

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

Daniela Vega¹ Hannah V. Ceballos¹ Javier S. Vera¹ Santiago Rodriguez¹
 Alejandra Perez¹ Angela Castillo¹ Maria Escobar¹
 Dario Londoño^{1,2} Luis A. Sarmiento^{1,2} Camila I. Castro¹
 Nadiezhda Rodriguez^{1,2} Juan C. Briceño¹ Pablo Arbelaez¹

¹ Universidad de los Andes, Colombia ² Fundación Santa Fe de Bogotá, Colombia
 {d.veгаа, h.ceballos, j.verar, s.rodriгuezr2, a.perezr20, a.castillo13, mc.escobar11,
 d.londono25, ansarmie, cami-cas, narodrig, jbriceno, pa.arbelaez}@uniandes.edu.co

Abstract

Prenatal diagnosis of Congenital Heart Diseases (CHDs) holds great potential for Artificial Intelligence (AI)-driven solutions. However, collecting high-quality diagnostic data remains difficult due to the rarity of these conditions, resulting in imbalanced and low-quality datasets that hinder model performance. Moreover, no public efforts have been made to integrate multiple sources of information, such as imaging and clinical data, further limiting the ability of AI models to support and enhance clinical decision-making. To overcome these challenges, we introduce the Congenital Anomaly Recognition with Diagnostic Images and Unified Medical records (CARDIUM) dataset, the first publicly available multimodal dataset consolidating fetal ultrasound and echocardiographic images along with maternal clinical records for prenatal CHD detection. Furthermore, we propose a robust multimodal transformer architecture that incorporates a cross-attention mechanism to fuse feature representations from image and tabular data, improving CHD detection by 11% and 50% over image and tabular single-modality approaches, respectively, and achieving an F1-score of $79.8 \pm 4.8\%$ in the CARDIUM dataset. We will publicly release our dataset and code to encourage further research on this unexplored field.

1. Introduction

Congenital Heart Diseases (CHDs) are structural abnormalities of the heart and blood vessels that develop during fetal growth and are the leading cause of infant mortality [20]. Prenatal detection through ultrasound and echocardiographic imaging is crucial to improving clinical outcomes. Yet, detection rates can be as low as 30%, particularly in low- and middle-income countries, due to limited access to

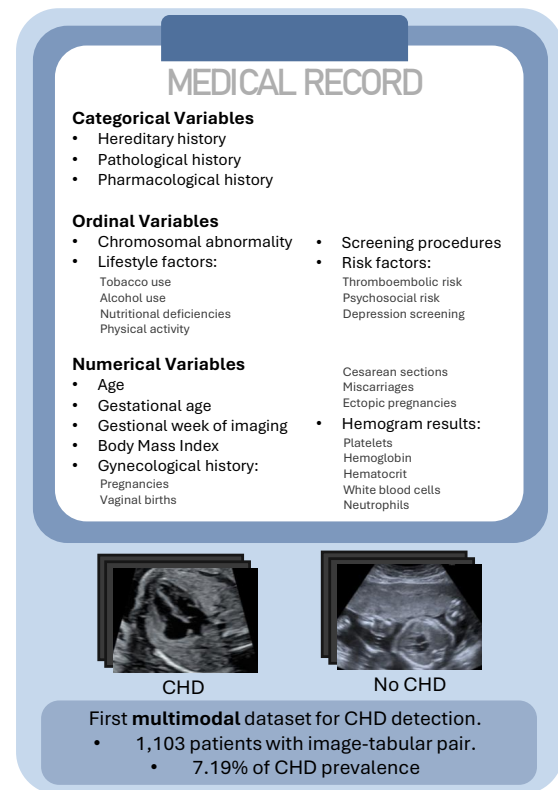


Figure 1. **Overview of the CARDIUM dataset.** The CARDIUM dataset includes diagnostic images from 1,103 patients and 26 physiological variables from the mother’s clinical record.

specialists and equipment [2], [9].

Artificial Intelligence (AI) offers the potential to reduce these disparities and improve prenatal diagnosis [12] by assisting specialists in recognizing cardiac abnormalities. However, the unique characteristics of these conditions, along with the sensitivity involved in working with fetal

data, introduce significant challenges.

First, CHDs are extremely rare, affecting approximately 8 in every 1,000 live births globally each year, which makes it difficult to collect extensive and diverse datasets [24]. Moreover, the small size of the fetal heart and the fetus’s constant movement make it challenging to acquire clear diagnostic images [14]. As a result, existing datasets are often imbalanced and of low quality, which limits the ability of AI models to learn robust and generalizable patterns. Integrating clinical data could help compensate for the scarcity and imbalance of imaging datasets; however, such approaches remain largely unexplored.

Second, fetal data is highly sensitive, requiring strict regulations and extensive approvals for collection and sharing. Consequently, creating publicly available CHD datasets is very challenging. Nevertheless, access to public datasets is essential for meaningful progress in automated CHD detection, as it ensures the reproducibility of AI models, encourages collaboration, and accelerates the development of more effective diagnostic methods.

