HSDreport: Heart Sound Diagnosis with Echocardiography Reports

Zihan Zhao[1,2*], Pingjie Wang[1,2*], Liudan Zhao[1,4*], Yuchen Yang[2,3],
Ya Zhang[1,2], Kun Sun[4], Xin Sun[4], Xin Zhou [4],
Yu Wang[1,2†], Yanfeng Wang[1,2†]
[1]Shanghai Jiao Tong University, [2]Shanghai AI Laboratory,
[3]University of Science and Technology of China,
[4]Xinhua Hospital Affiated to Shanghai Jiao Tong University School of Medicine
{zihanzhao,pingjiewang,sjdyszld,yuwangsjtu,wangyanfeng622}@sjtu.edu.cn

Abstract

Heart sound auscultation holds significant importance in the diagnosis of congenital heart disease. However, existing methods for Heart Sound Diagnosis (HSD) tasks are predominantly limited to a few fixed categories, framing the HSD task as a rigid classification problem that does not fully align with medical practice and offers only limited information to physicians. Besides, such methods do not utilize echocardiography reports, the gold standard in the diagnosis of related diseases. To tackle this challenge, we introduce HSDreport, a new benchmark for HSD, which mandates the direct utilization of heart sounds obtained from auscultation to predict echocardiography reports. This benchmark aims to merge the convenience of auscultation with the comprehensive nature of echocardiography reports. First, we collect a new dataset for this benchmark, comprising 2,275 heart sound samples along with their corresponding reports. Subsequently, we develop a knowledge-aware query-based transformer to handle this task. The intent is to leverage the capabilities of medically pre-trained models and the internal knowledge of large language models (LLMs) to address the task's inherent complexity and variability, thereby enhancing the robustness and scientific validity of the method. Furthermore, our experimental results indicate that our method significantly outperforms traditional HSD approaches and existing multimodal LLMs in detecting key abnormalities in heart sounds.

1 Introduction

Heart sound auscultation is an essential part of clinical medicine and is extensively utilized to screen for congenital heart disease (CHD) due to its cost-effectiveness. CHD is the most common congenital abnormality, with 13.3 million patients worldwide in 2019 (Roth et al., 2020; of the Report et al., 2022). For newborn patients, around one-third will pass away in their first year and risk for mortality (Pan et al., 2022). So massive promotion of heart sound auscultation is crucial. However, its effectiveness heavily depends on the clinician's expertise and the human ear's acoustic range (Montinari et al., 2018; Mangione and Nieman, 1997). Consequently, the task of heart sound diagnosis (HSD) has been highlighted (Clifford et al.; Oliveira et al., 2021; Yaseen et al., 2018).

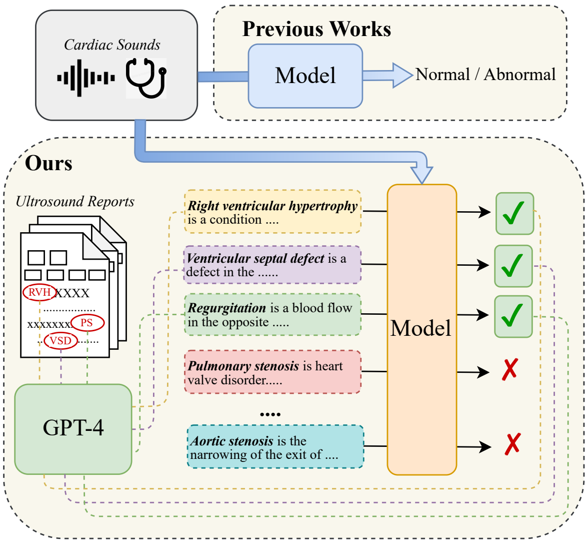
Cardiac Sounds
Previous Works
Model
Normal / Abnormal
Ours
Ultrosound Reports
RVH
XXXX
xxxxxxx
PS
VSD
GPT-4
Right ventricular hypertrophy is a condition ....
Ventricular septal defect is a defect in the ......
Regurgitation is a blood flow in the opposite .....
Pulmonary stenosis is heart valve disorder.....
....
Aortic stenosis is the narrowing of the exit of ....
Model
✓
✓
✓
✗
✗

Figure 1: Comparison between previous works and ours. In previous works, heart sound diagnosis (HSD) was treated as a multi-class problem with two to five categories. Our new benchmark, starting from echocardiography reports, treats HSD as a twelve-category multi-label task. Furthermore, we have developed a knowledge-aware, query-based transformer approach that utilizes medical descriptions to address this novel benchmark.

However, existing HSD datasets are limited to a few diseases and are strictly categorized as multi-class classification problems, which do not well align with medical practice (multi-label classifica-

[*]Equal contribution
[†]Corresponding author

Findings of the Association for Computational Linguistics: EACL 2024, pages 11369–11378
November 12-16, 2024 ©2024 Association for Computational Linguistics

tion). Besides, such datasets overlook echocardiography due to its difficult-to-obtain nature (Wang et al., 2021), which provides a more comprehensive set of diagnostic information yet. Hence, we aim to combine the convenience of auscultation with the comprehensive capabilities of echocardiography, enabling the screening of a broader population and providing physicians with more extensive reference information.

In this paper, we propose a new benchmark named HSDreport, which includes 2, 275 segments of heart sounds and their corresponding echocardiography. Each heart sound segment spans approximately 75 seconds across five body sites. Since our dataset is derived from practical applications, it presents unique two challenges compared to previous HSD datasets. Firstly, as echocardiography is composed by physicians based on images from ultrasound scans instead of directly derived from heart sounds, it poses a challenge to discard information that is difficult to discern from the heart sounds alone. Secondly, due to the natural language form of the report and the varying writing styles of different physicians, noise is introduced into the report. To address these issues, we leverage the strong semantic understanding capabilities and extensive internal knowledge of large language models (LLMs) (Achiam et al., 2023) to extract abnormalities, ultimately constructing a 12-category multi-label benchmark.

