Generative Diffusion Model Bootstraps Zero-Shot Classification of Fetal Ultrasound Images in Underrepresented African Populations

Fangyijie Wang[1,2(✉)], Kevin Whelan[2], Guénolé Silvestre[1,3], and Kathleen M. Curran[1,2]

[1] Science Foundation Ireland Centre for Research Training in Machine Learning, Dublin, Ireland
fangyijie.wang@ucdconnect.ie
[2] School of Medicine, University College Dublin, Dublin, Ireland
[3] School of Computer Science, University College Dublin, Dublin, Ireland

Abstract. Developing robust deep learning models for fetal ultrasound image analysis requires comprehensive, high-quality datasets to effectively learn informative data representations within the domain. However, the scarcity of labelled ultrasound images poses substantial challenges, especially in low-resource settings. To tackle this challenge, we leverage synthetic data to enhance the generalizability of deep learning models. This study proposes a diffusion-based method, **F**etal **U**ltrasound **LoRA** (**FU-LoRA**), which involves fine-tuning latent diffusion models using the LoRA technique to generate synthetic fetal ultrasound images. These synthetic images are integrated into a hybrid dataset that combines real-world and synthetic images to improve the performance of zero-shot classifiers in low-resource settings. Our experimental results on fetal ultrasound images from African cohorts demonstrate that FU-LoRA outperforms the baseline method by a 13.73% increase in zero-shot classification accuracy. Furthermore, FU-LoRA achieves the highest accuracy of 82.40%, the highest F-score of 86.54%, and the highest AUC of 89.78%. It demonstrates that the FU-LoRA method is effective in the zero-shot classification of fetal ultrasound images in low-resource settings. Our code and data are publicly accessible on GitHub.

Keywords: Diffusion Model · Synthetic Ultrasound Data · Fetal Anatomical Planes Classification · Low-resource Settings

1 Introduction

Ultrasound imaging is widely used for the diagnosis, screening and treatment of many diseases because of its portability, low cost and non-invasive nature [28]. In recent decades, ultrasound screening has become a common method used for

D. Link-Sourani et al. (Eds.): PIPPI 2024, LNCS 14747, pp. 143–154, 2025.
https://doi.org/10.1007/978-3-031-73260-7_13

prenatal evaluation of fetal growth, fetal anatomy, and estimation of gestational age (GA), as well as monitoring pregnancy [15,24]. After 14 weeks of gestation, GA is estimated using standard measurements such as head circumference, biparietal diameter, occipito-frontal diameter, transcerebellar diameter, lateral ventricles, abdominal circumference, and femur length [2]. These fetal biometric measurements are performed following a standardized procedure through the identification of the standard sonographic plane [25]. An automated classification system can assist sonographers in promptly and accurately identify maternal-fetal standard planes. Recent deep learning methods have demonstrated significant potential for analyzing fetal ultrasound images. However, existing developments have mainly concentrated on applications in high-resource settings (HRS), where there is access to extensive clinical imaging datasets utilized for training deep learning models [17]. In low-resource settings (LRS), the scarcity of clinical images remains a significant challenge for the generalizability of deep learning models. LRS is characterized by a lack of adequate healthcare resources and systems that fail to meet recognized global standards [17].

Utilization of diffusion models for synthetic image generation has the potential to enrich medical imaging datasets, especially in scenarios where data are limited in LRS and where increased diversity is essential in existing medical imaging modalities [10]. The diffusion model has been used in various medical image enhancement applications, such as Computed Tomography (CT) [4], Positron Emission Tomography (PET) [7], Magnetic resonance imaging (MRI) [18], X-ray [11], and ultrasound [8]. Lee et al. [13] propose an information maximizing generative adversarial network (GAN) to generate synthetic examples of fetal brain ultrasound images from twenty-week anatomy scans. Lasala et al. [12] introduce an approach that leverages class activation maps (CAM) as a prior condition to generate standard planes of the fetal head using a conditional GAN model. However, these studies only investigate the effectiveness of synthetic fetal head ultrasound images, and their methods are image-to-image based generation approaches.