To address these limitations, we propose two key contributions in this paper. First, we introduce the **Congenital Anomaly Recognition with Diagnostic Images and Unified Medical records (CARDIUM)** dataset, the first publicly available multimodal dataset for prenatal CHD detection. This dataset combines echocardiographic and ultrasound images with maternal clinical data, enabling a more comprehensive analysis of CHD risk, while facilitating open research and fair comparisons between methods. Second, we present the CARDIUM model, a multimodal transformer that achieves promising results on our dataset, establishing a baseline for future studies and encouraging advancements in prenatal CHD diagnosis. We will make our dataset and code publicly available to promote open research.

2. Related Work

2.1. Deep Learning Algorithms for Congenital Heart Disease Detection

The rapid advancements of deep learning have led to significant progress in AI-based methods for prenatal CHD detection. For instance, Arnout *et al.* [2] trained a ResNet, achieving an AUC of up to 99% across four datasets. Qiao *et al.* [22] used a residual CNN, reaching 93% accuracy in four-chamber fetal images. Moreover, Nurmani *et al.* [19] employed a DenseNet21, achieving 92% inter-patient and 100% intra-patient accuracy. Despite these promising results, all methods rely on private datasets and, except for [2], lack publicly available code, hindering reproducibility and fair comparison. Furthermore, none incorporate multimodal data, limiting their ability to replicate real-world clinical practice [13]. Our approach addresses these limitations by introducing the first public multimodal dataset

for CHD detection, along with an open-source multimodal baseline model.

2.2. Multimodal Models

In clinical practice, physicians rely on multiple data types, including medical images and clinical records, to make accurate diagnoses. Some multimodal models combining imaging and tabular data have been explored for other diagnostic tasks. Hager *et al.* [10] combine imaging and tabular data in a contrastive multimodal learning (MMCL) framework, achieving AUCs of 73.76% for predicting coronary artery disease risk and 76.60% for predicting myocardial infarction risk on the UK Biobank dataset [6]. More recently, Du *et al.* [8] introduced Tabular-Image Pre-training (TIP), which improves on MMCL by combining image–tabular contrastive learning, masked tabular reconstruction, and image–tabular matching, achieving AUCs of 86.43% and 85.58% on the same datasets. Despite these promising results, multimodal approaches for CHD detection remain largely unexplored. Moreover, both methods struggle with class imbalance, limiting their clinical applicability in scenarios like CHD diagnosis, where positive cases are far less common than negative ones. Although there are other studies on multimodal diagnostic models, these do not use tabular information and images as input modalities, and some require further adjustments to be comparable. In this context, the CARDIUM model emerges as the first multimodal approach for CHD detection, incorporating strategies to address class imbalance and enhance robustness for clinical use.

2.3. Datasets

Automated diagnosis of CHD remains constrained by limited and inaccessible datasets. ImageCHD [26] is the first open-access dataset for CHD classification; however, it is restricted to postnatal cases, underscoring the need for prenatal recognition datasets. Moreover, ImageCHD relies exclusively on imaging data, whereas real-world diagnoses also incorporate clinical information. CARDIUM represents the first multimodal dataset specifically designed for prenatal CHD classification, promoting the development and evaluation of novel algorithms and enabling significant advances in this field.

3. CARDIUM Dataset

We present the CARDIUM dataset, the first multimodal dataset for prenatal CHD detection. CARDIUM combines the mother’s clinical record with echocardiographic and ultrasound images, providing complementary diagnostic modalities that collectively contribute to a holistic understanding of the fetus’s physiological state. The dataset was constructed through a retrospective study on Colombian women, with data collected between 2013 and 2024.

