN that included target localization and segmentation.

All of the works mentioned above did not consider the resource constraints and training time. To the best of our knowledge, this is the first trial to employ ensemble transfer learning for fetal head segmentation and to develop a lightweight model with low resources and less training time with respect to model accuracy.

2.2. Fetal Head Measurement

Various methods have been proposed to derive accurate geometric measurements from segmentation masks, such as the head circumference and radii. In general, most methods consider various elliptical models for representing the fetal head shape. Zhang et al. [47] proposed a method that estimates the HC from ultrasound images without segmentation. Their technique uses a regression CNN, for which they tested four networks of varying complexity and three regression losses. It is the first direct measurement of fetal head circumference without segmentation. Region-proposal CNN for head localization and centering, and regression CNN for precise HC delineation are proposed by Fiorentino et al. [48]. Then, distance fields are used to train the regression CNN. In order to make the network task of directly regressing the HC line easier, a distance field is used

to smooth the HC line. Skeika et al. [43] used their own designed algorithm to calculate HC from the predicted mask. Zeng et al. [41] used fitted ellipses to calculate HC biometric measurements based on the following formula:

$$HC = 2\pi \times Semi\,Axis\,b + (4 \times (Semi\,Axis\,a - Semi\,Axis\,b)) \tag{1}$$

where *SemiAxis a* and *SemiAxis b* are the major and minor axes of the ellipse. Qiao and Zulkernine [36], and Li et al. [49] used the direct least square fitting of ellipses to measure the HC. A second-order polynomial, such as the following, can be used to express a generic conic:

$$F(\vec{a}, \vec{x}) = \vec{a} \cdot \vec{x} = ax^2 + bxy + cy^2 + dx + ey + f = 0 \tag{2}$$

where $\vec{a} = [a, b, c, d, e, f]^T$ and $\vec{x} = [x^2, xy, y^2, x, y, 1]^T$ Aji et al. [38] used an ellipse fitting method comparable to the ElliFit method, in which the median value of the largest area's edge points is generally sought. Following these two operations, five ellipse parameters are acquired and used for elliptical representation. Once these two numbers have been calculated, they are multiplied by the pixel size of the input image. After obtaining the parameters, it is possible to approximate HC by computing the ellipse border using the following formula:

$$HC = 0.5 \times \pi \times (a + b) \tag{3}$$

where *a* and *b* are the major and minor axes of the ellipse.

In this work, we propose a geometry fitting framework for computing fetal head measurements, composed of the following processing steps: smoothing, parameterization, resampling, the linear least square minimization process for fitting an explicit model, and the accurate geometric distance between points. The model is parameterized in a way that the Jacobian and the geometric parameters of the best-fit ellipse can be computed in closed-form.

2.3. GA and EFW Calculation

In general, the starting day of the last menstrual period (LMP) is used to calculate gestational age (GA). However, in around 40% of pregnancies, the LMP is unknown or unreliable [50]. Ultrasound provides more reliable information on GA and is primarily acknowledged as the preferred approach. Ultrasound can determine GA more accurately than physical examination in most pregnancies. During the first trimester, the gestational sac mean diameter and crown–rump length (CRL) are used to determine GA. Measurements of the fetal head, torso, and extremities are most frequently used in the second and third trimesters. A combination of BPD, HC, abdominal circumference (AC), and femur length (FL) are typically measured parameters that are used to calculate the GA [51]. Many other variables have been examined and linked to GA, but few increase the accuracy of GA estimation [52].

In fetal medicine, the ultrasound estimation of fetal weight (EFW) is essential for prenatal care. EFW helps the physician to determine whether fetuses are the proper size for their gestational age (GA), small (SGA), or large (LGA) [53]. The EFW is calculated from the HC, BPD, FL, and AC measurements. The formulas of Hadlock et al. [54] were the most accurate, with the lowest Euclidean distance and the highest absolute mean error being less than 10%. Hadlock et al. [54] (Equation (4)) used HC, AC, and FL measurements with or without BPD. They found a robust connection between birth weight and EFW based on HC, AC, and FL measurements [55].

$$log_e(EFW) = 1.326 - 0.00326 \times AC \times FL + 0.0107 \times HC + 0.0438 \times AC + 0.158 \times FL. \tag{4}$$

where *AC*, *FL*, and *HC* are the measurements that are mentioned in the previous paragraph. To the best of our knowledge, this is the first trial study to employ a machine learning regression model to predict fetal GA and EFW based on the fetal head, without the need for other measures such as AC and FL.

3. Materials and Methods

3.1. Methodology

Figure 2 illustrates the workflow of a full end-to-end pipeline that was proposed to achieve the main contribution of this paper. The pipeline components are demonstrated in three main blocks, as seen in Figure 2. These blocks can be subdivided as follows:

1. Automatic segmentation: takes as an input a ultrasound image, and gives an output binary mask representing the fetal head.
   (a) Eight segmentation models are fine-tuned independently using the pretrained CNN EfficientNetB0 as the feature extractor.
   (b) The segmentation predictions of these models are integrated through ETLM.
2. Measurements extraction: from an automatically computed and smoothed binary mask, we fit an analytic explicit ellipse model that we use for computing the geometric measurements of interest, such as semi-axis and head orientation.
   (a) Image post-processing and smoothing.
   (b) Fetal head measurement.
3. GA and EFW Prediction: from measurements and manual annotations, we fit a regression model that is able to predict GA and EFW, which we validate clinically.
   (a) Generate new GA and EFW dataset and labeling.
   (b) Trained multiple regression models on the new dataset.
   (c) Clinical and longitudinal study validation.

In the following, we firstly describe the dataset used in this study and detail the various components of the framework.
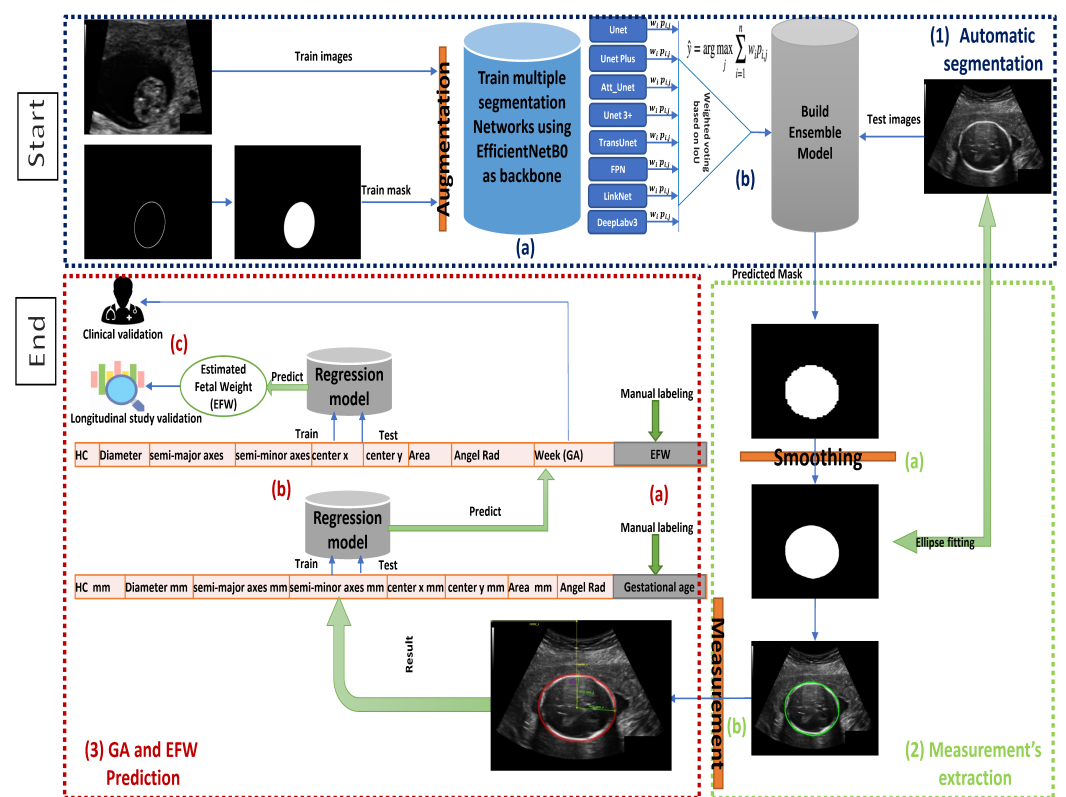

Figure 2. Workflow of the pipeline followed in this paper: Block (1) for fetal head segmentation. Block (2) for smoothing and measuring. Block (3) for fetal GA and weight prediction red.

3.2. Dataset

The dataset on which the suggested approach was evaluated is available on Grand Challenge HC18 (https://hc18.grand-challenge.org/, accessed on 21 May 2022). Table 1

shows the distribution of the dataset during various trimesters of pregnancy. The dataset consists of ultrasound images, a training set of 999 images, a CSV file containing the HC and pixels size of each image, a test set of 335 images, and a CSV file containing only the pixel size of each image. These images were taken from 551 women throughout their first, second, and third trimesters of pregnancy. The images were acquired from the Radboud University Medical Center's Department of Obstetrics in Nijmegen, Netherlands, using the Voluson E8 and the Voluson 730 (General Electric, Boston) [10]. All data were collected and anonymized by qualified sonographers following the Declaration of Helsinki. The local ethics commission authorized the data collection and usage for research purposes (CMO Arnhem-Nijmegen). Each image was 800×540 pixels in size, with pixel sizes varying from 0.052 to 0.326 mm due to sonographer modifications to accommodate varying fetus sizes. The sonographer manually marked each image by drawing an ellipse corresponding to the skull portion. The unique issues in the images are depicted in Figure 1. The difficulties included the head being in a variable location in the image, incomplete ellipse, and the fetal head's dimensions fluctuating over the gestational trimesters.

Table 1. Distribution of dataset during the various trimesters of pregnancy.

Trimesters of Pregnancy	Training Sets	Testing Sets
First trimester	165	55
Second trimester	693	233
Third trimester	141	47
Total	999	335

We augmented the dataset to increase the network's resilience, prevent overfitting of the training data, and improve the network's generalization ability. Nine images were generated for each image and mask in the training set using [56]. The final augmented training set includes: (1) Center Crop, (2) Random Rotate, (3) Grid Distortion, (4) Horizontal Flip, (5) Vertical Flip, (6) Random Brightness, (7) Sharpen, (8) Affine Transformation, (9) Fancy PCA, and (10) Original Image. The total number of training sets became 9990 images and 9990 masks.

