LEARNING SUBJECTIVE IMAGE QUALITY ASSESSMENT FOR TRANSVAGINAL ULTRASOUND SCANS FROM MULTI-ANNOTATOR LABELS

Author(s) Name(s)

Author Affiliation(s)

ABSTRACT

This paper proposes a novel AI model that automatically assesses the quality of transvaginal ultrasound (TVUS) images, offering support to sonographers, especially those still learning, in acquiring high-quality scans for gynecological pathology diagnosis. Addressing the challenge of varying interpretations by different medical professionals, this model approaches the issue as a multi-annotator noisy label problem. Our novel machine learning architecture first aggregates quality assessments from multiple raters using a weighted ensemble algorithm to estimate consensus labels. The model then employs a multi-axis vision transformer to enhance the process of image quality evaluation. We evaluated the model on a new multi-annotator TVUS dataset, where our model successfully predicted image quality with an accuracy of 80%. This development represents an exciting first step in empowering sonographers to assess scan quality on the spot, reduce the need for repeated imaging, and improve the diagnosis of gynecological pathology.

***Index Terms*—** Image Quality Assessment, Multi-Rater, Transvaginal Ultrasound

1. INTRODUCTION

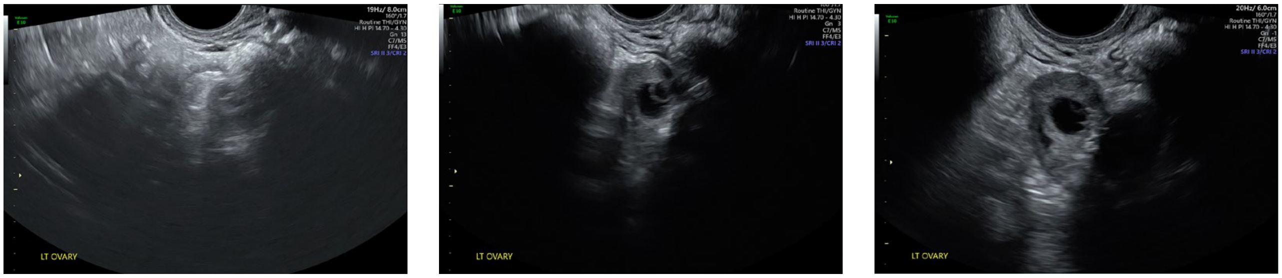
Transvaginal Ultrasound (TVUS) is the first-line tool in the diagnosis of many gynaecological conditions and has, in recent years, emerged as a promising, non-invasive diagnostic tool for endometriosis, with an average sensitivity rate of 79% in recent findings [1]. It has gained recognition for its potential to reduce unnecessary laparoscopy, which is considered the gold standard for diagnosing endometriosis , but presents drawbacks including dependency on surgical skill, potential for injury, and a small but relevant mortality rate [2]. However, obtaining high-quality TVUS images, which are crucial for diagnosis, requires a high level of sonographic expertise. Currently, this level of expertise is lacking within the field, with few people posessing the skills required to perform TVUS to diagnose endometriosis. [3].

In this context, an AI model capable of assessing the quality of TVUS images would be of significant value. It could assist non-specialist sonographers in assessing the quality of TVUS images during the scan, thereby providing high quality scans that could facilitate an accurate diagnosis of endometriosis.This would reduce the need for patients to return for subsequent imaging sessions, improve the patient experience and lead to better outcomes [1]. However, developing an AI model for the analysis of TVUS images poses unique challenges due to the involvement of multiple professionals. These images are initially captured by sonographers and later interpreted by radiologists or sonologists for diagnosis. Consequently, the model needs to account for the varying perspectives of these professionals. To the best of our knowledge, no multi-annotator dataset for TVUS images currently exists for the development of AI models.

To achieve this objective, we introduce a novel AI model for TVUS image quality assessment, specifically tailored to accommodate data with multi-annotator labels and a new multi-annotator TVUS image quality assessment dataset. The training process of our model unfolds in three stages : 1) Fine-tuning a pre-trained multi-axis vision transformer using a small subset of multi-annotated TVUS images, applying majority voting to reconcile label discrepancies. 2) Identifying noisy labels, evaluating the accuracy of each annotator's annotations, and generating refined labels. 3) Further fine-tuning the multi-axis vision transformer with these enhanced labels to improve model performance. The new TVUS dataset consisting of 150 TVUS images from 50 unique patients, with each image evaluated by six medical professionals, including two sonographers, two radiologists, and two gynecologists. Furthermore, we have adopted a novel grading system for appraising TVUS image quality, as proposed by Deslandes et al. (2023) [4], which considers various subjective factors such as the visibility of anatomic structures, interpretability of the scan, and the reliability of the scan for diagnostic purposes.