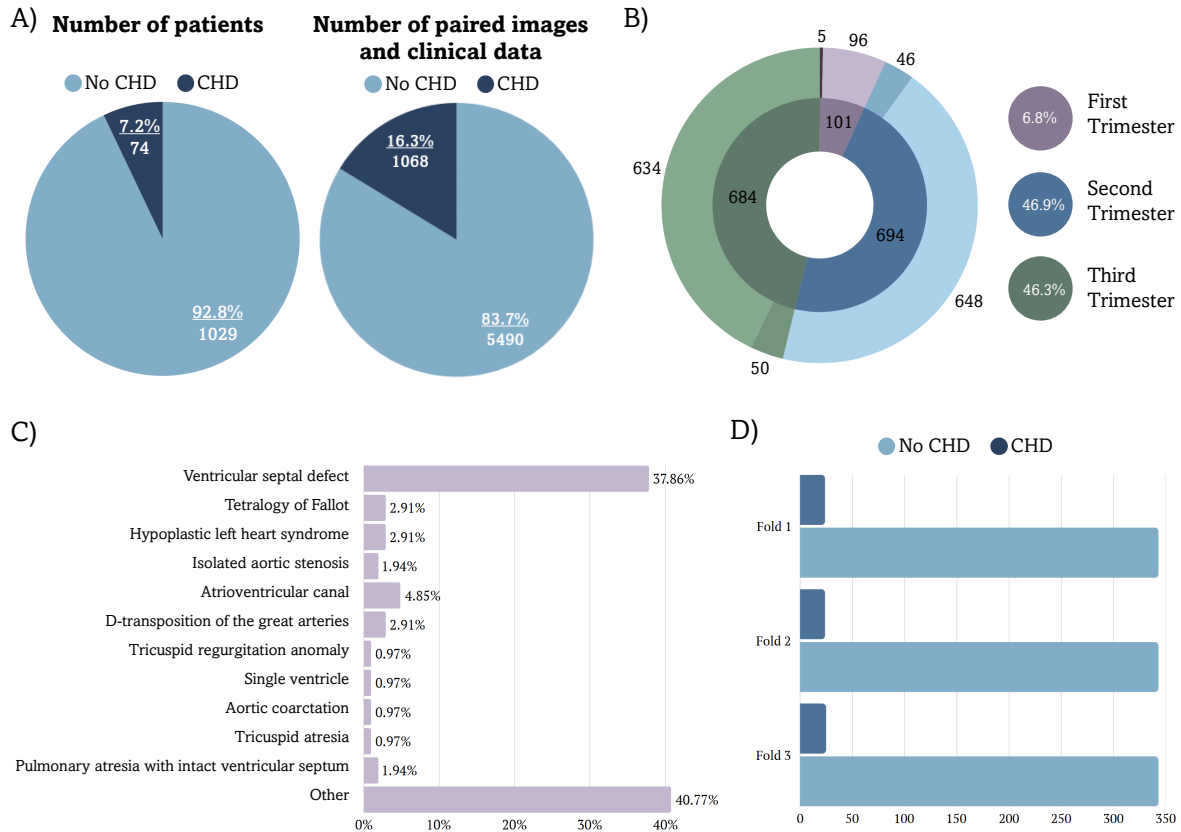


Figure 2. **CARDIUM dataset statistics.** (A) Number of patients with and without CHD (left) and number of images corresponding to patients with and without CHD (right). (B) Trimester distribution: The inner circle represents the overall number of patients with images from the first, second, and third trimesters, while the outer circle distinguishes between positive (darker) and negative (lighter) cases. On the right, we present the percentage of the total dataset corresponding to each gestational period. (C) Distribution of different types of CHDs present in our dataset. (D) Number of patients with and without CHD per fold.

3.1. Image Collection

We acquired 2D echocardiographic and ultrasound images using Voluson E6/E8/E10 systems (GE Healthcare, Austria), following established protocols [18]. For each examination, a CHD specialist captured the standard four cardiac views included in routine fetal ultrasound evaluations: the four-chamber view, the three-vessel trachea view, the left ventricular outflow view, and the right ventricular outflow view. These views provide different perspectives of the fetus’s heart, offering crucial anatomical insights for CHD detection.

After image acquisition, an expert echographer reviewed all images and discarded those that were considered inconsistent or of very low quality. We retained all images approved by the echographer, including multiple images of the same view, although not all examinations contained all four views. As a result, each patient had more than one image.

Color and power Doppler with high-definition flow enhanced image quality and vascular detail.

3.2. Medical Records Collection

We extracted categorical and numerical variables from the mother’s medical records by converting event-based notes into a tabular format. We selected categories that capture essential maternal and fetal health indicators, providing relevant physiological context. We also confirmed that the chosen categories were available across most clinical records. Figure 1 showcases all the variables included in the dataset, along with their corresponding data type (categorical, ordinal, and numerical).

Since the available clinical data and ultrasound images were not collected on the same day, we consolidated all available medical records from the duration of each pregnancy to capture a broader clinical context. Specifically, we aggregate all clinical events for a given patient into a single tabular entry, which often contains multiple values for most numerical variables. For variables that remain stable throughout pregnancy, such as gynecological history, we retained only one value, as these do not change across events.

We also retained a single value per patient for ordinal variables. In binary fields, which include chromosomal abnormalities, screening procedures, and lifestyle factors, we assigned a value of 1 ("yes") if any record indicated a positive case. For ordinal scale fields, such as risk factors, we selected the highest reported level across all records (low, intermediate, or high). This approach ensured that clinically relevant risks were not underestimated due to variability in timing or documentation. For categorical variables, we included all categories recorded across the available medical records. Pathological, hereditary, and pharmacological histories contained 74, 43, and 50 unique categories, respectively.

Although some clinical variables were recorded after the ultrasound images were acquired, all data were collected during the same pregnancy, ensuring they reflected a consistent clinical context. Most variables, such as pathological history, hereditary history, and risk factors, remain stable throughout pregnancy or are more reliably documented during later visits. Including these data provides a comprehensive and accurate clinical profile that closely reflects the type of information typically available alongside ultrasound and echocardiographic imaging.

3.3. Dataset Statistics

Figure 1 provides an overview of our dataset, highlighting selected variables and imaging examples. The study involved a population of 1,103 patients with either obstetric echocardiographic or ultrasound images and associated clinical records available at the Fetal-Maternal Medicine Unit. In cases where multiple images were available for a single patient (e.g., from different visits), we linked all images to a single tabular record that consolidated information from all clinical events. All patients were required to be over 18 years old, and those with twin pregnancies were excluded from the study. The cohort had a mean age of 34.86 ± 4.92 years. Given the relatively low prevalence of CHD, gathering a sufficient number of positive cases required a significant effort. However, we achieved a CHD prevalence of 7.19%, which is higher than the approximately 1% observed in the general population [24]. Furthermore, since each patient could have more than one image, the total number of images is 6558, with 16.3% corresponding to positive patients and 83.7% corresponding to negative patients. These statistics are depicted in Figure 2A.

