This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

MultiASNet: Multimodal Label Noise Robust Framework for the Classification of Aortic Stenosis in Echocardiography

Victoria Wu, Andrea Fung, Bahar Khodabakhshian, Baraa Abdelsamad, Hooman Vaseli, Neda Ahmadi, Jamie A.D. Goco, Michael Y. Tsang, Christina Luong, Purang Abolmaesumi *Senior Member, IEEE*, Teresa S.M. Tsang

***Abstract*— Aortic stenosis (AS), a prevalent and serious heart valve disorder, requires early detection but remains difficult to diagnose in routine practice. Although echocardiography with Doppler imaging is the clinical standard, these assessments are typically limited to trained specialists. Point-of-care ultrasound (POCUS) offers an accessible alternative for AS screening but is restricted to basic 2D B-mode imaging, often lacking the analysis Doppler provides. Our project introduces MultiASNet, a multimodal machine learning framework designed to enhance AS screening with POCUS by combining 2D B-mode videos with structured data from echocardiography reports, including Doppler parameters. Using contrastive learning, MultiASNet aligns video features with report features in tabular form from the same patient to improve interpretive quality. To address misalignment where a single report corresponds to multiple video views, some irrelevant to AS diagnosis, we use cross-attention in a transformer-based video and tabular network to assign less importance to irrelevant report data. The model integrates structured data only during training, enabling independent use with B-mode videos during inference for broader accessibility. MultiASNet also incorporates sample selection to counteract label noise from observer variability, yielding improved accuracy on two datasets. We achieved balanced accuracy scores of 93.0% on a private dataset and 83.9% on the public TMED-2 dataset for AS detection. For severity classification, balanced accuracy scores were 80.4% and 59.4% on the private and public datasets, respectively. This model facilitates reliable AS screening in non-specialist settings, bridging the gap left by Doppler data while reducing noise-related errors. Our code is publicly available at github.com/DeepRCL/MultiASNet.**

***Index Terms*— Aortic stenosis, cardiac imaging, multimodal learning, sample selection, ultrasound.**

This work was supported in part by the Canadian Institutes of Health Research (CIHR) and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

V. Wu and A. Fung are joint first authors. T.S.M. Tsang and P. Abolmaesumi are joint senior authors.

V. Wu, B. Khodabakhshian, B. Abdelsamad, H. Vaseli and P. Abolmaesumi are with the Department of Electrical and Computer Engineering, the University of British Columbia, Vancouver, BC, Canada.

A. Fung and J.A.D. Goco are with the School of Biomedical Engineering, the University of British Columbia, Vancouver, BC, Canada.

M.Y. Tsang, C. Luong and T.S.M. Tsang are with the Faculty of Medicine, the University of British Columbia, Vancouver, BC, Canada.

I. Introduction

Aortic Stenosis (AS) is a life-threatening cardiac disease characterized by restricted movement of the aortic valve. For patients with untreated severe AS, average life expectancy is just 2 to 3 years [1]. Early detection is crucial but often missed in primary care settings. Previous work suggests that in primary care and resource-limited environments, non-cardiologists can achieve more widespread screening and accurate detection of by performing point-of-care ultrasound (POCUS) [2]. POCUS does not require the help of a specialist, but its assessment of AS involves only simple B-mode videos. These are used to assess the appearance and motion of the aortic valve (AV). In contrast, cardiologists also assess key blood flow parameters using Doppler imaging. Following clinical guidelines [3], three Doppler parameters are used to characterize AS severity: aortic valve area, peak valvular jet velocity, and mean pressure gradient. Doppler imaging is vital to a cardiologist's echo diagnosis, but POCUS users typically lack the necessary operator expertise [4], [5] to acquire high-quality images and produce accurate measurements. As such, machine learning solutions should follow standard usage of POCUS involving only the assessment of simple B-mode videos. In real-world practice, cardiologists interpret an ensemble of cine series by integrating B-mode videos with patient risk factors, clinical observations, and Doppler-based measurements to make reliable diagnoses. This information is summarized in the echocardiography report, which includes key Doppler metrics and clinical assessments of valve function. Integrating this information during training may help align the video embedding space with essential clinical details, leading to better representations. Thus, while only B-mode videos should be required by POCUS users at test time, other clinically-relevant data from the echo report can be leveraged during model training to learn a comprehensive diagnostic model.

Learning a robust multimodal model for AS diagnosis requires important consideration of issues related to multimodal alignment and the accuracy of diagnostic labels. Common approaches to multimodal alignment, often between text and image, use contrastive learning techniques. In particular, numerous works have built upon the well-known contrastive pre-

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

training approach from CLIP [6]. However, applying the CLIP-based approach to AS diagnosis presents two key challenges. Firstly, the effectiveness of CLIP alignment techniques is largely unknown for video and/or tabular data [7]–[10], which are the modalities of greatest relevance for AS detection and classification, as report data often contains echo measurement numbers in tabular form. Secondly, CLIP-based approaches assume perfect alignment between multimodal data, as in image captioning. However, these methods are less effective with 'noisy' multimodal data, where there is semantic mismatch within the paired data [11]. In echo, misalignment arises between video-report pairs because each cine corresponds only to specific parts of the report, depending on the imaging technique, view, and quality. Each report typically contains information derived from multiple cines. To handle noisy multimodal data, the CLIP-based approach can be adapted to align echo cines with only the relevant report features. This machine learning task is further complicated by another source of noise within AS diagnosis data that is commonly referred to as label noise. This is a case of label error that occurs mainly due to the presence of inter-observer variability in AS classification labels [12]. In deep learning (DL), models are prone to overfitting to noisy labels, ultimately leading to poor generalizability on test datasets [12]. A potential solution is to employ sample selection to learn from samples with lower losses, thus discarding high-loss cases [13]. Low loss samples tend to be less noisy and cleaner for training the model, preventing the model from overfitting on noisy labels [14]. To our knowledge, this strategy has yet to be explored for automated AS classification using deep learning.