3.3. Ensemble Transfer Learning Model (ETLM)

3.3.1. Transfer Learning

Transfer learning is the capacity of a system to recognize and apply information from one area to another. Transfer learning has three levels. First, full-adaptation uses a pre-trained network's weights and updates during training. Second, partial-adaptation starts with a pre-trained network but freezes the first few layers' weights and updates the final layers during training. Third, zero-adaptation uses a pre-trained model to establish the weights for the whole network without updating any layers [57].

This work took weights from a lightweight network (EfficientNet) and then fine-tuned them on prenatal ultrasound images. Because the dataset consists of medical images, the full-adaptation approach was used. To ensure that the best model was selected for low cost and efficiency, the lightweight EfficientNet [58] versions from B0 to B3 were utilized. EfficientNetB0 was selected based on the obtained result. EfficientNetB0 was used as the backbone (encoder) for different segmentation networks; therefore, the last block, which includes the dense layer, was removed, as seen in Figure 3.
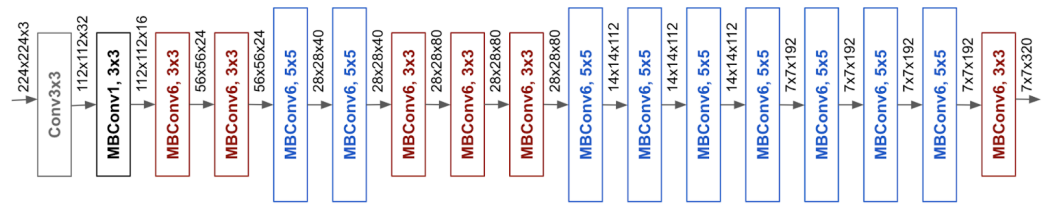
Figure 3. The architecture of EfficientNetB0.

3.3.2. Ensemble Learning

Many artificial intelligence applications have significantly benefited from the use of ensemble learning, a machine-learning approach that uses numerous base learners to construct an ensemble learner for improved generalization of the learning system. A voting ensemble (sometimes known as a "majority voting ensemble") is a type of ensemble machine learning model that incorporates predictions from several other models to arrive at a final prediction [59]. When applied effectively, it can help models perform better than any of the individual models. Voting ensembles combine the results of numerous models to arrive at a final result. For example, the predictions for each label are added together, and the label with the most votes is predicted. Almost the same results were obtained across all segmentation models in our study. Therefore, using a voting ensemble is practical when two or more models perform well on a predictive modeling task.

The models must all agree on most of their predictions for the ensemble to work. Hence, each model's contribution is proportionate to its capacity or competence in a weighted average or weighted sum ensemble. A weighted average forecast begins by assigning each ensemble member a fixed weight coefficient [60]. A percentage of the weight may be represented as a floating-point number in the range of 0 to 1. Consider a case of three-segmentation models with three fixed weights of 0.2/0.3/0.4, where larger weights indicate a better performing model. It is possible to achieve the ideal average weight using classification accuracy or negative error, depending on the competence of each model. In this work, we used Intersection Over Union (IoU) to determine the optimal average weight for each of our eight segmentation models. The following equation is the base of weighted voting ensemble learning:

$$\hat{y} = \arg\max_j \sum_{i=1}^{n} W_i P_{i,j} \tag{5}$$

where $P_{i,j}$: predicted class membership probability of the i classifier for class label j and W_i: optimal weighting parameter.

The weighted voting method was applied to eight segmentation models to find the final prediction's optimal average weight. The segmentation models include UNet [22], UNetPlus [61], AttUNet [62], UNet 3+ [63], TransUNet [64], Feature Pyramid Network (FPN) [65], LinkNet [65], and DeepLabv3 [66]. All models were trained on the same parameter. Further, the hyperparameter tuning method [67] was applied to select a set of optimal hyperparameters, including optimizer, learning rate, loss function, and trainable parameters for the eight models, as seen in Table 2.

Table 2. The selected segmentation models and their details.

Model Name	Backbone	Output Function	Normalization	One-Hot Encoding	Optimizer	Loss Function	Batch Size	Epoch	Input Size	Trainable Params
UNet										2,776,114
UNet_plus										2,389,042
Att_UNet										2,614,725
UNet 3+	EfficientNetB0	Softmax	0 to 1	0 = black pixel 1 = weight pixel	RMSprop + Scheduler Learning Rate Step Decay	Categorical Focal Jaccard loss	32	100	64 × 64 128 × 128	3,183,330
TransUNet										2,218,322
FPN										4,911,614
LinkNet										6,049,342
DeepLabv3										4,027,810

3.3.3. Image Pre-Processing

As seen in Table 2, three image preprocessing steps were applied to eliminate undesirable distortions and to highlight certain image features. The three steps can be summarized as follows:

1. Normalization: the ultrasound image intensity range is 0 to 255. Therefore, we applied a normalization technique for shifting and rescaling values to fit in a range between 0 and 1. The Normalization Formula is as follows:

$$Z = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{6}$$

where Z: the normalized value in the image, X: the original value in the image, X_{min}: the minimum value in the image, and X_{max} the maximum value in the image .

2. Resizing: The original image and mask size is 800×540 pixels; the images and masks were resized into two different sizes, and the difference between the two inputs, 64×64 and 128×128, is compared to evaluate the lightweight models and to use low-cost resources. In addition, while the original mask intensity was only two values, 0 and 255, after mask resizing, the intensity of the masks randomly ranged between 0 and 255. Therefore, the threshold of the resized masks had to be set to the original intensity, where 0 represents black pixels, and 255 represents white pixels. Finally, Softmax [68] was used as the output function; therefore, we had to encode the mask values to 0 for black and 1 for white pixels.

3. One-Hot encoding: One-hot encoding is not often used with numerical values (images). In this study, because the output function is Softmax and the loss function is categorical focal Jaccard loss, it is recommended that one-hot encoding be used. The class representing white pixels is $(0, 1)$, and the class representing black pixels is $(1, 0)$.

3.3.4. Hybrid Loss Function and Optimizer

As part of the ensemble transfer learning process, selecting the appropriate loss functions increased segmentation accuracy during subsequent inference time. Therefore, various loss functions were used for medical image segmentation [69]. This work used hyperparameter tuning to comprise the best loss function based on the IoU score. The optimal loss function was the categorical focal Jaccard loss (CFJL), which is a combination of the categorical focal loss (CFL) [70] and Jaccard loss (JL) [71], as defined below:

$$CFL(GT, PR) = -GT \cdot \alpha \cdot (1 - PR)^{\gamma} \cdot log(PR) \tag{7}$$

$$JL(A, B) = 1 - \frac{A \cap B}{A \cup B} \tag{8}$$

$$CFJL = CFL + JL \tag{9}$$

Among the different optimizers, Adam and RMSProp [72] achieve accurate segmentation. The result demonstrates that the loss value of the Adam and RMSProp optimizers was lower than the others. However, using RMSProp with schedule learning rate and step decay that drops the learning rate (LR) by a factor every few epochs, it outperformed Adam. The step decay learning rate was defined as below:

$$LR = Initial\ LR \times drop^{\left\lfloor \frac{epoch}{epochDrop} \right\rfloor} \tag{10}$$

3.4. Measurements Extraction

3.4.1. Post-Processing

Multi-smoothing and edge detection techniques were applied as post-processing to correct the defective segmented mask and improve the segmentation results. The aim was to smooth and sharpen the ellipse of the contour. Among various smoothing techniques, we employed a median filter combined with morphological image processing in our scenario, where the median filter is a non-linear digital filter that suppresses pulsed (non-stationary random process) interference by eliminating all suspicious readings. The filter calculates the median output value from a set of input data (see Equation (11)) [73].

Morphological image processing is a technique that deals with the shape or morphology of picture features. Morphological operations are well suited to the processing of binary images, since they rely solely on the relative ordering of pixel values, rather than their numerical values. Greyscale images can also be subjected to morphological techniques in which the light transfer functions are unknown, and where the absolute pixel values are of no or small importance. In our scenario, a pixel is in the neighborhood if its Euclidean distance from the origin is less than the ideal value of 25 [74]. This combination of median filter and morphological process provided the best result. Figure 2 illustrates the predicted mask before and after the smoothing.

$$\hat{f}(x,y) = \underset{(s,t) \in S_{xy}}{\text{median}}\{g(s,t)\} \tag{11}$$

where $g(s,t)$ is noise, and the median filtering method is to sort the pixels in the sliding filter window, then the output pixel value $\hat{f}(x,y)$ of the filtering result is the median value of the sequence [75].

3.4.2. HC Measurements

After the post-processing stage, the predicted mask is ready for measurements, which are obtained through fitting an ellipse model to the extracted contour. The task of fitting an ellipse model on top of scatter measurements is still considered a challenging problem by the computer vision and computational geometry community [76]. In our case, we started from the assumption that the contours extracted from generated masks are closed and smooth. To enforce this assumption, we used the preprocessing method described in [77], consisting of smoothing, parametrization and resampling, in a way where the input for the fitting procedure is a uniform angular parametrization of a given contour composed of a list of points $\mathbf{x_i} = (u(\theta_i), v(\theta_i))^T$ and a 1-to-1 mapping between angles θ_i and samples $\mathbf{x_i}$ in pixel units. Then, we used a non-linear least squares minimization process for fitting an explicit model $\mathbf{x} = \mathbf{x}(\theta)$ based on angular parametrization:

$$\mathbf{x}(\theta) = \mathbf{c} + A\mathbf{r}(\theta), \tag{12}$$

where $\mathbf{c} = (c_u, c_v)^T$ is the barycenter of the ellipse, $\mathbf{r}(\theta) = (\cos\theta, \sin\theta)^T$ is the angular unit vector, and $A = [(a_{uu}, a_{uv}), (a_{vu}, a_{vv})]$ is a 2×2 matrix mapping the unit circle to ellipse. The proposed explicit model has various advantages. First, it depends on six parameters $\text{Œ} = \{c_u, c_v, a_{uu}, a_{uv}, a_{vu}, a_{vv}\}$ all having the same dimensions (in pixels), and this makes it easy to define meaningful geometric bounds for the minimization process; second, the cost function can be computed with respect to the real geometric distance between points;

finally, the Jacobian of cost function and the geometric parameters of the best-fit ellipse can be computed in closed-form. In our case, we considered as the cost function the square geometric distance weighted with the curvature computed for each point and regularized with the Tikhonov term for avoiding that the Jacobian matrix becomes singular during the minimization process:

$$C(\text{Œ}) = \sum_i w_i \|\mathbf{x}_{\text{Œ}}(\theta_i) - \mathbf{x_i}\|^2 + \tau \|\text{Œ}\|^2, \tag{13}$$

where $w_i = \kappa_i = \frac{1}{R_i} = \frac{1}{\|\mathbf{x_i} - \mathbf{c}\|}$ is an estimate of the curvature of the ellipse in the point $\mathbf{x_i}$ and τ is a small regularization constant (in our experiments $\tau = 10^{-8}$). Hence, the fitting problem can be stated as finding the set of parameters Œ for minimizing the cost function:

$$\text{Œ}_{\text{opt}} = \arg\min_{\text{Œ}} C(\text{Œ}), \tag{14}$$

which can be solved using standard methods, like the Levenberg–Marquardt (LM) [78] or Trust Regions (RTS) [79]. In our experiments, we tried both methods as implemented in the Python scipy module, without noticeable differences in the fitting accuracy. As initial values for the minimization process, we used the parameters extracted from the bounding box of the contour.