In this study, we aim to address the data scarcity of fetal ultrasound images in LRS. We propose a novel text-to-image generation approach that utilizes the diffusion model to enhance the performance of convolutional neural networks (CNN) for the classification of five common maternal fetal ultrasound planes. The proposed approach involves the following steps: i) Fine-tuning the diffusion model on a dataset collected in HRS to enable the pre-trained model to learn distinct features in fetal ultrasound data, ii) Creating hybrid datasets by integrating real-world data with synthetic data using the fine-tuned text-to-image diffusion model.

Our contributions are: i) proposing a novel approach to improve zero-shot image classification accuracy in low-resource settings using synthetic data generated from latent diffusion models, ii) publishing the first LoRA model for the generation of synthetic fetal ultrasound images, and iii) releasing a synthetic dataset of 5000 images for further research in fetal ultrasound field.

2 Methods

Preliminaries. Latent Diffusion Model (LDM) [20], stands out as one of the most successful diffusion models available within the current open-source community. Structurally, LDM is a Denoising Diffusion Probabilistic Model (DDPM) [5] implemented in the latent space to efficiently decrease computational costs.

Training the LDM involves a process that contains a diffusion (or forward) process and a sampling (or reverse) process. Given an image $x_0 \in X$, the diffusion process gradually adds random Gaussian noise to the input image x_0 at diffusion step t (T denotes the total number of diffusion steps) to produce x_t, following a Markov Chain by: $q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right)$ where $\beta_t \in [0,1]$ represents the variance schedule across diffusion steps, $\boldsymbol{I}$ is the identity matrix, x_t and x_{t-1} are adjacent image status. Accordingly, a noisy target x_t distribution from the data x_0 is represented as: $x_t = \sqrt{\alpha_t}x_0 + \sqrt{(1-\alpha_t)}\epsilon$ where $\alpha_t = \Pi_{s=1}^{t}(1-\beta_s)$. Then a U-Net [21] is trained to approximate the reverse denoising process: $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \beta_t \boldsymbol{I}\right)$ where μ_θ is a parameterized mean with the noise predictor ϵ_θ, which consists of a U-Net (θ denotes model parameters):

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right) \tag{1}$$

The U-Net (ϵ_θ) is trained with the mean square loss $L := E_{t,x}||\epsilon - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)||^2$ where $\epsilon \sim N(0, \boldsymbol{I})$. In the sampling process, LDM learns the Markov chain to convert the Gaussian noise distribution $x_T \sim N(0, \boldsymbol{I})$ into the target distribution x_0 by the iterative denoising steps: $\boldsymbol{x}_{t-1} = \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) + \beta_t \boldsymbol{z}$ where $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$. We follow Rombach's work [20] to control the generation procedure by integrating the conditioning input, text c, to the noise predictor U-Net $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$. Therefore, $\boldsymbol{\epsilon}_\theta$ is defined as $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, c, t)$ optimized according to the objective loss L where c represents the text prompts with CLIP (Contrastive Language-Image Pre-Training) encoding [19].

The Low-Rank Adaption (LoRA) technique [6] is designed for the efficient fine-tuning of large language models and can also be utilized for fine-tuning generative models. Therefore, we apply LoRA technique to the cross-attention layers within the U-Net architecture ($\boldsymbol{\epsilon}_\theta$) [21]. This integration accelerates the training process, leading to decreased computational demands and a reduced model size. The formula of LoRA is defined as:

$$h = W_0 x + \frac{\alpha}{r}\Delta W x = W_0 x + \frac{\alpha}{r} BAx \tag{2}$$

where W_0 is a pre-trained weight matrix of U-Net within diffusion models. W_0 can be decomposed into two smaller matrices, denoted as A and B, with a reduced rank r compared to the original matrix. $\frac{\alpha}{r}$ represents the merging ratio and ranges from 0 to 1. During the training process, W_0 remains frozen, while the matrices A and B are equipped with trainable parameters. Matrix A is initialized using Gaussian distribution, while matrix B is initialized with zeros.

