VAWKUM Transactions on Computer Sciences
http://vfast.org/journals/index.php/VTCS@ 2025, ISSN(e): 2308-8168, ISSN(p): 2411-6335
Volume 13, Number 1, January-June 2025 *pp: 40-53*

Automated Fetal Femur Segmentation and Length Measurement in Ultrasound Images: A Key Tool for Accurate Gestational Age Assessment

Hafiz Muhammad Danish [1*], **Zobia Suhail** [1], **Faiza Farooq** [2], **Reyer Zwiggelaar** [3]

[1]Department of Computer Science, University of the Punjab, Lahore, Pakistan; [2]Department of Radiology, University of Lahore Teaching Hospital, Lahore, Pakistan; [3]Department of Computer Science, Aberystwyth University, Aberystwyth, UK

Keywords: *Bilateral filter, Speckle noise reduction, K-means clustering, Fetal femur segmentation, 2D ultrasound images, Medical imaging.*

Journal Info:
Submitted:
March 28, 2025
Accepted:
March 7, 2025
Published:
April 15, 2025

Abstract

Accurate gestational age (GA) estimation in the second and third trimesters is crucial for effective prenatal care. It is typically determined by measuring fetal femur length (FL) in ultrasound (US) images. However, manual FL measurements are time-consuming and require expertise, leading to the need for automation. This study aims to develop an automated multi-step approach to improve FL measurement accuracy while addressing common US challenges such as speckle noise, shadows, and low signal-to-noise ratio (SNR). The proposed method includes image acquisition, preprocessing for contrast enhancement, speckle noise reduction using a bilateral filter, k-means clustering for initial femur segmentation, and morphological analysis to isolate the femur for precise FL measurement.The method had a Dice similarity coefficient of 93.18 $\pm$ –9.54% and a mean difference of measurement of 0.062 cm with a 95% limits of agreement ranging between -1.06 cm and 1.19 cm, which confirms the method as accurate and clinically reliable. These findings show that the suggested approach will improve the accuracy of the FL measurements, decrease the number of staff members who have to perform their tasks manually, and increase the accuracy of the GA estimates, thus becoming a useful instrument in prenatal care.

***Correspondence author email address:** muhammad.danish@pucit.edu.pk
DOI: 10.21015/vtcs.v13i1.2101

1 Introduction

In obstetrics, ultrasound (US) imaging has a wide range of clinical uses because it has many benefits compared to other imaging techniques such as computed tomography (CT), x-rays, and magnetic resonance imaging (MRI)

[1]. Prenatal care is popular in the US because it is radiology free, safe, painless, economical, and portable and can be carried everywhere [2, 3]. Also, it provides real-time imaging, which supports dynamic evaluation of fetal development and maternal wellbeing. Nevertheless, US imaging is not without its drawbacks, which can be classified as speckle noise, attenuation, low signal-noise ratio (SNR), signal dropout, boundary ambiguities, artifacts, and low penetration, among other factors, as presented in the literature [4]. These difficulties may impede the proper diagnosis, and the experience of competent clinicians is required to interpret images successfully and reduce errors to a minimum mistake [5].

Prenatal care can also be characterized as accurate fetal growth estimation, growth pattern monitoring, identification of abnormalities, and evaluation of pathological and physiological conditions [6]. To do this, the most important biometric measurements like Abdominal Circumference (AC), Head Circumference (HC), Bi-parietal Diameter (BPD), and Femur Length (FL) are commonly taken by the obstetricians as the metrics of fetal size, weight, and gestational age (GA) [7, 8]. The measurements are also performed conventionally by manual procedures that the experienced radiologists mark the endpoints of the femur or draw an ellipse around the HC to derive the measurements accurately [9]. Nonetheless, manual measurements are prone to low SNR, operator effects, and variations in sonographer expertise which may affect the reliability and diagnostic accuracy[10]. Moreover, manual evaluation will add a larger workload and can result in more inconsistencies especially in resource-constrained or rural settings where they lack trained professionals [11].

To address these limitations a number of automated and semi-automated methods of segmentation and measuring the femur have been investigated. The classical methods are thresholding when using various levels of intensity and thresholding when using binary masks fixedly [12]. Such techniques are however limited by certain grayscale intensity settings which may result in inconsistency between imaging machines. Deep-learning models have demonstrated potential on semantic segmentation in measurement of the femur over recent years but are impeded by the requirement of massive amounts of training data and the scarcity of open repositories [13]. With these drawbacks in mind, the need to have robust and traditional segmentation methods that can meet these data difficulties and offer credible outputs is still imminent. Fully automating these measurements with state-of-the-art image processing methods has potential to improve efficiency and reduce observer variability and also provide more reliable fetus assessments.

Automating these measurements through advanced image processing techniques holds promise for improving efficiency, reducing observer variability, and ensuring more consistent fetal evaluations.

To this end, our study provides a fetal femur segmentation and length measurement algorithm in US images which is an automated algorithm that is specifically developed to enhance the accuracy and consistency of GA estimation. The approach will solve typical imaging issues, minimize the use of manual measurements by certified radiologists, and simplify the clinical process, which will eventually enable obstetricians to use GA evaluation as a reliable tool.

Our study makes the following key contributions:

- We come up with a multi-step algorithm to improve the segmentation and measurement of the femur in US images. Our method combines image pre-processing, bilateral filtering-based speckle noise reduction, and k-means based clustering, and morphological analysis to enhance the accuracy of segmentation.
- We select a new data collection of 100 images of a woman in the US over a 22-38 gestation period. Both pictures are carefully annotated by two trained obstetricians, which means that they provide solid ground truth to test algorithms against.
- We test our algorithm as rigorously as possible against the state-of-the-art algorithms on the basis of region-based and distance-based evaluation measures, proving its high level of approval and precision of fetal biometrics analysis.

The remainder of this paper is organized as follows: Section 2 reviews related work on femur segmentation; Section 3 details the proposed segmentation and measurement algorithm; Section 4 discusses experimental results, and Section 5 concludes with our findings and suggests future research directions.