Contributions This paper represents the first exploration of multi-annotator subjective image quality assessment (IQA) for TVUS scans. This paper offers two principal contributions: First, we introduce and implement a novel approach for training an AI model on subjective IQA using a dataset annotated by multiple annotators. This method is designed to leverage the diverse perspectives of various annotators to enhance the model's assessment capabilities. Second, we establish the first dataset for multi-annotated subjective IQA specifically tailored for TVUS scans, facilitating the diagnosis

Some author footnote.



Grade 1 Grade 2 Grade 3

Fig. 1. TVUS images of each image quality grade. The anatomy in the Grade 1 image is occluded and not clearly recognisable. In the Grade 2 image, the target anatomy feature's are more prominent. The Grade 3 image anatomy is confidently recognisable with high image clarity.

of endometriosis. This dataset is a pioneering resource in the field. The encouraging outcomes highlight the model's potential to assist sonographers in capturing high-quality TVUS images, which in turn can provide invaluable support to gynecologists in the accurate diagnosis of endometriosis.

2. RELATED WORK

IQA for Ultrasound Images has been a topical area of research. Highlighting recent developments, Zhang et al. [5] demonstrated that CNNs outperform traditional methods in evaluating ultrasound image quality, using a dataset of high-quality images degraded by various techniques. Luo et al. [6] introduced a multitask learning-based IQA scheme for fetal sonography, ensuring essential anatomical structures are clear, achieving top performance in assessing fetal head, abdomen, and heart sections. Asch et al. [7] presented a machine learning algorithm to estimate ventricular contraction in echocardiography, achieving clinical-level sensitivity and specificity. Last but not least, Schneider et al. [8] developed an algorithm for echocardiogram image quality assessment with 83.2% accuracy, coupled with a feedback system for echocardiographers to improve image quality. However, exiting methods have not explored IQA tasks within a multi-annotator context. Furthermore, there is a lack of datasets specifically collected for subjective, multi-annotator IQA of TVUS. Addressing these gaps, this paper introduces a model designed for multi-annotator IQA tasks along with a multi-annotator TVUS dataset.

Multi-annotator classification involves using the input of multiple annotators and assigning the correct label to a sample. There a few common methods to reach a consensus label such as majority voting or weighted voting. These approaches are often too simplistic produce high quality estimates of the true label. Goh et al [9] proposed *CROWDLAB* a method which combines trained classifier class prediction probabilities with multiannotator ratings to produce new consensus labels. This approach produces superior performance when compared to other methods for inferring consensus labels from multiannotated data. This method will be adopted as solution to the multirater aspect of this problem.

3. METHODOLOGY

The proposed multiaxis vision transformer model is pictured in the figure. The method consists of 3 distinct aspects: a relabeling model, a weighted ensemble algorithm and a quality prediction model. The relabeling model is pre-trained on ImageNet-1K and then finetuned using on a dataset of multiannotated TVUS images. The weighted ensemble algorithm uses the class prediction probabilities from the relabeling model and the multiannotator's labels to improve the consensus labels of each sample. Finally, a quality prediction model is pretrained on ImageNet-1K and then finetuned using the improved consensus labels.

Formally, let $D_B = \{(x_i, y_i)\}_{i=1}^{N}$ represent the TVUS dataset with N TVUS images $x \in \chi \subset \mathbb{R}^{H \times W \times D}$ paired with the quality label $y \in \{0, 1, 2\}$ where H, W and D are height, width and depth of the TVUS image respectively. Let $D_E = \{(p_i, a_i)\}_{i=1}^{N}$ denote the class prediction probabilities dataset, with N-class categorical probabilities where $p \in P = (c_0, c_1, c_2)$, $c_0 + c_1 + c_2 = 1$ and c_j represents the probability of class j for the sample. The prediction probabilities are paired with multiannotator labels $a \in A = (l_1, l_2, ..., l_k)$ where l is the label of an annotator and K is the number of annotators. Once the weighted ensemble algorithm is used for relabeling, let $D_R^t = \{(x_i^t, y_i^t)\}_{i=1}^{N}$ be the improved "true" label dataset, with the same images as D_B but new labels. Let ω represent the pretrained weights from training on ImageNet-1K. Let $f_M : a \to \hat{y}$ be the majority voting algorithm which returns the set of the most commonly occurring labels for a sample. The label improvement model f_{θ_L}: $\chi^t \to \Delta$, where $\Delta \subset [0, 1]^3$ is the probability simplex, is initialised by ω and trained with dataset D_B. The weighted ensemble algorithm $f_W : (p, A) \to \{0, 1, 2\}$ uses dataset D_E, where p is the class prediction probabilities from the well-trained model f_{θ_L}. The final quality prediction model $f_{\theta_F} : \chi^t \to \Delta$ is initalised by ω and trained with D_R^t.