FU-LoRA. Figure 1 shows the two steps involved in our FU-LoRA method: (1) Fine-tuning the pre-trained diffusion model using the LoRA method on a small fetal ultrasound dataset from HRS. (2) Employing the fine-tuned LoRA model for training downstream tasks in LRS. This approach integrates synthetic images to enhance generalization and performance of deep learning models. We conduct three fine-tuning sessions for the diffusion model to generate three LoRA models with different hyper-parameters: $\alpha \in [8, 32, 128]$, and $r \in [8, 32, 128]$. The merging rate $\frac{\alpha}{r}$ in Eq. 2 is fixed to 1. The purpose of this operation is to delve deeper into LoRA to uncover optimizations that can improve the model's performance and evaluate the effectiveness of parameter r in generating synthetic images.

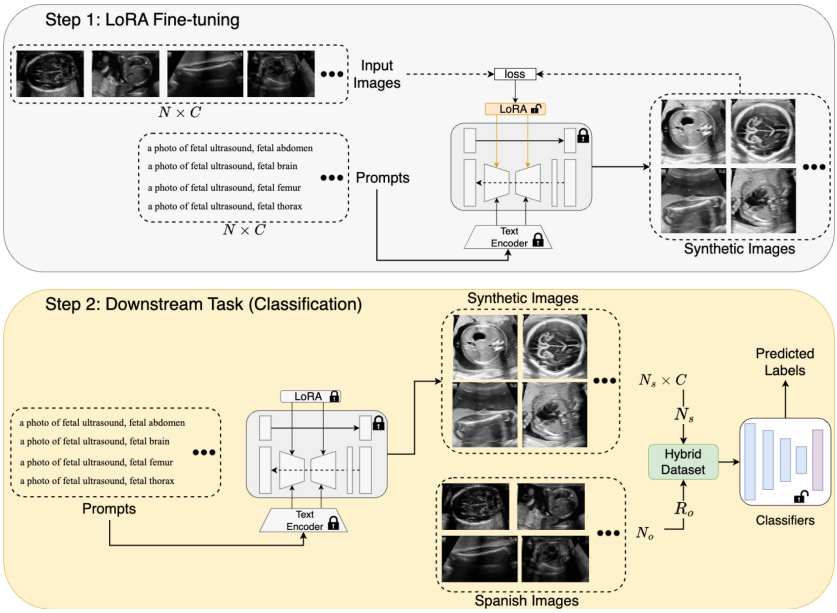

Fig. 1. The overview of our proposed FU-LoRA method. N: number of Spanish data per plane for training LoRA; C: number of planes; N_s: number of synthetic images; N_o: total number of Spanish images; R_o: random subset of N_o.

Downstream Classification. We aim to evaluate the quality and utility of the synthetic data generated by FU-LoRA. Therefore, we use them to train five classifiers, incorporating DenseNet169, ResNet18, EfficientNet b0, MobileNet v2, and Vision Transformer (ViT b16), which are subsequently tested on African data. The final fully connected layers of these models are substituted with new fully connected layers having an output dimension of 5. Initially, all models are pre-trained on ImageNet [23] and subsequently fully re-trained throughout the entire network.

Evaluation Metrics. The generation performance of LoRA with different values of rank r in Eq. 2 is assessed by evaluating the diversity of the generated samples using t-distributed stochastic neighbor embedding (t-SNE) [16]. t-SNE is a statistical technique used for visualizing high-dimensional data, assigning each data point a specific location on a two or three-dimensional map. We extract ultrasound image features using the InceptionNet [27] pre-trained on ImageNet and visualize them by t-SNE in a two-dimensional map (see Sect. 4). In the African dataset, the fetal thorax plane contains 40 images. To ensure fair evaluation, we randomly select 35 images from the Spanish, Synthesis, and African datasets for four anatomical planes: Abdomen, Brain, Femur, and Thorax. Thus, a total of 140 images are gathered for qualitative analysis.