2 Related Works

In US image segmentation, numerous studies have investigated automated and semi-automated approaches, broadly categorizing these into traditional computer vision algorithms and deep learning methods, as extensively reviewed by Nobel et al. [14] and Meiburger et al. [15]. Traditional segmentation techniques include edge detection, active contour models, k-means clustering, and thresholding, all of which rely on predefined rules and mathematical operations to delineate structures within US images. For femur segmentation and length measurement, Wang et al. [16] introduced two automated methods for GA estimation. Their first approach used entropy-based segmentation to identify femur candidates in preprocessed US images, followed by morphology-based object detection. Where this technique did not turn out, a second technique used was edge detection and morphological operations to locate and measure FL. Testing on a dataset of 90 scans [5] of the US gave a Dice similarity coefficient of $73.95 \pm 14.56\%$ which indicates moderate segmentation accuracy.

Ponomarev et al. [12] suggested an automated profile of the femoral segmentation technique on the basis of an intensity thresholding in several levels. They used the technique of varying intensity cutoffs and shape descriptors to cluster and choose femur objects and they got a Dice similarity coefficient of $77.40 \pm 15.35\%$ on the same dataset [5]. Although their approach proved the efficiency of the intensity-based segmentation in regulating the image variability in the US, it also had significant shortcomings. The method was very sensitive to differences in intensities hence it was not strong to speckle noise, artifacts, and shadowing as seen in US imaging. Also, the use of fixed intensity thresholds also decreased the ability to adapt to various imaging settings, and thus its ability to generalize across various datasets.

The article by Hermawati et al. [17] presents a semi-automatic method of femur segmentation based on localizing region-based active contour (LRAC). Their approach combined speckle noise reduction and refinement to increase the detection of their femur regions with an average FL error of 4.61% per cent on a small dataset of only 11 US images. Although it was good at raising the accuracy of segmentation, it had its limitations in generalizability because it used a manual initialization and had a constrained dataset. On the same note, morphological operators and fixed thresholding were used by Thomas et al. to refine the shape of the femur and eliminate noise in the femur shape refinement process [18]. Even though this conventional method enhanced the precision of the segmentation, it had disadvantages of long processing time and sensitivity to changes in intensity, which disqualified it in real-time applications.

In more recent activities, Zhu et al. [13] tried a comparative analysis of both traditional machine learning and deep learning methods with the implementation of SegNet to perform femur segmentation. They have shown in their research not only the high accuracy potential of deep learning models but also indicated that large amounts of annotated training data are required, especially in obstetrics and gynecology. Despite the promising results of deep learning methods, the scarcity of annotated datasets in this area contributes to the further development and refinement of conventional methods of segmentation, and the reported accuracies of segmentation in the literature usually fall within the 73-78% range.

To address the limitations in the earlier researches, our study is aimed at coming up with a fully automated and adaptive femur segmentation method that would increase the level of accuracy and robustness in the US imaging. Our method is based on a combination of advanced image processing techniques to enhance the reliability of segmentation under various imaging scenarios as compared to the traditional methods that depend on a set of predetermined intensity thresholds or manual initiation.

3 Proposed Algorithm

Our proposed methodology is presented in this section. All the steps are summarized in Fig.1 and the methods described in details in the further parts.

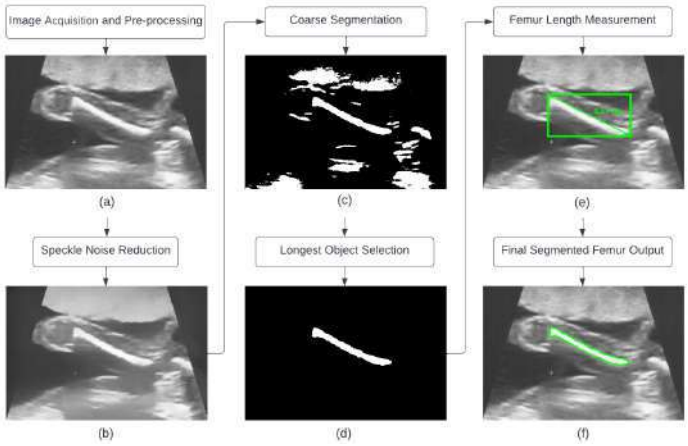


Figure 1. Steps of the proposed femur segmentation and length measurement algorithm. (a) Image acquisition and pre-processing, (b) Speckle noise reduction, (c) Coarse Segmentation, (d) Longest object selection, (e) Femur length (FL) measurement, and (f) Final segmented femur output.

3.1 Image Acquisition and Pre-processing

The first format that the US scans are stored under is the Digital Imaging and communications in medicine (DICOM). The DICOM images at this point are obtained in the US machine and then converted to Portable Network graphics (PNG) format. PNG is chosen due to the fact that it does not degrade image quality which can be used to view, process, and annotate US images. The processed image shown in Fig. 1 (a) in flowchart is the pre-processed image which is used as input in the next processing stages.

3.2 Speckle Noise Reduction

The next process is bilateral filtering that is used to smooth the US image effectively by removing noise, as suggested in the literature [19]. It uses a straightforward but powerful non-linear method which combines spatial and range filtering to improve the denoising process. The intelligent balance of noise reduction and edge preservation makes it a very efficient tool to remove the noise one would want to and keep the structure relevant, important details intact, so that the fine edges are not blurred. The bilateral filtering of an image, denoted by the letter I, and dimension of an image I with dimensions $M \times N$, is mathematically expressed through the use of the following equation:Eq. (1).

$$BF(I)_{i,j} = \frac{\sum_{m=-k}^{k} \sum_{n=-k}^{k} W_{i,j,m,n} \cdot I_{i+m,j+n}}{\sum_{m=-k}^{k} \sum_{n=-k}^{k} W_{i,j,m,n}} \tag{1}$$

where $BF(I)_{i,j}$ is the filtered value at pixel position (i, j), $W_{i,j,m,n}$ represents bilateral weights for spacial offset (m, n) at pixel location (i, j), and $I_{i+m,j+n}$ is value of the neighboring pixel at offset (m, n). $W_{i,j,m,n}$ is the product of spatial weight $S_{i,j,m,n}$ and range weight $R_{i,j,m,n}$ defined in Eq. (2).