To utilize the unique characteristics in our HSDreport, we have found that existing heart sound models (Chen et al., 2023; Cheng and Sun, 2023; Guo et al., 2023) fail to tackle the multi-label classification problem. This is because these models are primarily designed for multi-class classification and are limited to a few fixed classes. Hence, we innovatively introduce a knowledge-aware query-based transformer for HSDreport to align the audio and text modalities (Kritharoula et al., 2023; Kim et al., 2022; Xiao et al., 2022; Zhang et al., 2023) and leverage the extensive knowledge from medical pre-trained models by feeding abnormalities in textual expression. Moreover, we replace the typical phrase-based category input with detailed medical descriptions, enabling the model to acquire a more comprehensive understanding of medical diseases, their symptoms, and their characteristics, thereby facilitating a holistic view of the disease and enhancing classification accuracy. Considering the diversity in medical expressions, we further employ LLMs to provide multidimensional and varied descriptions of the same disease, thus aligning our approach more closely with medical practice and enhancing its robustness.

Our contributions are summarized as follows:

- **A benchmark for heart sound diagnosis.** We propose HSDreport, a practical and challenging benchmark for heart sound diagnosis. HSDreport combines the auscultation with echocardiographic analysis for a more accurate heart sound diagnosis, aiming to combine the convenience of auscultation with the comprehensiveness of echocardiographic analysis, which is not present in previous datasets and is highly significant for medical practice.

- **A knowledge-aware approach for HSD.** We propose a knowledge-aware query-based transformer for HSD task. In addition, we use abnormality descriptions as inputs to enhance the model's accuracy in abnormality understanding. Furthermore, during inference, we comprehensively utilize the multidimensional diverse descriptions generated by the LLM, thereby further enhancing model robustness.

- **Comprehensive performance evaluation.** We perform comparative analysis with state-of-the-art heart sound models on our comprehensive benchmark, showcasing notable improvements across all metrics. This study is the first to demonstrate the feasibility of using heart sounds to infer information from echocardiography, also proving our model's ability to effectively utilize knowledge to address multiple challenges.

2 Related Works

2.1 Heart Sound Diagnosis

The current datasets on heart sounds primarily target binary classification and multi-class classification for a limited number of categories. (Clifford et al.) categorized heart sound recordings into three classes based on diagnostic outcomes: normal, abnormal, and uncertain. (Oliveira et al., 2021) approached heart sounds from the perspective of murmurs, classifying them into a binary category of presence or absence of murmurs, and additionally provided information on the location and timing of these murmurs. (Yaseen et al., 2018) collected heart sound recordings and corresponding annotations available on the internet to construct a five-category dataset. Regarding these datasets, the prevailing methodologies primarily consist of audio feature extraction, a main network, and a classification network. (Chen et al., 2023) initially applied noise reduction to the audio, then used Short-Time Fourier Transform (STFT) to obtain the spectrogram of the audio. Their main network was built on a CNN and incorporated an attention module at the end. (Guo et al., 2023) employed a combination of high-order spectral estimation and STFT for audio feature extraction, and designed a dual-stream CNN as the main network. (Cheng and Sun, 2023) simplified pre-processing, automatically extracting features using a one-dimensional convolution and built their main network based on a transformer architecture (Vaswani et al., 2017). In contrast to previous datasets, our dataset initiates with echocardiography reports, which serve as the gold standard for HSD, containing a wealth of information essential for medical practitioners. In this paper, we process these reports into a 12-category multi-label task. Methodologically, we depart from the traditional paradigms of heart sound classification models by constructing a knowledge-aware, query-based transformer, which significantly enhances the model's performance.

2.2 Query-based Transformer

Query-based transformers (Ma et al., 2023) (Dan and Roth) (Lopez-Avila and Suárez-Paniagua) use adjustable query embeddings to make predictions and benefit from global attention, enabling them to gather information from an entire input. This allows them to outperform convolutional networks in terms of results. (Carion et al., 2020) first introduced the query-based transformer in the object detection task and viewed it as a direct set prediction problem. (Li et al., 2022) introduces the concept of incorporating noised ground-truth boxes as positional queries in denoising training, an approach that has been shown to accelerate detection speeds. In addition to detection targets, (Cheng et al., 2022) employs mask attention for segmentation by utilizing predicted masks as attention masks, which enhances query refinement more efficiently than other query-based models. In the HSD task, we are the first to introduce the query-based transformer as the principal architecture. Unlike the models in other tasks, our queries consist of medical descriptions rather than words. Additionally, during the inference stage, we proposed a method that comprehensively utilizes multi-dimensional and diversified descriptions.

3 HSDreport: New Benchmark for Heart Sound Diagnosis

3.1 Background

The use of heart sounds as a cost-effective method for detecting congenital heart disease (CHD) is widely recognized. However, the current HSD dataset is limited to distinguishing between normal and abnormal conditions, which provides limited information in clinical practice. Echocardiography, on the other hand, offers more comprehensive information but is costly and difficult to obtain. Given the correlation between heart sounds and echocardiography, and to leverage the affordability of heart sound acquisition along with the comprehensiveness of echocardiography, we propose the HSDreport. This initiative aims to utilize paired heart sounds and echocardiography reports to extract key information from echocardiography reports based on heart sounds.

