Final Project - Fetal Head Circumference Prediction

Nitay Levi Daniel Rivni Ori Elimelech

August 15, 2024

Abstract

Accurate prediction of fetal head circumference (HC) is crucial for assessing fetal growth and development during prenatal care. Traditional methods based on ultrasound image segmentation often suffer from variability due to factors like operator experience, fetal position, and equipment quality, leading to potential inaccuracies in HC estimation. In this study, we explore the application of advanced machine learning techniques, particularly convolutional neural networks (CNN), to enhance the precision and reliability of HC predictions from ultrasound images.

We experimented with both classification and regression models, utilizing various data augmentation strategies such as RandomResizedCrop, GaussianBlur, and downsampling. While classification models initially offered some improvements, they were ultimately outperformed by regression models, which provided more accurate and consistent results, particularly when using key anatomical measurements like Biparietal Diameter (BPD) and Occipitofrontal Diameter (OFD).

To address the challenge of limited data, we incorporated Generative Adversarial Networks (GAN) to generate synthetic ultrasound images, which significantly enhanced model performance, especially in regression tasks. The integration of GANs introduced greater data diversity, improving model generalization and reducing overfitting.

Our results demonstrate the superior accuracy of regression models over classification approaches in predicting HC and highlight the effectiveness of GAN-Based Data Generation. This study emphasizes the importance of selecting the appropriate model architecture and preprocessing techniques in medical imaging tasks and suggests future directions for refining GAN-generated data and expanding the dataset to further improve predictive performance. These advancements have the potential to enhance prenatal care by providing more reliable and accurate HC measurements, ultimately improving maternal and fetal health outcomes.

Keywords: Head circumference, Convolutional neural networks, Biparietal Diameter, Occipitofrontal Diameter, Generative Adversarial Networks, fully convolutional networks.

Introduction

1.1 Background

Before birth examinations are vital for monitoring fetal growth, screening for fetal abnormalities, and reducing the frequency of congenital impairments. Among the diagnostic tools available, ultrasound imaging is preferred due to its non-invasive nature, lack of radiation, and ability to provide real-time visualizations of the fetus. This makes it particularly suitable for repeated use throughout pregnancy. This technology enables healthcare providers to obtain detailed visualizations of the fetus, which are crucial for assessing various aspects of fetal development.

A key measurement obtained through ultrasound is the fetal HC, an important indicator of fetal growth and brain development. Accurate measurement of HC is essential for determining gestational age, estimating the expected date of delivery, and making informed decisions regarding pregnancy management. Deviations from normal HC measurements can signal potential developmental issues, prompting further investigation and early intervention if necessary. The ability to precisely monitor HC can significantly impact both immediate clinical decisions and long-term health outcomes for the child.

However, despite its widespread use, traditional methods for estimating HC, particularly those relying on manual or semi-automated segmentation techniques, are prone to inconsistencies. These inconsistencies can arise from various sources, including the operator's level of experience, the fetus's position during the ultrasound, and the technical quality of the ultrasound equipment itself[1]. Such factors can introduce variability in measurements, leading to either overestimation or underestimation of HC. These measurement inaccuracies are not trivial; they can affect the assessment of fetal growth progressions and the timing of delivery, potentially resulting in suboptimal clinical outcomes. For instance, an overestimated HC might lead to unnecessary concerns about macrocephaly, while an underestimation could delay the diagnosis of conditions like microcephaly, which requires timely intervention.

Recent advances in medical imaging and machine learning offer new opportunities to improve the precision and reliability of these measurements. CNNs, a class of deep learning models specifically designed for image analysis, have emerged as powerful tools in medical imaging. These models excel in processing and classifying complex image data, making them well-suited for tasks such as identifying subtle anatomical features in ultrasound images that might be overlooked by human observers or traditional algorithms. CNNs like EfficientNet, MobileNet, and InceptionNet have been successfully applied in various medical imaging tasks[2], achieving high levels of accuracy and demonstrating the potential to significantly enhance clinical diagnostics.

By training these models on large datasets of labeled ultrasound images, researchers can develop algorithms capable of delivering more precise and consistent HC estimates. These advanced techniques promise to reduce the variability inherent in manual measurements, thus providing clinicians with more reliable data to base their decisions on. The integration of machine learning into prenatal care could lead to more accurate assessments of fetal growth, better prediction of delivery dates, and improved monitoring of fetal health. Ultimately, this technology holds the potential to improve health outcomes for both mothers and their babies by enabling earlier and more accurate detection of potential issues, allowing for timely and appropriate interventions.