3 Materials and Experiments

Datasets: We utilize two public datasets and one synthetic dataset in this study: i) the Spanish dataset is acquired from two centers in Spain [1], ii) the African dataset is acquired from Algeria, Egypt, Malawi, Ghana, and Uganda [26], and iii) the synthetic fetal ultrasound dataset is generated by the FU-LoRA method. The variability of the fetal head across various pregnancy trimesters poses a challenge in developing a robust deep learning model.

The Spanish dataset[1] in HRS includes 1,792 patient records in Spain [1]. All images are acquired during screening in the second and third trimesters of pregnancy using six different machines operated by operators with similar expertise. We randomly selected 20 Spanish ultrasound images from each of the five maternal-fetal planes (Abdomen, Brain, Femur, Thorax, and Other) to fine-tune the LDM using LoRA technique, and 1150 Spanish images (230 $\times$ 5 planes) to create the hybrid dataset. In summary, fine-tuning the LDM utilizes 100 images including 85 patients. Training downstream classifiers uses 6148 images from 612 patients. Within the 6148 images used for training, a subset of 200 images is randomly selected for validation purposes. The hybrid dataset employed in this study has a total of 1150 Spanish images, representing 486 patients.

The African dataset[2] in LRS contains 450 images (125 patients) from five African countries [26] collected with five different ultrasound machines during the second and third pregnancy trimesters. This dataset only contains four maternal-fetal standard planes, including Abdomen, Brain, Femur, and Thorax. A total of 217 images from 61 patients can be used for training, while 233 images from 66 patients are allocated for testing purposes. To evaluate generalization of classification models with synthetic data, the 217 training images are only used for ablation studies as detailed in Sect. 4. The 233 testing images, on the other hand, are utilized for evaluating the performance of downstream classifiers, as discussed in Sect. 4.

[1] https://zenodo.org/records/3904280.
[2] https://zenodo.org/records/7540448.

We create the synthetic dataset comprising 5000 fetal ultrasound images (500 × 2 samplers × 5 planes) accessible to the open-source community. The generation process utilizes our LoRA model Rank $r = 128$ with Euler [9] and UniPC [14] samplers known for their efficiency. Subsequently, we integrate this synthetic dataset with a small amount of Spanish data to create a hybrid dataset, shown in Fig. 1.

Traditional Data Augmentation: The traditional data augmentation techniques are employed to train downstream classifiers for performance comparison with our methods. These techniques are: rotation by an angle from $[-25°, 25°]$, horizontal flipping with 50% probability ($P(\cdot) = 0.5$), vertical flipping with 10% probability ($P(\cdot) = 0.1$), and pixel normalization to float precision in $[-1, 1]$ range. All images are resized to 512 × 512 pixels for training, validation and testing purposes.

Implementation Details: The hyper-parameters of LoRA models are defined as follows: batch size to 2; training epochs to 1; LoRA learning rate to $1e{-}4$; total training steps to 10000 (100 images × 100 steps × 1 epoch); LoRA dimension to 128; mixed precision selection to fp16; learning scheduler to constant; and input size (resolution) to 512. The model is trained on a single NVIDIA RTX A5000, 24 GB with 8-bit Adam optimizer on PyTorch. The training hyper-parameters of downstream classification models are: epochs to 20, batch size to 24, and learning rate to $1e - 3$. The loss function computes the cross-entropy loss between input logits and target. All training processes are conducted on a single NVIDIA RTX 4090, 24 GB with stochastic gradient descent (SGD) optimizer on PyTorch. The input pixels of all images are converted from integer in the range $[0, 255]$ to single float precision in $[-1, 1]$ for training and testing the downstream classification models.