Given an input heart sound vector $\mathbf{h} \in \mathbb{R}^{\mathbf{t}}$, we aim to predict the key abnormalities $E = \{E_1, E_2, ..., E_k\}$ extracted from the corresponding echocardiography report. The relationship can be modeled as:

$$\hat{E}_i = f_i\left(\mathbf{h}\right), i \in [1, k]$$

Here $\hat{E}_i$ represents the predicted value of abnormality E_i, and f_i is a predictive function for each E_i that maps from the heart sound input $\mathbf{h}$ to the output space of abnormalities.

3.2 Data Collection

HSDreport includes 2, 275 participants and 2, 275 auscultation recordings. Specifically, digital auscultation recordings of heart sounds from patients aged $\leq$18 years who had undergone echocardiography, are collected. The auscultation protocol consists of recordings over 5 body sites: aortic region (right 2nd intercostal space), pulmonic region (left 2nd intercostal space, parasternal), Erb's point (left 3rd intercostal space aka left lower sternal border), tricuspid region (left 4th intercostal space, parasternal), mitral region (left 5th intercostal space, midclavicular). To detect sufficient cardiac cycles, at least 15s of heart sound using direct skin contact is obtained per site. The electronically amplified stethoscope (Littmann 3200, 3M) is used for data acquisition. During the examination the participant is seated, laid down, or

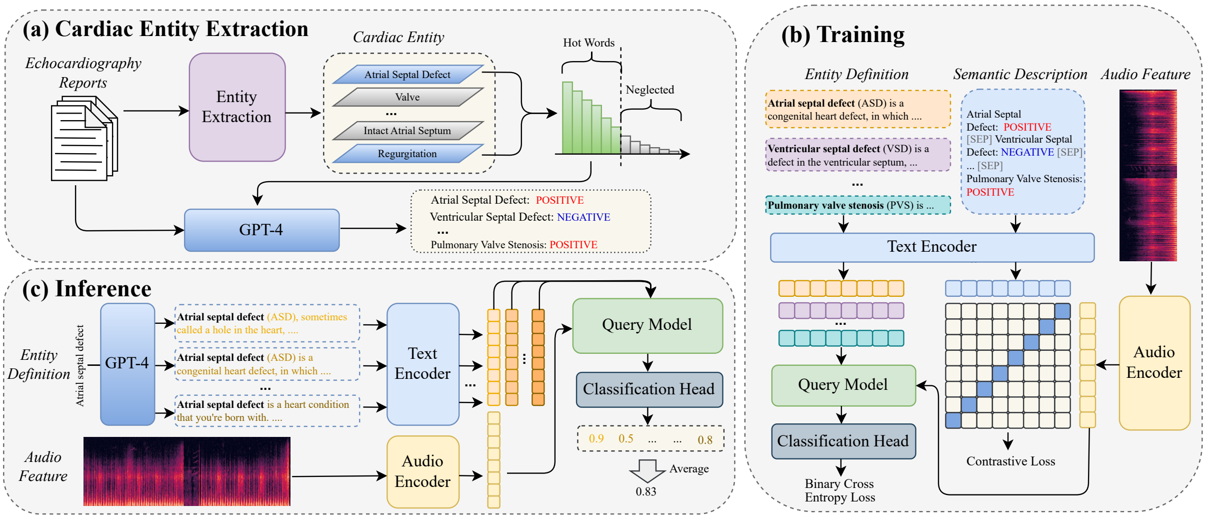
(a) Cardiac Entity Extraction
Echocardiography Reports
Entity Extraction
Cardiac Entity
Atrial Septal Defect
Valve
...
Intact Atrial Septum
Regurgitation
Hot Words
Neglected
GPT-4
Atrial Septal Defect: POSITIVE
Ventricular Septal Defect: NEGATIVE
...
Pulmonary Valve Stenosis: POSITIVE
(b) Training
Entity Definition
Semantic Description
Audio Feature
Atrial septal defect (ASD) is a congenital heart defect, in which ....
Ventricular septal defect (VSD) is a defect in the ventricular septum, ...
...
Pulmonary valve stenosis (PVS) is ...
Atrial Septal Defect: POSITIVE [SEP] Ventricular Septal Defect: NEGATIVE [SEP] ... [SEP] Pulmonary Valve Stenosis: POSITIVE
Text Encoder
...
Query Model
Classification Head
Binary Cross Entropy Loss
Contrastive Loss
Audio Encoder
(c) Inference
Entity Definition
Atrial septal defect
GPT-4
Atrial septal defect (ASD), sometimes called a hole in the heart, ...
Atrial septal defect (ASD) is a congenital heart defect, in which ....
...
Atrial septal defect is a heart condition that you're born with. ....
Text Encoder
...
Query Model
Classification Head
0.9 0.5 ... ... 0.8
Average
0.83
Audio Feature
Audio Encoder

Figure 2: Framework overview of HSDreport, which consists of **(a) Cardiac Entity Extraction** to filter the hot words from the abnormal cardiac entities and verify the existence of them with GPT-4, **(b) Training** stage to train our model with extracted entity definitions and semantic descriptions paired by the heart sounds, and **(c) Inference** stage to obtain the diagnosis for a specific entity with various definitions derived from GPT-4.

held to the most comfortable position. A complete transthoracic echocardiography is available for all subjects using standard views and techniques according to established guidelines. The report is written by professional ultrasound technicians.