Despite these promising developments, there is a clear need for continued research to refine these techniques and ensure their broad applicability in clinical practice. The ongoing challenge is to further enhance the accuracy and generalizability of these models, particularly in diverse clinical settings with varying ultrasound equipment and patient populations. Moreover, integrating these advanced models into routine clinical workflows requires careful consideration of factors such as ease of use, interpretability of results, and the ability to provide real-time feedback to clinicians. Addressing these challenges will be critical to fully realizing the benefits of machine learning in prenatal care and ensuring that this technology can be widely adopted to improve maternal and fetal health outcomes.

1.2 Related work and motivation

In the field of fetal HC prediction, choosing the right approach for modeling network output is crucial for achieving optimal performance. While regression is often the straightforward choice because it directly handles numerical values, classification models have shown unexpected effectiveness in various applications, as highlighted in [3]. Their study demonstrates that classification networks can outperform regression models in terms of performance, even when the target values are inherently numerical.

Currently, the common approach for HC prediction is through image segmentation. However, accurate segmentation and measurement remain challenging due to several factors, including image artifacts, incomplete ellipse fitting, fluctuations, and the varying pixel sizes of fetal ultrasound images throughout all trimesters of pregnancy. This is further complicated by earlier approaches that relied on morphological methods, such as using Hough transforms and least squares for ellipse fitting, which were time-consuming and sensitive to noise [4]. As machine learning techniques like K-means clustering and Random Forests became more popular, these were initially employed to detect the outer edge of the fetal skull before applying ellipse detection algorithms. However, these traditional methods faced significant challenges with noisy or low-contrast images, especially in the early and late stages of pregnancy. Despite these advancements, these methods still struggled with issues like inaccurate skull edge detection, particularly when the fetal skull is not well-defined.

In recent years, deep learning has increasingly been applied to fetal HC measurement, with approaches like fully convolutional networks (FCN) and U-Net showing promise in segmenting fetal heads and predicting bounding boxes for accurate measurement. These methods, while improving on tra-

ditional machine learning approaches, still face challenges due to unclear ultrasound images and the complex structure of the uterine wall and amniotic fluid, which can resemble the texture of the fetal HC.

Given these challenges, our work initially adopted a classification approach, motivated by its potential for improved accuracy and robustness in handling the variability and complexities of ultrasound imaging. This aligned with the findings by Ferber[3] that classification models can sometimes outperform regression models. However, as our research progressed, we found that switching to a regression approach yielded better performance, providing more accurate and consistent measurements and addressing some limitations of the classification method. This transition emphasizes the importance of flexibility in model selection and adapting methodologies based on empirical results.

The paper by Jing Zhang [5] focuses on the direct estimation of fetal HC using CNN without relying on traditional segmentation methods. Unlike approaches that require manual or automated segmentation of the fetal head, this study leverages CNN to directly predict HC from ultrasound images. The authors argue that CNN can effectively learn to identify and measure the HC, thus bypassing the need for segmented images. Their results demonstrate promising accuracy, indicating that direct estimation can be a viable alternative to segmentation-based methods.

This study is particularly relevant to our work as it supports the shift towards regression models for HC prediction, highlighting the potential benefits of direct measurement approaches in reducing reliance on segmentation and improving prediction accuracy.

Materials and methods

2.1 Data sources

The data utilized in this study comes from the publicly available HC18 dataset, which focuses on the automatic measurement of fetal HC using ultrasound images. This dataset includes ultrasound images collected from 551 pregnant women, totaling 1,334 two-dimensional images. The dataset is divided into 999 labeled training images and 335 unlabeled test images. Additionally, there is a file that provides the pixel dimensions for each image in both the training and test sets. The original size of each image is 540 by 800 pixels.

To address the limitations of the relatively small dataset, we supplemented our data with additional images generated using StyleGAN2-ADA [6], a generative model. This addition increased our dataset to approximately 4,000 images, enhancing the diversity and robustness of our training data. However, it's important to note that some of the generated images were of lower quality and may not have contributed equally to the overall dataset

We trained a Nested U-Net model to predict masks identifying the regions where the fetal head is located within the images (average dice coefficient: 96%, mean absolute error: 0.018). These masks are crucial for obtaining accurate ellipse measurements, which are necessary for calculating the fetal HC and diameters (BPD and OFD). Additionally, we developed a pixel size model to predict the pixel dimensions of the ultrasound images (mean absolute error: 0.0094). This model allows us to convert the pixel measurements into real-world units, enabling the calculation of the final circumference for the GAN-generated images and the unlabeled test set images.