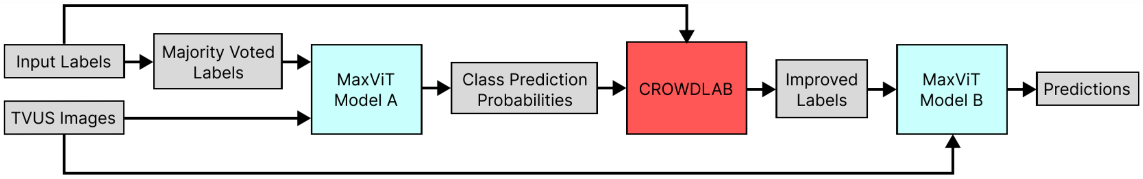


Fig. 2. The architecture of the proposed MaxViT model.

```
Algorithm 1: Proposed algorithm
  Input: TVUS images $X$, multi-annotator labels $A$
  Output: Predicted quality of TVUS image $p$
  Initialize the model with the pre-trained weights $\omega$.
  $\hat{Y} = f_M(A)$ and $D_B = (X, \hat{Y})$
  while training has not converged do
      $p = f_{\theta_L}(D_B, \omega)$
      $l(D_B, W) = -\sum_{i=1}^{N} y_i \log(p_i)$
  end
  $C = f_W(p, A)$ and $D_R^t = (X, C)$
  while training has not converged do
      $p = f_{\theta_F}(D_R^t, \omega)$
      $l(D_R^t, W) = -\sum_{i=1}^{N} y_i^t \log(p_i^t)$
  end
```

3.1. Multi-annotator Consensus Labeling

The first round of training is preceded by majority vote to generate a consensus label for each sample; tiebreaks are resolved by randomly selecting one of the tied labels. Formally, let $C(0)$, $C(1)$, $C(2)$ be the number of elements of a which are equal to 0, 1 and 2 respectively. Let $z = max(C(0), C(1), C(2))$, if $C(y) = z$, it is in V. Let v be the majority vote label. If $|V| > 1$, $v = V_1$, else $v = R(V)$ where $R(V)$ randomly returns an element of set V. The model is then trained on this set of sample, consensus label pairs. Once the training is complete, following [9], the weighted ensemble algorithm is used to generate new consensus labels for each TVUS image. The model's weights are reset and then trained with the new consensus labels. The weighted ensemble algorithm is formally defined in Equation 1. Where $\hat{p}_{Aj}$ are class prediction probabilities, representing each annotators label as a probabilistic prediction and ω_j, ω_M are weights for the relative trustworthiness of each annotator and the classifier respectively.

$$\hat{p}_{CR}(Y_i|X_i, \{Y_{ij}\}) = \frac{\omega_M \cdot \hat{p}_M(Y_i|X_i) + \sum_{j \in J_i} \omega_j \cdot \hat{p}_{Aj}(Y_i|\{Y_{ij}\})}{\omega_M + \sum_{j \in J_i} \omega_j} \quad (1)$$

3.2. Model Training

The multi-axis vision transformer uses pre-trained weights from ImageNet-1K. We then finetune this model on our dataset. The limited size of the dataset made pretraining a necessity. Formally, we can formalize the pre-training process as $p_i^t = f_{\theta_F}(x_i^t)$, for all $x_i^t \in \chi^t$, where the p_i^t is the prediction given the data sample x_i^t.

We adopt cross-entropy as our objective function. Formally:

$$l(D_R^t, W) = -\sum_{i=1}^{N} y_i^t \log(p_i^t). \quad (2)$$

The model is optimized through minimising the $l(D_R^t, W)$ objective function.