We collected data from various stages of pregnancy, as shown in Figure 2B. This figure illustrates the number of patients, both CHD-positive and CHD-negative, in each trimester, along with the percentage of the total dataset corresponding to each gestational period. Additionally, Figure 2C presents the distribution of CHD types in the CARDIUM dataset. The dataset contains images from 11 of the most frequent CHD types worldwide, with a 12th cat-

egory labeled "Other" for less frequent conditions [3].

We divided the dataset into three cross-validation folds to ensure robust evaluation and better assess the model's generalization. Stratified sampling preserved the CHD and non-CHD proportions across folds, as shown in Figure 2D, ensuring each fold accurately reflects the overall distribution in the CARDIUM dataset.

3.4. Data Privacy and Ethical Approval

To ensure patient privacy, we implemented strict anonymization protocols by assigning unique anonymized IDs and removing all sensitive information from the tabular data. Images were securely stored on the REDCap platform, which provides robust data protection and adheres to ethical and legal standards. The research protocol received approval from the Institutional Review Board (IRB) in accordance with international ethical guidelines.

3.5. Tabular Data Preprocessing

We establish a dataset preprocessing pipeline with two key components: numerical and categorical data refinement and categorical variable encoding.

3.5.1. Numerical Data Refinement

For numerical data refinement, we standardize the units of all numerical variables to ensure consistency across medical records. After unit standardization, we rectify any out-of-bounds values and apply z-score normalization to all numerical features (mean of 0 and standard deviation of 1).

3.5.2. Categorical Data Refinement

We first correct typographical errors using a combination of automated scripts and manual review. We also review categorical values and standardize the names of diseases and medications, as naming conventions often vary between clinical records despite referring to the same underlying category (e.g., *progesterone*, *progesterone intravaginal*, *Progen*).

We then group categorical variables based on semantic similarity to reduce the number of unique entries in the pathological, hereditary, and pharmacological history fields. For example, terms such as *vaginitis*, *candidiasis*, and *acute vaginitis* were all grouped under the broader category of *vaginal infections*. Finally, we combine categories with fewer than four occurrences into an "Others" label.

3.6. Evaluation Metrics

Given the imbalanced nature of the CARDIUM dataset, we propose evaluating the model's performance using a three-fold cross-validation strategy. During training, we treat each image as an individual sample to maximize the available data. However, for inference, we compute metrics on a per-patient basis by averaging the outputs from all corresponding images, aligning the evaluation process more

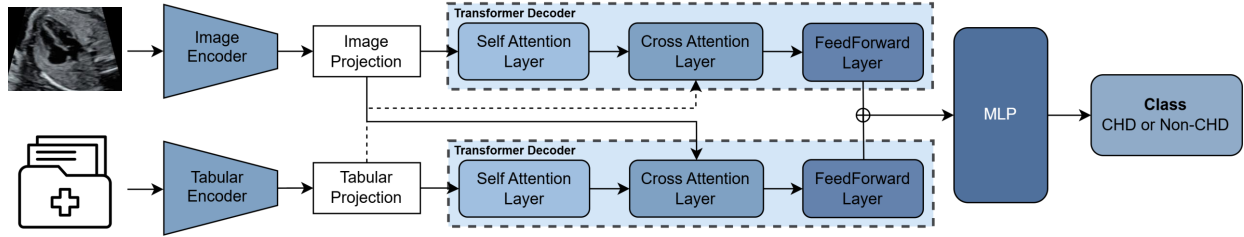


Figure 3. **Overview of the CARDIUM model.** We process image and tabular data through modality-specific encoders, E_I and E_T , to obtain distinct embeddings. We pass these embeddings through transformer decoder layers, where modality fusion occurs in the cross-attention layer. Then, we concatenate the fused representations and process them through a Multi-Layer Perceptron (MLP) to classify cases as CHD or non-CHD.

closely with clinical practice. To assess the model’s ability to identify CHD cases, we report the F1-score, precision, and recall for the CHD class, along with the Area Under the Receiver Operating Characteristic Curve (AUC) to measure overall performance.

4. CARDIUM Multimodal Model Architecture

The CARDIUM model is a novel multimodal framework that leverages a dual cross-attention mechanism to capture intricate dependencies between imaging and clinical data. Figure 3 illustrates the CARDIUM model architecture.

Given an image $I \in \mathbb{R}^{C \times H \times W}$ and encoded tabular data $T \in \mathbb{R}^n$, where n represents the number of tabular features, we process each type of data through its corresponding encoder. The image E_I and tabular E_T encoder transform their inputs into modality-specific embeddings $z_I \in \mathbb{R}^{1 \times D_I}$ and $z_T \in \mathbb{R}^{1 \times D_T}$, respectively. Here, D_I and D_T denote the embedding dimensions for each modality. We map both representations into a shared representation space with dimension D and input them into the multimodal fusion architecture for inter-modality learning. Full implementation details can be found in Appendix A.1.

4.1. Image Module

4.1.1. Image Encoder

We experiment with various architectures for the image encoder, including both Convolutional Neural Network (CNN)-based and transformer-based models. The best results are achieved by fine-tuning a Vision Transformer (ViT) [7] model pre-trained on ImageNet. This ViT configuration consists of twelve layers and six attention heads, with dropout rates for the transformer path and classification head set at 0.3 and 0.2, respectively.