4 Results

Synthetic Images: The process of generating synthetic images from text is commonly known as text-to-image generation. We utilize two sampling methods: Euler [9] and UniPC [14]. During the process, the weights of LoRA are fixed at 1.0, and the number of sampling steps is maintained at 20. As a result, we generate a hybrid dataset of 5000 images by utilizing the LoRA Rank 128 model to generate 2500 synthetic images using Euler and UniPC samplers, respectively.

Figure 2 presents the samples of the synthetic fetal ultrasound images generated with our proposed FU-LoRA method in Fig. 1 by providing text prompt inputs for each maternal plane. It shows that the fetal abdominal images depict the stomach situated within the abdomen. The skull is prominently visible in the brain images as a bright, elliptical structure. Consequently, the femur images illustrate a single fetal femur with high brightness. The fetal thorax images feature the right atrium, left atrium, and both ventricles.

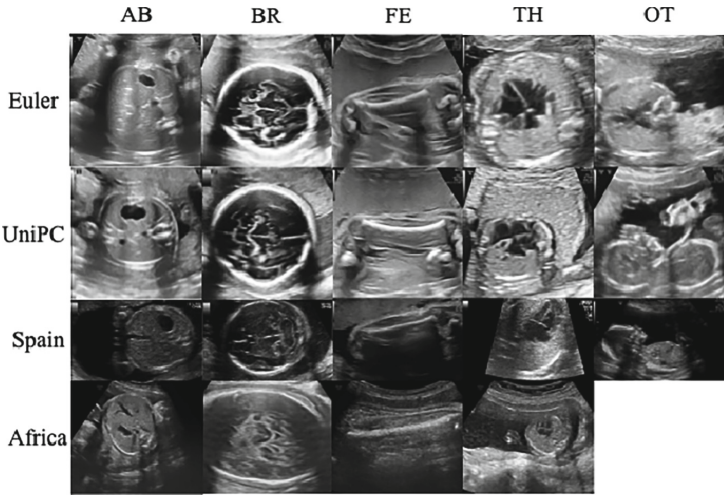

Fig. 2. Examples of synthetic images generated using the FU-LoRA method are presented. The first two rows display images generated by the Euler sampler and the UniPC sampler, while the third and fourth rows show Spanish and African images utilized for fine-tuning the diffusion model. AB: Abdomen; BR: Brain; FE: Femur; TH: Thorax; OT: Other.

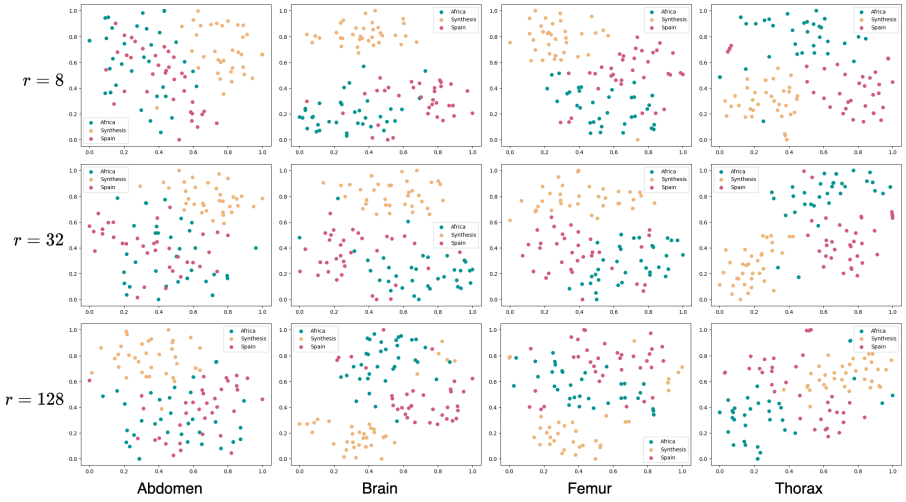

Fig. 3. An overview of the feature embeddings of synthetic images generated by LoRA models (rank $r \in [128, 32, 8]$) using Spanish and African images through t-SNE visualization. The feature embeddings are extracted using a pre-trained InceptionNet model. From left to right, the t-SNE visualization is applied to Abdomen, Brain, Femur, and Thorax plane. Green: Africa; Yellow: Synthesis; Red: Spain. (Color figure online)