4. EXPERIMENTS

4.1. Dataset

The dataset contains 150 ultrasounds images from 50 unique patients. Each patient has provided a TVUS image of their left ovary, right ovary and uterus. Each image is annotated by 6 medical professionals: 2 sonographers, 2 radiologists and 2 gynae sonologists. The medical professional's used the grading system [4] to determine the quality of each image. We adapted the system to remove the intermediary grades of 2 and 4. Therefore, the annotators graded the images 0, 1, 2 and 3, corresponding to the grades 0, 1, 3 and 5 from the grading system. All images that were graded 0 or 1 were merged together into one class. This resulted in each image being graded either 0, 1 or 2 representing a very poor, suboptimal or optimal image. For training, 40 patients and their associated 120 images were used. To validate our model, we selected the remaining 10 patients and used their 30 images as the validation set. This set of patients was selected so that the distribution of image quality would be the same across the training and validation sets. The dataset is imbalanced with 351 class 2 annotations, 331 class 1 annotations and 218 class 0 annotations; class 0 is underrepresented making up 24.20% of annotations. This imbalance is exacerbated when observing the majority voted labels where there are 27 class 0 labels, 52 class 1 labels and 71 class 2 labels. Class 0 only makes up 18.00% of the majority voted labels. Class 2 labels are over represented making up 47.33% of the majority voted labels.

Model	**Consenus Label**	**Accuracy**	**Macro Average R**	**Class 0**			**Class 1**			**Class 2**		
				P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Resnet50	M	0.57	0.59	0.67	0.86	0.75	0.38	0.33	0.35	0.62	0.57	0.59
	WE	0.63	0.65	0.50	0.75	0.60	0.67	0.50	0.57	0.67	0.71	0.69
Resnet101	M	0.53	0.53	0.56	0.71	0.63	0.33	0.22	0.27	0.60	0.64	0.62
	WE	0.60	0.69	0.50	**1.00**	0.67	0.64	0.58	0.61	0.64	0.50	0.56
MaxViT	M	0.63	0.64	**1.00**	0.71	**0.83**	0.45	0.56	0.50	0.64	0.64	0.64
	WE	**0.80**	**0.77**	0.75	0.75	0.75	**0.88**	**0.64**	**0.74**	**0.78**	**0.93**	**0.85**

Table 1. Performance of the proposed algorithm compared with majority voting on the TVUS dataset using Resnet50, Resnet101 and MaxViT. WE is weighted ensemble, M is majority voting, R is recall, P is precision, F1 is F1-score.

Each image in the dataset was resized to a size of 224x224 and normalised. Each image was also augmented using AutoAugment using the ImageNet augmentation policy.

4.2. Implementation Details

To generate the improved labels, the multiaxis vision transformer is trained for 100 epochs on the TVUS dataset. The training used a batch size of 4 with AdamW optimiser and a learning rate of 5e-5. The MaxViT network loads pretrained weights from ImageNet-1k as do Resnet50 and Resnet101. Early stopping was used, so if the validation set accuracy did not improve from the global maximum for 10 epochs the training would terminate. The saved checkpoint with the greatest validation set accuracy was then combined with the weighted ensemble algorithm to generate the improved labels. The MaxViT network was then then fine-tuned for 100 epochs with early stopping using a patience of 10, batch size of 4 with AdamW optimiser and a learning rate of 5e-5. We evaluated our method using accuracy and macro average recall and per-class precision, recall and F1-score.

4.3. Overall Model Performance

The validation accuracy of the proposed algorithm was 0.80 and the macro average recall was 0.77. Resnet50 had an accuracy of 0.63 and macro average recall of 0.65. Resnet101 had an accuracy of 0.60 and macro average recall of 0.60. Each model's accuracy and macro average recall improved when changing from majority vote to our proposed algorithm.

4.4. Analyses

Despite the imbalanced dataset each model's accuracy and macro average recall are similar. This indicates that performance is not biased towards any class. When interrogating the performance of MaxViT with weighted ensemble we can see that the F1 scores are 0.75, 0.74 and 0.85 for class 0, class 1 and class 2 respectively, this further reinforces the balanced performance across all classes. MaxViT had the most significant improvement in both accuracy and macro average recall when using weighted ensemble compared to majority voting. The class 0 F1-score decreased from 0.83 to 0.75, however, class 1 increased significantly from 0.50 to 0.74 and class 2 increased from 0.64 to 0.85. This indicates that the labels produced by the weighted ensemble algorithm lead to the training of a model which could better identify class 1 and class 2 images. This improvement in class 1 and class 2 F1-scores between majority voting and weighted ensemble labels can be seen both in Resnet50 and Resnet101 as well.