Figure 1: Image samples.

2.2 Data preprocessing

As part of the preprocessing step, we resized all images from 800 × 540 pixels to 224×224 and normalized them for training our CNN models. Additionally, we categorized the images into classes based on a 10mm range of fetal HC. For example, class 1 contained images with HC measurements between 65-75 mm. The first and last classes are special cases: we grouped images with HC measurements under 65mm into a single class (class 0) and those with measurements greater than 325mm into the final class (class 27).

Furthermore, we split the dataset into about 80% (812 training images, 269 test images, and 3168 gan images) for training and about 20% for testing, ensuring that this division was uniformly balanced across both the original and StyleGAN2-generated images. Out of the testing split, we selected only the already labeled training data belonging to real ultrasound images for validation (187 images). This approach helped maintain the diversity and consistency of the data across different sources, enhancing the robustness of our model training and evaluation process.

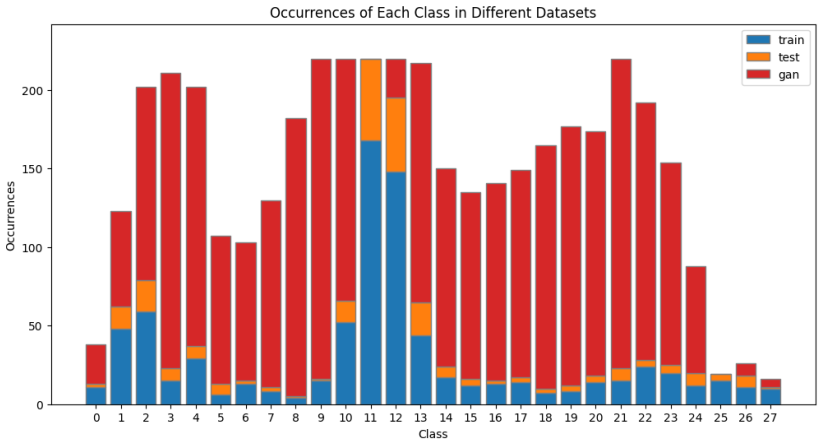

Figure 2: Training data (80%).

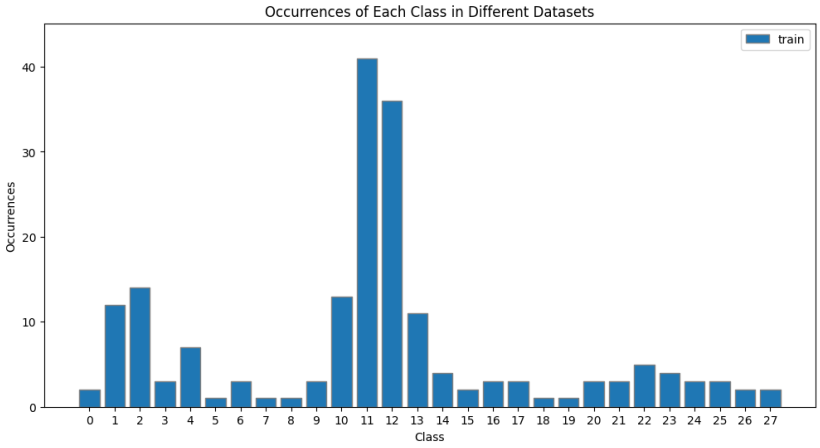

Figure 3: Validation data (20%).

In the evaluation of BPD and OFD, we used a similar cross-validation strategy; however, we did not include the GAN images and had a different number of classes. Specifically, OFD was classified into 14 classes ranging from 30 mm to 160 mm, while BPD was classified into 13 classes ranging from 20 mm to 140 mm.