The t-SNE visualization in Fig. 3 offers a qualitative assessment of the overall quality of the generated images. Within this two-dimensional embedded space, the real (green and red) and synthetic (yellow) data are positioned in distinct regions in Abdomen, Brain and Femur planes with rank $r = 8$. There are obvious separations in Brain and Thorax planes with rank $r = 32$. The lack of separation in the feature space in Abdomen, Femur and Thorax with rank $r = 128$. Compared with the other two LoRA models with lower rank r as shown in the second and third rows of Fig. 3, the model with rank $r = 128$ generates data points that are located closer to those regions of real-world data.

Table 1. An overview of the classification performance of multiple deep learning models on the African fetal ultrasound dataset. ES: Spanish dataset; HB: Hybrid dataset (Spanish and Synthetic images); DA: Traditional data augmentation techniques.

Data	Model	Training Data			ACC	Recall	Precision	F-score	AUC
		# ES	# Synthetic	# Patients					
ES	DenseNet169	6148	0	612	61.37	62.38	96.46	73.31	80.50
ES	MobileNet v2	6148	0	612	68.67	68.86	96.50	78.07	83.99
ES	EfficientNet b0	6148	0	612	63.09	62.64	96.94	74.34	81.15
ES	ResNet18	6148	0	612	54.94	51.23	90.95	61.51	76.09
ES	ViT b16	6148	0	612	68.24	67.05	96.58	77.53	83.46
ES+DA	DenseNet169 [26]	6148	0	612	68.67	67.67	92.43	75.87	82.52
ES+DA	MobileNet v2	6148	0	612	69.96	69.05	95.27	78.95	83.93
ES+DA	EfficientNet b0	6148	0	612	59.66	59.26	94.72	70.91	79.09
ES+DA	ResNet18	6148	0	612	63.09	61.10	91.52	71.01	79.86
ES+DA	ViT b16	6148	0	612	74.68	72.80	97.05	82.15	86.23
HB	DenseNet169	1150	5000	486	75.54	73.40	**98.66**	83.40	77.04
HB	MobileNet v2	1150	5000	486	67.81	65.18	96.52	76.63	75.78
HB	EfficientNet b0	1150	5000	486	72.96	69.81	97.50	79.45	77.65
HB	ResNet18	1150	5000	486	64.81	60.15	96.92	69.95	76.88
HB	ViT b16	1150	5000	486	72.10	70.28	96.81	80.57	76.26
HB+DA	DenseNet169	1150	5000	486	80.26	77.61	89.11	82.59	87.99
HB+DA	MobileNet v2	1150	5000	486	77.68	74.29	86.69	79.36	86.11
HB+DA	EfficientNet b0	1150	5000	486	72.96	68.75	92.65	75.91	83.87
HB+DA	ResNet18	1150	5000	486	**82.40**	**81.56**	89.15	84.90	89.29
HB+DA	ViT b16	1150	5000	486	81.97	79.92	95.36	**86.54**	**89.78**

Effectiveness Evaluation: We evaluate the model's classification performance using synthetic data to classify fetal anatomical planes. Notably, African data are used only to test classification models. The quantitative results of five classification models are provided in Table 1. It is observed that utilizing the $HB+DA$ data model yields highly comparable accuracy for anatomical plane classifiers.

Table 2. Effectiveness of utilizing various data models to train ViT b16 to identify maternal-fetal standard planes in African populations. ES: Spanish dataset; AF: African dataset; HB: Hybrid dataset. ACC: Accuracy. AUC: Area under the ROC Curve.