To further evaluate the MaxViT's performance, we can compare the model's accuracy to the annotator's using the weighted ensemble algorithm's labels. The accuracy of the annotators was 0.97, 0.90, 0.73, 0.70, 0.60 and 0.40. MaxViT's 0.80 accuracy outperforms 4 of the 6 annotators. These results indicate that MaxViT has reached a performance that is very close to human level on this dataset.

5. CONCLUSION

In this paper, we introduce an AI model that harbours the potential to enrich the diagnostic process for endometriosis by enhancing the quality of TVUS. The distinctive architecture of our AI model effectively utilises a pre-trained multiaxis vision transformer in conjunction with annotations from diverse healthcare professionals. This approach yields accurate evaluations of TVUS scan quality. With a prediction accuracy of 80.00% , our model shows potential to benefit non-specialist sonographers, as it empowers them to consistently generate high-quality TVUS scans. This work paves the way for future research to the rapidly evolving research of AI-assisted ultrasound technology.

6. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by Human Research Ethics Committee (HREC) of University of Adelaide(Date 01-03-2020/No. H-2020-051) and the Southern Adelaide Clinical Human Research Ethics Committee (SAC HREC) (Date 01-11-2021/No. 111.20).

7. REFERENCES

[1] Devashana Gupta, M Louise Hull, Ian Fraser, Laura Miller, Patrick MM Bossuyt, Neil Johnson, Vicki Nisenblat, Cochrane Gynaecology, and Fertility Group, "Endometrial biomarkers for the non-invasive diagnosis of endometriosis," *Cochrane Database of Systematic Reviews*, vol. 2016, no. 4, 2016.

[2] C Chapron, L Cravello, N Chopin, G Kreiker, B Blanc, and JB Dubuisson, "Complications during set-up procedures for laparoscopy in gynecology: open laparoscopy does not reduce the risk of major complications," *Acta obstetricia et gynecologica Scandinavica*, vol. 82, no. 12, pp. 1125–1129, 2003.

[3] Margaret Ann Fraser, Sugandha Agarwal, Innie Chen, and Sukhbir Sony Singh, "Routine vs. expert-guided transvaginal ultrasound in the diagnosis of endometriosis: a retrospective review," *Abdominal imaging*, vol. 40, pp. 587–594, 2015.

[4] Alison Deslandes, Jodie Avery, Hsiang-Ting Chen, Mathew Leonardi, Steven Knox, Catrina Panuccio, M. Louise Hull, and George Condous, "A quantitative grading system for the assessment transvaginal ultrasound image quality," *15th World Congress on Endometriosis*, 2023.

[5] Siyuan Zhang, Yifan Wang, Jiayao Jiang, Jingxian Dong, Weiwei Yi, Wenguang Hou, and Hocine Cherifi, "Cnn-based medical ultrasound image quality assessment," *Complex.*, vol. 2021, jan 2021.

[6] Hong Luo, Han Liu, Kejun Li, and Bo Zhang, "Automatic quality assessment for 2d fetal sonographic standard plane based on multi-task learning," *CoRR*, vol. abs/1912.05260, 2019.

[7] Federico M Asch, Nicolas Poilvert, Theodore Abraham, Madeline Jankowski, Jayne Cleve, Michael Adams, Nathanael Romano, Ha Hong, Victor Mor-Avi, Randolph P Martin, and Roberto M Lang, "Automated echocardiographic quantification of left ventricular ejection fraction without volume measurements using a machine learning algorithm mimicking a human expert," *Circ. Cardiovasc. Imaging*, vol. 12, no. 9, pp. e009303, Sept. 2019.

[8] Matthias Schneider, Philipp Bartko, Welf Geller, Varius Dannenberg, Andreas König, Christina Binder, Georg Goliasch, Christian Hengstenberg, and Thomas Binder, "A machine learning algorithm supports ultrasound-naïve novices in the acquisition of diagnostic echocardiography loops and provides accurate estimation of LVEF," *Int. J. Cardiovasc. Imaging*, vol. 37, no. 2, pp. 577–586, Feb. 2021.

[9] Hui Wen Goh, Ulyana Tkachenko, and Jonas Mueller, "Crowdlab: Supervised learning to infer consensus labels and quality scores for data with multiple annotators," 2023.