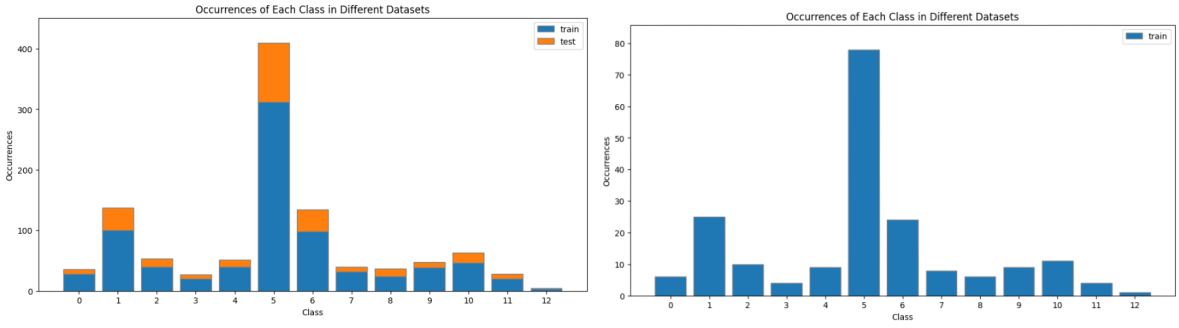

Figure 4: BPD class distribution.

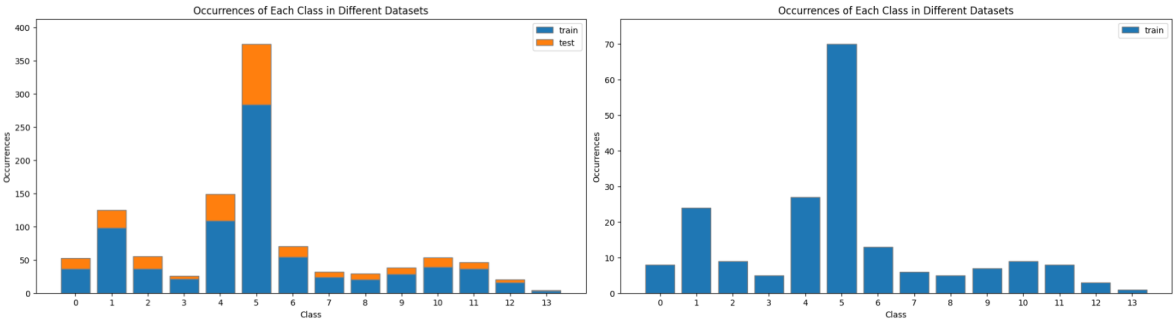

Figure 5: OFD class distribution.

Empirical evaluation

Prediction of fetal HC is a challenging task due to the limited availability and variability of data. To overcome these challenges and improve our model's performance, we implemented a series of strategies, both successful and unsuccessful. This section outlines the history of our empirical efforts and the insights gained from each approach we took.

All models were implemented with Pytorch and Timm. They were trained with a batch size of 5, a learning rate between 5e-4 to 1e-4, and AdamW as the optimizer. We used transfer learning and fine-tuned the models to achieve faster and better convergence. Also, we relied on CosineAnnealingLR to gradually reduce the learning rate during training.

In oversampling and downsampling, we used only the 999 training images from the HC18 dataset since adding 355 test images from the HC18 dataset wouldn't significantly improve results if the original 999 images didn't yield high scores. In the other evaluations, we added both the 999 training images and 355 test images from the HC18 dataset (in GAN-Based Data Generation the GAN images were also added).

Oversampling and Downsampling: Initially, to address the imbalance distribution between classes, we attempted to duplicate samples in smaller classes until they all reached the same number of samples as the largest class. The duplicated images did not introduce new information, and the training loss was significantly higher than the validation loss, showing a high likelihood of overfitting in classes where we duplicated hundreds of images. We also tried downsampling the number of samples in each class to 70, however, this approach did not yield any improvement either.

Classification with U-Net: Next, we tried to utilize a Nested U-Net model, a neural network architecture designed for image segmentation, to create masks of ultrasound images. These masks were intended to help the CNN model better identify and measure the HC by highlighting the relevant regions of the images. We tested two approaches: the first approach was to use the predicted mask as input for the CNN model (Figure 6) or to use the mask to keep only the pixels where the head is (Figure 7). Despite the theoretical benefits, this method did not yield improved results. We assume that the area around the fetal head in the original ultrasound image is beneficial data for the CNN model to classify more accurately, and removing it only deteriorates our results.

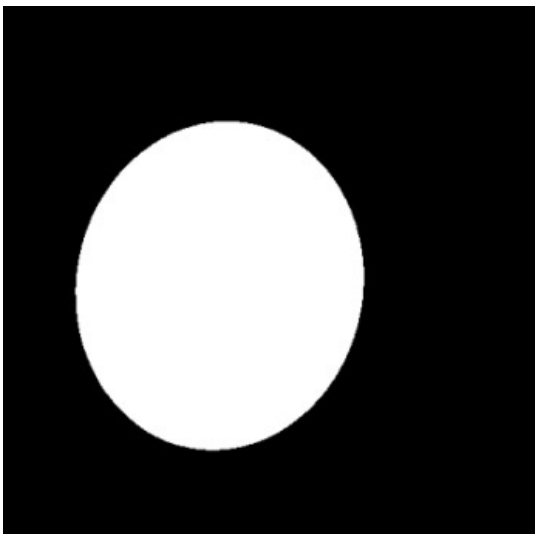

Figure 6: Mask example.