Once the parametric representation of the ellipse was recovered, the geometric measurements can be computed in a closed form. Specifically, the semi-axes length and vectors can be computed by finding the extrema of the square distance between the ellipse and the center of the ellipse. According to the parametric model:

$$\theta_{ext} = \arg\min_{\theta} \|\mathbf{x}(\theta) - \mathbf{c}\|^2 = \arg\min_{\theta} \|A\mathbf{r}(\theta)\|^2, \tag{15}$$

leading to the equation $A\,\dot{\mathbf{r}} \cdot A\mathbf{r} = 0$, with solution

$$\theta_{ext} = \frac{1}{2} \arctan \frac{(a_{uv}^2 + a_{vv}^2) - (a_{uu}^2 + a_{vu}^2)}{2(a_{uu}a_{uv} + a_{vu}a_{vv})} + k\frac{\pi}{2} \tag{16}$$

from which the semi-axes vectors can be directly computed. As seen in Figure 4, the measurements of interest include:

1. center x: represents the length in millimeters between the image's beginning pixel on the *x*-axis and the ellipse's middle pixel.
2. center y: represents the length, in millimeters, between the image's beginning pixel on the *y*-axis and the ellipse's middle pixel.
3. semi-axes a: Once the ellipse's center is determined, the semi-axes determine the radius's maximum value based on the distance between the ellipse's middle and its farthest point.
4. semi-axes b: Once the ellipse's center is determined, the semi-axes determine the radius's minimum value based on the distance between the ellipse's middle and its nearest point.
5. angle: contains the radian value of the angle formed by the center y and the semi-axis b.
6. area: is the size of the area in millimeters that represent the fetal head.

From previous values, the equivalent diameter, biparietal diameter (BPD), occipitofrontal diameter (OFD), and HC were calculated based on the following formula (Equation (20) [6]):

$$Equivalent\ diameter = semi\ axes\ a + semi\ axes\ b \tag{17}$$

$$BPD = semi\ axes\ b * 2 \tag{18}$$

$$OFD = semi\ axes\ a * 2 \tag{19}$$
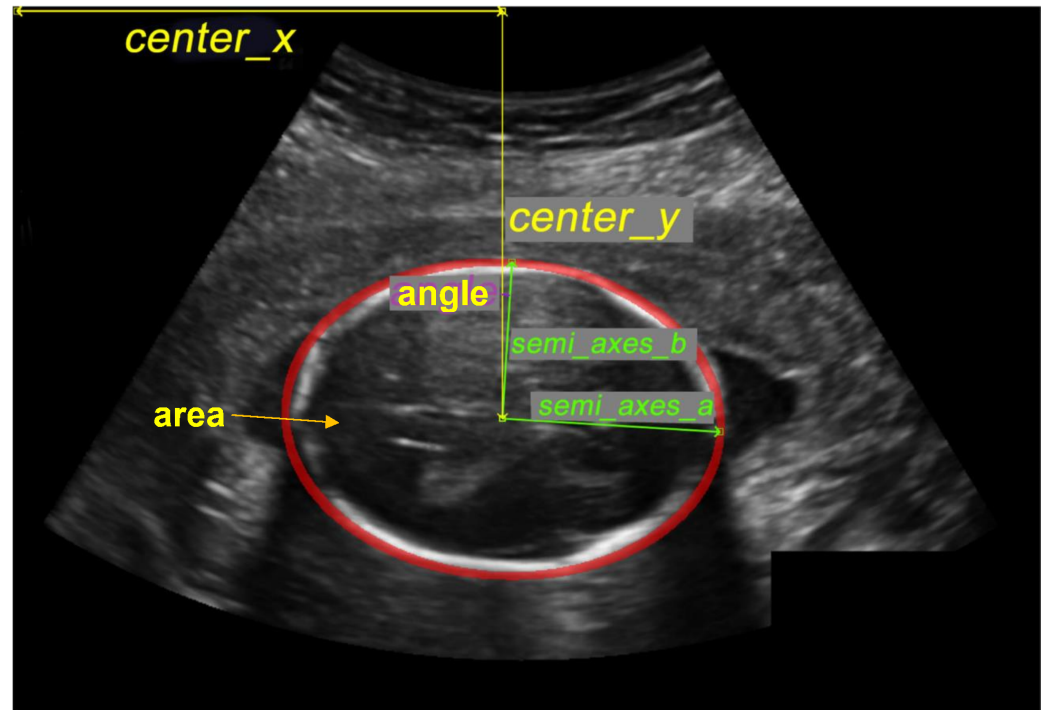
$$HC = \pi(BPD + OFD)/2 \tag{20}$$

Figure 4. Illustration of the fetal head measurement.

To ensure that the formula that we obtained from [6] for calculating HC is more accurate than that formally used in [41], the mean difference was calculated to compare both the formulas with the HC ground truth, which was given for the whole training set. Table 3 shows that our HC measurement is the closest to the HC ground truth.

Table 3. Comparing two HC measurement formulas with the HC ground truth using the mean difference.

Formula	Mean HC of the GT	Mean HC by Each Formula	Mean Difference
Our formula	174.3831 mm	174.2411 mm	**−0.14203**
Other Formula		178.3705 mm	3.9874

3.5. GA and EFW Prediction

After completing the segmentation and fetal head measurements in the previous section, eight values (features) that represent the fetal head were obtained. These values are needed to generate a new dataset for fetal GA and EFW prediction.

3.5.1. Fetal Gestational Age Dataset

In the domain of fetal size and dating, Altman and Chitty [80] proposed a new formula for calculating the gestation age based on HC; later Loughna et al. [6] proved that this formula is only accurate when the fetal age is between 13 to 25 weeks. Therefore, this study used the formula recommended by Altman and Chitty [80] to label the new dataset manually, but only included GA from 13 to 25 weeks. Finally, the new dataset was used

to train multi-regression models and predict GA from 10 to 40 weeks to overcome the limitation of the original formula:

$$log_e(GA) = 0.010611 \times HC - 0.000030321 \times HC^2$$
$$+0.43498 \times 10^{-7} \times HC^3 + 1.848$$

$$GA = \exp\left(log_e(GA)\right) \tag{21}$$

Table 4 shows that we generated the new dataset from both the training and testing images. The new dataset was split into three partitions for training (13–25 weeks), validation (10–40 weeks), and testing ($GA < 13$ and $GA > 25$ weeks). The purpose of the validation set was to select the optimum regression model. The test set is used to compare the efficiency of the selected model, with results being obtained by an expert doctor. The mean square error (MSE) was used to evaluate different regression models, and Pearson's r [81], to measure the statistical association between the predicted results by the regression models and the physician results based on test dataset GA prediction.

Table 4. Fetal gestational age dataset.

	GA Validation	**GA Training**	**GA Testing**
Dataset	(10–40) weeks	(13–25) weeks	GA < 13 GA > 25
Training	999	692	307
Testing	335	232	103
Total	1334	924	410

3.5.2. Fetal Weight Dataset

Estimated Fetal Weight (EFW) is calculated based on Hadlock's formula [54], which required a pre-knowledge of HC, BPD, AC, and FL. In addition, Salomon et al. [53] proposed a polynomial formula to find a new reference chart for EFW calculation which only required the knowledge of GA. This new formula (see Equation (22)) is used to estimate fetal weight in grams based on fetal GA from 20 to 36 weeks. This formula was used to label the new dataset manually, but only fetal weights for GAs between 20 to 36 weeks were used in this dataset. The new dataset was then used to train multi-regression models and predict the EFW from 10 to 40 weeks to overcome the limitations of the original formula:

$$EFW = -26256.56 + 4222.827 \times GA - 251.9597 \times GA^2$$
$$+6.623713 \times GA^3 - 0.0628939 \times GA^4 \tag{22}$$

Table 5 shows that we generated the new dataset from both training and testing images. The new dataset was split into three partitions for training (20–36 weeks), validation (10–40 weeks), and testing ($GA < 20$ and $GA > 36$ weeks). The purpose of the validation set is to select the optimum regression model for fetal weight prediction. The test set is used to compare the efficiency of the selected model, with results being obtained from longitudinal reference [82]. The mean square error (MSE) was used to evaluate different regression models, and Pearson's r [81] was used to measure the statistical association between predicted results by the regression models and longitudinal reference [82].

Table 5. Estimated fetal weight dataset.

	EFW Validation	EFW Training	EFW Testing
Dataset	(10–40) weeks	(20–36) weeks	GA < 20 GA > 36
Training	999	551	448
Testing	335	175	160
Total	1334	726	608

4. Experiments

In this section, the experiment set up is identified, and the three levels of evaluation for the segmentation model and the two levels of evaluation for the GA and EFW predictions are explained.

4.1. Training

This study's experiments were performed on a graphics workstation, with Intel(R) Core(TM) i9-9900K CPU @ 3.60 GHz, NVIDIA GeForce RTX 2080 Ti 11 GB, and 64 G RAM. The popular Tensorflow 2.6.0 and Keras 2.4.0 were chosen for the deep learning framework. All segmentation models were trained using the same hyperparameter settings as seen in Table 2; each model was trained for 100 epochs, and the training time was reported. The input size of model training for the first experiment was 64×64, and the second was 128×128.

