Heart Murmur Classification Report

Ternary Classification based on M2D and Transformers

Name	:	Yue Zhang
ID	:	202364820921
Class	:	Data Science Class 2
College	:	School of Future Technology

May 29, 2025

Abstract

Cardiac auscultation serves as a non-invasive and cost-efficient approach for the preliminary screening of cardiovascular diseases (CVDs). In this work, we present a ternary heart murmur classification framework that combines pretrained Masked Modeling Duo (M2D) representations with a multi-head Transformer-based classifier. Evaluated on the CirCor DigiScope dataset, our proposed model achieves superior performance compared to previous state-of-the-art (SOTA) approaches, notably in terms of weighted accuracy (W.acc) and unweighted average recall (UAR), highlighting its effectiveness for clinical audio intelligence applications.

1 Introduction

Cardiovascular diseases (CVDs) remain the leading cause of global mortality. Heart murmurs, which often signal underlying valvular abnormalities, are traditionally diagnosed via auscultation—a process that heavily relies on the clinician's expertise. However, the subjectivity and variability of human interpretation pose limitations on consistency and diagnostic accuracy. Recent advances in deep learning have opened new avenues for robust audio-based diagnostics, particularly through the use of pretrained self-supervised models. In this report, we introduce a fully automated ternary murmur classification framework that integrates M2D representations with a Transformer-based classifier, aiming to enhance reliability and reproducibility in clinical auscultation analysis.

2 Related Work

Previous works on heart murmur classification can be grouped into two broad categories.

(1) Conventional Deep Learning Methods. Early approaches adopted convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to classify phonocardiograms. For example, CNN-based models were trained directly on spectrograms extracted from heart sound recordings, while bidirectional LSTMs were employed to capture temporal dynamics. However, these methods often require large annotated datasets and suffer from overfitting on small clinical samples.

(2) Self-Supervised Pretrained Audio Models. Recent advances introduced self-supervised learning (SSL) models such as wav2vec 2.0, Audio Spectrogram Transformer (AST), and Masked Modeling Duo (M2D), which are pretrained on large-scale unlabeled audio corpora like AudioSet. Among them, M2D has shown state-of-the-art performance on the CirCor DigiScope dataset without fine-tuning the encoder, thanks to its task-agnostic feature learning.

Our work builds upon the M2D framework and proposes several key enhancements:

- **Multi-Site Token Encoding:** Incorporating auscultation site structure into token design.
- **Attention-based Global Pooling:** Allowing the model to learn dynamic importance across sites.
- **Multi-Head Distillation:** Encouraging diverse yet consistent predictions via inter-head KL regularization.

This design leads to improved performance on challenging classes, especially "Unknown", where label ambiguity is high.

3 Methodology

3.1 Feature Extraction via Multi-Site Embedding with Pretrained M2D

To transform raw auscultation audio signals into semantically meaningful representations, we design a structured feature extraction pipeline termed **M2D-Embed**, leveraging the pretrained Masked Modeling Duo (M2D) model [choi2022m2d]. Given a subject's recordings from up to four auscultation sites (AV, MV, PV, TV), the pipeline proceeds as follows:

(1) **Audio Standardization:** Each raw waveform $x \in \mathbb{R}^T$ is resampled to 16 kHz and padded or truncated to a fixed length of $T = 160{,}000$ samples (i.e., 10 seconds) to form $\tilde{x}$.

(2) **Spectrogram Encoding:** Using the pretrained M2D encoder f_θ, each $\tilde{x}$ is mapped to a frame-level embedding $f_\theta(\tilde{x}) \in \mathbb{R}^{L \times D}$, where L is the temporal resolution and $D = 768$ is the embedding size.

(3) **Clip-Level Pooling:** We average over all frames to obtain a site-level embedding, computed as:

$$\mathbf{z}_s = \frac{1}{L} \sum_{t=1}^{L} \mathbf{h}_t, \quad \text{where } \mathbf{H} = f_\theta(\tilde{x}_s) \in \mathbb{R}^{L \times D} \tag{1}$$

where $s \in \{\text{AV}, \text{MV}, \text{PV}, \text{TV}\}$ denotes the site.

(4) **Multi-Site Fusion:** All site-level embeddings for a patient are concatenated to form:

$$\mathbf{X}_i \in \mathbb{R}^{4 \times 768} \tag{2}$$

Missing sites are replaced with zero vectors, and a binary mask $\mathbf{M}_i \in \{0, 1\}^4$ is generated to indicate valid sites.