Recent DL models for AS diagnosis [15]–[18] show that video-based approaches using B-mode echo cine series can capture key features related to aortic valve motion and changes in pixel intensity reflecting calcification [15], [17]. These methods align with the use of POCUS for AS screening, where only B-mode videos are available. While prior work has focused on video-based models, no existing methods have explored incorporating tabular patient information and Doppler-derived measurements alongside B-mode videos for AS diagnosis. Prior models ignore vital clinical information that could help to bridge the gap between video-based models and real-world clinical decision-making. We hypothesize that incorporating structured data from echo reports into video-based models can enhance the quality of learned embeddings. Previous DL studies for AS classification have also largely overlooked the problem of label noise. Some methods, such as ProtoASNet [17] and RT4U [19], incorporate uncertainty estimation, which can be associated with label noise. However, these approaches do not explicitly address label noise in training.

We introduce MultiASNet, a multimodal DL framework for AS classification that uses video-tabular alignment during training to improve the representation of unimodal video embeddings. A key feature of our design is that it performs inference without requiring structured reports as input. To our knowledge, this is the first method to leverage structured tabular report data during training with ultrasound video while remaining video-only at inference, enhancing robustness without adding deployment complexity. While prior work has integrated structured data with medical images [8], most focus on static modalities such as X-ray or MRI. Our approach instead addresses the unique temporal and dynamic challenges of ultrasound video. In addition, we use a distinctive multimodal fusion-alignment architecture that integrates cross-attention and contrastive learning mechanisms, and we explore sample selection to mitigate the impact of discordant labels in AS training.

Our contributions are as follows:

- We aim to enable safer, broader use of POCUS by non-cardiologists for accurate AS detection. In contrast to prior work, we incorporate clinically relevant information from echo reports to guide video-based feature learning, using a transformer-based tabular encoder [20] to extract rich, structured feature representations from report data.
- We introduce a multimodal integration approach that uses cross-attention to align structured report data features with video features. This method transforms report data into a video-relevant representation by focusing on relevant features in video-report pairs. By addressing noisy and misaligned data, our approach improves CLIP-based alignment between the two modalities.
- We employ sample selection to avoid overfitting to noisy labels and show that incorporating this method improves performance.
- We achieve state-of-the-art study-level balanced accuracy and average F1 scores, outperforming existing methods.

II. Related Work

A. Multimodal Learning

Many self-supervised methods bypass the need for multimodal data at test time [8], [9], [21]–[23], but few address the combination of tabular and medical imaging data. Among these methods, CLIP is a prominent self-supervised framework aligning image and text encoder outputs into a shared latent space, using contrastive learning for image-text pairs [6]. While optimized for natural images, CLIP's efficacy diminishes with specialized datasets, such as medical images, due to limited representation in pretraining data. Consequently, MedCLIP [24] and BiomedCLIP [25] adapt CLIP for medical contexts by leveraging medical image-text datasets, improving performance in tasks like zero-shot prediction and biomedical vision-language retrieval.

Extensions of CLIP, including VideoCLIP [26] and XCLIP [27], incorporate temporal information, making them suitable for video-text tasks. Hager et al. [8] modify CLIP for image-tabular data by integrating an MLP-based tabular encoder, though this simple encoder may limit network's abilities to detect complex patterns in the data. Christensen et al. [21] train CLIP further on over 200,000 echo image-report pairs, developing EchoCLIP to predict clinical measures like left ventricular ejection fraction (LVEF) and pulmonary arterial pressure (PAP). However, EchoCLIP and similar models assume perfectly aligned image-text pairs, which does not hold for many echo datasets. Subsequent work from EchoPrime [28] introduce a video-based model using contrastive learning to align

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

echo video-report pairs. Their approach incorporates attention-based learning to weigh relevant echo views, but relies on coarse-grained alignment with full anatomical descriptions done by expert clinicians. For instance, while the PLAX view may be indicated as aligned with the aortic valve, it only partially visualizes the structure, missing key details like valve area or leaflets. Fine-grained alignment with view-specific descriptions could enable more meaningful representations.

B. Tabular Data Models

Traditional models, such as Random Forest and XGBoost [29], have long been effective for tabular data due to their capacity to manage heterogeneous features, outliers, and the structured nature of such data [30]. However, recent advancements in DL models for tabular data, especially those leveraging attention mechanisms, show competitive or even superior performance in many cases [31]. For example, TabNet [32] uses sequential attention to improve interpretability, while TabTransformer [33] and Tabbie [34] adapt transformers to handle categorical and numerical features in tabular data. Yet, these approaches can be limited in their ability to capture nuanced interactions across both feature types in complex datasets.

To address these limitations, FTTransformer [31] represents each feature as a token, providing comprehensive embeddings for both categorical and numerical data. While recent developments include large foundation models like TabuLa-8b [35] that exhibit robust few-shot capabilities, they require extensive computational resources.

C. Sample Selection

On noisy datasets, DL models initially learn easy patterns commonly found in clean samples. However, as training progresses through epochs, they tend to overfit to noisy samples, leading to error propagation. Some sample selection methods mitigate this by focusing on samples with lower losses, effectively learning from cleaner data. Jiang et al. [13] introduce a novel technique to learn to select and weight samples for a student model dynamically, focusing on cleaner samples in early training stages and incorporating more difficult samples later, ensuring the student model learns robustly in the presence of noisy or mislabeled data. Sample selection techniques are particularly useful in real-world and medical settings, as data can be noisy and are often subject to observer variability. Xue et al. [36] apply sample selection and noise-robust techniques to a variety of different medical datasets, showing improvement when dealing with noisy medical datasets.

In this work, we introduce sample selection for AS detection and classification to address label noise arising from inter-observer variability.