$$W_{i,j,m,n} = S_{i,j,m,n} \cdot R_{i,j,m,n} \tag{2}$$

(a) (b) (c) (d)

Figure 2. Fetal femur image in 4 clusters. (a) Cluster 1 with centroid value 68, (b) Cluster 2 with centroid value 128, (c) Cluster 3 with centroid value 178, and (d) Cluster 4 with centroid value 218.

where $S_{i,j,m,n}$ and $R_{i,j,m,n}$ measure proximity and intensity similarity between the pixels respectively by using the Gaussian functions defined in Eqs. (3) and (4).

$$S_{i,j,m,n} = e^{-\frac{m^2+n^2}{2\sigma_s^2}} \tag{3}$$

$$R_{i,j,m,n} = e^{-\frac{(I_{i+m,j+n}-I_{i,j})^2}{2\sigma_r^2}} \tag{4}$$

A larger σ_r value smoothes the image aggressively, while a smaller value preserves subtle intensity variations. A larger σ_s produces more global smoothing, while a smaller value produces stronger local smoothing. We have determined the optimal values of $\sigma_r = 0.1$ and $\sigma_s = 9$ through rigorous experimentation, as elaborated in the upcoming section. The obtained denoised US image is depicted in Fig. 1 (b).

3.3 Coarse Segmentation

K-means clustering [20] is adopted for coarse segmentation. It is an iterative machine-learning algorithm used for partitioning n observations into K clusters by using Eq. (5), where c_j is the j^{th} cluster and μ_j is the centroid of the cluster c_j.

$$J = \sum_{j=1}^{k} \sum_{i \in c_j} \|x_i - \mu_j\|^2 \tag{5}$$

An overview of how K-means clustering is applied:

1. Pick K cluster centers randomly from the denoised US image shown in Fig. 1 (b), and our experimental results show that K = 4 provides us with the best segmentation scores.
2. Assign each pixel to a cluster that minimizes the distance between cluster center and pixel value.
3. Compute centroids again by averaging all of the pixels in the cluster.
4. Repeat steps 2 and 3 until no pixels change clusters. The fetal femur in 4 clusters is displayed in Fig. 2.
5. The fetal femur is the brightest region in the US image [18]. Therefore, we selected the cluster with the highest intensity centroid, as depicted in Fig. 1 (c).

3.4 Longest Object Selection

A contour detection algorithm is executed to find all the contours in the binary image obtained in the previous step, displayed in Fig. 1 (c). After that, the perimeter of each contour is calculated, and the results are sorted in descending order. The most extended object as a femur [21] is selected from the top of the sorted contour list displayed in Fig. 1 (d). Let $O = \{o_1, ..., o_n\}$ be the list of all contours; then the most extended object c^* selection is formulated as Eq. (6).

$$c^* = c_{i=0} \in SortDesc(O) \tag{6}$$

3.5 Femur Length Measurement

It is observed that the length of the femur is approximately equal to the diagonal of the bounding rectangle. Therefore, a very simple Eq. (7) is applied to estimate FL, where h is the height and w is the width of the bounding rectangle. Fig. 1 (e) displays the bounding rectangle surrounding the segmented fetal femur and the corresponding FL measured in centimeters (cm).

$$FL = \sqrt{h^2 + w^2} \tag{7}$$

3.6 Final Segmented Femur Output

Finally, the segmented femur object is represented on the original US image, as shown in Fig. 1 (f). Our newly devised algorithm autonomously calculates the FL at 3.57cm, closely matching the manually determined FL measurements of 3.59cm conducted by medical professionals.

3.7 Gestational Age Estimation

For estimating GA, we employed the equation proposed by Honarvar et al. [22], as shown in Eq. 8. In this formula, the FL measured in centimeters, predicted by our proposed pipeline, is used to calculate GA in weeks.

$$\text{GA (weeks)} = 0.262 * \text{FL (cm)} + 2 * \text{FL (cm)} + 11.5 \tag{8}$$

4 Experiments and Results

4.1 Evaluation Dataset

The study was approved by the Ethical Committee of The University of Lahore Teaching Hospital, Lahore, Pakistan (Reg. No. ERC/108/23/08). All human volunteers participated in accordance with the 2008 Helsinki ethical standards, with their personal information kept confidential and informed written consent obtained from each volunteer. US images were acquired by a skilled obstetrician using a Canon Aplio 300 US machine equipped with a PVT-375BT convex probe operating at a center frequency of 3.5 MHz. A total of 100 mid-sagittal fetal femur planes spanning GA from 22 to 38 weeks were collected in DICOM format, each image sized at 900 $\times$ 600. The values of grayscale lie between 0 and 255. Two radiologists were hired in two separate stages in the process of annotation where they formed part of the team of more than 10 years of experience. The initial stage was a training session that was aimed at introducing the sonographers to the GIMP 2.10.34 annotation tool and the need to make the marking process consistent. The sonographers were asked to outline the region of the femur separately without any idea of what the other sonographers had noted. During the second stage, they conducted a micro-reviewing of the documents and compared the annotations of each other.

4.2 Evaluation Metrics

To assess the quality of segmentation and measurements, three evaluation standards are considered. Firstly, region-based metrics evaluate precision, sensitivity, specificity, and Dice similarity. Secondly, distance-based metrics assess local variability between the algorithm's results and expert annotations. Thirdly, Bland-Altman plots analyze the difference between automatically measured FL and expert-provided measurements.