During the data collection process, investigators have taken the following steps to improve the quality of the data collected. Firstly, all investigators remove data they deem inappropriate, including low-quality data where heart sounds are difficult to hear, data without corresponding echocardiography, and data from patients over 18 years old. In the next step, investigators review the echocardiography entries, screening all entries for incorrectly entered or measured values, inconsistent data, or outliers, and delete such entries as appropriate. Despite these steps taken to ensure data quality, it is important to note that the heart sounds are recorded in an ambulatory environment. Therefore, a variety of noisy sources are inevitably present, including stethoscope rubbing noise, and background sounds such as speaking, crying, or laughing. From another perspective, however, this dataset is more closely aligned with medical practice.

3.3 Data Processing

A typical echocardiography report is composed of three primary sections: numerical indices, description, and diagnosis. Echocardiography reports present challenges for direct learning by models, thus necessitating an approach where we extract all abnormalities from an echocardiography as task annotations through a series of steps, detailed further below and in Figure 1(a). Firstly, this report is produced by physicians who interpret various imaging modalities. Although various numerical indices such as blood flow velocities and thicknesses are closely related to heart sounds—where different velocities and thicknesses may correspond to variations in heart sounds—it is challenging to directly measure specific numerical indices solely through heart sounds given the current data size. Therefore, we have initially excluded all numerical information from the report. We believe this approach does not result in the loss of critical information, as physicians tend to describe any abnormal indices using natural language within the description or diagnosis sections.

Secondly, since the description and diagnosis sections of echocardiography reports are written by physicians using natural language without a fixed template, these reports contain significant noise. A typical characteristic is that most of the content in the description sections pertains to normal findings, with abnormalities being quite sparse. Consequently, it is challenging for models to directly learn from the description sections. Moreover, the diagnosis sections only contain the final disease diagnoses without the associated abnormalities, resulting in incomplete information. Therefore, we believe it is necessary to integrate both the description and diagnosis sections and denoise them to extract all key abnormalities for model learning.

To achieve this, we first employ a medical entity extraction method to extract all medical entities from the reports and rank these entities by their frequency of occurrence. Subsequently, we remove non-abnormality entities and select the 12 most frequently occurring abnormalities to represent the potential abnormal information in an echocardiography report. The remaining abnormalities, which appear fewer than 20 times, are excluded from model training due to their infrequency.

The final step involves utilizing a LLM to identify the presence of 12 categories of abnormalities in each report. This step is necessary due to the inherent natural language characteristics of description and diagnosis, which often result in irregular forms of abnormalities. Consequently, the internal knowledge and natural language understanding capabilities of the LLM are leveraged to accurately detect the existence of these 12 types of abnormalities. For any abnormalities not mentioned by the physician, we assume their absence.

4 Methodology

4.1 Training

In the development of the HSDreport, we diverge from the conventional architectural paradigm prevalent in previous heart sound models, which typically entails feature extraction and a linear classification layer. Instead, we innovatively constructed our model based on a query-based transformer. This approach facilitates the integration of a medically pre-trained text encoder to encode each abnormal entity. We argue that this method will enhance alignment between the two modalities. Contrary to the traditional use of query-based transformers, where words are employed as categories, we employ medical descriptions rather than words due to the frequent obscurity and variability in the expression of medical phrases in practice. This innovative substitution allows the model to gain a more detailed understanding of the symptoms and characteristics of medical conditions, thereby offering a holistic view of the disease. So for each abnormal entity E_j, the following applies:

$$\mathbf{t}_j = f_{\text{text}}\left(f_{\text{definition}}\left(E_j\right)\right) \in \mathbb{R}^d$$

Here j represents the j^{th} class. For the i^{th} sample's heart sound $\mathbf{h}_i \in \mathbb{R}^{t\times 1}$, we first extract their spectrograms, then obtain their features through a pre-trained audio encoder:

$$\mathbf{H}_i = f_{\text{audio}}\left(f_{\text{feature}}\left(\mathbf{h}_i\right)\right) \in \mathbb{R}^{x_h \times d}$$

Then we employ the semantically enriched text features $\mathbf{t}_j$ as the query, which, together with the heart sound features $\mathbf{H}_i$, are fed into the query-based transformer. After processing through a classification head composed of linear layers, the predicted values $\hat{E}_{i,j}$ are obtained:

$$\hat{E}_{i,j} = f_{\text{linear}}\left(f_{\text{transformer}}\left(\mathbf{H}_i, \mathbf{t}_j\right)\right) \in \mathbb{R}$$

Then we can compute the binary cross-entropy loss (BCE) as follows:

$$\mathcal{L}_{BCE} = -\tfrac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K}\left[s_{i,j}\log\left(\hat{E}_{i,j}\right) + (1-s_{i,j})\log\left(1-\hat{E}_{i,j}\right)\right]$$

Here $s_{i,j}$ is the label. To further align the audio and textual modalities, we choose to employ the semantic description R_i, composed of the entity names in textual form, annotations, and the delimiter [SEP]. The rationale for using names instead of medical descriptions is that concatenating all medical descriptions results in excessive length. After R_i is processed through the text encoder to obtain $\mathbf{r}_i \in \mathbb{R}^d$, we compute the contrastive loss between $\mathbf{r}_i$ and the audio features $\mathbf{H}_i$:

$$\mathcal{L}_{Con} = -\tfrac{1}{N}\sum_{i=1}^{N}\left(\log\frac{e^{\langle \mathbf{H}_i, \mathbf{r}_i\rangle/\tau}}{\sum_{k=1}^{N} e^{\langle \mathbf{H}_i, \mathbf{r}_k\rangle/\tau}} + \log\frac{e^{\langle \mathbf{r}_i, \mathbf{H}_i\rangle/\tau}}{\sum_{k=1}^{N} e^{\langle \mathbf{r}_i, \mathbf{H}_k\rangle/\tau}}\right\rangle \tag{1}$$