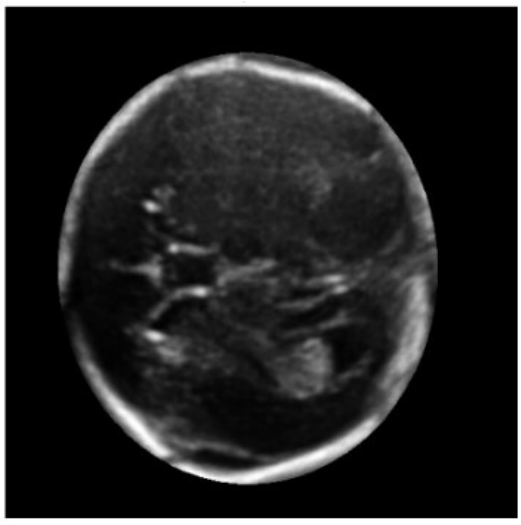

Figure 7: Only head example.

Diameter-Based Measurement: The second approach we attempted is measuring the diameters along and across key anatomical landmarks, specifically the BPD and OFD (Figure 8). The intention was to simplify the prediction task by focusing on these critical dimensions, allowing us to calculate the circumference with a few more multiplications between the predictions. Unfortunately, this method didn't yield any better results.

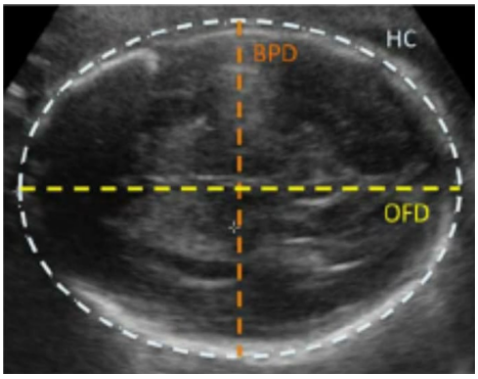

Figure 8: BPO/OFD Measurements.

Analyzing Data from the Second Trimester: We also tried to focus on weeks in the middle of pregnancy (18 - 25 weeks) when the fetal head is more grown and has more features for the model to detect. To do that, we kept only samples with a circumference between 15-24 cm (classes 18-25 in the diagram above). However, this strategy did not lead to any noticeable improvements. The reduced dataset might have limited the model's exposure to diverse cases, which is crucial for generalization. This experiment reinforced the importance of a comprehensive dataset that includes a wide range of cases to ensure the model's robustness and generalizability.

GAN-Based Data Generation: Later, we introduced additional synthetic data using GAN. By generating additional images with StyleGAN2, we balanced the quantities in each group more effectively. While this approach didn't improve the results for classification, it did lead to noticeable improvements in a regression model. The synthetic data provided more diversity and made it harder for the model to overfit by learning more robust features for HC prediction. To predict the synthetic image circumference, we trained a Nested U-net model to predict a mask where the head is in the image and a regression model to predict image pixel size. Using both predictions gave us a rough estimation of the image circumference. Again, due to the lack of real ultrasound images, the generated images sometimes lack crucial features in the fetal head (especially in earlier weeks of pregnancy). In addition, the pixel size predictions weren't very accurate in images of the final weeks of pregnancy. So, the performance improvement wasn't as significant as we had hoped for.

Transition to Regression Models: Finally, we transitioned from classification to regression models, aiming for more precise and consistent predictions. This shift resulted in better performance percentages. The regression approach allowed for finer granularity in predictions, which better suited the continuous nature of HC measurements. By treating HC prediction as a regression problem, we captured the constant and variable nature of the measurements more accurately, leading to more precise outcomes. Regression models are inherently better suited for tasks where the output variable is continuous, providing more nuanced and accurate predictions compared to classification models.