Data	Model	Training Data				ACC	Recall	Precision	F-score	AUC
		# ES	# AF	# Synthetic	# Patients					
AF	ViT b16	0	217	0	61	91.41	88.71	93.23	90.44	93.81
ES	ViT b16	6148	0	0	612	68.24	67.05	96.58	77.53	83.46
ES+AF	ViT b16	6148	217	0	673	**92.27**	**91.53**	96.67	**93.87**	**95.40**
HB	ViT b16	6148	0	5000	612	78.54	78.14	**97.44**	85.85	88.76

Among the five classifiers, ViT achieves the highest F-score of 86.54% and AUC of 89.78% on African data. More importantly, we observe that $HB + DA$ [ViT] achieves a significant improvement in AUC of 7.2% compared to the model $ES + DA$ [DenseNet169] used by other researchers [26]. The average F-score across all models trained with $HB + DA$ is 81.86%, whereas HB, ES and $ES + DA$ data models have average F-score of 78%, 72.95% and 75.78%, respectively. Compared to the ES [MobileNet] and $ES + DA$ [ViT] data models, the HB [DenseNet169] increases the best classification accuracy by 6.87% and 0.86%. Notably, the $HB + DA$ [ResNet18] has 13.73% higher accuracy than the baseline model $ES + DA$ [DenseNet169] [26].

Ablation Studies: To verify the effectiveness of our proposed method, FU-LoRA, we compare our hybrid data model with the other three data models. Because the ViT b16 model achieves the highest F-score and AUC in Table 1, it is selected to train and test with the same experimental settings for ablation studies. Their quantitative results are given in Table 2. Our observations are: (1) Our HB data model has better zero-shot classification performance than the ES data model across all metrics using 1150 Spanish images, a significantly smaller quantity than the original training set. (2) Compared to AF data model, HB exhibits a reduction in F-score of 4.6% and AUC of 5%, respectively, in the absence of target domain data during training. (3) Compared to the best case scenario of having $ES + AF$ for training, our HB data model has acceptable differences across all evaluation metrics. Our key finding reveals that in the context of zero-shot classification, our HB data model exhibits superior performance and generalization compared to the ES data model when African data is not included for training classifiers.

5 Limitations

Our proposed method showcases promising results in generating realistic and anatomically meaningful synthetic images based on textual information. Nevertheless, our current work still has limitations. This study integrates text prompts

to generate accurate and medically significant images. Integrating more detailed texts could improve the diffusion model. Moreover, Fig. 3 shows the separation in the feature embeddings of the Brain and Femur planes with rank $r = 128$. It suggests that synthetic images may lack semantically relevant features. Higher-quality synthetic images could be generated using alternative fine-tuning methods such as DreamBooth [22] and Textual Inversion [3]. However, our study focuses on the feasibility of the most efficient fine-tuning method, LoRA, which involves the fewest images. We aim to address these limitations in future work.

6 Conclusion

In this study, we have effectively fine-tune the LDM to produce synthetic fetal ultrasound images that accurately reproduce the characteristics observed in real-world images. Our proposed method, FU-LoRA, requires only 100 fetal ultrasound images from 85 patients for the effective process of fine-tuning. Additionally, we demonstrate that FU-LoRA method can facilitate zero-shot classification for standard fetal ultrasound planes in obstetric ultrasound within low-resource settings. Using a hybrid dataset comprising synthetic images generated by FU-LoRA for training deep learning models, we achieve the highest F-score of 86.54% and the highest AUC of 89.78% in classifying African standard fetal planes. We publicly release a synthetic dataset to address a significant challenge in ultrasound image analysis: the scarcity of extensive annotated fetal ultrasound data, all while safeguarding privacy. In conclusion, our results highlight the potential of the FU-LoRA method for zero-shot learning in fetal ultrasound imaging analysis within low-resource settings.