D. Aortic Stenosis Detection and Severity Classification

POCUS-assisted screening of AS can be augmented with DL to ensure affected patients are referred for echocardiographic evaluation in a timely manner. Significant progress has been made in advancing DL techniques for AS diagnosis. Several B-mode echo image-based and video-based models have been proposed for AS detection and severity estimation.

Huang et al. [37] employ a semi-supervised multi-task learning framework to simultaneously classify views and assess AS severity in echocardiographic images. The approach involves training two distinct encoders for view classification and AS diagnosis. To derive a study-level prediction, a weighted average of the image-level predictions is calculated, with greater emphasis placed on the PLAX and PSAX views. Further evaluations of the model's performance are conducted by Wessler et al. [38]. Subsequently, Huang et al. [39] introduce an updated dataset, TMED-2, to further advance research in this domain. Additionally, Huang et al. [40] propose a method using multi-instance learning to make predictions based on multiple images within a study. They design an attention mechanism based on the echo view to evaluate the significance of each image.

In addition to the image-based models presented above, video-based models have shown increased accuracy in AS classification. Ginsberg et al. [41] use an R(2+1)D based model trained on PLAX and PSAX views. They incorporate uncertainty-awareness and multi-tasking in the model, achieving comparable classification performance while being able to detect out-of-distribution samples. While Ginsberg et al. assume that all frames within an echo video should be treated with the same importance, Ahmadi et al. [15] further improves the model by adding Temporal Deformable Attention (TDA), which learns which frames within the echo video are more informative. Temporal coherence loss is used to force consecutive frames to have similar embeddings, leading to a higher classification accuracy. Vaseli et al. [17] also build their model based on R(2+1)D, but incorporated prototypical neural networks to provide explainable classifications based on learned class prototypes. A subsequent study by Huang et al. [18] incorporates spectral Doppler images alongside B-mode ultrasound videos, achieving a 10% accuracy improvement compared to B-mode only approaches. However, this method requires high-quality Doppler image acquisition, making this approach impractical for POCUS applications.

III. Method

MultiASNet is a multimodal deep learning framework for AS assessment, combining ultrasound video and tabular clinical data. It extracts video and tabular embeddings, aligns them via cross-attention, and projects them into a shared space for video-only inference. AS predictions are obtained by computing a weighted average of the frame-level outputs, resulting in a single classification for each video. Training is guided by multiple loss functions and a sample selection strategy to reduce the effect of label noise. The following sections detail each component, as shown in Figure 1.

A. Video Embedding Extraction

We employ a transformer-based spatio-temporal architecture, as proposed by Ahmadi et al. [15] for our video encoder. We selected this method as it is a state of the art AS classification approach that uses attention to provide

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

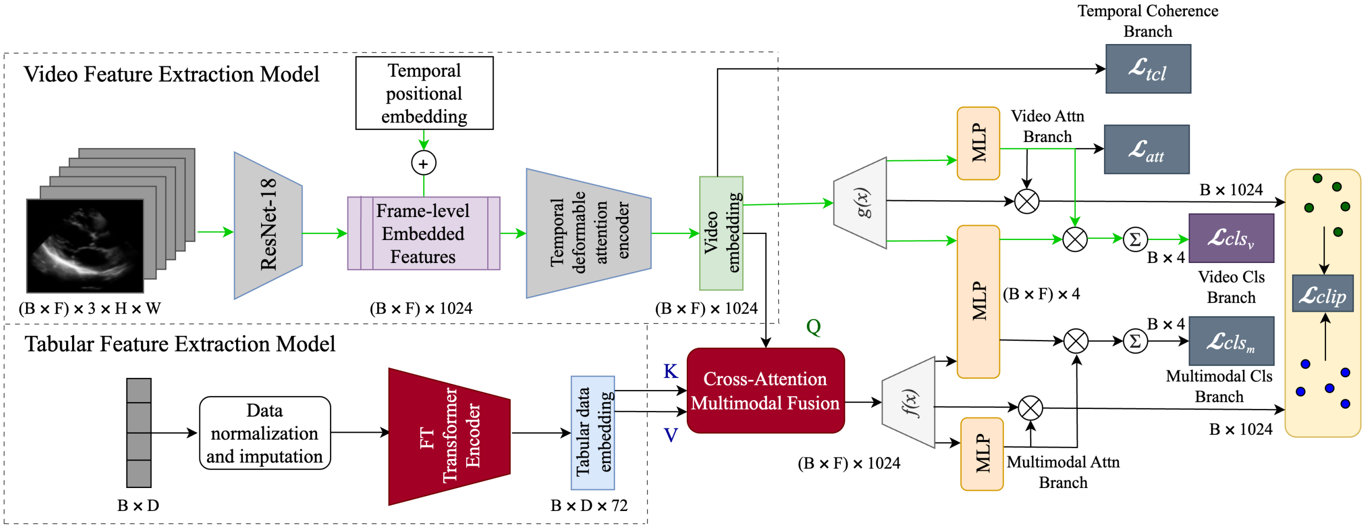

Fig. 1. An overview of our proposed architecture. Our model uses an FT Transformer Encoder [33] for the tabular data, combines the tabular and video embeddings through cross-attention, and uses a shared decoder for classification. At test time, we skip the cross-attention and only use the video pathway (highlighted in green). Our distinct contributions over current CLIP-based methods [6], [8], [42] are indicated in red.

temporally coherent embeddings for each video. The input to the encoder is a single echo video from the PLAX or PSAX view. Each image frame of the video input is passed through an image encoder to obtain per-frame feature embeddings. Then, a Temporal Deformable Attention encoder [43] is used to extract relevant temporal relationships between the image frame embeddings. Additional details about the video encoder can be found in [15].