Here the operation $\langle ., .\rangle$ first applies average pooling to $\mathbf{H_i}$, followed by the computation of the cosine similarity and τ represents the temperature. Finally, we sum the two losses to obtain the final loss, where the weight λ is a learnable parameter:

$$\mathcal{L} = \mathcal{L}_{BCE} + \lambda\mathcal{L}_{Con} \tag{2}$$

4.2 Inference

Considering the diverse nature of medical expressions, we utilize an LLM to generate multi-dimensional and diverse descriptions of the same disease. During the inference stage, we integrate these descriptions to align our approach more closely with medical practice and enhance its robustness. Specifically, for the j^{th} class of abnormal entity E_j, we generate p descriptions using the LLM and input them collectively into the model:

$$\hat{E}_{i,j} = f_{\text{linear}}\left(f_{\text{t}}\left(\mathbf{H}_i, \mathbf{t}_{j,1}, ..., \mathbf{t}_{j,p}\right)\right) \in \mathbb{R}^p$$

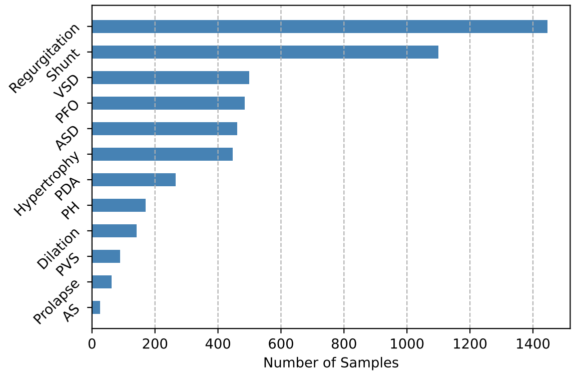

Figure 3: Data distribution. The total number of samples is 2275, and the numbers in the figure represent the number of positives in that category.

Here $\mathbf{H}_i$ represents the audio feature of the i^{th} sample, while $\mathbf{t}_{j,p}$ denotes the feature of the p^{th} description generated for class j. The function f_t refers to the query-based transformer. Subsequently, we compute the average of $\hat{E}_{i,j}$ to obtain the final result for that class.

5 Experiments

5.1 Experimental Setup

Datasets The dataset utilized in our study comprises 2, 275 heart sounds paired with the echocardiography reports, each approximately 75 seconds long, split into training and test sets respectively in a 9:1 ratio. The extracted abnormal entities from the paired reports are filtered and categorized into 7 diseases: Atrial Septal Defect (ASD), Ventricular Septal Defect (VSD), Pulmonary Valve Stenosis (PVS), Patent Ductus Arteriosus (PDA), Patent Foramen Ovale (PFO), Aortic Stenosis (AS), Pulmonary Hypertension (PH), and 5 symptoms: Prolapse, Regurgitation, Shunt, Hypertrophy, and Dilation. The distribution is demonstrated in Figure 3.

Baselines We adopt three state-of-the-art heart sound auscultation methods as our baselines: STFT-HSC (Chen et al., 2023), DS-CNN (Guo et al., 2023) (both CNN-based), and CTENN (Cheng and Sun, 2023) (transformer-based). To enable them to adapt to our multi-label classification task, we have modified their classification heads and retrained them following the instructions. Due to the existing LLMs' capability for audio processing, we also adopt GPT-4o as an LLM baseline.

Training Stage At the training stage, we utilize the filter banks of heart sounds as the audio features, with a frame length of 100ms and a frame shift of 40ms. Chunk dropping, speed perturbation, clipping, noising, amplifying, and SpecAugment (Park et al., 2019) are applied with the SpeechBrain library as data augmentation. The definition for each entity is derived from Wikipedia. For the text encoder and audio encoder, we adopt pre-trained PubMedBERT and ResNet50 as the initialization. We use an AdamW optimizer with a learning rate of 5e-5 and a weight decay of 0.02. We train on an A100 GPU for 100 epochs with 20 for warming up, and the cosine learning rate schedule is applied. λ in Equation (2) is initialized to 1.

Inference Stage At the inference stage, we utilize GPT-4 to generate 100 descriptions for each class. We adopt precision, recall, and F1 scores as the evaluation metrics and set 0.5 as the discriminant threshold.

5.2 Results

5.2.1 Multi-label Classification

We conduct experiments on the proposed multi-label benchmark HSDreport and compare the diagnosis performance with the state-of-the-art approaches. The precision, recall, and F1 scores of each category and the averaged values are reported. As demonstrated in Table 1, Our approach attains the highest F1-scores in nearly all categories and exceeds the best baseline by 9.4%, demonstrating the effectiveness of the proposed knowledge-aware, query-based transformer for HSD task. Among the baselines, the transformer-based method (CTENN) outperforms those based on CNNs (STFT-HSC and DS-CNN). In addition, GPT-4o struggles with this task, achieving the lowest scores. We attribute this performance to the limited size of heart sound data in the public dataset and the absence of annotations similar to those in the HSDreport, which prevents GPT-4o from generalizing to our benchmark.