4.2. Tabular Module

4.2.1. Categorical Variables Encoding

To capture the relationship between categorical variables and the target outcome, we use Weight of Evidence (WoE)

encoding, a Bayesian encoding technique that is inherently target-aware [1]. This method assigns each category a numerical value based on the log-odds ratio of positive to negative class observations, effectively quantifying how informative a category is in predicting the target. To prevent data leakage, we employ a five-fold cross-encoding strategy, in which the encoding for each fold is computed using data from the remaining four folds. This ensures that the encoding of a category is not influenced by the target labels in the fold being evaluated. For further details on the encoding strategy, see Appendix B.

4.2.2. Tabular Encoder

To effectively encode tabular features, we modify the transformer architecture to process both numerical and encoded categorical data. In this approach, we treat each feature as an individual token. First, we project these tokens into a higher-dimensional space and process them through a two-layer transformer encoder with eight attention heads. Next, we flatten the output and map it to an embedding space. This design allows the model to capture complex dependencies between features, resulting in a rich tabular representation suitable for multimodal fusion.

4.3. Multimodal Interaction Module

For multimodal representation learning, we employ a transformer decoder architecture to capture both intra- and inter-modality relationships through self-attention and cross-attention mechanisms. Let $z_I \in \mathbb{R}^{1 \times d}$ and $z_T \in \mathbb{R}^{1 \times d}$ denote the feature representations extracted from the image and tabular encoders, respectively. For each modality, we treat each batch of patients as a sequence, where each patient’s feature vector is treated as a token in this sequence. This design enables the model to learn contextual dependencies among patients within the batch, allowing it to assign adaptive importance to different features and effectively capture relationships across both modalities.

The interaction module consists of two parallel stacks of L transformer decoder layers, each comprising self-attention, cross-attention, and feedforward layers.

Table 1. CHD detection results on the CARDIUM dataset for modality-specific variants of our model.

| Images | Clinical Data | CHD F1 Score | CHD Precision | CHD Recall | AUC |
|--------|---------------|----------------------|----------------------|----------------------|----------------------|
| ✓ | ✓ | 0.798 ± 0.048 | 0.876 ± 0.173 | 0.757 ± 0.104 | 0.974 ± 0.012 |
| ✓ | | 0.689 ± 0.066 | 0.659 ± 0.135 | 0.742 ± 0.119 | 0.955 ± 0.0154 |
| | ✓ | 0.294 ± 0.019 | 0.192 ± 0.019 | 0.634 ± 0.049 | 0.794 ± 0.028 |

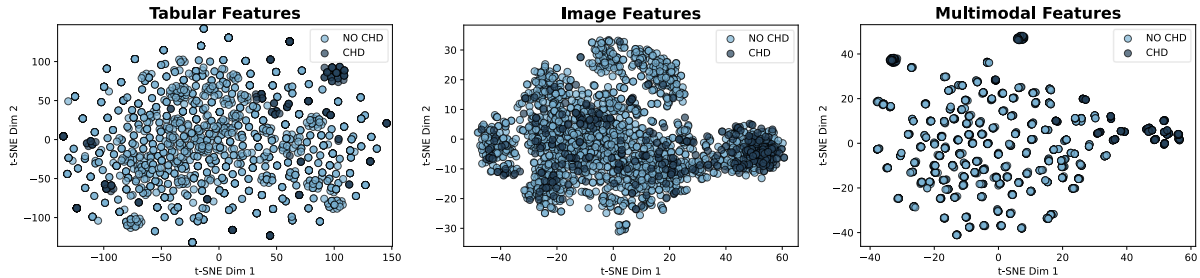


Figure 4. Feature distributions for CHD and non-CHD cases across intermediate outputs of our multimodal model on the CARDIUM dataset. (A) Tabular encoder. (B) Image encoder. (C) Final multimodal module. In plot (C), the point density appears lower compared to plots (A) and (B); however, this lower density is due to overlapping points. Feature distributions are visualized using t-SNE [25].

Each single-modality representation first undergoes self-attention, allowing the model to capture intra-modality dependencies. The self-attention mechanism enables the model to analyze complex relationships among features within a single modality and to capture dependencies between patients within the same modality.

Subsequently, the output of the self-attention layer interacts with the representation from the opposite modality through the cross-attention mechanism. Here, the key and value matrices are derived from the encoder’s output of the opposite modality, while the query originates from the self-attention output. The cross-attention mechanism allows for effective information exchange between modalities, helping each representation refine itself by using complementary features from the other modality. By dynamically re-weighting features, cross-attention highlights critical diagnostic patterns that may not be as apparent in a single modality.

We process the output from the cross-attention through a feedforward layer. Finally, we concatenate the refined features from both modalities and pass them through a three-layer Multilayer Perceptron (MLP) for classification.

5. Results and Discussion

5.1. Multimodal CHD Detection Results

Table 1 displays the overall performance of our model on the CARDIUM dataset, highlighting the effects of using one or both modalities. The results demonstrate that combining fetal echocardiography and ultrasound images with clinical data enhances performance by 11% compared to

using images alone and by 50% compared to using clinical data alone. These findings align with real-world clinical practice, where, although fetal echocardiography is the primary diagnostic tool, physicians benefit significantly from maternal-specific clinical information to improve diagnostic accuracy.