B. Tabular Embedding Extraction

We gather 35 clinically relevant tabular features from each echo study, as identified by expert cardiologists. These include demographic data, physiological metrics, and echo-derived parameters, which are processed using a transformer-based tabular encoder. The encoder tokenizes these input features into d-dimension vectors using a piece-wise linear encoding [31]. Then, the features are passed through transformer blocks. We first pretrain the tabular encoder on the AS classification task, then freeze its weights and use it to extract tabular embeddings for use in the subsequent multimodal model.

C. Tabular and Video Embedding Integration

Given a video data embedding z_v and a tabular data embedding z_t, we use cross-attention to combine the two into a single embedding z_m that captures video-relevant tabular features. The cross-attention equation is shown in Eq. (1). Before applying cross-attention, both the video and tabular embeddings are first projected to the same dimension using a linear layer to ensure the dimesions are properly aligned.The query vector Q_v, is obtained from the video data, while the key and value vectors K_t and V_t, respectively, are from the tabular data. This setup allows the query vector Q_v, to emphasize the tabular features most relevant to the video. By attending to the key and value vectors K_t and V_t from the tabular data, the model focuses on the tabular information that complements the video content, enhancing alignment and relevance between video and tabular data for AS classification.

$$\text{CrossAttention}(Q_v, K_t, V_t) = \text{softmax}\left(\frac{Q_v K_t^T}{\sqrt{d_k}}\right) V_t \quad (1)$$

D. Embedding Space Mapping And Decoding

Following the cross-attention fusion of the tabular and video data, we use a projection head to project both z_m and z_v into the same embedding space, giving E_m and E_v. Because we only operate on video data at test time, the shared embedding space allows for the model to leverage the information learned from the video-relevant tabular data. We project the embeddings into the shared space as follows:

$$E_m = f(z_m), \qquad E_v = g(z_v), \quad (2)$$

where f and g are two separate projection heads learned during the end-to-end training process. E_m and E_v are both passed into a combined downstream classification head, which outputs one classification prediction per frame, indicated in Eq. (3) as $y_{m,i}$ and $y_{v,i}$, respectively, where i denotes the frame number. The final AS predictions, y_m and y_v, are calculated as a weighted sum of the per-frame output predictions, where weights $w_{m,i}$ and $w_{v,i}$ correspond to frame-level attention. These attention weights are obtained by applying a Multi-Layer Perceptron (MLP) to the video output embeddings of the temporal encoder module. This approach assigns greater weight to frames that are more relevant to the classification task, modulating the contributions of individual frames, $y_{m,i}$ and $y_{v,i}$, to amplify their influence on the final prediction. The equation for calculating the final predictions is given by Equation 4:

$$y_{m,i} = h(E_{m,i}), \qquad y_{v,i} = h(E_{v,i}). \quad (3)$$

$$y_m = \sum_i w_{m,i} \cdot y_{m,i}, \qquad y_v = \sum_i w_{v,i} \cdot y_{v,i}. \quad (4)$$

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

E. Loss Function

Our training loss consists of a weighted combination of cross-entropy loss applied to the modality-specific predictions y_m and y_v against the ground-truth labels, attention-based entropy loss, and temporal coherence loss [15]. The attention-based entropy loss promotes sparsity in the frame-level attention weights, highlighting the importance of individual frames, as shown in Equation 5. Specifically, the raw attention weights, $w \in \mathbb{R}^F$ where F is the number of frames in a video clip, are first normalized using a softmax function σ_F across the temporal dimension, producing $\hat{w}$. The entropy loss is then computed as

$$\hat{w} = \sigma_F(w), \quad \mathcal{L}_{attn} = -\sum_{\tau=1}^{F} \hat{w}_\tau log(\hat{w}_\tau). \tag{5}$$

Minimizing this loss reduces entropy in the attention distribution, ensuring that fewer frames receive higher weights, thereby improving interpretability and reducing noise from less relevant frames.

The temporal coherence loss encourages similar embeddings for adjacent frames and more distinct embeddings for frames with larger spatial differences, and is shown in Equation 6. It is computed using the inner product of adjacent frame embeddings s_τ and distant frame embeddings $d_{\tau,w}$, where a frame pair is considered distant if separated by more than $T = 3$ frames. By minimizing this loss, the model learns to preserve temporal structure while avoiding redundant representations:

$$\begin{aligned} L_{tcl} &= \frac{1}{F}(\sum_{\tau=1}^{F} -log(\frac{e^{s_\tau}}{e^{s_\tau} + \sum_w e^{d_{\tau,w}}})), \\ s_\tau &= \begin{cases} v_{f_1}^T v_{f_2} & \text{if } \tau = 1 \\ v_{f_{F-1}}^T v_{f_F} & \text{if } \tau = F \\ \frac{1}{2} v_{f_{\tau-1}}^T v_{f_\tau} + \frac{1}{2} v_{f_\tau}^T v_{f_{\tau+1}} & \text{otherwise} \end{cases} \\ d_{\tau,w} &= v_{f_\tau}^T v_{f_w} \quad if \ |\tau - w| > T. \end{aligned} \tag{6}$$

Additional details on these two loss components can be found in [15].