Ulteriorly, we split the evaluation metric into precision and recall, where recall is particularly important for medical diagnosis. Obviously, our approach exhibits the most superiority in the recall score for various diseases and symptoms. Such a superiority means that our method can reduce the false dismissal rate and avoid disease misdiagnosis

https://openai.com/index/hello-gpt-4o/

https://huggingface.co/NeuML/pubmedbert-base-embeddings

torchvision.models.resnet50

Method	ASD	VSD	PVS	PDA	PFO	AS	PH	Prolapse	Regurgitation	Shunt	Hypertrophy	Dilation	Average
							Precision						
GPT-4o	38.04	37.83	48.83	44.57	41.09	49.35	46.30	48.26	18.48	24.35	40.87	46.96	40.33
STFT-HSC	**88.54**	88.61	47.83	**94.96**	**89.47**	49.35	46.29	48.26	68.99	67.93	77.60	46.96	67.89
DS-CNN	62.28	82.17	47.81	85.33	84.56	49.35	46.30	48.26	67.13	70.89	**81.93**	46.94	64.41
CTENN	61.43	82.83	69.69	72.93	86.82	**99.78**	**80.03**	48.26	57.46	69.19	69.27	**57.11**	71.23
Ours	68.78	**90.28**	**84.31**	93.65	78.48	**99.78**	76.89	**86.39**	**73.40**	**72.35**	81.62	51.57	**79.79**
							Recall						
GPT-4o	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
STFT-HSC	52.73	80.71	50.00	54.00	64.37	50.00	49.77	50.00	69.98	67.23	66.00	50.00	58.73
DS-CNN	57.87	75.09	49.54	57.76	61.67	50.00	50.00	50.00	64.91	69.48	81.33	49.77	59.79
CTENN	61.14	78.98	73.41	60.78	65.32	**83.33**	55.65	50.00	57.85	68.88	65.06	**52.65**	64.42
Ours	**68.31**	**86.66**	**84.31**	**75.75**	**70.78**	**83.33**	**58.35**	**68.52**	**73.28**	**71.93**	**88.06**	51.25	**73.38**
							F1 Score						
GPT-4o	43.21	43.07	48.89	47.13	45.11	49.67	48.08	49.12	26.98	32.75	44.98	48.44	43.95
STFT-HSC	48.70	83.67	48.89	54.75	68.53	49.67	47.96	49.12	69.20	66.70	69.14	48.43	58.73
DS-CNN	58.53	77.62	48.66	60.78	64.81	49.67	48.08	49.11	65.38	68.70	81.62	48.31	60.11
CTENN	61.28	80.61	71.36	63.84	69.50	**89.89**	58.18	49.12	57.43	68.61	66.62	**53.35**	65.82
Ours	**68.53**	**88.27**	**84.31**	**81.78**	**73.57**	**89.89**	**61.80**	**74.33**	**73.34**	**71.64**	**84.20**	51.35	**75.25**

Table 1: The precision, recall, and F1 scores of GPT-4o, three state-of-the-art models for HSD, and our method on the multi-label benchmark. Our model achieves significant improvements in all three metrics.

effectively. It is also noticed that all of the methods are not able to recognize the existence of Dilation effectually, and we guess it's because the characteristics of Dilations may hardly be identified by heart sounds.

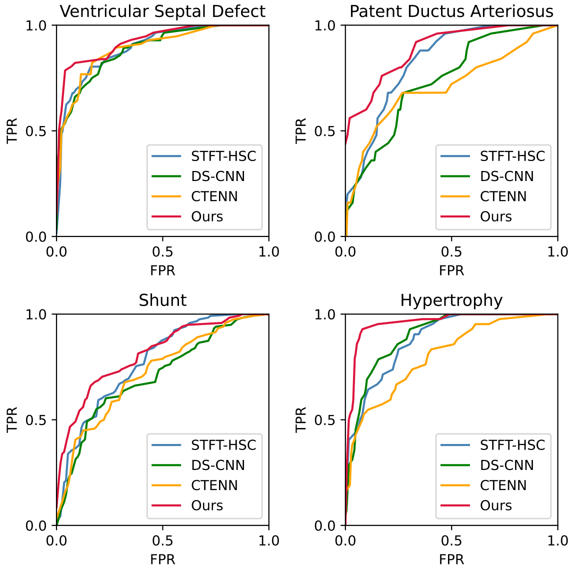

Figure 4: ROC curves of four classes in the benchmark: VSD, PDA, Shunt, and Hypertrophy.

5.2.2 ROC Curve

To further analyze the robustness under different discrimination thresholds, we plot the receiver operating characteristic (ROC) curve of diseases (VSD and PDA) and symptoms (Shunt and Hypertrophy) respectively. As shown in Figure , the area under the ROC curve (AUC) for our approach is notably superior and maintains the highest values compared to that of the established baselines, which not only signifies the discriminatory capacity of our model between positive and negative cases but also underscores its robustness and generalizability across different threshold settings.

5.3 Ablation Study

We conduct ablation studies for the training and inference stage respectively to verify the effectiveness and robustness of each counterpart in the proposed method.

5.3.1 Training Stage

We first establish the ablation by systematically disabling components during the training stage. This involves, specifically, replacing the pre-trained initializations of both the text encoder and audio encoder with random initialization procedures individually. Additionally, we excise the contrastive loss component from the overarching loss function outlined in Equation 1. Lastly, we substitute the comprehensive entity definitions with the entity words in our model's architecture. The outcomes of these ablation experiments are documented in Table 2, revealing the indispensable nature of each module in the diagnostic process as any module removal brings a distinct decline in performance. Among these, replacing the entity definition exerts

Method	ASD	VSD	PVS	PDA	PFO	AS	PH	Prolapse	Regurgitation	Shunt	Hypertrophy	Dilation	Average
Ours	**68.53**	**88.27**	**84.31**	**81.78**	73.57	**89.89**	61.80	**74.33**	73.34	71.64	84.20	51.35	**75.25**
- Text Encoder	68.50	84.46	83.69	79.03	72.22	**89.89**	61.09	67.40	**76.48**	72.09	86.41	**52.95**	74.52
- Audio Encoder	65.39	82.56	69.22	68.96	65.04	49.56	**64.97**	70.53	66.23	71.24	82.87	52.94	67.46
- $\mathcal{L}_{Con}$	66.78	87.40	67.56	78.19	**75.23**	83.11	64.04	67.40	74.44	**74.30**	85.94	47.72	72.68
- Entity Definition	55.28	82.29	58.18	74.82	64.90	48.77	61.85	63.40	72.20	73.46	**89.46**	52.26	66.41

Table 2: This table presents an ablation study to assess the validity of the text encoder, audio encoder, contrastive loss, and entity definition input during the training stage. F1 scores are reported for each category. The table demonstrates the efficacy of all modules.