Conclusion: Through these iterative steps, we gained valuable insights into the challenges and potential solutions for predicting fetal head circumference. While some strategies did not yield the desired improvements, each attempt provided critical lessons that informed subsequent approaches. Our experience emphasizes the importance of data quality, variability, and appropriate model selection in achieving accurate and reliable predictions in medical imaging. The classification approach was more prone to overfitting than the regression one and we assume this is due to the lack of differentiating feature between each class. The empirical approaches have highlighted the complexities of fetal head circumference prediction and the necessity for innovative solutions to address these challenges more effectively.

Results

4.1 RandomResizedCrop and Augmentation Effects

Table 1: RandomResizedCrop(img_size, scale=(0.8, 1.0)) + HorizontalFlip + Rotation(15):

Model	Final F1 Score (Weighted) (%)	Final Accuracy (%)
EfficientNet-B0	30.48	29.95
MobileNetV3	23.51	24.60
ResNet18	13.78	12.30
VGG16	33.86	34.22

The RandomResizedCrop, combined with HorizontalFlip and Rotation, aimed to introduce variability and robustness in the model by slightly altering the input images during training. VGG16 showed the highest performance with a weighted F1 score of 33.86% and an accuracy of 34.22%. The other models, such as EfficientNet-B0, MobileNetV3, and ResNet18, demonstrated significantly lower performance, with ResNet18 performing the worst. This suggests that while VGG16 might be more resilient to this type of augmentation, the other models, especially ResNet18, might not benefit as much from these specific augmentations, potentially due to their architectural characteristics or inability to generalize well from such augmented data.

4.2 Impact of Gaussian Blur

Table 2: GaussianBlur(kernel_size=(5, 5), sigma=(2.0, 5.0)):

Model	Final F1 Score (Weighted) (%)	Final Accuracy (%)
EfficientNet-B0	35.16	38.50
MobileNetV3	30.17	33.69
ResNet18	26.91	28.88
VGG16	30.66	37.97

Gaussian Blur was introduced to the images to simulate the effect of varying image quality or focus, which is common in real-world ultrasound data. The results indicate a general improvement across all models compared to the RandomResizedCrop approach. EfficientNet-B0 achieved the highest accuracy of 38.50% and a weighted F1 score of 35.16%, showing that it could leverage the GaussianBlur augmentation more effectively. VGG16 also showed competitive results with an accuracy of 37.97%. The overall improvement suggests that Gaussian Blur helps the models to generalize better by focusing on significant features of the fetal HC that remain distinguishable even with blurred images.

4.3 Downsampling Approach

Table 3: Downsampling (70 samples per class):

Model	Final F1 Score (Weighted) (%)	Final Accuracy (%)
EfficientNet-B0	33.38	35.83
MobileNetV3	32.80	34.22
ResNet18	27.93	34.76
VGG16	38.32	39.04

The downsampling strategy, where each class was reduced to 70 samples, aimed to balance the dataset by ensuring uniform class distribution. VGG16 again outperformed other models, achieving

an F1 score of 0.3832 and an accuracy of 39.04%. The performance increase, particularly for VGG16, suggests that the network benefits from more balanced data, possibly due to its deeper architecture which requires less variance in data distribution for effective training. However, the results also indicate that simply downsampling may not always be sufficient, as the overall accuracy and F1 scores remain modest across models, implying that other factors, such as data quality and inherent model biases, play a significant role in performance.

4.4 Mask classification with U-Net

Table 4: classification with U-Net on predicted head zone

Model	Final F1 Score (Weighted) (%)	Final Accuracy (%)
EfficientNet-B0	31.10	32.62
MobileNetV3	30.21	33.16
ResNet18	25.35	35.83
VGG16	37.35	40.11

Table 5: classification with U-Net on predicted mask

Model	Final F1 Score (Weighted) (%)	Final Accuracy (%)
EfficientNet-B0	20.65	24.60
MobileNetV3	27.13	32.09
ResNet18	20.07	31.55
VGG16	24.48	28.88

When using the predicted head zone, VGG16 showed the highest performance, with a weighted F1 score of 37.35% and an accuracy of 40.11%. Other models, such as EfficientNet-B0, MobileNetV3, and ResNet18, had lower F1 scores ranging from 25.35% to 31.10% and accuracies between 32.62% and 35.83%. These results suggest that focusing on the head zone can improve classification accuracy, although the overall performance remains moderate.

In contrast, using the entire predicted mask led to a significant drop in performance across all models. VGG16's F1 score decreased to 24.48% with an accuracy of 28.88%, and other models showed similarly reduced scores, indicating that the inclusion of irrelevant regions in the mask likely introduced noise that hindered classification accuracy.