In addition to these primary loss terms, we incorporate a sample selection mechanism to improve the reliability of the cross-entropy loss optimization. We apply sample selection by selectively updating the model weights based on the samples with the lowest cross-entropy loss. Specifically, during each training iteration, we exclude a small percentage of samples with the highest cross-entropy loss values from backpropagation, effectively preventing the model from learning from potentially noisy or mislabelled samples. This approach allows the model to focus on the remaining percentage of samples with the smallest loss values, enhancing generalization by reducing the risk of overfitting to noisy data. By training on cleaner samples, the model learns more robust patterns, improving its overall performance on the classification task. We calculate the video-based cross-entropy loss, $\mathcal{L}_{cls_v}$, based on the samples selected from the tabular branch, and the tabular cross-entropy loss, $\mathcal{L}_{cls_m}$, based on the samples selected from the video branch. Additionally, since frames within an echo video are highly similar, attention-based entropy loss, $\mathcal{L}_{att}$, is used to encourage variation in the important places across the image frames. Temporal coherence loss, $\mathcal{L}_{tcl}$, is used to help the model learn that nearby frames, which possess similar spatial features, should have similar embeddings, and vice versa for distant frames. In combination with these three loss terms, we also add the CLIP loss, $\mathcal{L}_{clip}$, between the weighted average embeddings of the video and video-relevant tabular features, CLIP($\sum_i w_{v,i} \cdot E_{v,i}$, $\sum_i w_{m,i} \cdot E_{m,i}$), aligning the embedding spaces E_m and E_v. The total loss is calculated as follows:

$$\mathcal{L} = \lambda_{clip}\mathcal{L}_{clip} + \mathcal{L}_{cls_v} + \mathcal{L}_{cls_m} + \lambda_{att}\mathcal{L}_{att} + \lambda_{tcl}\mathcal{L}_{tcl}. \tag{7}$$

IV. Experiments

A. Datasets

We evaluated our model on private and public 2D echo datasets. Our private dataset consisted of echo studies from a tertiary care hospital, collected with appropriate permissions. Using a view detection algorithm [44] and cardiologist verification, we extracted only PLAX and PSAX views without colour or spectral Doppler. Similar to previous studies [15], [17], [41], we also removed cases where all three Doppler markers indicated different severity grades, with no two markers agreeing. We removed all image annotations on the raw cines and trimmed each cine to one cardiac cycle. For data augmentation, we used random horizontal flipping, cropping, and rotation with respect to the origin of the ultrasound beam. AS severity labels were assigned based on clinical guideline recommendations [3], where assignment is based on aortic valve area, peak valvular jet velocity, and mean pressure gradient values. The final dataset included 2,627 reports and 9,297 videos. We had three AS label classes: no AS, early AS (mild), and significant AS (moderate, severe). Data were split using an 80%-10%-10% ratio for training, validation, and testing, with unique patients in each subset. We also used the TMED-2 public dataset from Tufts Medical Center [37], containing a fully-labelled set of 17,270 echo images. AS labels were created by an expert clinician following the same guidelines [3].

B. Training Strategies

Although CLIP-based models are typically recognized for their effectiveness on unseen data, zero-shot performance from existing pretrained models did not generalize well to the AS severity classification task, as shown in Table VIII. For instance, VideoCLIP [26], a large foundation model pretrained on diverse datasets, failed to translate its zero-shot capabilities to the task of aortic stenosis (AS) severity classification. Additionally, finetuning only the classification head of the pretrained VideoCLIP model was insufficient to achieve competitive performance. When pretrained and finetuned on the same private AS dataset, VideoCLIP achieved a balanced study-level accuracy of 76.5%. Similarly, all other CLIP-based baseline models exhibited limited zero-shot generalization and performed suboptimally when only the classification head was finetuned.

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

To overcome these challenges, our model is trained end-to-end, incorporating the CLIP loss as one component of the overall loss function. This approach was motivated by the fine-grained nature of our dataset, which requires the model to capture subtle variations in disease severity and anatomical features. Furthermore, the lack of large public datasets suitable for pretraining in this domain highlighted the need for a customized solution. In addition, we experimented with pretraining using CLIP loss followed by finetuning with additional loss functions on the same dataset. However, this method did not improve performance compared to training the model end-to-end while integrating the CLIP loss directly as part of the overall loss function.

C. Implementation Details

For the image encoder, we used ResNet-18 to obtain per-frame feature embeddings [15]. For the tabular data encoder, we first trained an FT-Transformer [33] model on our private dataset for AS classification over 10 epochs. After training, we froze the model weights and used it to extract tabular embeddings. This transformer model was selected because it generates informative embeddings of the tabular data while remaining lightweight enough to train end-to-end. In contrast, foundation models such as [45], though powerful, are typically much larger and more computationally intensive, making them less practical for application to hand-held POCUS devices. The FT-Transformer strikes a balance by providing high-quality embeddings without the overhead of larger models. As shown in Table I, it achieved the highest accuracy among the tabular models tested, while maintaining a relatively small parameter count. This makes it well-suited for POCUS settings, where both efficiency and reliability are critical.

We used a loss temperature of 0.1 and a class lambda of 0.5 for the embedding space CLIP loss. Based on the hyperparameters of Ahmadi et al. [15], we selected λ_{attn} as 0.05 and λ_{tcl} as 0.01. We empirically selected λ_{clip} as 0.5, based on a sweep over values in the range of 0.25 to 1. For sample selection, we do not propagate the loss from the 5% of cases with the highest classification cross-entropy loss values; the model weights were updated using 95% of the samples with the smallest loss values, based on empirical testing. Our model was trained for 100 epochs on one NVIDIA Tesla V100 16GB GPU, using a cosine annealing learning rate between 10^{-4} and 10^{-6} [46].

For evaluation on the public TMED-2 dataset, we first modified all video-based models, including our proposed MultiASNet, to handle image-only inputs, as TMED-2 lacks video sequences and structured tabular report data. Specifically, we removed the temporal attention module and attention weights from MultiASNet and replaced the video encoder with a WideResNet [37]. Each image embedding was then paired with a corresponding tabular embedding via cross-attention. During training, we optimized the model using classification cross-entropy and CLIP loss, omitting the attention-based entropy and temporal coherence losses.