Text Input	ASD	VSD	PVS	PDA	PFO	AS	PH	Prolapse	Regurgitation	Shunt	Hypertrophy	Dilation	Average
Only Entity	**69.27**	87.40	80.78	77.95	73.04	71.66	61.80	70.53	71.06	70.32	84.97	55.81	72.88
$N = 1$	68.12	87.57	76.87	79.92	70.52	**89.89**	61.09	65.77	**73.74**	71.28	**84.20**	**59.69**	74.05
$N = 10$	68.53	**88.27**	**84.31**	**81.78**	**73.57**	**89.89**	**61.80**	**74.33**	73.34	71.64	**84.20**	51.35	**75.25**
$N = 50$	68.12	87.57	**84.31**	79.92	73.04	**89.89**	**61.80**	67.40	73.34	**72.42**	**84.20**	58.68	75.05

Table 3: This table presents an ablation study aggregating descriptions generated by different quantities of GPT-4 during the inference stage. The F1 scores are reported for each category. The table demonstrates that using ten descriptions is the most effective method.

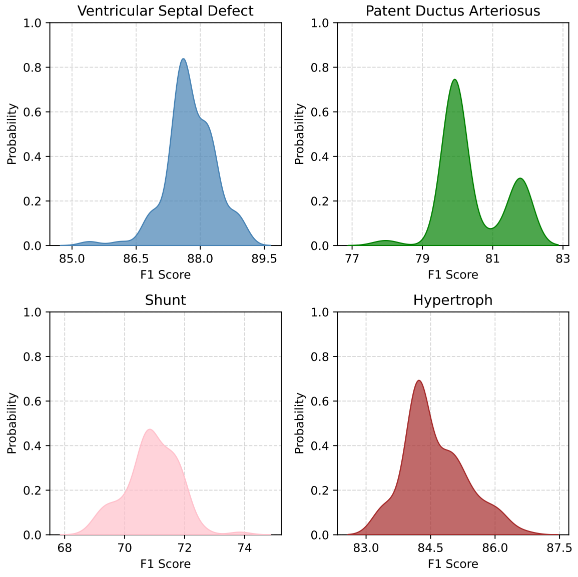

Figure 5: The distribution of F1 scores for VSD, PDA, Shunt, and Hypertrophy categories.

the most significant impact, substantiating our assertion that medical terminology is characterized by rare and diverse issues, with the lexical semantics being insufficient for models to comprehend diseases. Replacing the audio encoder also has a substantial effect, likely due to the complexity of audio features, which are difficult to learn from a random state given the current sample size. In contrast, replacing the text encoder and contrastive loss has a relatively minor impact.

5.3.2 Inference Stage

During the inference stage, we validate the outcomes under various quantities of GPT-4-generated entity definitions. Specifically, we substitute the text input for the inference stage with the entity words (which means $N = 0$) and varying numbers of generated entity definitions and document the performance as shown in Table 3. It is evident that using the entity definition as a text query consistently yields superior results compared to merely utilizing the entity words. This also substantiates the usefulness of medical descriptions in the HSD task, which possess more precise semantics. Additionally, such performance enhancement can be further amplified by increasing the number of entity definitions. This enhancement stabilizes when $N = 10$, which also serves as the default setting in our experiments. No further changes in outcomes were observed with an increase in the number of entity definitions beyond this point, likely because ten descriptions are sufficiently diverse and rich for most categories as generated by GPT-4.

To further analyze the generality of our model, we feed 100 definition descriptions generated by GPT-4 for the same diseases (VSD and PDA) and symptoms (Shunt and Hypertrophy) and record the distribution of the resulting F1 scores. As illustrated in Figure 5, the performance can be affected by varied definition descriptions, which also supports the necessity of our aggregation strategy.

6 Conclusion

This study introduces HSDreport, a new benchmark and method designed to revolutionize HSD by integrating the diagnostic gold standard of echocardiography with the accessibility of auscultation. By collecting a novel dataset of 2,275 heart sound samples paired with echocardiography, we set the stage for a more detailed approach to diagnosing heart conditions. The development of our knowledge-aware query-based transformer model marks a significant advancement in the field, effectively harnessing the potential of medically pre-trained models and the nuanced understanding of LLM.

Limitations

The dataset used in this study focuses exclusively on pediatric subjects, offering valuable insights specific to this demographic but may not be entirely representative of broader population dynamics. Moreover, our novel dataset is enriched with detailed ultrasonography report data, allowing for potential stratification of severity in abnormal conditions based on numerical indicators. However, the current methodology primarily addresses the presence rather than the severity of these abnormalities, suggesting an avenue for future enhancement. Lastly, due to the multiple inferences required by our approach, the inference speed is slightly slower compared to the baseline model, mainly because we prioritize accuracy over speed.