4.2.1 Region-based metrics

In the following region-based metrics [23], let O_{SR} denote the auto segmentation result and O_{GT} be the expert's provided ground truth. The precision measures the proportion of identifications made by the correct segmentation model, as expressed in Eq. (9). A higher value indicates a lower false detection rate.

$$Precision = \frac{|O_{GT} \cap O_{SR}|}{|O_{SR}|} \tag{9}$$

Sensitivity measures the true positive (TP) rate defined in Eq. (10), while specificity measures the true negative (TN) rate computed by using Eq. (11). Both sensitivity and specificity are typically reported as decimal values ranging from 0 to 1. Higher values indicate better performance of segmentation.

$$Sensitivity = \frac{|O_{GT} \cap O_{SR}|}{|O_{GT}|} \tag{10}$$

$$Specificity = \frac{|(O_{GT} \cup O_{SR})^c|}{|(O_{GT})^c|} \tag{11}$$

Dice similarity represents the mutual overlap between O_{SR} and O_{GT}, defined in Eq. (12). Its value ranges between 0 to 1, where 0 indicates no overlap, and 1 indicates a perfect match or complete similarity between O_{SR} and O_{GT}.

$$Dice = \frac{2|O_{GT} \cap O_{SR}|}{|O_{GT}| + |O_{SR}|} \tag{12}$$

4.2.2 Distance-based metrics

The distance-based metrics, as explained in [24], are incorporated into the evaluation to compare errors between contours of O_{SR} and O_{GT}, denoted as $C(O_{SR})$ and $C(O_{GT})$ respectively. $c_{O_{SR}}$ represents a contour element of $C(O_{SR})$ and $c_{O_{GT}}$ a contour element of $C(O_{GT})$. The Euclidean distance of a pixel p to $c_{O_{GT}}$ is computed by using Eq. (13).

$$d_E\left(p, C(O_{GT}\right) = \min_{cO_{GT} \in C(O_{GT})} \|p - c_{O_{GT}}\| \tag{13}$$

The Maximum Symmetric Contour Distance (MSD) can then be defined as Eq. (14). It quantifies the maximum distance from a point in one set to the closest point in the other set. Its higher value indicates the greater dissimilarity or separation between the two sets being analyzed.

$$MSD\left(O_{GT}, O_{SR}\right) = \max\left(\max_{c_{O_{GT}} \in C(O_{GT})} d_E\left(c_{O_{GT}}, C\left(O_{SR}\right)\right), \max_{c_{O_{SR}} \in C(O_{SR})} d_E\left(c_{O_{SR}}, C(O_{GT})\right)\right) \tag{14}$$

The Average Symmetric Contour Distance (ASD) is the average of all distances between O_{SR} and O_{GT} and can be computed by using Eq. (15). An ideal segmentation would yield a contour distance of 0 mm.

$$ASD\left(O_{GT}, O_{SR}\right) = \frac{1}{|C\left(O_{GT}\right)| + |C\left(O_{SR}\right)|} \times \left(\sum_{c_{O_{GT}} \in C(O_{GT})} d_E\left(c_{O_{GT}}, C(O_{SR})\right) + \sum_{c_{O_{SR}} \in C(O_{SR})} d_E\left(c_{O_{SR}}, C\left(O_{GT}\right)\right)\right) \tag{15}$$

The Root Mean Square Symmetric Contour Distance (RMSD) is similar to ASD as defined in Eq. (16). The lower RMSD value indicates a better match between the contours.

$$RMSD\left(O_{GT}, O_{SR}\right) = \sqrt{\frac{1}{|C\left(O_{GT}\right)| + |C\left(O_{SR}\right)|}} \times \sqrt{\left(\sum_{c_{O_{GT}} \in C(O_{GT})} d_E\left(c_{O_{GT}}, C(O_{SR})\right) + \sum_{c_{O_{SR}} \in C(O_{SR})} d_E\left(c_{O_{SR}}, C\left(O_{GT}\right)\right)\right)} \tag{16}$$

4.2.3 Bland–Altman plots

Bland–Altman plots [25] are a widely used graphical method for evaluating the agreement between two measurement techniques. They are especially effective when comparing methods that aim to measure the same variable. In these plots, the differences between the two measurements are plotted on the y-axis, while the x-axis represents the average of the measurements from both methods. This approach helps to visually assess the level of agreement and detect any systematic bias or trends in the measurements, providing a clear understanding of how well the two methods align.

4.3 Adjustable Parameter Configurations

In an effort to optimize the parameters that were under control, several experiments were run on our test data. We first used our segmentation algorithm to the unfiltered images of the US. Various values of K were experimented with respect to coarse segmentation and K = 4 showed higher results in region based measurements and K = 5 better results in distance based measurements. The results of the performance without denoising are in Table 1.

Table 1. Segmentation results of our algorithm without using bilateral filtered denoised images.

Metric	K-means Clustering		
	K = 3	K = 4	K = 5
Precision (%)	53.23 ± 35.51	**79.94 ± 17.89**	70.21 ± 18.43
Sensitivity (%)	65.50 ± 43.66	**83.40 ± 17.62**	70.71 ± 18.68
Specificity (%)	99.48 ± 00.30	99.76 ± 00.31	**99.92 ± 00.12**
Dice (%)	59.04 ± 39.36	**84.01 ± 17.80**	78.71 ± 17.38
MSD (mm)	8.19 ± 11.47	3.70 ± 7.98	**3.30 ± 7.49**
ASD (mm)	4.90 ± 09.94	1.23 ± 4.51	**1.04 ± 3.81**
RMSD (mm)	4.37 ± 07.03	0.87 ± 3.19	**0.74 ± 2.70**

The next step involved denoising of our dataset before the algorithm implementation using values with σ_s ranging from 3 to 15 aσ_r values from 0.01 to 0.16.This method led to better performance than where there was no denoising. Table 1, (table 2) shows that the highest Dice similarity of 93.18% was obtained, using with σ_s = 9 and σ_r = 0.10 as the mean of sigma r in denoising and coarse segmentation with K = 4, respectively.