4.2. Segmentation Models Evaluation

Three levels of evaluation were conducted to quantitatively analyze and evaluate the segmentation model's performance, as seen in Table 6.

Table 6. Evaluation levels for segmentation model.

	Level 1 Segmentation Evaluation	Level 2 Post-Processing Evaluation	Level 3 Measurement Evaluation
Total	Training 80% Validation 20%	Validation 100%	Validation 100%
Augmented	7992 1998		
Training Set		999	
Testing Set			335

4.2.1. Level 1: Segmentation Evaluation

Eight indices (Equations (23)–(30)) were used to evaluate segmentation model performance. These indices included area under the curve (AUC), accuracy (ACC), mean intersection over union (mIoU), precision (Pre), recall, dice similarity coefficient (DSC), mean squared error (MSE), and mean pixel accuracy (mPA), as defined below:

$$True\ Positive\ Rate(TPR) = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

$$False\ Positive\ Rate(FPR) = \frac{False\ Positive(FP)}{False\ Positive(FP) + True\ Negative(TN)}$$

$$AUC = TPR - FPR \tag{23}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{24}$$

$$mIoU(U, V) = \sum_{i=1}^{C} \frac{|U \cap V|}{|U \cup V|} = \sum_{i=1}^{C} \frac{TP}{TP + FP + FN} \tag{25}$$

$$Pre = \frac{TP}{TP + FP} \tag{26}$$

$$Recall = \frac{TP}{TP + FN} \tag{27}$$

$$DSC = \frac{2 * Pre * Recall}{Pre + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{28}$$

$$MSE = \sum_{i=1}^{C} (G_i - P_i)^2 \tag{29}$$

$$mPA = \frac{\sum_{i=1}^{C} TP}{\sum_{i=1}^{C} TP + FP} \tag{30}$$

4.2.2. Level 2: Post-Processing Evaluation

This study compared predicted masks using different models with ground truth masks to evaluate the predicted mask in terms of quality assessment and smoothing (post-processing). For this purpose, five indices [83] (Equations (31)–(35)) were used, including mean Hausdorff distance (mHD), mean surface distance (MSD), relative volume difference (RVD), mean structural similarity index (MSSIM), and peak signal-to-noise ratio (PSNR):

$$mHD(P, G) = \frac{\frac{1}{P} \sum_{p \in P} max\, d(p, g) + \frac{1}{G} \sum_{g \in G} max\, d(p, g)}{2} \tag{31}$$

$$MSD = \frac{1}{2} \left[\hat{d}(S_p, S_g) + \hat{d}(S_g, S_p) \right] \tag{32}$$

$$RVD(P, G) = \frac{|G| - |P|}{|P|} \tag{33}$$

$$MSSIM(G, P) = \frac{1}{M} \sum_{j=1}^{M} SSIM(g_j, p_j) \tag{34}$$

$$PSNR(P, G) = 10\, log\left(\frac{255^2}{MSE(P, G)} \right) \tag{35}$$

4.2.3. Level 3: Measurement Evaluation

To ensure the set of values obtained through the measurement algorithm, three indices (Equations (28), (31) and (36)) were used to evaluate the test dataset, including mHD, DCS, and mean absolute difference (MAD), as defined below:

$$MAD = \frac{\sum |HC_p - HC_g|}{n} \tag{36}$$

4.3. Evaluation of GA and EFW Prediction

Regression models were used for the estimated fetal GA and EFW predictions. MSE (Equation (29)) was used to evaluate and select the best regression model. Pearson's r [81] (Equation (37)) was used to evaluate the predicted value (GA and EFW) by calculating the statistical association between our model, the medical doctor, and the longitudinal reference.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{37}$$

where r is correlation coefficient, x_i is the values of the x-variable in a sample, $\bar{x}$ is the mean of the values of the x-variable, y_i is the values of the y-variable in a sample, $\bar{y}$ is the mean of the values of the y-variable.

5. Results and Discussion

The first part of this section presents the obtained results for the different models' segmentation efficiency, mask quality assessment (post-processing), measurement performance, and a comparison with the previous state-of-the-art. The second part presents the obtained result for the fetal GA and EFW regression models' efficiency and clinical validation.

5.1. Segmentation Performance

Figure 5 shows that all models obtained a validation score above 0.98 IoU during training. The FPN reached 0.9861 IoU, which is slightly better than other models. It is a 0.04 IoU improvement, compared to the lower performing model LinkNet, which has a 0.9825 IoU. UNet3+ obtained the second-best value but took a long time to train, as seen in Table 7. Therefore, UNet3+ was excluded from the weighted voting algorithm. LinkNet, Deeplab3,TransUNet, and UNet Plus obtained low scores of 0.982, 0982, and 0.983, respectively; therefore, they were excluded during the weighted voting algorithm. The FPN, UNet, and AttUNet models obtained the highest IoU score with a low training time. These models were used to perform weighted voting and to select the optimum weight for our ETLM. Table 7 reports eight indices that are used to evaluate each model's segmentation performance [84].
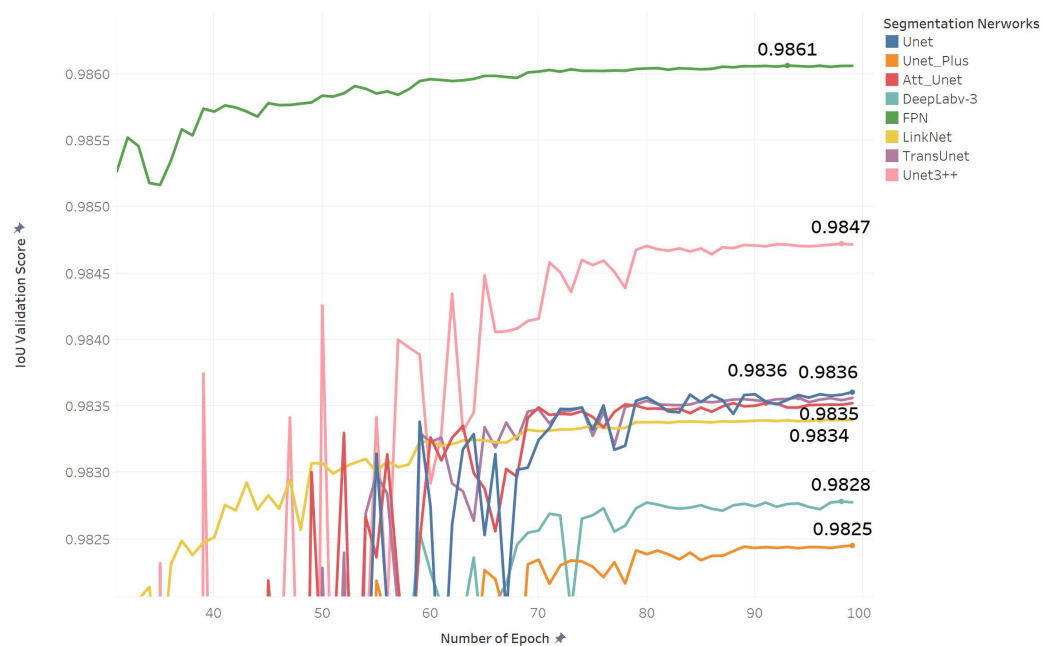

Figure 5. Segmentation networks performance based on IoU validation score during training with 128×128 input size.

The overall result proves that transfer learning using EfficientNetB0 achieved promising results, despite a low input size and less training time. Therefore, this study proves that transfer learning can develop a lightweight model, which was a challenge for medical image segmentation tasks. With an input size of 128×128 and no augmentation, results may vary from one model to another. It can be seen from the two indices, mIoU and MSE, that the FPN and AttUNet achieved the best result with the average training time. Further, with input size 64×64 and augmentation, ETLM outperformed all other models in terms of ACC, mIoU, Pre, Recall, DSC, AUC, and mPA. In the case of input size 128×128 with augmentation, ETLM outperformed all other models in terms of ACC,

mIoU, Pre, Recall, AUC, MSE, and mPA. Finally, as seen in Table 7, all indices reported during validation showed that ensemble learning could add slight improvements to the segmentation model and predict image masks. However, these new predicted masks had to be post-processed for edge smoothing and required quality assessment tests, as discussed in the following subsection.

Table 7. Level one: performance evaluation and comparison of segmentation results for all models with various input sizes, and with and without augmentation.

Model Trained with Input Size	Network	Augmentation	ACC	mIoU	Pre	Recall	DSC	AUC	MSE	mPA	Time (min)
128 × 128	UNet	No	0.9855	0.9667	0.9780	0.9740	0.9854	0.9820	0.014	0.9711	0:11:02
	UNet_plus		0.9852	0.9662	0.9710	**0.9815**	0.9851	0.9842	0.014	0.9665	0:10:43
	Att_UNet		0.9862	0.9680	0.9769	0.9787	**0.9862**	0.9841	**0.013**	0.9721	0:11:35
	UNet 3+		0.9856	0.9671	0.9770	0.9766	**0.9856**	0.9830	0.014	0.9693	0:25:20
	TransUNet		0.9852	0.9662	0.9783	0.974	0.9852	0.9821	0.014	0.9756	0:12:08
	FPN		**0.9866**	**0.9693**	0.9790	0.9778	0.9860	0.9840	**0.013**	0.9730	0:13:29
	LinkNet		0.9857	0.9673	0.9770	0.9760	0.9856	0.9830	0.014	0.9692	0:12:14
	Deeplabv3		0.9852	0.9660	**0.9791**	0.9727	0.9845	0.9817	0.014	**0.9763**	0:11:04
64 × 64	UNet	Yes	0.9917	0.9810	0.9870	0.9859	0.9916	0.9900	0.008	0.9870	0:39:00
	UNet_plus		0.9898	0.9767	0.9843	0.9833	0.9896	0.9880	0.010	0.9815	0:38:18
	Att_UNet		0.9919	0.9815	0.9881	0.9863	0.9919	0.9900	0.008	0.9875	0:40:38
	UNet 3+		0.9920	0.9816	0.9883	0.9862	0.9919	0.9904	0.007	0.9892	1:16:44
	TransUNet		0.9913	0.9802	0.9873	0.9851	0.9912	0.9896	0.008	0.9873	0:44:44
	FPN		0.9926	0.9831	0.9887	0.9878	0.9925	0.9913	**0.007**	0.9886	0:48:51
	LinkNet		0.9912	0.9800	0.9868	0.9854	0.9911	0.9896	0.008	0.9860	0:46:13
	Deeplabv3		0.9908	0.9790	0.9869	0.9838	0.9903	0.9889	0.009	0.9842	1:07:17
	ETLM		**0.9928**	**0.9841**	**0.9892**	**0.9881**	0.9934	**0.9918**	0.008	**0.9904**	NA
128 × 128	UNet	Yes	0.9928	0.9820	0.9888	0.9886	0.9928	0.9917	0.007	0.9898	0:37:15
	UNet_plus		0.9923	0.9807	0.9879	0.9879	0.9922	0.9911	0.007	0.9877	0:35:10
	Att_UNet		0.9928	0.9819	0.9887	0.9885	0.9927	0.9916	0.007	0.9891	0:38:59
	UNet 3+		0.9933	0.9832	0.9900	0.9890	**0.9933**	0.9921	0.006	0.9908	1:40:12
	TransUNet		0.9928	0.9819	0.9890	0.9884	0.9927	0.9916	0.007	0.9892	0:38:29
	FPN		0.9939	0.9846	0.9908	0.9899	0.9938	0.9928	0.006	0.9905	0:42:47
	LinkNet		0.9927	0.9817	0.9892	0.9879	0.9926	0.9914	0.007	0.9886	0:36:30
	Deeplabv3		0.9926	0.9828	0.9886	0.9878	0.9923	0.9913	0.007	0.9884	0:43:11
	ETLM		**0.9942**	**0.9853**	**0.9913**	**0.9903**	0.9908	**0.99316**	**0.005**	**0.9914**	NA