TABLE I
AS CLASSIFICATION ACCURACY ON TEST SET USING TABULAR FEATURES FOR DIFFERENT TABULAR TRANSFORMER MODELS

AS Classification	Num params	Accuracy
TabNet [32]	$8K$	96.28
TabPFN [47]	$7.2M$	98.89
FT-Transformer [31]	$\mathbf{3.8M}$	**99.61**

We conducted two experiments to assess generalization and adaptability. In the first, a generalization experiment, we trained all models on our private dataset and tested them directly on the public TMED-2 dataset using only PLAX and PSAX images, since tabular echo reports are not available in TMED-2. No retraining was performed. All other models shown in Table IV were also trained on our private dataset and then tested on the PLAX and PSAX images of the public dataset. In the second experiment, we adopted a ”label as a feature” approach [8] to compensate for the lack of structured echo report data. Here, the ground-truth label was used during training as a pseudo-tabular feature, transformed into an informative embedding and aligned with the image features using cross-attention and CLIP loss. At test time, only the TMED-2 image data was used.

D. Subgroup Data Imbalance

In echocardiography, there are various patient subpopulations with different presentations of AS. We found in our evaluations that MultiASNet performed worse on cases belonging to rarer subpopulations, such as patients with bicuspid valves, compared to the more common tricuspid valve cases. To account for subgroup data imbalance, we examined the effectivness of a SOTA method for handling subgroup imbalance, known as Deep Feature Reweighting (DFR) [48], [49]. This approach [48] assumes that the feature extractor learns sufficient core features and claims that the problem of imbalanced training lies in the classifier placing high weight on spurious features learned, which ultimately leads to the poor performance on minority subgroups. Based on this concept, DFR reduces overfitting to spurious correlations by retraining the classifier layers of the model with an independent and balanced dataset. In our experiments, we identified biases against rare patient subpopulations and examined the effectiveness of DFR in mitigating these biases in order to make the model more fair across subpopulations. We demonstrated that the model can identify bicuspid valve-relevant features, but its original predictions underrepresent this class due to the imbalance during training. Specifically, we constructed the DFR dataset by selecting a balanced subset of the training dataset, with an equal number of bicuspid and tricuspid valve cases. Since the original dataset contained only a few bicuspid valve samples, we included all available training samples from this subgroup. We then retrained only the final classification layer of MultiASNet on this balanced dataset, keeping the feature extractor frozen. The results of applying DFR are discussed in Section V-A.2.

V. Results

A. Evaluations on Private Dataset

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

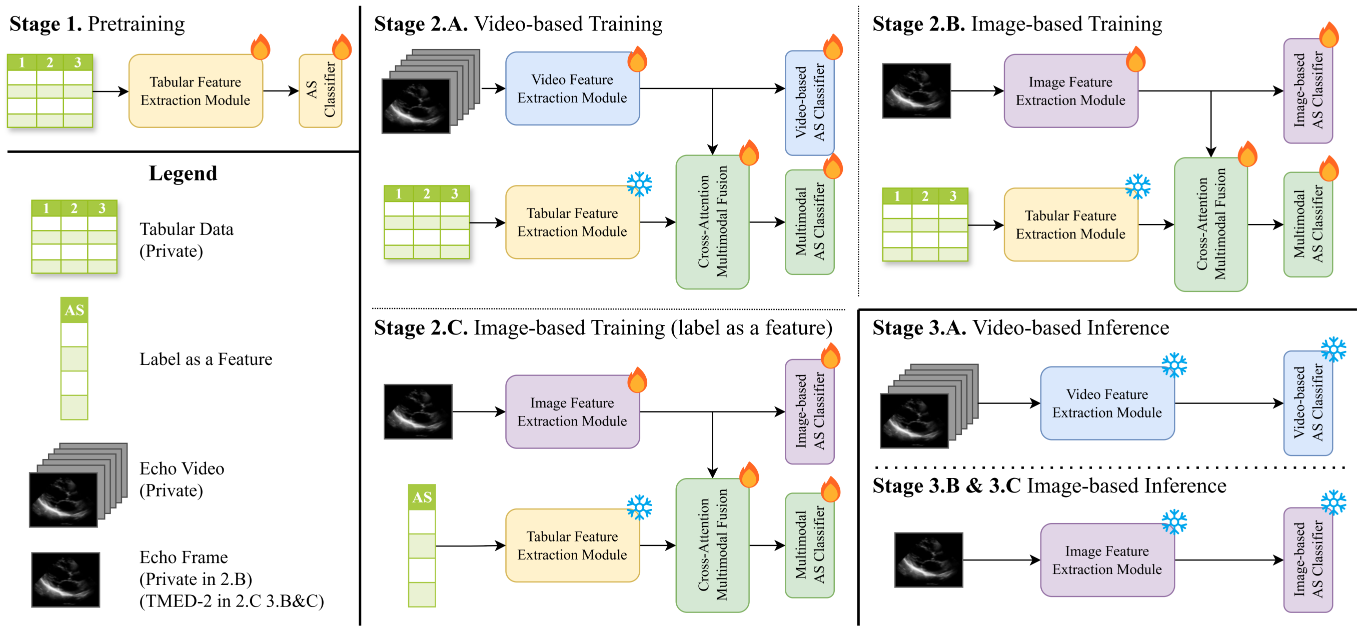
Stage 1. Pretraining
1 2 3
Tabular Feature Extraction Module
AS Classifier
Legend
Tabular Data (Private)
AS
Label as a Feature
Echo Video (Private)
Echo Frame (Private in 2.B) (TMED-2 in 2.C 3.B&C)
Stage 2.A. Video-based Training
Video Feature Extraction Module
Video-based AS Classifier
Tabular Feature Extraction Module
Cross-Attention Multimodal Fusion
Multimodal AS Classifier
Stage 2.B. Image-based Training
Image Feature Extraction Module
Image-based AS Classifier
Tabular Feature Extraction Module
Cross-Attention Multimodal Fusion
Multimodal AS Classifier
Stage 2.C. Image-based Training (label as a feature)
Image Feature Extraction Module
Image-based AS Classifier
Tabular Feature Extraction Module
Cross-Attention Multimodal Fusion
Multimodal AS Classifier
Stage 3.A. Video-based Inference
Video Feature Extraction Module
Video-based AS Classifier
Stage 3.B & 3.C Image-based Inference
Image Feature Extraction Module
Image-based AS Classifier

Fig. 2. Overview of the training and inference pipeline for MultiASNet. The process involves three stages: (1) pretraining the Tabular Feature Extraction Module on private data, (2) multimodal training using paired echo data and tabular features—via video-based (Stage 2.A), image-based (Stage 2.B), or image-based with "Label as a Feature" (Stage 2.C) pathways—and (3) inference using echo data alone. The video model is trained and evaluated on private data. The image model is either (B) trained on private data and evaluated on the public TMED-2 dataset (generalization experiment), or (C) trained and evaluated on public data with the label included as a tabular feature. Cross-attention-based fusion integrates visual and tabular modalities in all branches.