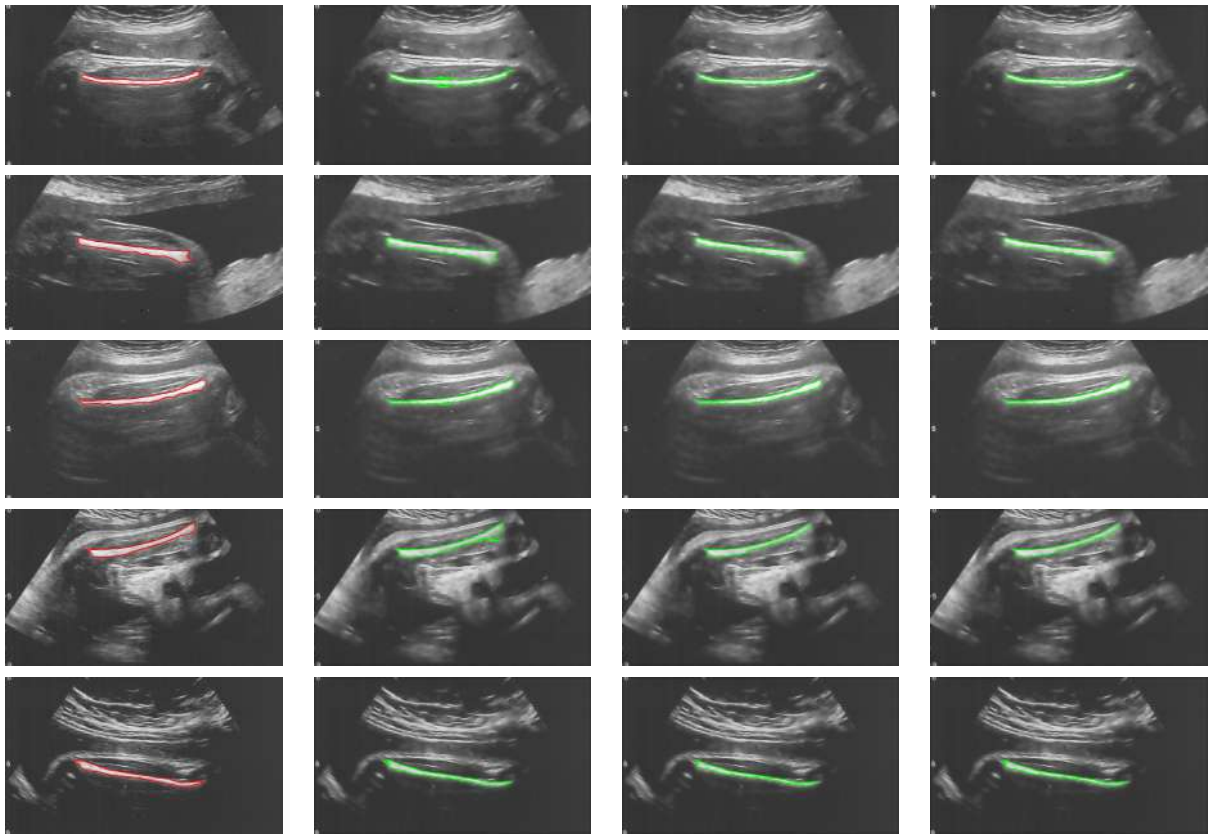
4.4 Performance Evaluation

Our algorithm used optimal values of the parameters used to denoise the dataset: the values of both parameters were set of σ_r = 0.1 and σ_s = 9. We then performed a comparative analysis with the annotations of medical professionals and also the auto-segmented results of the results of the segmentation using various values of K. Fig. 3 denotes the best five results of the femur segmentation, with successful segmentations being noticed. Also, Fig.4 to show the two poorest results in which the femur gets wrongly attached to brighter muscular regions that do not belong to the anatomical part of the femur, causing segmentation results to be mistaken.

Summarization of the evaluation of our proposed technique is in Table 3. At K = 4, the model gave a precision of 90.15 ± 13.72%, sensitivity of 93.18 ± 9.64%, and a dice of 93.18 ± 9.54% which were all better than the denoising results. Specificity was found to be 99.87 ± 0.19%, favorably (19) /100, which is a little less than that of the specificity with K = 5. Furthermore, the distance-based metrics in the proposed approach improved, with the MSD of 1.19 ± 3.15 mm, ASD of 0.18 ± 1.40 mm, and RMSD of 0.12 ± 0.99 mm.However, it is noteworthy that specificity is higher for both the noised and denoised datasets when K = 5. Based on these findings, we can conclude that reducing speckle noise before applying K-means clustering significantly impacts the results. Mean and standard deviation comparisons of precision, sensitivity, specificity, and Dice similarity are presented in Fig. 5.

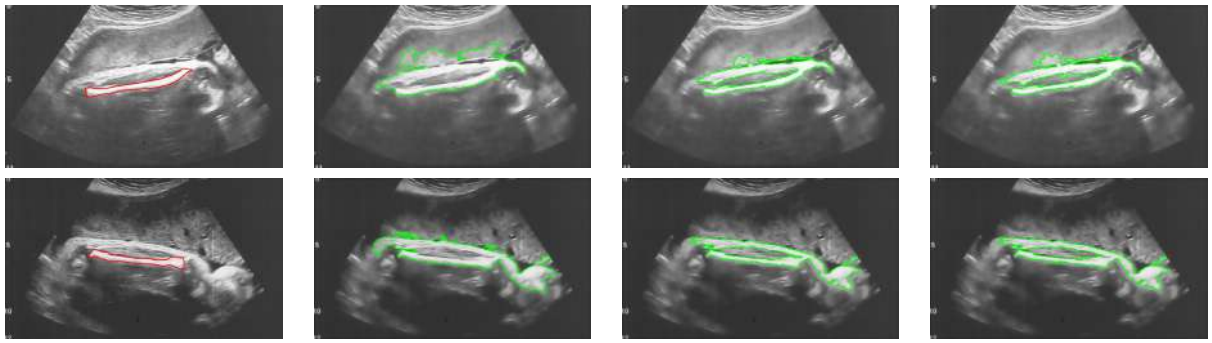
Table 2. Mean Dice (%) comparison with different values of controllable parameters.