Ethical Considerations

This study was conducted with a strong commitment to ethical standards in medical research, ensuring the protection and confidentiality of participant data and compliance with relevant regulations. Here we outline the ethical considerations addressed in this study.

Informed Consent All patients, or their legal guardians in the case of minors, provided informed consent for the use of their medical data in this research. Prior to data collection, participants were adequately informed about the nature of the study, the type of data to be collected (heart sounds and ultrasound reports), and the intended use of this data in research. This process was conducted in accordance with the Declaration of Helsinki regarding ethical principles for medical research involving human subjects.

Data Confidentiality and Security Rigorous measures were taken to ensure the confidentiality and security of the data collected. Personal identifiers were removed from all datasets to achieve anonymization. Additionally, all digital data were stored in an encrypted environment to prevent unauthorized access, ensuring that the privacy of the participants was maintained throughout the study.

Compliance with Regulatory Standards The study strictly adhered to national laws and regulations concerning medical research and data protection. This adherence was continuously monitored by our legal and ethical advisory board to ensure ongoing compliance throughout the study's duration.

Ethical Review and Oversight The research protocol was thoroughly reviewed and approved by an independent ethics committee. This committee provided continuous oversight and guidance to ensure that all aspects of the study were conducted ethically and that the welfare of the participants was prioritized at all times.

Acknowledgement

This work is supported by National Key R&D Program of China (No. 2022ZD0162101), National Natural Science Foundation of China (No. 62106140) and STCSM (No. 21511101100, No. 22DZ2229005)

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Junxin Chen, Zhihuan Guo, Xu Xu, Li-bo Zhang, Yue Teng, Yongyong Chen, Marcin Woniak, and Wei Wang. 2023. A robust deep learning framework based on spectrograms for heart sound classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299.

Jiawen Cheng and Kexue Sun. 2023. Heart sound classification network based on convolution and transformer. *Sensors*, 23(19):8168.

Gari D Clifford, Chengyu Liu, Benjamin Moody, David Springer, Ikaro Silva, Qiao Li, and Roger G Mark. Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016. In *Computing in cardiology conference (CinC)*, pages 609–612. IEEE.

Soham Dan and Dan Roth. On the effects of transformer size on in-and out-of-domain calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2096–2101.

Zhihuan Guo, Junxin Chen, Tongyue He, Wei Wang, Haider Abbas, and Zhihan Lv. 2023. DS-CNN: Dual-stream convolutional neural networks based heart sound classification for wearable devices. *IEEE Transactions on Consumer Electronics*.

Suyoun Kim, Ke Li, Lucas Kabela, Ron Huang, Jiedan Zhu, Ozlem Kalinli, and Duc Le. 2022. Joint audio/text training for transformer rescorer of streaming speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 5717–5722, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anastasia Kritharoula, Maria Lymperaiou, and Giorgos Stamou. 2023. Large language models and multi-modal retrieval for visual word sense disambiguation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13053–13077, Singapore. Association for Computational Linguistics.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. 2022. DN-DETR: Accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627.

Alejo Lopez-Avila and Víctor Suárez-Paniagua. Combining denoising autoencoders with contrastive learning to fine-tune transformer models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhengrui Ma, Shaolei Zhang, Shoutao Guo, Chenze Shao, Min Zhang, and Yang Feng. 2023. Non-autoregressive streaming transformer for simultaneous translation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Salvatore Mangione and Linda Z Nieman. 1997. Cardiac auscultatory skills of internal medicine and family practice trainees: a comparison of diagnostic proficiency. *Jama*, 278(9):717–722.

Maria Rosa Montinari, Simona Giardina, Pierluca Minelli, and Sergio Minelli. 2018. History of music therapy and its contemporary applications in cardiovascular diseases. *Southern medical journal*, 111(2):98–102.

The Writing Committee of the Report et al. 2022. Report on cardiovascular health and diseases in china 2021: an updated summary. *Biomedical and Environmental Sciences*, 35(7):573–603.

Jorge Oliveira, Francesco Renna, Paulo Dias Costa, Marcelo Nogueira, Cristina Oliveira, Carlos Ferreira, Alípio Jorge, Sandra Mattos, Thamine Hatem, Thiago Tavares, et al. 2021. The circor digiscope dataset: from murmur detection to murmur classification. *IEEE journal of biomedical and health informatics*, 26(6):2524–2535.

Feixia Pan, Weize Xu, Jiabin Li, Ziyan Huang, and Qiang Shu. 2022. Trends in the disease burden of congenital heart disease in china over the past three decades. *Journal of Zhejiang University. Medical Sciences*, 51(3):267–277.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Gregory A Roth, George A Mensah, Catherine O Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M Baddour, Noël C Barengo, Andrea Z Beaton, Emelia J Benjamin, Catherine P Benziger, et al. 2020. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study. *Journal of the American college of cardiology*, 76(25):2982–3021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jing Wang, Xiaofeng Liu, Fangyun Wang, Lin Zheng, Fengqiao Gao, Hanwen Zhang, Xin Zhang, Wanqing Xie, and Binbin Wang. 2021. Automated interpretation of congenital heart disease from multi-view echocardiograms. *Medical image analysis*, 69:101942.

Shaoning Xiao, Long Chen, Kaifeng Gao, Zhao Wang, Yi Yang, Zhimeng Zhang, and Jun Xiao. 2022. Rethinking multi-modal alignment in multi-choice VideoQA from feature and sample perspectives. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8188–8198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yaseen, Gui-Young Son, and Soonil Kwon. 2018. Classification of heart sound signal using multiple features. *Applied Sciences*, 8(12):2344.

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2023. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542.