4.5 Evaluation of Diameter-Based Measurements

Table 6: OFD:

Model	Final F1 Score (Weighted) (%)	Final Accuracy (%)
EfficientNet-B0	47.46	51.28
MobileNetV3	47.81	50.26
ResNet18	43.12	52.31
VGG16	48.49	56.92

In the approach where the BPD and OFD were used as key measurements, the models showed a noticeable performance improvement. When focusing on OFD (Table 6), VGG16 had the best accuracy (56.92%) and a weighted F1 score of 48.49%. However, when BPD was used (Table 7), EfficientNet-B0 excelled with an accuracy of 65.13% and an F1 score of 61.06%. This notable improvement, especially with BPD, highlights the significance of using critical anatomical measurements in predicting fetal HC. It indicates that these dimensions provide crucial information that the models can leverage to make more accurate predictions. The higher performance with BPD suggests it may be a more reliable metric for HC prediction compared to OFD, possibly due to its stability and clear anatomical relevance.

Table 7: BPD:

Model	Final F1 Score (Weighted) (%)	Final Accuracy (%)
EfficientNet-B0	61.06	65.13
MobileNetV3	58.58	63.59
ResNet18	45.05	57.44
VGG16	52.21	52.31

4.6 Analyzing Data from the Second Trimester

Table 8: Analyzing Data from the Second Trimester (Weeks 18-26)

Model	Final F1 Score (Weighted) (%)	Final Accuracy (%)
EfficientNet-B0	40.16	41.89
MobileNetV3	42.89	45.95
ResNet18	30.09	37.84
VGG16	38.55	3.92

Focusing the dataset on the second trimester (18-26 weeks) aimed to provide the models with more consistent and distinguishable features by narrowing the range of fetal head sizes. The results, however, were mixed. MobileNetV3 outperformed the other models with an F1 score of 42.89% and an accuracy of 45.95%, while VGG16 and EfficientNet-B0 also showed moderate improvements. This suggests that focusing on the central data provided a moderate improvement, particularly for models like MobileNetV3, which may have benefited from the reduced variability. However, the overall modest gains indicate that while the second-trimester data provides a stable prediction ground, the loss of data diversity might have hindered the models' ability to generalize across a broader spectrum of fetal development stages.

4.7 Impact of GAN-Based Data Generation

Table 9: Without GAN Classification

Model	Final F1 Score (Weighted) (%)	Final Accuracy (%)
EfficientNet-B0	33.01	37.97
MobileNetV3	33.17	36.36
ResNet18	23.97	32.09
VGG16	39.57	40.64

Table 10: With GAN Classification

Model	Final F1 Score (Weighted) (%)	Final Accuracy (%)
EfficientNet-B0	31.87	35.29
MobileNetV3	39.53	39.57
ResNet18	29.16	36.90
VGG16	32.46	35.29

We used StyleGAN2 to generate synthetic data to balance the dataset and introduce more diversity. The results varied between classification and regression tasks. For classification (Tables 9 and 10), adding GAN images led to mixed results. While VGG16 showed slight improvement without GAN (F1 score of 39.57%), MobileNetV3 benefited from GAN augmentation, showing an improvement in both F1 score (39.53%) and accuracy (39.57%). This suggests that while GAN-generated data can add

value, it might not always lead to significant improvements, likely due to the synthetic images lacking critical features or introducing noise that doesn't align well with the real data distribution.

Table 11: Without GAN Regression

Model	Accuracy (10 mm) (%)	MAE (mm)
EfficientNet-B0	57.22	11.1570
MobileNetV3	70.59	8.3421
ResNet18	49.20	12.4314
VGG16	52.41	13.6040

Table 12: With GAN Regression

Model	Accuracy (10 mm) (%)	MAE (mm)
EfficientNet-B0	78.07	7.3534
MobileNetV3	77.01	7.3789
ResNet18	67.38	9.6480
VGG16	55.08	12.0434

In regression tasks (Tables 11 and 12), the introduction of GAN-generated data significantly improved performance, especially for EfficientNet-B0, which achieved an accuracy of 78.07% and a lower MAE of 7.3534 mm. This indicates that for continuous prediction tasks like regression, GANs can be particularly effective in providing the necessary data diversity, thereby enhancing the model's ability to generalize and make more accurate predictions. The reduction in MAE across models, especially with EfficientNet-B0 and MobileNetV3, emphasizes the potential of GANs to augment real data effectively, leading to better performance in regression-based approaches.