5.2. Measurements Performance

5.2.1. Post-Processing Evaluation

Figure 6 presents samples of original images, predicted masks after post-processing, ground truths, and ellipse fitted masks; however, it is challenging to identify differences and similarities by looking at the predicted masks and ground truths. Therefore, this study performed a mask quality assessment test shown in Table 8, to prove that the promised result obtained during the level one evaluation is realistic and reliable.

Table 8 shows a comparison between two distinct groups of predicted masks; the first group was predicted using various segmentation networks trained with a 64 × 64 input size. The other group used networks that trained with a 128 × 128 input size. In both cases, the results indicate that ETLM is more like the ground truth mask, where minimum mHD, MSD, RVD, and maximum MSSIM and PSNR were obtained using masks predicted by ETLM with post-processing. However, some results may vary slightly, as seen in the case of the 128 × 128 FPN, which obtained minimum mHD, but the ETLM performance was best in other indices. The RVD is always negative, as seen in Table 8, which means that in all cases, the predicted mask size (fetal head contour) was bigger than the ground truth in the masks predicted by different networks. However, ETLM minimized this difference to

0.0016 to achive the best similarity with the ground truth. Overall, the level two evaluation proved that the predicted masks obtained by this study's ETLM are remarkably close to the ground truth, with a difference of 0.011, as reported by MSSIM (see Figure 6).
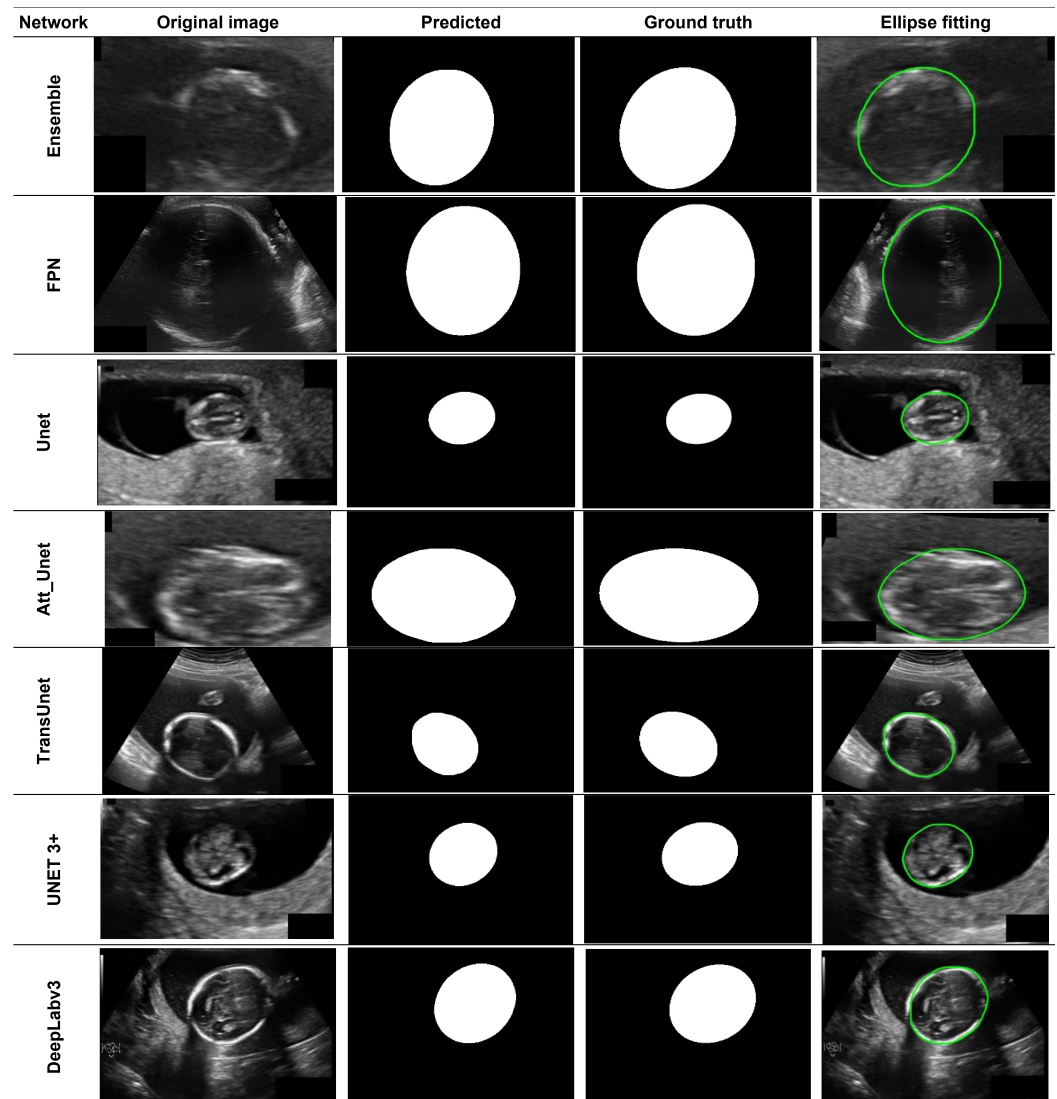

Figure 6. Qualitative comparison of segmentation performance of networks on a fetal head ultrasound image. The predicted mask, ground truth, and original image boundaries are shown. The predicted masks using different networks and the proposed ETLM are shown in the first row.

Table 8. Level two: predicted mask (post-processed) quality assessment for models with various input sizes.

Model Trained with Input Size	Network	Original Training Images	mHD (mm)	MSD (mm)	RVD	MSSIM	PSNR
	ETLM		**0.927634**	**0.0034989**	**−0.00387**	**0.98108–0.98255**	**25.142206**
	FPN		1.186636	0.0049680	−0.01237	0.97322–0.97544	23.47897
64 × 64	UNet	Yes	1.118771	0.0048532	−0.01213	0.97352–0.9757270	23.5358
	Att_UNet		1.512662	0.0049149	−0.01222	0.973301–0.9755263	23.505971
	Trans_UNet		1.118771	0.0049047	−0.01208	0.97344304–0.97563850	23.50993
	ETLM		0.753095	**0.0018117**	**0.001639**	**0.989922–0.990706**	**28.247806**
	FPN		**0.625412**	0.0020034	−0.00264	0.9888480–0.9896689	27.536022
128 × 128	UNet	Yes	1.250824	0.0020566	−0.00196	0.98856–0.989421	27.484995
	Att_UNet		0.988862	0.0020950	−0.00177	0.988375–0.989247	27.41142
	Trans_UNet		0.753095	0.0020579	−0.00243	0.988523–0.98937365	27.43699

5.2.2. Fetal Head Measurement Evaluation

Fetal head measurements were evaluated on the testing dataset, which consisted of 355 images. Unfortunately, the ground truth for this dataset is not available to the public; therefore, the measurement evaluation result was obtained by submitting measurement values to the dataset website (https://hc18.grand-challenge.org/ 11 August 2022)and obtaining the mHD, MAD, and DCS, as shown in Table 9.

Table 9. Level three: measurement evaluation based on testing dataset.

Model Trained with Input Size	Network	Original Testing Images	mHD (mm)	MAD (mm)	DCS
128 × 128	ETLM	Yes	1.6715	1.8735	0.9716

5.3. Comparative Analysis

Table 10 provides a comprehensive comparison between our ETLM and the published results reported in the literature. First, the ETLM outperformed the state-of-the-art models in the segmentation task regarding ACC, mIoU, Pre, and mPA. Second, the results of this study are better than [32,36,39,43,47], in terms of MAD, and better than [32,39,42] in terms of mHD. However, the result in this study is inferior to the results found in [41,49] because the models used in those studies were heavy and trained for more than 30 h with high input resolution, making the models very expensive in terms of required resources and time. Finally, a model weight comparison showed that the lightweight ETLM used in this study is superior, because promising results with very low resolution (128 × 128) and less training time (2 h) were achieved. This study proves that ensemble and transfer learning overcomes medical image segmentation challenges such as low image intensity, the need for expensive resources, long training time, and heavy model deployment.

Table 10. Comprehensive comparison with state-of-the-art models.

