CARDIUM: Congenital Anomaly Recognition with Diagnostic Images and Unified Medical records

Supplementary Material

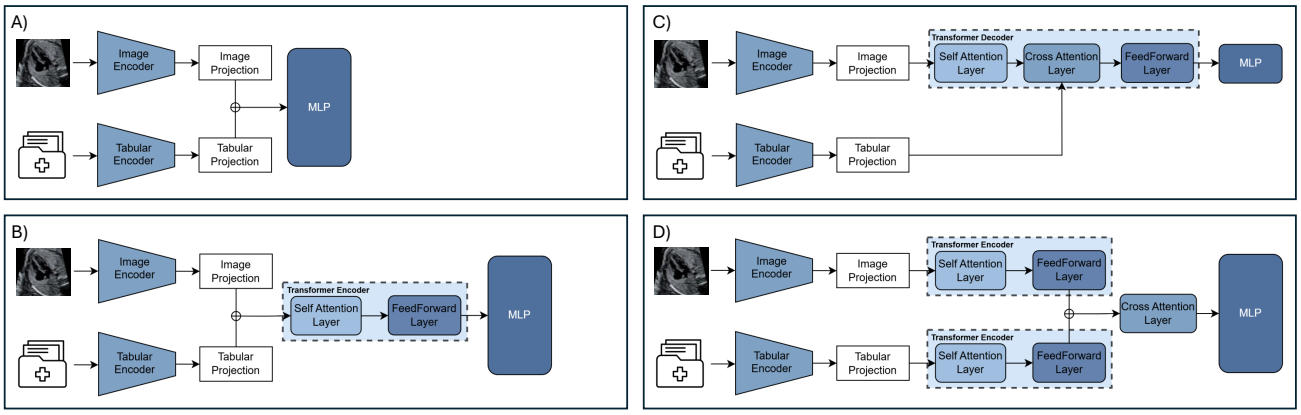

Figure A. Comparison of multimodal fusion strategies. (A) *MLP-Fusion*: concatenate modality features, then process them with an MLP. (B) *Transformer Encoder Fusion*: concatenate features, then process them with a transformer encoder. (C) *Transformer Decoder Fusion*: process image features with a decoder, then integrate tabular features through cross-attention. (D) *Transformer Encoder with Cross-Attention Fusion*: each modality is encoded separately, then fused via cross-attention.

A. Implementation Details

A.1. Training and architecture of CARDIUM model

We train our model on an NVIDIA Quadro RTX 8000 and optimize parameters of the tabular, image, and multimodal module using Weights & Biases [4]. To address class imbalance, we employ loss weighting, image data augmentation, weighted random sampling, and hard positive mining (i.e., oversampling false negative examples). This last strategy was applied exclusively to the tabular encoder, where we apply a weighted random sampler on the trained loader every 20 epochs to oversample false negative examples. We train tabular and image encoders separately, freeze them, and then transfer the weights to the fusion module. We train our multimodal model for 100 epochs with binary cross-entropy loss, AdamW optimizer, and learning rate of 5×10^{-7}. The optimal multimodal parameters consist of eight-layer decoders with two attention heads and dropout rates of 0.4.

A.2. Training on the External Ultrasound Fetal Dataset

To adapt our model for the external fetal ultrasound dataset, which is designed for image-only multiclass classification, we modify the classification head to output predictions for six classes and replace the binary cross-entropy loss with cross-entropy loss. Additionally, we optimize key hyperparameters to better suit the dataset's larger size and more balanced class distribution. Specifically, we adjust the learning rate from 5×10^{-7} to 4×10^{-5} and reduce the dropout rates from 0.4 to 0.1. To evaluate the performance of our model pretrained on the CARDIUM dataset, we load the model's pretrained weights and modify the classification head, initializing it from scratch. We then finetune the model on the fetal dataset. Since we perform three-fold cross-validation, we finetune the best model for each fold, and during inference, we average the predictions from the three models to obtain the final prediction.

A.3. Training TIP and MMCL on the CARDIUM Dataset

We evaluate the performance of TIP and MMCL on the CARDIUM dataset, using the same fold and split distribution as the CARDIUM model to ensure a fair comparison. TIP was fine-tuned using publicly available pre-trained weights, originally trained on the UK Biobank [6], which includes cardiac MRI images and clinical data. We followed the authors' recommended hyperparameters during fine-tuning. Since MMCL does not provide pre-trained weights, we trained it from scratch using the authors' suggested hyperparameters.

A.4. Training with Half the Data

To train on half of the CARDIUM dataset, we split the training set in half while maintaining the same three-fold cross-validation setup, ensuring that each fold has a reduced training split. Additionally, we preserve the class and trimester distribution in the reduced training set to maintain consistency in data composition and allow for a fair comparison. The test split in each fold remained the same as in the original dataset, ensuring consistency in evaluation across all folds.

B. Mathematical Formulation of Weight of Evidence Encoding

For encoding categorical variables, we use Weight of Evidence (WoE) encoding combined with a five-fold cross-validation strategy. This technique can be summarized as follows,

$$\text{WoE}_k(X) = \log\left(\frac{P(X \mid Y = 1, D_{-k})}{P(X \mid Y = 0, D_{-k})}\right) \tag{1}$$

where $\text{WoE}_k(X)$ denotes the Weight of Evidence value for category X in fold k; $P(X \mid Y = 1,\ D_{-k})$ is the probability of observing X among positive samples in the data excluding fold k; $P(X \mid Y = 0,\ D_{-k})$ is the probability of observing X among negative samples in the data excluding fold k; and D_{-k} represents the dataset excluding fold k.

C. Architecture of the Different Multimodal Fusion Strategies

The different multimodal fusion strategies implemented are depicted in Figure A. The MLP Fusion strategy takes the output of each modality encoder, concatenates the features, and then processes them with an MLP. The Transformer Encoder Fusion strategy concatenates the modality features and processes them with a transformer encoder. The resulting output is then passed through an MLP. The Transformer Decoder Fusion strategy processes the image features with a transformer decoder and integrates the tabular features through the cross-attention layer. The output is then processed by an MLP. Finally, the Transformer Encoder with Cross-Attention Fusion strategy processes the features of each modality separately with its own transformer encoder. The outputs of these encoders are fused using a cross-attention layer and then processed with an MLP.