Figure 4 further illustrates the impact of multimodal integration on feature representation. Plots (A) and (B) display the feature distributions from the tabular and image encoders, respectively, while plot (C) presents the fused multimodal representation. After fusion, class clusters become more compact and distinct, enhancing the model’s ability to differentiate between CHD and non-CHD cases. This visualization highlights the transformer decoder’s effectiveness in capturing both intra- and inter-modality relationships, as well as the complementarity between data modalities. Consequently, multimodal fusion creates a more discriminative and structured feature space, enhancing CHD detection accuracy.

5.2. Trimestral Model Performance

We evaluate our model’s performance separately on data from the first, second, and third trimesters. This allows us to assess the model’s ability to detect CHD at different stages of pregnancy. As shown in Table 2, the model performs best with data from the third trimester and shows the most difficulties with data from the first trimester. However, it is important to note that only five CHD-positive cases are available in the first trimester, making it difficult to draw definitive conclusions about the model’s effectiveness at this early stage.

Table 2. Comparison of our model’s performance on data collected during the first, second, and third trimesters of pregnancy.

| Trimester | CHD F1 Score | CHD Precision | CHD Recall |
|-----------|---------------|---------------|---------------|
| First | 0.222 ± 0.314 | 0.333 ± 0.471 | 0.167 ± 0.236 |
| Second | 0.603 ± 0.092 | 0.701 ± 0.212 | 0.556 ± 0.101 |
| Third | 0.732 ± 0.072 | 0.825 ± 0.127 | 0.669 ± 0.074 |

These findings align with clinical expectations, as CHD detection improves in later gestational stages when cardiac anomalies become more visible [21] [11]. However, strong performance during early stages remains critical, given the significant impact of early diagnosis on the baby’s prognosis. The promising results from the second trimester underscore the potential of such tools for early CHD detection and highlight the need for additional early-stage data to improve the model’s ability to identify CHD during the initial phases of fetal development. Furthermore, the model achieves higher overall performance when evaluated on the full dataset (79.8% ± 4.8%), emphasizing the importance of comprehensive data for robust CHD detection.

5.3. Performance on Image Only Data

We evaluate the CARDIUM model’s ability to detect CHD using only images from patients without available clinical records in the hospital’s database. For this evaluation, we collected ultrasound and echocardiographic images from 11 patients with CHD and 113 patients without CHD, resulting in 144 CHD images and 767 non-CHD images. We performed inference on these images, achieving an F1-score of 0.8528 ± 0.106. This result demonstrates that the CARDIUM model can detect CHD effectively in unimodal contexts.

5.4. Generalization Experiments

To evaluate our multimodal model’s generalization capability, we compare its performance with state-of-the-art methods using a publicly available ultrasound fetal dataset [5], which includes maternal-fetal screening images from six anatomical planes. We use the same training/test split as proposed in [5] to ensure a fair and consistent comparison. We evaluate performance using ViT-Small, our multimodal approach trained from scratch, and our multimodal model pre-trained on CARDIUM. Implementation details and the modifications made to adapt our model for a unimodal multiclass classification task are provided in Appendix A.2.

The results for ViT-Small, our multimodal approach trained from scratch, and our multimodal model pre-trained on CARDIUM are summarized in the top section of Table 3. These results show a gradual improvement in F1-score, with the multimodal approach outperforming ViT-Small and further improvements resulting from pre-training on CARDIUM. This behavior suggests that our multi-

Table 3. Generalization Results on the Fetal-Planes-DB dataset [5].

| Model | F1 Score |
|--|--------------|
| ViT Small | 0.900 |
| CARDIUM model (ours) | 0.914 |
| CARDIUM model (ours) pretrained on CARDIUM dataset | 0.918 |
| MedMamba-B [27] | 0.933 |
| VMamba-B [15] | 0.927 |
| Swin Transformer-B [16] | 0.854 |
| ConvNext-B [17] | 0.855 |
| EfficientNetV2-B [23] | 0.885 |

modal framework enhances image representations, improving classification even in unimodal settings. Additionally, pre-training on CARDIUM consistently increased performance on an external ultrasound dataset, highlighting the dataset’s rich and transferable features.

Moreover, the results show that although MedMamba-B achieved the best results, our approach outperforms leading methods such as Swin Transformer, EfficientNet V2, and ConvNext, indicating effective generalization across distinct ultrasound datasets and tasks.

Table 4. Comparison of our multimodal model with multimodal state-of-the-art approaches on the CARDIUM dataset.

| Model | F1 Score |
|-----------------------------|----------------------|
| CARDIUM model (ours) | 0.798 ± 0.048 |
| TIP [8] | 0.459 ± 0.027 |
| MMCL [10] | 0.349 ± 0.090 |

5.5. Comparison with SOTAs

We compare our model’s performance with two state-of-the-art multimodal methods for binary classification using tabular and imaging data: MMCL [10] and TIP [8]. Both methods were evaluated on our dataset using the same data split used in CARDIUM model. The implementation details applied to each model are described in Appendix A.3.

Table 4 presents the results of TIP and MMCL evaluated on CARDIUM. Both models underperformed compared to our model, which may be attributed to the significant class imbalance present in our dataset. As noted in MMCL [10], contrastive learning struggles in scenarios involving imbalanced binary classifications, and both TIP and MMCL rely on contrastive learning strategies. These results emphasize that our multimodal approach, along with the strategies we employ to handle class imbalance, is highly effective, providing a distinct advantage in real-world clinical situations where negative cases are much more common than positive ones.