		σ_s = 3	σ_s = 5	σ_s = 7	σ_s = 9	σ_s = 12	σ_s = 15
K = 3	σ_r = 0.01	47.76	46.22	49.85	51.64	49.87	51.64
	σ_r = 0.04	45.27	48.34	49.85	51.64	49.93	49.13
	σ_r = 0.07	57.63	57.65	58.67	59.73	59.75	57.01
	σ_r = 0.10	60.45	60.79	61.82	63.80	61.53	60.70
	σ_r = 0.13	60.04	60.56	61.89	62.48	61.43	60.04
	σ_r = 0.16	50.21	49.69	52.33	50.23	48.34	46.78
K = 4	σ_r = 0.01	81.61	79.63	86.65	83.63	81.61	84.58
	σ_r = 0.04	82.61	84.63	86.65	83.63	80.61	79.58
	σ_r = 0.07	87.61	87.65	88.70	89.73	89.75	87.71
	σ_r = 0.10	89.75	89.79	91.82	**93.18**	91.83	90.80
	σ_r = 0.13	89.80	89.82	91.17	92.18	91.13	90.07
	σ_r = 0.16	90.09	89.39	92.41	90.11	88.72	87.34
K = 5	σ_r = 0.01	67.92	66.45	74.29	72.87	69.45	74.62
	σ_r = 0.04	68.24	68.49	74.29	72.87	69.46	68.75
	σ_r = 0.07	78.83	76.88	77.42	78.94	78.98	76.85
	σ_r = 0.10	79.12	79.76	80.09	82.72	81.39	79.69
	σ_r = 0.13	78.70	78.87	80.08	81.61	81.04	79.69
	σ_r = 0.16	80.26	78.69	81.88	79.87	77.13	76.36



(a) (b) (c) (d)

Figure 3. The best segmentation results of our algorithm executed on a denoised dataset with different values of k clusters. (a) Doctor's marked, (b) k = 3, (c) k = 5, and (d) k = 4 (Our proposed).



(a) (b) (c) (d)

Figure 4. Two examples of poor femur segmentation. (a) Doctor's marked, (b) k = 3, (c) k = 5, and (d) k = 4 (our proposed).

From the figure, it is evident that denoising consistently improves all evaluation metrics. The highest values for precision, sensitivity, and Dice score, along with the smallest standard deviation range, are observed when K = 4.

Table 3. Segmentation results of our algorithm using bilateral filtered denoised images.

Metric	K-means Clustering		
	K = 3	K = 4	K = 5
Precision(%)	58.56 ± 36.28	**90.15 ± 13.72**	75.50 ± 17.27
Sensitivity(%)	68.48 ± 41.08	**93.18 ± 9.64**	76.17 ± 17.39
Specificity (%)	99.57 ± 00.38	99.87 ± 00.19	**99.92 ± 00.12**
Dice(%)	63.80 ± 38.28	**93.18 ± 9.54**	82.72 ± 15.16
MSD (mm)	8.13 ± 10.69	**1.19 ± 3.15**	2.26 ± 4.87
ASD (mm)	0.74 ± 02.69	**0.18 ± 1.40**	0.63 ± 3.07
RMSD (mm)	3.15 ± 06.58	**0.12 ± 0.99**	0.44 ± 2.17

4.5 Performance Comparison

We compared the performance of our proposed segmentation method with existing femur segmentation techniques, including those by [16], [12], and [13]. The results demonstrate that our algorithm outperformed the others, achieving a precision of 90.15%, a specificity of 99.87%, and a Dice similarity of 93.18%. The region-based and distance-based evaluation metrics are summarized in Table 4, with the best results highlighted in bold. Although the Zhu method [13] achieved a higher sensitivity, its precision, specificity, and Dice similarity were lower compared to ours. Additionally, our method outperformed the others in terms of MSD and RMSD (distance-based metrics), except for the ASD score, where the Zhu method had a better ASD score than all others. A comparative analysis of the mean and standard deviation of our results is also presented in Fig. 6 using bar charts. From this figure, it is evident that our method achieved the highest precision and Dice similarity. However, in terms of specificity, Zhu et al. [13] reported the highest value.

Indeed, directly comparing these methodologies may not be entirely equitable or fair due to differences in the datasets used. The variability in dataset composition highlights the challenge posed by the scarcity of publicly available repositories containing fetal femur US images, which hinders researchers' access to standardized datasets for consistent evaluation and benchmarking. Moreover, unlike the semi-automatic methods commonly found in the literature, our fully automated approach is specifically designed to support obstetricians in accurate femur segmentation and length measurement. The proposed algorithm offers consistent, accurate, and reliable GA estimation, making it especially valuable in resource-limited settings where experienced radiologists are scarce. By reducing dependence on specialist expertise, this solution has the potential to improve prenatal care accessibility,

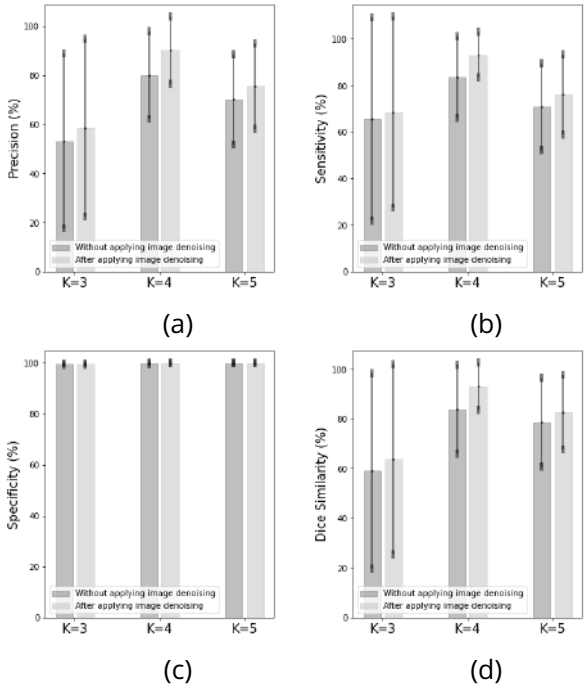


Figure 5. The mean and standard deviation comparison with the K-means clustering algorithm in terms of (a) Precision, (b) Sensitivity, (c) Specificity, and (d) Dice Similarity.