Formally, the full dataset becomes:

$$\mathcal{D} = \{(\mathbf{X}_i, \mathbf{M}_i, y_i)\}_{i=1}^{N}$$

where $y_i \in \{0, 1, 2\}$ represents the ternary murmur class (Absent, Present, Unknown). This structured design enables the downstream model to explicitly reason over missing-site uncertainty, while benefiting from high-level M2D representations without requiring any fine-tuning.

This pipeline processes 1568 subjects, automatically skipping those with invalid metadata or malformed audio, and serializes the outputs into NumPy archives for subsequent training.

3.2 M2DTransformer Architecture and Training Procedure

To perform murmur classification over variable-site M2D embeddings, we propose a specialized encoder called **M2DTransformer**, which combines multi-head self-attention, attention-based pooling, and multi-head classification with distillation.

Architecture Overview. Given a 4-token sequence $\mathbf{X} \in \mathbb{R}^{4 \times 3840}$ representing the embeddings of auscultation sites, we apply a linear projection $\mathbf{W}_p \in \mathbb{R}^{3840 \times 512}$ to map each token to a 512-dimensional space. A mask embedding $\mathbf{E}_m \in \mathbb{R}^{2 \times 512}$ is then added to encode missing sites:

$$\tilde{\mathbf{x}}_i = \mathbf{W}_p \mathbf{x}_i + \mathbf{E}_m[m_i], \quad i = 1, \ldots, 4$$

The resulting tokens are passed through a 4-layer Transformer encoder with $n = 4$ heads and feedforward width 1024. Let $\mathbf{H} \in \mathbb{R}^{4\times512}$ denote the encoder output.

Attention Pooling. We employ a learnable global attention pooling layer to compress token-wise representations:

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{H}\mathbf{w}) \in \mathbb{R}^4, \quad \mathbf{z} = \sum_{i=1}^{4} \alpha_i \mathbf{h}_i$$

where $\mathbf{w} \in \mathbb{R}^{512\times1}$ is a learnable query vector.

Multi-Head Distilled Classification. To enhance ensemble diversity and reduce overfitting, we branch out $K = 3$ classification heads $f^{(k)} : \mathbb{R}^{512} \to \mathbb{R}^3$ and compute logits:

$$\hat{\mathbf{y}}^{(k)} = f^{(k)}(\mathbf{z}), \quad \hat{\mathbf{y}} = \frac{1}{3}\sum_{k=1}^{3} \hat{\mathbf{y}}^{(k)}$$

Loss Function. We combine two objectives:

- **Focal Loss with Label Smoothing:** to mitigate class imbalance:

$$p_c = \frac{e^{\hat{y}_c}}{\sum_j e^{\hat{y}_j}}, \quad \mathcal{L}_{\text{main}} = -\sum_{c=1}^{C}(1-p_c)^{\gamma} \cdot y_c^{\text{smooth}} \cdot \log p_c$$

- **Distillation Loss:** a KL divergence between each head and the ensembled logits:

$$\mathcal{L}_{\text{distill}} = \frac{1}{K}\sum_{k=1}^{K} \text{KL}(\text{softmax}(\hat{\mathbf{y}}^{(k)}) \parallel \text{softmax}(\hat{\mathbf{y}}))$$

The total objective is:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{main}} + (1-\alpha) \cdot \mathcal{L}_{\text{distill}}, \quad \alpha = 0.7$$

Training Strategy. We train the model using AdamW with cosine annealing, batch size 32, and a learning rate of 10^{-4}. To enhance generalization, we apply SpecAugment during training by randomly masking frequency bins and time steps on the token axis:

$$x[b, :, f_s : f_s + f_l] = 0 \quad \text{(frequency masking)}, \quad x[b, t_s : t_s + t_l, :] = 0 \quad \text{(time masking)}$$

where f_l, t_l are random lengths and f_s, t_s are sampled starting indices.

All experiments are conducted on NVIDIA GPU with early stopping based on weighted accuracy (W.acc) over the validation set.

3.3 Audio Representation Backbone: PortableM2D

To extract high-level semantic representations from raw audio signals, we adopt a pretrained audio transformer backbone termed **PortableM2D**. This model extends the original M2D [choi2022m2d] framework with increased flexibility in embedding strategies, patch-level inference, and CLAP-style projection capabilities.