Type of Comparison	Segmentation				Measurement			Model Weight				
Network	ACC	mIoU	Pre	mPA	DSC	MAD	mHD	Model Trained with Input Size	Bach Size	GPU RAM	Epochs	Training Time
ETLM [UNet: Att_UNet: FPN] [0.3: 0.3: 0.4]	**0.9942**	**0.9853**	**0.9913**	**0.9914**	0.9716	1.87	1.67	**128 × 128**	**32**	**11 GB**	**300**	**2:01 h**
VNet-c [43]	0.9888	0.9594	0.9767	NA	0.9791	1.89	NA	512 × 512	4	6 GB	300	53:35 h
VSAL [42]	NA	NA	N/A	0.990	0.9710	NA	3.234	256 × 256	4	24 GB	100	17:30 h
SAPNet [49]	NA	0.9646	NA	0.9802	0.9790	1.81	**1.22**	480 × 320	10	11 GB	700	NA
Regression CNN [47]	NA	NA	NA	NA	0.9776	1.90	1.32	800 × 800	16	NA	1500	NA
DAG V-Net [41]	NA	NA	NA	NA	**0.9793**	**1.77**	1.27	768 × 512	2	11 GB	20	30 h
MTLN [32]	NA	NA	NA	NA	0.9684	2.12	1.72	800 × 540	NA	11 GB	200	15 h
UNet [36]	NA	NA	NA	NA	0.9731	2.69	NA	216 × 320	4	32 GB	100	NA
DSCNN [40]	NA	NA	NA	NA	0.9689	NA	NA	NA	NA	NA	NA	NA
MS-LinkNet [39]	NA	NA	NA	NA	0.9375	2.27	3.70	NA	10	11 GB	150	18 h

5.4. GA and EFW Prediction Performance

For fetal GA and EFW prediction, we trained 17 regression models on each dataset independently. Because the dataset contains large numerical values, a log transformation was applied to both datasets before training, making the highly skewed distributions less skewed. The performance of each model was evaluated using MSE, and the result was reported in Table 11. This task aimed to address the limitation of both formulas (see Equations (21) and (22)) used to estimate the GA and EFW. Therefore, the regression model was used to predict GA for the fetus when the GA of the fetus was $13 > GA > 25$, and the EFW for the fetus when the GA of the fetus was $20 > GA > 36$. In both cases, the ground truth was non-existent because both formulas had limitations, and a GA and EFW could not be calculated in the mentioned periods; therefore, the following steps were taken:

1. Validation of predicted GA: 50 random samples images taken from the testing set ($13 > GA > 25$) were given to a senior attending physician with 21 years of experience in maternal-fetal medicine, to estimate GA. We used Pearson's r to measure the strength of a linear association between the physician prediction and the model prediction for the same sample set. Because we do not have any pre-knowledge of the dataset in terms of ethnicity or location, the GA may vary based on these factors; therefore, in this work, we tried to predict the GA in the 50th percentile, and considered the median.

2. Validation of predicted EFW: In the case of EFW, the senior physician could not estimate the EFW based on fetal head images and required more factors such as FL, AC, and CRL. Therefore, a growth chart taken from a longitudinal reference was used for estimated fetal weight, regardless of fetal sex [82]. Then, Pearson's r was used to measure the strength of the linear association between the longitudinal reference and the model prediction for the same sample set that fell in the range of $20 > GA > 36$. This study tried to predict the EFW in the 50th percentile and considered the median for the above mentioned reason.

Table 11, shows that most regression models achieved a promising result in GA and EFW datasets based on MSE. In the GA validation dataset, polynomial regression and Deep NN achieved a lower MSE of 0.0003 and 0.00072, respectively. However, to ensure the reliability of each model, all models were used to predict the 50th percentile of GA. The predicted GA was then compared with the physician's estimations using Pearson's r. After comparing the predicted GA with the physician's estimation, Table 11 shows that Deep NN and polynomial regression outperformed all regression models for predicting the GA, with Pearson's r of 0.9978 and 0.9958, respectively.

For Fetal EFW, LinearSVR, XGBRFRegressor, and linear regression achieved the lower MSE in the EFW validation dataset, as reported in Table 11. Nonetheless, all the models were used to predict the 50th percentile of EFW in the test dataset to ensure the reliability of each model's prediction. Then, it was compared with the longitudinal reference table, as seen in Appendix A1. As a result, Pearson's r showed that LinearSVR outperformed all the models and predicted the EFW in the 50th percentile with the highest association with the longitudinal reference (r = 0.9989). In addition, XGBRFRegressor showed a low MSE during validation, and a low association with the longitudinal reference.

Overall, most regression models could predict the GA and EFW in the 50th percentile, as seen from Pearson's results in Table 11. It is concluded that the regression models in this study address the limitations of the formulas currently used to calculate GA and EFW in the specific period. Without limitation, these models only required measurement of the fetal head to calculate GA and EFW from the 10th week to the 40th week. This study is the first work that utilized machine learning to predict the GA and EFW based on fetal head images. A sample of model prediction for GA and EFW was added to (Supplementary File S1 and File S2), respectively.

Table 11. Result and validation of multiple regression models for GA and EFW prediction.

	Fetal GA Prediction in the 50th Percentile (13 > GA > 25) Week		EFW Prediction in the 50th Percentile (20 > GA > 36) Week	
Regression model	MSE	Pearson's r	MSE	Pearson's r
Polynomial Regression	**0.00033**	**0.9958**	9.08723	0.9422
Linear Regression	0.00205	0.9899	**0.00035**	**0.9988**
Random Forest Regressor	0.00842	0.9511	6.54380	0.9844
XGBRFRegressor	0.02268	0.9505	**0.00018**	0.9847
Neural network	0.01392	0.9805	0.00256	0.9946
KNeighbors Regressor	0.00921	0.9582	0.00214	0.9841
SGDRegressor	0.00219	0.9901	0.00146	0.9968
AdaBoostRegressor	0.01086	0.9505	0.00100	0.9843
BaggingRegressor	0.01081	0.9832	0.00281	0.9964
StackingRegressor	0.00824	0.9506	6.93890	0.9843
LinearSVR	0.00199	0.9901	**0.00054**	**0.9989**
LGBMRegressor	0.01011	0.9514	7.72867	0.9843
Lasso	0.08300	NA	0.17339	0.8507
VotingRegressor	0.00248	0.9909	0.00031	0.8507
BayesianRidge	0.00206	0.9899	0.00035	0.9988
Deep NN	**0.00072**	**0.9978**	0.00068	NA

6. Strength and Limitations

Including US machines in various medical settings is advised; however, this is not always feasible, due to the cost of purchasing multiple devices or portability concerns. Mobile Health companies such as Clarius (Clarius Mobile Health Corp., Vancouver, BC, Canada) [85] developed portable pocket handheld ultrasound scanners that represent a promising tool in regional anesthesia procedures and vascular accesses [86]. Furthermore, these portable devices are still examined for extensive imaging, such as prenatal scans, which require a lightweight AI system to maintain high accuracy and low resource. Therefore, in this work, we deployed lightweight architectures that can be used in portable

devices without client-server communication. The architectures resulted in fast training on low-end machines and fast inference without the need for complex client-server architecture that would pose issues for data privacy and security limitations related to image resolution that can affect measurement accuracy. In addition to fetal head segmentation, a regression model was employed to predict GA and EFW in the 50th percentile in all trimesters based on fetal head features, which current methods cannot do. Furthermore, the framework in this study can be extended to build a fully automatic AI system in the client-server to provide a detailed report for any fetal head ultrasound images.

Despite the study's strengths, the framework still has some constraints that will need to be overcome in the future. First, downsampling the original images reduced the measurement accuracy. For example, subsampling images from 128×128 to 64×64 reduced the PSNR value by 3.1 and mHD by 0.17 mm, as seen in Table 8. Second, fetal GA and EFW may vary slightly from one group to another, based on ethnicity and gender. This study did not have this information, so the 50th percentile was predicted as the median. Moreover, the clinical appliance has to be decided by medical personnel, since the existing differences between the actual image and the one generated by the proposed model could be substantial in the medical field.

7. Conclusions and Future Work

This work proposed a new pipeline that utilized transfer learning and ensemble learning to build ensemble models called ETLM. Eight segmentation networks were evaluated to build an ensemble model based on the weighted voting method for fetal head segmentation. These segmented masks were used to accurately measure HC, BPD, OFD, and other values in ultrasound images. Masks segmented by each model went through a quality assessment test to ensure the efficiency of ETLM, and were compared with other independent models. Our experimental results show that the proposed pipeline achieved comparable performance to state-of-the-art models in segmentation and measurement. Further, regression models showed that of the features obtained from the segmented fetal images to build a new dataset for GA and EFW, only fetal head images were required to predict GA and EFW. The results of this study were validated with the assistance of an expert physician and longitudinal reference. This study is the first work that provides a completed approach from image segmentation to GA and EFW prediction. Future work will include a full adoption of transfer learning based on a model trained on ultrasound images, regardless of the domain of the images. Further, a traditional machine learning classifier will be used to find the best features to reduce ultrasound images' intensity and noise. Finally, the cavum septum pellucidum and the lateral ventricle will be segmented, measured, and compared with the ultrasound machine.

Future work will include a full adoption of transfer learning based on a model trained on ultrasound images, regardless of the domain of the images. Further, a traditional machine learning classifier will be used to find the best features that will reduce the intensity and the noise in the ultrasound images. Finally, we will segment and measure the cavum septum pellucidum and the lateral ventricle, and compare our results with the ultrasound machine.

Author Contributions: M.A. (Mahmood Alzubaidi) and M.A. (Marco Agus): Conceptualization, Formal Analysis, Methodology, Writing. U.S.: Review and Editing. M.M.: Validation. K.A. and M.H.: Supervision, Review, Editing. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Appendix A. Longitudinal Reference

Table A1. Growth chart for estimated fetal weight regardless of fetal sex.