1) Quantitative Assessment: Accurate detection of AS is a critical first step in clinical workflows, particularly in POCUS, where identifying the presence of AS determines whether a patient should be referred for Doppler imaging. MultiASNet bridges a critical gap in non-specialist settings by reliably detecting AS from echo video data. Specifically, MultiASNet achieved a balanced accuracy of 93.0% for binary AS detection.

In Table II, we compare the performance of MultiASNet against baseline models for AS severity classification. Compared to all unimodal and multimodal baselines, our model achieves the best study-level balanced accuracy and F1-score. The multimodal baselines [8], [22], [24]–[27], [42] were pretrained on private training data to align tabular-image pairs with CLIP loss, after which the image encoders were fine-tuned to perform AS classification using the same image dataset. For models that used text-based encoders instead of tabular encoders, the tabular data were first converted into a text format to serve as input. The multimodal image-based baselines performed worse than the unimodal video-based models but better than the unimodal image-based baseline. We suspect that our addition of sample selection and multimodal approaches to the model by Ahmadi et al. [15] has the greatest impact at the study level because the noisy labels and tabular data are both captured at this level. The AS labels are derived from measurements in the echo reports, which are recorded at the study level, where each report corresponds to multiple cines and a single diagnosis is made based on the aggregated data.

2) Qualitative Assessment: The model was trained on a dataset where 90% of the cines represent patients with a tricuspid aortic valve (TAV). Among failure cases for severity classification, we observe that our model performs disproportionately worse on cines with bicuspid aortic valve (BAV), which account for 17% and 38% of all mild and moderate misclassifications in the test set, respectively. After excluding all BAV cases from the test set, all metrics improved on a cine and study level VII. These results suggest that the model overfits to TAV cases.

To account for subgroup data imbalance, we implemented DFR [48], [49], as specified in Section IV-D. We retrained the classification head of MultiASNet with a new DFR dataset, consisting of a balanced subset of the validation dataset. Results in Table IX show that DFR substantially improved bicuspid valve subgroup performance on a study level. In Table III, we further observed that MultiASNet tended to be overly confident in its incorrect predictions. Applying DFR reduced the model's confidence in these incorrect predictions, while preserving its confidence in correct ones. As a result of using DFR, the model developed higher confidence in correct predictions than in incorrect ones.

B. Evaluations on TMED-2 Public Dataset

In evaluating generalization performance for AS detection and severity classification on the TMED-2 dataset, we chose baselines that align with our approach and dataset compatibility. Specifically, we compare against [37], which is the foundational work on the Tuft's dataset, providing a directly comparable framework for our cross-dataset setting. We exclude newer contributions such as [40] and [18], as these employ multi-instance learning for study-level accuracy, an approach that does not translate well across datasets

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

TABLE II

AS SEVERITY CLASSIFICATION PERFORMANCE OF UNIMODAL AND MULTIMODAL MODELS ON CINE- AND STUDY-LEVEL FOR A PRIVATE DATASET USING MEAN BALANCED ACCURACY AND F1 SCORE METRICS AVERAGED ACROSS THREE RUNS, REPORTED AS "MEAN (STANDARD DEVIATION)".

Training Modalities	Method	Num Trainable Parameters	Cine-level		Study-level	
			bACC↑	F1 ↑	bACC↑	F1 ↑
Image	Huang et al. [37]	23M	62.5(0.8)	0.62(0.01)	71.7(2.1)	0.71(0.03)
Image	Huang et al. [40]	2.3M	–	–	71.6(3.2)	0.72(0.04)
Video	Ginsberg et al. [41]	30M	75.3(0.7)	0.74(0.01)	77.9(1.6)	0.77(0.02)
Video	Vaseli et al. [17]	8.1M	75.2(1.0)	**0.75(0.01)**	78.9(1.4)	0.79(0.01)
Video	Ahmadi et al. [15]	21.3M	73.5(0.7)	0.74(0.01)	76.9(1.7)	0.77(0.02)
Image+Tabular	Hager et al. [8]	25.6M	65.8(1.3)	0.66(0.01)	71.7(2.6)	0.72(0.03)
Image+Tabular	Ebrahimi et al. [22]	44.9M	57.8(6.2)	0.58(0.06)	63.0(1.1)	0.63(0.01)
Image+Text	Christensen et al. [42]	36.7M	70.9(0.6)	0.71(0.00)	74.7(0.2)	0.74(0.00)
Image+Text	Wang et al. [24]	136M	69.4(1.2)	0.69(0.01)	75.7(1.0)	0.76(0.01)
Image+Text	Zhang et al. [25]	195M	63.7(1.1)	0.63(0.01)	69.7(1.4)	0.69(0.02)
Video+Text	Ni et al. [27]	197M	66.7(3.9)	0.67(0.04)	72.8(3.6)	0.73(0.04)
Video+Text	Xu et al. [26]	208M	70.5(2.4)	0.70(0.03)	76.5(1.9)	0.76(0.02)
Video+Tabular	Ours	28M	**75.5(0.3)**	**0.75(0.01)**	**80.4(0.5)**	**0.80(0.01)**

TABLE III

THE MEAN PREDICTION CONFIDENCE WITH VERSUS WITHOUT DFR ON INCORRECT AND CORRECT SEVERITY CLASSIFICATIONS OF BICUSPID VALVE CASES (AVERAGED ACROSS 3 RUNS).