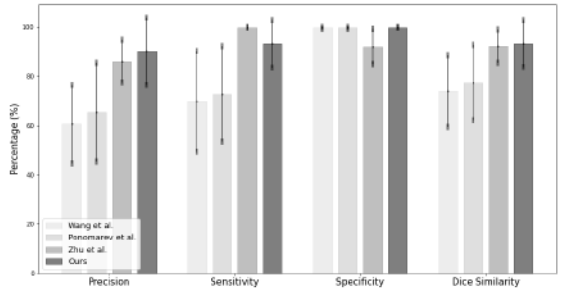


Figure 6. The mean and standard deviation comparison with existing work regarding Precision, Sensitivity, Specificity, and Dice.

promote early detection of fetal growth issues, and contribute to better maternal and fetal health outcomes in communities where resources are limited or in rural areas.

4.6 Femur Length Measurement Evaluation

We conducted a Bland-Altman analysis [25] by calculating the difference and average between the automatically computed FL and the manually measured FL provided by doctors. The scatter plot of the difference against the average revealed a mean difference of 0.062 cm, indicating a slight bias towards one of the methods. The comprehensive assessment of our agreement, depicted in Fig. 7, showed that only 2 measurements fall beyond the 95% agreement limits, with the majority falling between an upper limit of 1.19 cm and a lower limit of -1.06cm.

Table 4. Femur segmentation performance comparison with existing work.

Metric	Method			
	Wang et al. [16]	Ponomarev et al. [12]	Zhu et al. [13]	Ours
Precision (%)	60.56 ± 15.88	65.44 ± 19.98	86.06 ± 08.73	**90.15 ± 13.72**
Sensitivity (%)	69.84 ± 20.36	72.79 ± 19.40	**99.86 ± 00.13**	93.18 ± 09.64
Specificity (%)	99.66 ± 00.37	99.70 ± 00.39	92.11 ± 07.03	**99.87 ± 00.19**
Dice (%)	73.95 ± 14.56	77.40 ± 15.35	92.20 ± 06.71	**93.18 ± 09.54**
MSD (mm)	6.02 ± 7.29	6.39 ± 9.53	1.97 ± 1.03	**1.19 ± 3.15**
ASD (mm)	1.04 ± 1.29	1.23 ± 2.30	**0.05 ± 0.12**	0.18 ± 1.40
RMSD (mm)	1.77 ± 2.41	2.04 ± 3.76	0.19 ± 0.21	**0.12 ± 0.99**

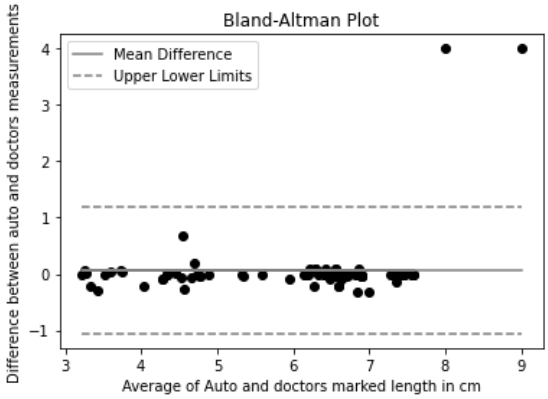


Figure 7. Bland-Altman Plot between differences and average measurements.

4.7 Computational Complexity Analysis

The proposed algorithm is implemented in Python and is available upon request from the corresponding author. All experiments were conducted on a system equipped with an Intel Core i5 processor, 8GB RAM, and a 64-bit Windows 11 operating system. Our execution time analysis revealed that, on average, our method takes approximately 8.45 seconds to segment and measure the length of the femur. In comparison, the [16] and [12] methods require an average of 2.28 seconds and 24.2 seconds, respectively.

5 Conclusion

This paper introduces an automated technique for segmenting fetal femurs and measuring their length using US images. Our method consists of several key steps, including speckle noise reduction via a bilateral filter, coarse segmentation using k-means clustering, selection of the longest object, and precise measurement of FL. Evaluation on a dataset containing 100 mid-sagittal femur planes, spanning GA from 22 to 38 weeks, demonstrated exceptional results with 90.15% precision, 93.18% sensitivity, 99.87% specificity, and 93.18% Dice similarity. These outcomes, coupled with comparisons against existing techniques, underscore the effectiveness of our approach. Given the limited ratio of obstetricians to the population, our method serves as a valuable tool for radiologists, enabling efficient femur segmentation and accurate length measurements. Moreover, it supports obstetricians in diagnostic and clinical decision-making processes. Future work will extend beyond FL estimation in obstetrics by incorporating larger datasets with more cases. Additionally, deep learning-based techniques will be explored to automate the estimation of other parameters, such as biparietal diameter (BPD), AC, and HC, thereby improving GA calculation.

Acknowledgments

We appreciate and acknowledge all the hard work of Dr. Amna Arshad, Itefaq Trust Hospital, Lahore, Pakistan; Dr. Saima, Saima Medical Center, Lahore, Pakistan, and Dr. Sumaiya, The University of Lahore Teaching Hospital, Lahore, Pakistan. They helped us to understand how to identify the femur from US images and how to use a US machine to calculate FL. They also provided us with a dataset of 100 annotated scans.

Author Contributions

Hafiz Muhammad Danish: Writing- Original draft preparation, Methodology, Software. **Zobia Suhail**: Writing-Original draft preparation, Conceptualization, Visualization. **Faiza Farooq**: Data curation, Supervision. **Reyer Zwiggelaar**: Supervision, Software, Validation.

Compliance of Ethical Standards

This research received no funding, and the authors declare that they have no conflict of interest. The code and dataset used in the experiments are available upon request from the corresponding author. All procedures involving human participants were conducted in accordance with the ethical principles of the 2008 Helsinki Declaration. The dataset was fully anonymized to protect participant confidentiality, and informed written consent—provided in both English and Urdu—was obtained from each volunteer during their initial hospital visit. The study was approved by the Ethics Committee of The University of Lahore Teaching Hospital, Lahore, Pakistan, under registration number ERC/108/23/08.

References

[1] F. Recker, U. Gembruch, and B. Strizek, "Clinical ultrasound applications in obstetrics and gynecology in the year 2024," p. 1244, 2024.

[2] Y. E. E. Salama, A. E.-M. M. Zakaria, and Y. M. S. K. Diab, "Accuracy of ultrasound in the diagnosis of lateral and posterior invasion in patients with placenta accreta spectrum disorders," *Al-Azhar International Medical Journal*, vol. 4, no. 5, p. 20, 2023.