Gestational Age (Weeks)	Estimated Fetal Weight (g) by Percentile								
	2.5	5	10	25	50	75	90	95	97.5
14	70	73	78	83	90	98	104	109	113
15	89	93	99	106	114	124	132	138	144
16	113	117	124	133	144	155	166	174	181
17	141	146	155	166	179	193	207	217	225
18	174	181	192	206	222	239	255	268	278
19	214	223	235	252	272	292	313	328	340
20	260	271	286	307	330	355	380	399	413
21	314	327	345	370	398	428	458	481	497
22	375	392	412	443	476	512	548	575	595
23	445	465	489	525	565	608	650	682	705
24	523	548	576	618	665	715	765	803	830
25	611	641	673	723	778	836	894	938	970
26	707	743	780	838	902	971	1038	1087	1125
27	813	855	898	964	1039	1118	1196	1251	1295
28	929	977	1026	1102	1189	1279	1368	1429	1481
29	1053	1108	1165	1251	1350	1453	1554	1622	1682
30	1185	1247	1313	1410	1523	1640	1753	1828	1897
31	1326	1394	1470	1579	1707	1838	1964	2046	2126
32	1473	1548	1635	1757	1901	2047	2187	2276	2367
33	1626	1708	1807	1942	2103	2266	2419	2516	2619
34	1785	1872	1985	2134	2312	2492	2659	2764	2880
35	1948	2038	2167	2330	2527	2723	2904	3018	3148
36	2113	2205	2352	2531	2745	2959	3153	3277	3422
37	2280	2372	2537	2733	2966	3195	3403	3538	3697
38	2446	2536	2723	2935	3186	3432	3652	3799	3973
39	2612	2696	2905	3135	3403	3664	3897	4058	4247
40	2775	2849	3084	3333	3617	3892	4135	4312	4515

References

1. Mayer, D.P.; Shipilov, V. Ultrasonography and magnetic resonance imaging of uterine fibroids. *Obstet. Gynecol. Clin. N. Am.* **1995**, *22*, 667–725. [CrossRef]
2. Griffin, R.M. Fetal Biometry. *WebMD* 2020. Available online: https://www.webmd.com/baby/fetal-biometry (accessed on 11 August 2022).
3. Whitworth, M.B.L.; Mullan, C. Ultrasound for fetal assessment in early pregnancy. *Cochrane Database Syst. Rev.* **2015**, *2015*, CD007058 . [CrossRef] [PubMed]
4. Alzubaidi, M.; Agus, M.; Alyafei, K.; Althelaya, K.A.; Shah, U.; Abd-Alrazaq, A.; Anbar, M.; Makhlouf, M.; Househ, M. Toward deep observation: A systematic survey on artificial intelligence techniques to monitor fetus via ultrasound images. *iScience* **2022**, *25*, 104713. [CrossRef] [PubMed]
5. Halle, K.F.; Fjose, M.; Kristjansdottir, H.; Bjornsdottir, A.; Getz, L.; Tomasdottir, M.O.; Sigurdsson, J.A. Use of pregnancy ultrasound before the 19th week scan: An analytical study based on the Icelandic Childbirth and Health Cohort. *BMC Pregnancy Childbirth* **2018**, *18*, 512. [CrossRef]
6. Loughna, P.; Chitty, L.; Evans, T.; Chudleigh, T. Fetal Size and Dating: Charts Recommended for Clinical Obstetric Practice. *Ultrasound* **2009**, *17*, 160–166. [CrossRef]
7. Jatmiko, W.; Habibie, I.; Ma'sum, M.A.; Rahmatullah, R.; Satwika, I.P. Automated Telehealth System for Fetal Growth Detection and Approximation of Ultrasound Images. *Int. J. Smart Sens. Intell. Syst.* **2015**, *8*, 697–719. [CrossRef]
8. Schmidt, U.; Temerinac, D.; Bildstein, K.; Tuschy, B.; Mayer, J.; Sütterlin, M.; Siemer, J.; Kehl, S. Finding the most accurate method to measure head circumference for fetal weight estimation. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **2014**, *178*, 153–156. [CrossRef]
9. Noble, J.A. Ultrasound image segmentation and tissue characterization. *Proc. Inst. Mech. Eng. Part H J. Eng. Med.* **2010**, *224*, 307–316. [CrossRef]
10. van den Heuvel, T.L.A.; de Bruijn, D.; de Korte, C.L.; Ginneken, B. Automated measurement of fetal head circumference using 2D ultrasound images. *PLoS ONE* **2018**, *13*, e0200412. [CrossRef]

11. Espinoza, J.; Good, S.; Russell, E.; Lee, W. Does the Use of Automated Fetal Biometry Improve Clinical Work Flow Efficiency? *J. Ultrasound Med.* **2013**, *32*, 847–850. [CrossRef]

12. Ciurte, A.; Bresson, X.; Cuadra, M.B. A semi-supervised patch-based approach for segmentation of fetal ultrasound imaging. In Proceedings of the Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012, Barcelona, Spain, 2–5 May 2012; pp. 5–7.

13. Ponomarev, G.V.; Gelfand, M.S.; Kazanov, M.D. A multilevel thresholding combined with edge detection and shape-based recognition for segmentation of fetal ultrasound images. In Proceedings of the Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012, Barcelona, Spain, 2–5 May 2012; pp. 17–19.

14. Stebbing, R.V.; McManigle, J.E. A boundary fragment model for head segmentation in fetal ultrasound. In Proceedings of the Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI, Barcelona, Spain, 2–5 May 2012; pp. 9–11.

15. Perez-Gonzalez, J.L.; Muńoz, J.C.B.; Porras, M.C.R.; Arámbula-Cosío, F.; Medina-Bańuelos, V. Automatic Fetal Head Measurements from Ultrasound Images Using Optimal Ellipse Detection and Texture Maps. In Proceedings of the VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina, 29–31 October 2014; Braidot, A., Hadad, A., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 329–332. [CrossRef]

16. Shrimali, V.; Anand, R.S.; Kumar, V. Improved segmentation of ultrasound images for fetal biometry, using morphological operators. In Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MI, USA, 3–6 September 2009; pp. 459–462. [CrossRef]

17. Rueda, S.; Fathima, S.; Knight, C.L.; Yaqub, M.; Papageorghiou, A.T.; Rahmatullah, B.; Foi, A.; Maggioni, M.; Pepe, A.; Tohka, J.; et al. Evaluation and Comparison of Current Fetal Ultrasound Image Segmentation Methods for Biometric Measurements: A Grand Challenge. *IEEE Trans. Med. Imaging* **2014**, *33*, 797–813. [CrossRef]

18. Jardim, S.M.; Figueiredo, M.A. Segmentation of fetal ultrasound images. *Ultrasound Med. Biol.* **2005**, *31*, 243–250. [CrossRef]

19. Ahmad, M.; Qadri, S.F.; Ashraf, M.U.; Subhi, K.; Khan, S.; Zareen, S.S.; Qadri, S. Efficient Liver Segmentation from Computed Tomography Images Using Deep Learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 2665283. [CrossRef]

20. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]

21. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]

22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Strasbourg, France, 27 September–1 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]

23. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571. [CrossRef]

24. Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* **2020**, *53*, 5455–5516. [CrossRef]

25. Torres, H.R.; Morais, P.; Oliveira, B.; Birdir, C.; Rüdiger, M.; Fonseca, J.C.; Vilaça, J.L. A review of image processing methods for fetal head and brain analysis in ultrasound images. *Comput. Methods Programs Biomed.* **2022**, *215*, 106629. [CrossRef]

26. Mayer, C.; Joseph, K.S. Fetal growth: A review of terms, concepts and issues relevant to obstetrics. *Ultrasound Obstet. Gynecol.* **2013**, *41*, 136–145. [CrossRef]

27. Dudley, N.J. A systematic review of the ultrasound estimation of fetal weight. *Ultrasound Obstet. Gynecol.* **2005**, *25*, 80–89. [CrossRef]

28. Carneiro, G.; Georgescu, B.; Good, S.; Comaniciu, D. Detection and Measurement of Fetal Anatomies from Ultrasound Images using a Constrained Probabilistic Boosting Tree. *IEEE Trans. Med. Imaging* **2008**, *27*, 1342–1355. [CrossRef]

29. Lu, W.; Tan, J.; Floyd, R. Automated fetal head detection and measurement in ultrasound images by iterative randomized hough transform. *Ultrasound Med. Biol.* **2005**, *31*, 929–936. [CrossRef]

30. Zhang, L.; Ye, X.; Lambrou, T.; Duan, W.; Allinson, N.; Dudley, N.J. A supervised texton based approach for automatic segmentation and measurement of the fetal head and femur in 2D ultrasound images. *Phys. Med. Biol.* **2016**, *61*, 1095–1115. [CrossRef] [PubMed]

31. Li, J.; Wang, Y.; Lei, B.; Cheng, J.Z.; Qin, J.; Wang, T.; Li, S.; Ni, D. Automatic Fetal Head Circumference Measurement in Ultrasound Using Random Forest and Fast Ellipse Fitting. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 215–223. [CrossRef] [PubMed]

32. Sobhaninia, Z.; Rafiei, S.; Emami, A.; Karimi, N.; Najarian, K.; Samavi, S.; Reza Soroushmehr, S.M. Fetal Ultrasound Image Segmentation for Measuring Biometric Parameters Using Multi-Task Deep Learning. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 6545–6548. [CrossRef]

33. Cerrolaza, J.J.; Sinclair, M.; Li, Y.; Gomez, A.; Ferrante, E.; Matthew, J.; Gupta, C.; Knight, C.L.; Rueckert, D. Deep learning with ultrasound physics for fetal skull segmentation. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 564–567. [CrossRef]