Conclusion The results demonstrate the varying effectiveness of different models and data augmentation strategies in predicting fetal head circumference. VGG16 consistently performed well across multiple strategies, particularly when downsampling and GaussianBlur were applied, while EfficientNet-B0 excelled in regression tasks with GAN augmentation. The experiments highlight the importance of choosing the right combination of model architecture and data preprocessing techniques to achieve optimal results in medical imaging tasks. Further exploration into more advanced augmentation techniques and hybrid models could yield even better predictive performance in future studies.

Summary

This study focused on enhancing the accuracy and reliability of fetal HC predictions using ultrasound imaging and advanced machine-learning techniques. Recognizing the challenges posed by data variability and the limitations of traditional segmentation methods, we explored a variety of approaches including classification and regression models, augmented with different data preprocessing strategies.

We started by experimenting with classification models combined with various augmentation techniques such as RandomResizedCrop, GaussianBlur, and downsampling. While these methods improved performance to some extent, they also highlighted the limitations of classification approaches in handling the complexities of HC prediction. The transition to regression models marked a significant improvement, with anatomical measurements like BPD and OFD proving to be particularly effective.

Additionally, the integration of GAN for synthetic data augmentation was essential in improving the performance of regression models. The GAN-generated data enhanced model generalization, leading to more accurate and consistent predictions, particularly with EfficientNet-B0.

Overall, the study demonstrates the importance of selecting appropriate model architectures and preprocessing techniques in medical imaging tasks. The shift from classification to regression models, combined with advanced augmentation methods, proved crucial in achieving better predictive accuracy in HC estimation.

Conclusion and future work

The research presented in this paper has successfully demonstrated the potential of using advanced machine learning techniques, specifically, regression models and GAN-based data generation, to improve the prediction of fetal HC from ultrasound images. By systematically evaluating various approaches, we found that regression models, particularly when supplemented with anatomical measurements like BPD, offered superior accuracy and consistency compared to traditional classification models.

One of the key contributions of this work is the application of GANs to generate synthetic data, which significantly enhanced the performance of our models, particularly in regression tasks. This approach not only mitigated the challenges posed by limited data availability but also introduced greater diversity into the training process, helping to reduce overfitting and improve model generalization.

Looking forward, there are various directions for future research. First, further refinement of GAN-generated images could help bridge the gap between synthetic and real ultrasound images, potentially leading to even greater improvements in model accuracy. Exploring more sophisticated GAN architectures or combining GANs with other generative models could be a promising direction.

Second, expanding the dataset to include a more diverse range of fetal development stages and ultrasound image qualities could help improve the robustness and generalizability of the models. Additionally, integrating other types of medical imaging data, such as 3D ultrasound or MRI, could provide a more comprehensive understanding of fetal development and further enhance prediction accuracy.

Lastly, the deployment of these advanced models in clinical settings requires thorough validation to ensure their reliability and safety. Collaborations with medical professionals and further research into model interpretability and explainability will be crucial for the successful integration of these technologies into clinical practice.

In conclusion, this study lays the groundwork for future advancements in the prediction of fetal head circumference using machine learning, with the potential to significantly impact prenatal care and improve outcomes for both mothers and their babies.

References

[1] G. Dubey, S. Srivastava, A. K. Jayswal, M. Saraswat, P. Singh, and M. Memoria, "Fetal ultrasound segmentation and measurements using appearance and shape prior based density regression with deep cnn and robust ellipse fitting," *PubMed Central*, 2024.

[2] F. A. Mohammed, K. K. Tune, B. G. Assefa, M. Jett, and S. Muhie, "Medical image classifications using convolutional neural networks: A survey of current methods and statistical modeling of the literature," *MDPI*, 2024.

[3] P. Ferber, M. Helmert, and J. Hoffmann, "Neural network heuristics for classical planning: A study of hyperparameter space," *ECAI 2020*, 2020.

[4] X. Wang, W. Wang, and X. Cai, "Automatic measurement of fetal head circumference using a novel gcn-assisted deep convolutional network," *Computers in Biology and Medicine*, vol. 145, p. 105515, 2022.

[5] J. Zhang, C. Petitjean, P. Lopez, and S. Ainouz, "Direct estimation of fetal head circumference from ultrasound images based on regression cnn," *MIDL 2020*, 2020.

[6] T. Karras and J. Hellsten, "stylegan2-ada-pytorch." https://github.com/NVlabs/stylegan2-ada-pytorch, 2021.