1) Mel-Spectrogram Frontend

Raw waveform $x \in \mathbb{R}^T$ is transformed into log-mel spectrograms via a learnable frontend:

$$S = \log(\text{MelSpec}(x) + \epsilon) \in \mathbb{R}^{B \times 1 \times 80 \times T'}$$

where $T' = \frac{T}{\text{hop\_size}}$, and ϵ ensures numerical stability.

2) Patch-wise Transformer Encoding

Spectrograms are embedded into tokens using a learnable patch embedding module:

$$z_{ij} = \mathbf{W} \cdot S[i : i + P, j : j + Q], \quad \mathbf{z}_{\text{patch}} \in \mathbb{R}^{N \times D}$$

where (P, Q) is the patch size (e.g., 16×16), and $D = 768$ is the transformer embedding size.

The transformer encoder applies standard block-wise modeling:

$$\mathbf{z}' = \text{Transformer}(\mathbf{z}_{\text{patch}}) \in \mathbb{R}^{N \times D}$$

3) Positional Adaptation & CLS Aggregation

Positional embeddings are dynamically truncated to fit varying temporal lengths. A CLS token is appended, and the output is aggregated via:

$$\mathbf{z}_{\text{scene}} = \mathbf{z}'_{\text{CLS}} \quad \text{or} \quad \mathbf{z}_{\text{mean}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}'_i$$

4) Output Modes and Generalization

The PortableM2D supports multiple modes:

- **Scene-level Embeddings:** $\mathbf{z}_{\text{scene}} \in \mathbb{R}^D$ (used for classification)
- **Timestamp-level Embeddings:** $\mathbf{Z}_{\text{ts}} \in \mathbb{R}^{T \times D}$ for fine-grained temporal modeling
- **CLAP Audio Embedding:** $\mathbf{z}_{\text{CLAP}} = \text{Proj}(\mathbf{z}_{\text{scene}})$ for contrastive language-audio alignment

Optional classification heads (e.g., MLP + BatchNorm) can be appended for supervised downstream tasks, or the output can be directly used for clustering, retrieval, or multimodal matching.

5) Design Benefits

The backbone preserves several desirable properties:

- Token-level semantics and temporal alignment
- High-dimensional contextualized embeddings
- Compatibility with CLAP-style fine-tuning
- Frozen-patch-encoder option for efficient adaptation

Overall, PortableM2D provides a unified, scalable interface for audio understanding across classification and retrieval settings, and serves as the representation engine throughout our pipeline.

Loss Design and Optimization Objective

To handle class imbalance and promote output consistency, we combine two loss components:

- **Focal Loss with Label Smoothing:** Focal loss dynamically down-weights easy examples, while label smoothing prevents overconfidence in predictions:

$$p_c = \frac{e^{\hat{y}_c}}{\sum_j e^{\hat{y}_j}}, \quad \mathcal{L}_{\text{main}} = -\sum_{c=1}^{C} (1 - p_c)^{\gamma} \cdot y_c^{\text{smooth}} \cdot \log p_c$$

  where $\gamma = 1.5$ and y_c^{smooth} is the smoothed target label.

- **Distillation Loss:** To reduce variance across classification heads, we introduce a KL-divergence loss between each head and the average ensemble:

$$\mathcal{L}_{\text{distill}} = \frac{1}{K} \sum_{k=1}^{K} \text{KL}(\text{softmax}(\hat{\mathbf{y}}^{(k)}) \parallel \text{softmax}(\hat{\mathbf{y}}))$$

  The final loss is:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{main}} + (1 - \alpha) \cdot \mathcal{L}_{\text{distill}}, \quad \alpha = 0.7$$

This hybrid loss balances classification accuracy with consistency among prediction heads, improving generalization under ambiguous label conditions.

4 Results

4.1 Classification Performance

The model was trained and evaluated using an 80/20 stratified split on the CirCor DigiScope dataset. Class-wise label distributions were preserved. The table below summarizes the performance on the test set:

Metric	**Value**
Accuracy	0.880
Macro F1	0.780
Recall - Absent	0.914
Recall - Present	0.833
Recall - Unknown	0.643
UAR	**0.797**
W.acc	**0.842**

Table 1: Test set performance on CirCor DigiScope dataset

4.2 Comparison with SOTA (M2D)

To evaluate the effectiveness of our proposed approach, we compare it against the original M2D classifier reported in [choi2022m2d], as shown in Table **??**. Our method achieves higher recall for the "Absent" and "Unknown" classes, and also outperforms in both UAR and W.acc. Notably,

the recall for "Unknown" improved by 28.2%, suggesting better handling of ambiguous or noisy recordings.