Acknowledgments. This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Burgos-Artizzu, X.P., et al.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Sci. Rep. **10**(1), 10200 (2020). https://doi.org/10.1038/s41598-020-67076-5
2. Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E., Moccia, S.: A review on deep-learning algorithms for fetal ultrasound-image analysis. Med. Image Anal. **83**, 102629 (2023)
3. Gal, R., et al.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022). https://doi.org/10.48550/ARXIV.2208.01618. https://arxiv.org/abs/2208.01618
4. Gao, Q., Li, Z., Zhang, J., Zhang, Y., Shan, H.: CoreDiff: contextual error-modulated generalized diffusion model for low-dose CT denoising and generalization. IEEE Trans. Med. Imaging **43**(2), 745–759 (2024). https://doi.org/10.1109/TMI.2023.3320812

5. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
6. Hu, E.J., et al.: LoRA: low-rank adaptation of large language models. In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=nZeVKeeFYf9
7. Jiang, C., et al.: PET-diffusion: unsupervised PET enhancement based on the latent diffusion model. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14220, pp. 3–12. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43907-0_1
8. Jiménez-Gaona, Y., Carrión-Figueroa, D., Lakshminarayanan, V., José Rodríguez-Álvarez, M.: GAN-based data augmentation to improve breast ultrasound and mammography mass classification. Biomed. Signal Process. Control **94**, 106255 (2024). https://doi.org/10.1016/j.bspc.2024.106255
9. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022). https://openreview.net/forum?id=k7FuTOWMOc7
10. Kazerouni, A., et al.: Diffusion models in medical imaging: a comprehensive survey. Med. Image Anal. **88**, 102846 (2023)
11. Kim, B., Oh, Y., Ye, J.C.: Diffusion adversarial representation learning for self-supervised vessel segmentation. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=H0gdPxSwkPb
12. Lasala, A., Fiorentino, M.C., Bandini, A., Moccia, S.: FetalBrainAwareNet: bridging GANs with anatomical insight for fetal ultrasound brain plane synthesis. Comput. Med. Imaging Graph. **116**, 102405 (2024)
13. Lee, L.H., Noble, J.A.: Generating controllable ultrasound images of the fetal head. In: 17th IEEE International Symposium on Biomedical Imaging ISBI, pp. 1761–1764. IEEE (2020). https://doi.org/10.1109/ISBI45749.2020.9098578
14. Liu, E., Ning, X., Yang, H., Wang, Y.: A unified sampling framework for solver searching of diffusion probabilistic models. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=W2d3LZbhhI
15. Loughna, P., Chitty, L., Evans, T., Chudleigh, T.: Fetal size and dating: charts recommended for clinical obstetric practice. Ultrasound **17**(3), 160–166 (2009). https://doi.org/10.1179/174313409X448543
16. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(86), 2579–2605 (2008). http://jmlr.org/papers/v9/vandermaaten08a.html
17. Piaggio, D., Castaldo, R., Cinelli, M., Cinelli, S., Maccaro, A., Pecchia, L.: A framework for designing medical devices resilient to low-resource settings. Glob. Health **17**(1), 64 (2021)
18. Pinaya, W.H.L., et al.: Brain imaging generation with latent diffusion models. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., Yuan, Y. (eds.) DGM4MICCAI 2022. LNCS, vol. 13609, pp. 117–126. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-18576-2_12
19. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021). https://proceedings.mlr.press/v139/radford21a/radford21a.pdf

20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, June 2022
21. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
22. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22500–22510, June 2023
23. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
24. Salomon, L.J., et al.: on behalf of the ISUOG Clinical Standards Committee: practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultras. Obstet. Gynecol. **37**(1), 116–126 (2011). https://doi.org/10.1002/uog.8831
25. Salomon, L.J., et al.: ISUOG practice guidelines: ultrasound assessment of fetal biometry and growth. Ultras. Obstet. Gynecol. **53**(6), 715–723 (2019)
26. Sendra-Balcells, C., et al.: Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five African countries. Sci. Rep. **13**(1), 2728 (2023). https://doi.org/10.1038/s41598-023-29490-3
27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
28. Zaffino, P., Moccia, S., De Momi, E., Spadea, M.F.: A review on advances in intraoperative imaging for surgery and therapy: imagining the operating room of the future. Ann. Biomed. Eng. **48**(8), 2171–2191 (2020)