34. Budd, S.; Sinclair, M.; Khanal, B.; Matthew, J.; Lloyd, D.; Gomez, A.; Toussaint, N.; Robinson, E.C.; Kainz, B. Confident Head Circumference Measurement from Ultrasound with Real-Time Feedback for Sonographers. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, Shenzhen, China, 13–17 October 2019; Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 683–691. [CrossRef]
35. Chaurasia, A.; Culurciello, E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4. [CrossRef]
36. Qiao, D.; Zulkernine, F. Dilated Squeeze-and-Excitation U-Net for Fetal Ultrasound Image Segmentation. In Proceedings of the 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Virtual, 27–29 October 2020; pp. 1–7. [CrossRef]
37. Desai, A.; Chauhan, R.; Sivaswamy, J. Image Segmentation Using Hybrid Representations. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 1–4. [CrossRef]
38. Aji, C.P.; Fatoni, M.H.; Sardjono, T.A. Automatic Measurement of Fetal Head Circumference from 2-Dimensional Ultrasound. In Proceedings of the 2019 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), Surabaya, Indonesia, 19–20 November 2019; pp. 1–5. [CrossRef]
39. Sobhaninia, Z.; Emami, A.; Karimi, N.; Samavi, S. Localization of Fetal Head in Ultrasound Images by Multiscale View and Deep Neural Networks. In Proceedings of the 2020 25th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran. 1–2 January 2020; pp. 1–5. [CrossRef]
40. Brahma, K.; Kumar, V.; Samir, A.E.; Chandrakasan, A.P.; Eldar, Y.C. Efficient Binary Cnn For Medical Image Segmentation. In Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13–16 April 2021; pp. 817–821. [CrossRef]
41. Zeng, Y.; Tsui, P.H.; Wu, W.; Zhou, Z.; Wu, S. Fetal Ultrasound Image Segmentation for Automatic Head Circumference Biometry Using Deeply Supervised Attention-Gated V-Net. *J. Digit. Imaging* **2021**, *34*, 134–148. [CrossRef] [PubMed]
42. Xu, L.; Gao, S.; Shi, L.; Wei, B.; Liu, X.; Zhang, J.; He, Y. Exploiting Vector Attention and Context Prior for Ultrasound Image Segmentation. *Neurocomputing* **2021**, *454*, 461–473. [CrossRef]
43. Skeika, E.L.; Luz, M.R.D.; Fernandes, B.J.T.; Siqueira, H.V.; 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. [CrossRef]
44. Wu, L.; Xin, Y.; Li, S.; Wang, T.; Heng, P.A.; Ni, D. Cascaded Fully Convolutional Networks for automatic prenatal ultrasound image segmentation. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, Australia, 18–21 April 2017; pp. 663–666. [CrossRef]
45. Sinclair, M.; Baumgartner, C.F.; Matthew, J.; Bai, W.; Martinez, J.C.; Li, Y.; Smith, S.; Knight, C.L.; Kainz, B.; Hajnal, J.; et al. Human-level Performance On Automatic Head Biometrics In Fetal Ultrasound Using Fully Convolutional Neural Networks. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 17–21 July 2018; pp. 714–717. [CrossRef]
46. Al-Bander, B.; Alzahrani, T.; Alzahrani, S.; Williams, B.M.; Zheng, Y. Improving fetal head contour detection by object localisation with deep learning. In *Medical Image Understanding and Analysis*; Zheng, Y., Williams, B.M., Chen, K., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 142–150. [CrossRef]
47. Zhang, J.; Petitjean, C.; Lopez, P.; Ainouz, S. Direct estimation of fetal head circumference from ultrasound images based on regression CNN. In Proceedings of the Third Conference on Medical Imaging with Deep Learning, Montreal, QC, Canada, 6–8 July 2020; Arbel, T., Ben Ayed, I., de Bruijne, M., Descoteaux, M., Lombaert, H., Pal, C., Eds.; PMLR , 2020; Volume 121, pp. 914–922.
48. Fiorentino, M.C.; Moccia, S.; Capparuccini, M.; Giamberini, S.; Frontoni, E. A regression framework to head-circumference delineation from US fetal images. *Comput. Methods Programs Biomed.* **2021**, *198*, 105771. [CrossRef]
49. Li, P.; Zhao, H.; Liu, P.; Cao, F. Automated measurement network for accurate segmentation and parameter modification in fetal head ultrasound images. *Med. Biol. Eng. Comput.* **2020**, *58*, 2879–2892. [CrossRef]
50. Verburg, B.O.; Steegers, E.A.P.; De Ridder, M.; Snijders, R.J.M.; Smith, E.; Hofman, A.; Moll, H.A.; Jaddoe, V.W.V.; Witteman, J.C.M. 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*, 388–396. [CrossRef]
51. Mu, J.; Slevin, J.C.; Qu, D.; McCormick, S.; Adamson, S.L. In vivo quantification of embryonic and placental growth during gestation in mice using micro-ultrasound. *Reprod. Biol. Endocrinol.* **2008**, *6*, 34. [CrossRef]
52. Butt, K.; Lim, K. Determination of Gestational Age by Ultrasound: In Response. *J. Obstet. Gynaecol. Can.* **2016**, *38*, 432. [CrossRef]
53. Salomon, L.J.; Bernard, J.P.; Ville, Y. Estimation of fetal weight: Reference range at 20–36 weeks' gestation and comparison with actual birth-weight reference range. *Ultrasound Obstet. Gynecol.* **2007**, *29*, 550–555. [CrossRef]
54. Hadlock, F.P.; Harrist, R.; Sharman, R.S.; Deter, R.L.; Park, S.K. Estimation of fetal weight with the use of head, body, and femur measurements—A prospective study. *Am. J. Obstet. Gynecol.* **1985**, *151*, 333–337. [CrossRef]
55. Hammami, A.; Mazer Zumaeta, A.; Syngelaki, A.; Akolekar, R.; Nicolaides, K.H. Ultrasonographic estimation of fetal weight: Development of new model and assessment of performance of previous models. *Ultrasound Obstet. Gynecol.* **2018**, *52*, 35–43. [CrossRef]

56. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and Flexible Image Augmentations. *Information* **2020**, *11*, 125. [CrossRef]

57. Hesamian, M.H.; Jia, W.; He, X.; Kennedy, P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *J. Digit. Imaging* **2019**, *32*, 582–596. [CrossRef]

58. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Chaudhuri, K., Salakhutdinov, R., Eds.; PMLR: Baltimore, MA, USA , 2019; Volume 97, pp. 6105–6114.

59. Yang, Y.; Lv, H. Discussion of Ensemble Learning under the Era of Deep Learning. *arXiv* **2021**, arXiv:2101.08387.

60. Polikar, R. Ensemble learning. In *Ensemble Machine Learning: Methods and Applications*; Zhang, C., Ma, Y., Eds.; Springer US: Boston, MA, USA, 2012; pp. 1–34. [CrossRef]

61. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R., Bradley, A., Papa, J.P., Belagiannis, V., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–11. [CrossRef]

62. Oktay, O.; Schlemper, J.; Le Folgoc, L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Y Hammerla, N.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.

63. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059. [CrossRef]

64. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306. https://doi.org/10.48550/ARXIV.2102.04306.

65. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]

66. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587. https://doi.org/10.48550/ARXIV.1706.05587.

67. O'Malley, Tom and Bursztein, Elie and Long, James and Chollet, François and Jin, Haifeng and Invernizzi, Luca and others. et al. KerasTuner. 2019. Available online: https://github.com/keras-team/keras-tuner (accessed on 1 April 2022).

68. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.

69. Asgari Taghanaki, S.; Abhishek, K.; Cohen, J.P.; Cohen-Adad, J.; Hamarneh, G. Deep semantic segmentation of natural and medical images: A review. *Artif. Intell. Rev.* **2021**, *54*, 137–178. [CrossRef]

70. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]

71. Bertels, J.; Eelbode, T.; Berman, M.; Vandermeulen, D.; Maes, F.; Bisschops, R.; Blaschko, M.B. Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, Shenzhen, China, 13–17 October 2019; Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 92–100. [CrossRef]

72. Zou, F.; Shen, L.; Jie, Z.; Zhang, W.; Liu, W. A Sufficient Condition for Convergences of Adam and RMSProp. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 11119–11127. [CrossRef]

73. Ning, C.; Liu, S.; Qu, M. Research on removing noise in medical image based on median filter method. In Proceedings of the 2009 IEEE International Symposium on IT in Medicine & Education, Albuquerque, NM, USA, 2–5 August 2009; Volume 1, pp. 384–388. [CrossRef]

74. van der Walt, S.; Schönberger, J.L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J.D.; Yager, N.; Gouillart, E.; Yu, T. Scikit-image: Image processing in Python. *PeerJ* **2014**, *2*, e453. [CrossRef] [PubMed]

75. Song, Y.; Liu, J. An improved adaptive weighted median filter algorithm. *J. Phys. Conf. Ser.* **2019**, *1187*, 042107. [CrossRef]

76. Hu, C.; Wang, G.; Ho, K.C.; Liang, J. Robust Ellipse Fitting With Laplacian Kernel Based Maximum Correntropy Criterion. *IEEE Trans. Image Process.* **2021**, *30*, 3127–3141. [CrossRef]

77. Al-Thelaya, K.; Agus, M.; Gilal, N.; Yang, Y.; Pintore, G.; Gobbetti, E.; Calí, C.; Magistretti, P.; Mifsud, W.; Schneider, J. InShaDe: Invariant Shape Descriptors for visual 2D and 3D cellular and nuclear shape analysis and classification. *Comput. Graph.* **2021**, *98*, 105–125. [CrossRef]

78. Gavin, H.P. *The Levenberg-Marquardt Algorithm for Nonlinear Least Squares Curve-Fitting Problems*; Duke University: Durham, NC, USA, 2019; pp. 1–19.

79. Voglis, C.; Lagaris, I. A rectangular trust region dogleg approach for unconstrained and bound constrained nonlinear optimization. In Proceedings of the WSEAS International Conference on Applied Mathematics, Corfu Island, Greece, 17–19 August 2004; Volume 7.

80. Altman, D.G.; Chitty, L.S. New charts for ultrasound dating of pregnancy. *Ultrasound Obstet. Gynecol.* **1997**, *10*, 174–191. [CrossRef]

81. Sedgwick, P. Pearson's correlation coefficient. *BMJ* **2012**, *345*, e4483. [CrossRef]

82. Kiserud, T.; Piaggio, G.; Carroli, G.; Widmer, M.; Carvalho, J.; Neerup Jensen, L.; Giordano, D.; Cecatti, J.G.; Abdel Aleem, H.; Talegawkar, S.A.; et al. The World Health Organization Fetal Growth Charts: A Multinational Longitudinal Study of Ultrasound Biometric Measurements and Estimated Fetal Weight. *PLoS Med.* **2017**, *14*, e1002220. [CrossRef]

83. Samajdar, T.; Quraishi, M.I. Analysis and Evaluation of Image Quality Metrics. In *Information Systems Design and Intelligent Applications*; Mandal, J.K., Satapathy, S.C., Kumar Sanyal, M., Sarkar, P.P., Mukhopadhyay, A., Eds.; Springer: New Delhi, India, 2015; pp. 369–378. [CrossRef]

84. Qadri, S.F.; Shen, L.; Ahmad, M.; Qadri, S.; Zareen, S.S.; Khan, S. OP-convNet: A Patch Classification-Based Framework for CT Vertebrae Segmentation. *IEEE Access* **2021**, *9*, 158227–158240. [CrossRef]

85. Lundin, A. Clarius Mobile Health Makes Leadership Changes to Accelerate Growth. *AXIS Imaging News*, 10 March 2022.

86. Strumia, A.; Costa, F.; Pascarella, G.; Del Buono, R.; Agrò, F.E. U smart: Ultrasound in your pocket. *J. Clin. Monit. Comput.* **2021**, *35*, 427–429. [CrossRef]