Table 2: Comparison with SOTA M2D Model on CirCor Dataset

Metric	SOTA (M2D)	Ours
Recall - Absent	0.868	**0.914**
Recall - Present	**0.911**	0.833
Recall - Unknown	0.361	**0.643**
UAR	0.713	**0.797**
W.acc	0.832	**0.842**

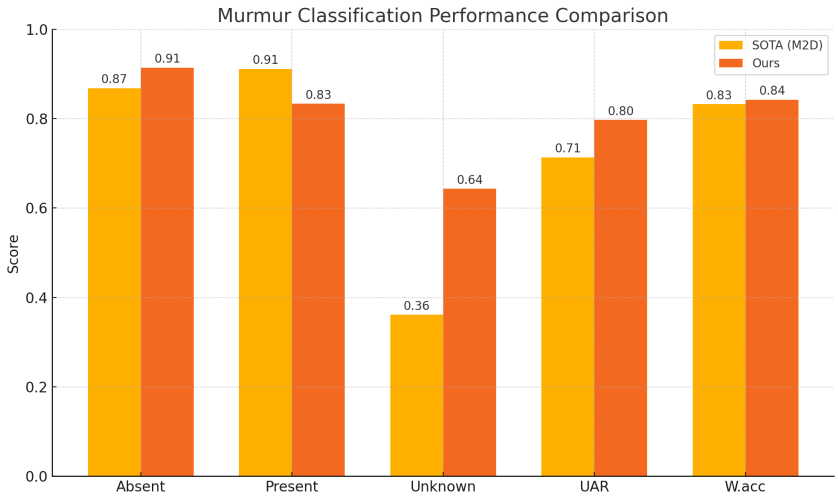

Figure 1: Visual comparison of our model and SOTA across key evaluation metrics.

4.3 Confusion Matrix

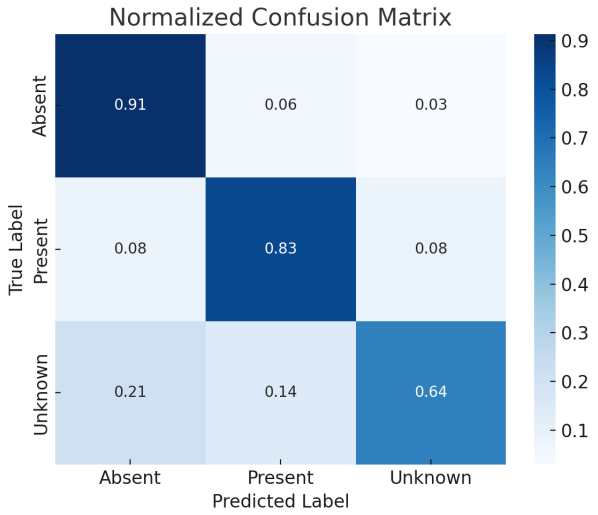

Figure 2: Normalized confusion matrix showing class-level predictions.

5 Conclusion

This work introduces a novel Transformer-based framework for ternary heart murmur classification, leveraging fixed M2D embeddings extracted from multi-site auscultation recordings. By integrating multi-site token-level encoding, attention-based global pooling, and multi-head output distillation, our system is designed to handle structural variability and enhance prediction consistency without requiring encoder fine-tuning.

Experimental evaluations on the CirCor DigiScope dataset demonstrate the superiority of our model over prior state-of-the-art baselines, achieving notable gains in both UAR and weighted accuracy. In particular, the recall for the "Unknown" murmur category improved significantly, highlighting the model' s robustness in the face of ambiguous or noisy recordings—a common challenge in real-world clinical audio data.

Despite its strong performance, the current approach may underperform on rare murmur types or unseen recording environments. Future research directions include integrating dynamic fine-tuning strategies, improving temporal resolution through token-level alignment, exploring ensemble-based uncertainty quantification, and employing attention heatmap visualizations to improve model interpretability and clinician trust.