Table 5. Results of different modality integration strategies.

| Multimodal Module | F1 Score |
|---|----------------------|
| MLP Fusion | 0.454 ± 0.067 |
| Transformer Encoder Fusion | 0.686 ± 0.086 |
| Transformer Decoder Fusion | 0.607 ± 0.091 |
| Transformer Encoder with Cross Attention Fusion | 0.681 ± 0.048 |
| Double Transformer Decoder Fusion (ours) | 0.798 ± 0.048 |

5.6. Ablation Experiments

5.6.1. Ablation on Training on Half the Data

To assess the impact of data quantity on the performance of the CARDIUM model, we train the model using half of the CARDIUM dataset and evaluate it on the full test split. The results reveal a 13% decrease in F1-score when only half of the data is used for training, highlighting the critical role data quantity plays in AI model performance. The CARDIUM model demonstrates higher performance when trained on a larger dataset, underscoring the importance of continually increasing dataset size to enhance CHD detection accuracy. See Appendix A.4 for implementation details.

5.6.2. Ablation on Different Multimodal Modules

We implement and evaluate several multimodal fusion strategies. *MLP-Fusion* concatenates modality features and processes them with an MLP. *Transformer Encoder Fusion* concatenates features and processes them with a transformer encoder. *Transformer Decoder Fusion* processes image features with a transformer decoder and integrates tabular features via cross-attention. Finally, *Transformer Encoder with Cross-Attention Fusion* encodes each modality separately and fuses them using cross-attention. See Appendix C for further details.

Table 5 presents the performance of these strategies compared to our final architecture. Our model outperforms all other approaches by at least 11%, highlighting its effectiveness in capturing complex multimodal relationships. Self-attention enables the model to extract rich intra-modality dependencies, while the dual cross-attention strategy enhances feature representation through modality interaction, resulting in stronger fusion and improved performance.

5.6.3. Ablation on Different Image Encoders

We evaluate various image encoders to assess the quality of the extracted representations in multimodal training. We test *ResNet 18* and *ResNet 50* as CNN models, and *ViT Tiny* and *ViT Small* as transformer alternatives. We also use *Med-ViT*, a hybrid model that captures local and global features. Notably, *ViT Small* outperforms all others by at least 6%.

5.6.4. Ablation on Key Parameters

Finally, we evaluate the impact of loss factor and random weight sampling to address class imbalance. Implement-

ing a weighted random sampler significantly increases the model’s performance by 39.6% (from 36.1% to 75.7%). Furthermore, combining the sampling strategy with a loss factor of 1.2 applied to the positive class improves the F1-score by an additional 4.1%, resulting in a final metric of 79.8%. These results demonstrate the effectiveness of these strategies in managing imbalanced datasets.

6. Limitations

Although the CARDIUM dataset represents a significant advance in automatic prenatal CHD diagnosis, several limitations remain. The limited number of CHD-positive cases in the first trimester and the overall small size of the dataset restrict the model’s ability to detect early-stage CHDs and to generalize effectively. Expanding the dataset is crucial for improving diagnostic performance. Furthermore, while generalization results are promising, the dataset’s exclusive focus on data from Colombian women may introduce demographic and geographic biases, underscoring the need for broader testing across diverse populations. Finally, variability in image quality and differences in how clinical protocols are applied by different specialists may impact real-world deployment, highlighting the need for multi-center validation.

7. Conclusion

In this work, we introduce CARDIUM, the first publicly available multimodal dataset for prenatal CHD detection, which integrates echocardiographic and ultrasound images with maternal clinical data. This dataset addresses the limitations associated with private datasets and unimodal approaches, providing a solid foundation for automated CHD diagnosis. Additionally, we propose a multimodal transformer architecture that leverages self-attention to capture intra-modality dependencies and cross-attention to model interactions between imaging and tabular features. Our model achieves an F1-score of 79.8%, surpassing the image-only variation by 11% and the tabular-only variation by 50%, underscoring the advantages of multimodal integration for CHD detection. Moreover, our model generalizes well to an external ultrasound dataset, maintaining strong performance in unimodal multiclass classification. It also outperforms other multimodal state-of-the-art methods, which struggled to accurately detect CHD—likely due to the imbalanced nature of the dataset. These results demonstrate the robustness of our approach in imbalanced clinical scenarios.

8. Acknowledgments

Daniela Vega acknowledges the support of a UniAndes Google-DeepMind 2024 scholarship.