[3] T. Tzatzairis, K. Skarentzos, C. Grammatikos, C. Karamalis, K. Korakianitis, R. Kourempeles, and G. Drosos, "Ultrasound applications in pediatric orthopedics," *Archives of Bone and Joint Surgery*, vol. 12, no. 7, p. 457, 2024.

[4] M. H. Rahman, M. S. Hossain, and F. Islam, "Design and implementation of speckle noise reduction algorithm using 2d ultrasound image," *International Journal of Image, Graphics and Signal Processing*, vol. 3, pp. 31–47, 2023.

[5] S. Rueda, S. Fathima, C. L. Knight, M. Yaqub, A. T. Papageorghiou, B. Rahmatullah, A. Foi, M. Maggioni, A. Pepe, J. Tohka *et al.*, "Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge," *IEEE Transactions on Medical Imaging*, vol. 33, no. 4, pp. 797–813, 2013.

[6] J. Agbenyah, D. Sule, W. K. Antwi, V. Hewlett, K. Asare-Dompreh, and E. T. Matey, "Exploring the relationship between placenta thickness, gestational age and fetal weight, a community study in ghana," *Korean Journal of Physiology and Pharmacology*, vol. 28, no. 1, pp. 364–373, 2024.

[7] M. M. Gebreel, M. F. Mohamed, and S. A. H. A. Fattah, "A comparative study of the transverse cerebellar diameter of the fetus with the biparietal diameter and femur length in estimation of diagnostic accuracy of gestational age in the second and third trimesters of pregnancy," *Al-Azhar International Medical Journal*, vol. 5, no. 2, p. 10, 2024.

[8] A. Shafique, Z. Suhail, and H. M. Danish, "Hybrid technique for estimating fetal head circumference using ultrasound imaging," *International Journal of Innovations in Science and Technology*, vol. 6, pp. 1058–1075, 2024.

[9] Y. Zeng, P.-H. Tsui, W. Wu, Z. Zhou, and S. Wu, "Fetal ultrasound image segmentation for automatic head circumference biometry using deeply supervised attention-gated v-net," *Journal of Digital Imaging*, vol. 34, pp. 134–148, 2021.

[10] A. Grimwood, K. Thomas, S. Kember, G. Aldis, R. Lawes, B. Brigden, J. Francis, E. Henegan, M. Kerner, L. Delacroix *et al.*, "Factors affecting accuracy and precision in ultrasound guided radiotherapy," *Physics and Imaging in Radiation Oncology*, vol. 18, pp. 68–77, 2021.

[11] H. M. Danish, Z. Suhail, and F. Farooq, "Deep learning-based automation for segmentation and biometric measurement of the gestational sac in ultrasound images," *Frontiers in Pediatrics*, vol. 12, p. 1453302, 2024.

[12] G. V. Ponomarev, M. S. Gelfand, and M. D. Kazanov, "A multilevel thresholding combined with edge detection and shape-based recognition for segmentation of fetal ultrasound images," *Proceedings of challenge US: biometric measurements from fetal ultrasound images, ISBI*, vol. 2012, pp. 17–19, 2012.

[13] F. Zhu, M. Liu, F. Wang, D. Qiu, R. Li, and C. Dai, "Automatic measurement of fetal femur length in ultrasound images: a comparison of random forest regression model and segnet," *Mathematical Biosciences and Engineering*, vol. 18, no. 6, pp. 7790–7805, 2021.

[14] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: a survey," *IEEE Transactions on medical imaging*, vol. 25, no. 8, pp. 987–1010, 2006.

[15] K. M. Meiburger, U. R. Acharya, and F. Molinari, "Automated localization and segmentation techniques for b-mode ultrasound images: A review," *Computers in biology and medicine*, vol. 92, pp. 210–235, 2018.

[16] C.-W. Wang, H.-C. Chen, C.-W. Peng, and C.-M. Hung, "Automatic femur segmentation and length measurement from fetal ultrasound images," *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012*, pp. 21–23, 2012.

[17] F. Hermawati, H. Tjandrasa, G. P. Sari, A. Azis *et al.*, "Automatic femur length measurement for fetal ultrasound image using localizing region-based active contour method," in *Journal of Physics: Conference Series*, vol. 1230, no. 1. IOP Publishing, 2019, p. 012002.

[18] J. G. Thomas, R. A. Peters, and P. Jeanty, "Automatic segmentation of ultrasound images using morphological operators," *IEEE Transactions on Medical Imaging*, vol. 10, no. 2, pp. 180–186, 1991.

[19] R. Ruhela, B. Gupta, and S. Singh Lamba, "An efficient approach for texture smoothing by adaptive joint bilateral filtering," *The Visual Computer*, vol. 39, no. 5, pp. 2035–2049, 2023.

[20] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2023.

[21] C.-W. Wang, "Automatic entropy-based femur segmentation and fast length measurement for fetal ultrasound images," in *2014 International Conference on Advanced Robotics and Intelligent Systems (ARIS)*. IEEE, 2014, pp. 1–5.

[22] M. Honarvar, M. Allahyari, and S. Dehbashi, "Assessment of gestational age based on ultrasonic femur length after the first trimester: a simple mathematical correlation between gestational age (ga) and femur length (fl)," *International Journal of Gynecology & Obstetrics*, vol. 70, no. 3, pp. 335–340, 2000.

[23] J. K. Udupa, V. R. LeBlanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. M. Currie, B. E. Hirsch, and J. Woodburn, "A framework for evaluating image segmentation algorithms," *Computerized medical imaging and graphics*, vol. 30, no. 2, pp. 75–87, 2006.

[24] T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes *et al.*, "Comparison and evaluation of methods for liver segmentation from ct datasets," *IEEE transactions on medical imaging*, vol. 28, no. 8, pp. 1251–1265, 2009.

[25] M. A. Mansournia, R. Waters, M. Nazemipour, M. Bland, and D. G. Altman, "Bland-altman methods for comparing methods of measurement and response to criticisms," *Global Epidemiology*, vol. 3, p. 100045, 2021.