References

- [1] David Alvarez-Melis, Hal Daumé III, Jennifer Wortman Vaughan, and Hanna M. Wallach. Weight of evidence as a basis for human-oriented explanations. *CoRR*, abs/1910.13503, 2019. 5
- [2] R. Arnaout, L. Curran, Y. Zhao, J. C. Levine, E. Chinn, and A. J. Moon-Grady. Expert-level prenatal detection of complex congenital heart disease from screening ultrasound using deep learning. 2020. 1, 2
- [3] American Heart Association. Common Types of Heart Defects — heart.org. <https://www.heart.org/en/health-topics/congenital-heart-defects/about-congenital-heart-defects/common-types-of-heart-defects>, 2022. 4
- [4] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com. 1
- [5] X. P. Burgos-Artizzu, D. Coronado-Gutiérrez, B. Valenzuela-Alcaraz, E. Bonet-Carne, E. Eixarch, F. Crispi, and E. Gratacós. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Scientific Reports*, 10(1), 2020. 7
- [6] Clare Bycroft, Chris Freeman, Desislava Petkova, Gavin Band, Louise T. Elliott, Kevin Sharp, Alex Motyer, Damjan Vukcevic, Olivier Delaneau, Jonathan O’Connell, Adrian Cortes, Simon Welsh, Alexander Young, Marsha Effingham, Gil McVean, Sarah Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562:203–209, 2018. 2, 1
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 5
- [8] Siyi Du, Shaoming Zheng, Yinsong Wang, Wenjia Bai, Declan P. O’Regan, and Chen Qin. Tip: Tabular-image pre-training for multimodal classification with incomplete data. *arXiv preprint arXiv:2407.07582*, 2024. Preprint. 2, 7
- [9] O. Elshazali, M. Ibrahim, and A. Elseed. *Management of Congenital Heart Disease in Low-Income Countries: The Challenges and the Way Forward*. IntechOpen, 2022. 1
- [10] Paul Hager, Martin J. Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. *arXiv preprint arXiv:2303.14080*, 2023. Preprint. 2, 7
- [11] Shin Hashiramoto, Mayumi Kaneko, Hiroko Takita, Yuka Yamashita, Ryu Matsuoka, and Akihiko Sekizawa. Factors affecting the accuracy of fetal cardiac ultrasound screening in the first trimester of pregnancy. *Journal of Medical Ultrasonics*, 52:131–138, 2025. 7
- [12] P.-N. Jone et al. Artificial intelligence in congenital heart disease. *JACC: Advances*, 1(5):100153, 2022. 1
- [13] Felix Krones, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi. Review of multimodal machine learning approaches in healthcare. *Information Fusion*, 114:102690, 2025. 2
- [14] X. Liu, Y. Zhang, H. Zhu, B. Jia, J. Wang, Y. He, et al. Applications of artificial intelligence-powered prenatal diagnosis for congenital heart disease. *Frontiers in Cardiovascular Medicine*, 11, 2024. 2
- [15] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7
- [16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 7
- [17] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. 7
- [18] Anita J. Moon-Grady, Mary T. Donofrio, Sarah Gelehrter, Lisa Hornberger, Joe Kreeger, Wesley Lee, Erik Michelfelder, Shaine A. Morris, Shabnam Peyvandi, Nelangi M. Pinto, Jay Pruetz, Neeta Sethi, John Simpson, Shubhika Srivastava, and Zhiyun Tian. Guidelines and recommendations for performance of the fetal echocardiogram: An update from the american society of echocardiography. *Journal of the American Society of Echocardiography*, 36(7):679–723, 2023. 3
- [19] Siti Nurmaini, Radiyati Umi Partan, Nuswil Bernolian, Ade Iriani Sapitri, Bambang Tutuko, Muhammad Naufal Rachmatullah, Annisa Darmawahyuni, Firdaus Firdaus, and Johannes C. Mose. Deep learning for improving the effectiveness of routine prenatal screening for major congenital heart diseases. *Journal of Clinical Medicine*, 11(21), 2022. 2
- [20] G. Ottaviani and L. M. Buja. Congenital heart disease. *Cardiovascular Pathology*, pages 611–647, 2016. 1
- [21] Aura Iuliana Popa, Nicolae Cernea, Marius Cristian Marinaş, Maria Cristina Comănescu, Ovidiu Costinel Sîrbu, Dragoş George Popa, Larisa Pătru, Vlad Pădureanu, and Ciprian Laurenţiu Pătru. Ultrasound screening in the first and second trimester of pregnancy for the detection of fetal cardiac anomalies in a low-risk population. *Diagnostics*, 15(6), 2025. 7
- [22] Sibao Qiao, Shanchen Pang, Gang Luo, Silin Pan, Zengchen Yu, Taotao Chen, and Zhihan Lv. RLDS: An explainable residual learning diagnosis system for fetal congenital heart disease. *Future Generation Computer Systems*, 128:205–218, 2022. 2
- [23] M. Tan and Q. V. Le. Efficientnetv2: Smaller models and faster training. In *Lecture Notes in Computer Science*, 2021. 7
- [24] Denise Van Der Linde, Elisabeth EM Konings, Maarten A Slager, Maarten Witsenburg, Willem A Helbing, Johanna JM Takkenberg, and Jolien W Roos-Hesselink. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. *Journal of the American College of Cardiology*, 58(21):2241–2247, 2011. 2, 4

- [25] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. [6](#)
- [26] Xiaowei Xu, Tianchen Wang, Jian Zhuang, Haiyun Yuan, Meiping Huang, Jianzheng Cen, Qianjun Jia, Yuhao Dong, and Yiyu Shi. Imagechd: A 3d computed tomography image dataset for classification of congenital heart disease, 2021. [2](#)
- [27] Y. Yue and Z. Li. Medmamba: Vision mamba for medical image classification. In *Lecture Notes in Computer Science*, 2024. [7](#)