Model	Mean prediction confidence	
	Incorrect predictions	Correct predictions
MultiASNet	0.86(0.03)	0.76(0.08)
MultiASNet + DFR (MLP)	0.52(0.03)	0.66(0.03)

TABLE IV

AS SEVERITY CLASSIFICATION PERFORMANCE ON PATIENT-LEVEL FOR UNSEEN TMED-2 DATASET USING MEAN BALANCED ACCURACY AND F1 SCORE METRICS AVERAGED ACROSS THREE RUNS, REPORTED AS "MEAN (STANDARD DEVIATION)".

Method	Patient-level (N=3)	
	bACC↑	F1 ↑
Huang et al. [37]	55.5(8.7)	0.56(0.10)
Ahmadi et al. [15]	47.7(1.5)	0.45(0.04)
Christensen et al. [42]	54.5(3.1)	0.56(0.03)
Ours	**59.4(0.1)**	**0.59(0.02)**

TABLE V

PATIENT-LEVEL BALANCED ACCURACY FOR THE "LABEL AS A FEATURE" EXPERIMENT, WHERE BOTH TRAINING AND TESTING OCCUR ON THE TMED-2 DATASET.

Method	Study-level bACC↑
Huang et al. [37]	74.6
Ahmadi et al. [15]	73.0
Ahmadi et al.* [15]	**79.7**
Ours	**78.9**

*Indicates performance after discarding approximately 5% of the patients based on view relevance.

without fine-tuning. Our evaluation also includes [15], the backbone model for our architecture, allowing us to demonstrate that our modifications—adding tabular embeddings and implementing cross-attention during training—offer transferable performance gains. Our model, modified as described in Sec.IV-C , achieves 83.9% and 59.4% balanced accuracy for AS detection and severity classification, outperforming all baselines, including the best-performing multimodal baseline[21] (see Table IV). The slightly lower performance for the public dataset on severity classification may be due to differences in data distribution: the private dataset includes over 40% normal cases, whereas the public dataset contains far fewer normal cases and a much higher proportion of Significant AS cases, with severe cases accounting for well over 60%. This imbalance could contribute to the performance gap. Additionally, the private dataset consists of echo video data, which is better suited for AS classification, whereas the public dataset is image-based. Despite this, our approach effectively leverages multimodal information for cross-dataset generalization without fine-tuning.

For the "label as a feature" experiments, both training and testing were conducted on the TMED-2 dataset. We compare our model, MultiASNet, against prior approaches Huang et al. [37] and Ahmadi et al. [15]. As shown in Table V, MultiASNet achieves the highest patient-level balanced accuracy when evaluated across all patients in the test set. Notably, the version of the method by Ahmadi et al. that incorporates view relevance filtering achieves slightly higher performance (79.7%); however, this approach discards certain patients and views deemed insufficiently informative, limiting its generalizability. Without view relevance, their performance drops to 73.0%, underscoring the robustness of MultiASNet, which achieves 78.9% balanced accuracy without discarding any data.

VI. Ablation Study

Table VI shows that incorporating all proposed components yields the highest balanced accuracy and average F1 score for severity classification. Our model without any tabular data is the model proposed by Ahmadi et al. [15]. As shown in Figure 3, the left panel illustrates this baseline, while the right panel includes all our proposed components. Our model shows a more clearly defined and ordinal latent space, where class separation and disease progression are more evident. This

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

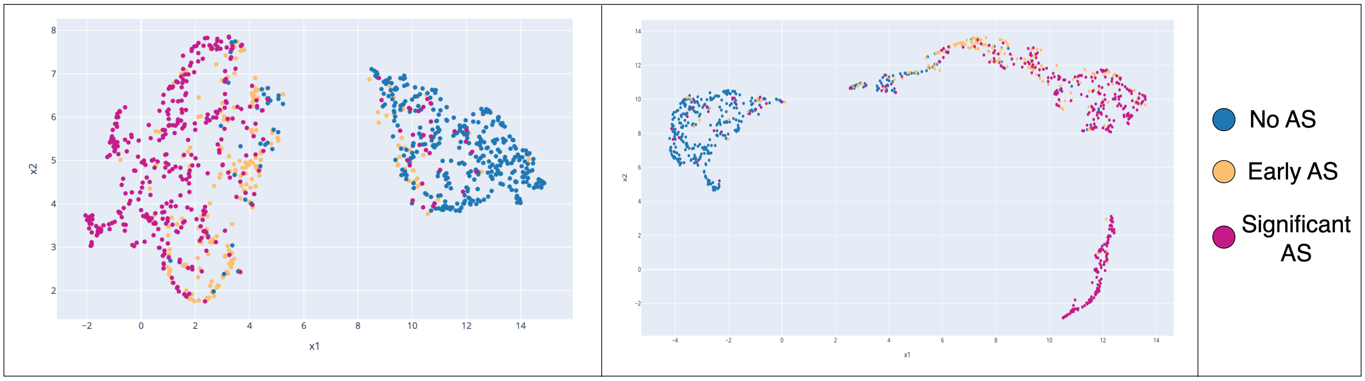

Fig. 3. Comparison of embedding spaces using Umap [50] for Ahmadi et al. [15] **(Left)** and MultiASNet **(Right)**. We show that our contributions lead to a more defined and ordinal embedding space.

suggests that the inclusion of additional tabular information and reduction of label noise contribute to a more informative and structured embedding space.

Adding tabular data to the model leads to notable improvements, and further gains are observed when these data are encoded using an FT-Transformer [33] to produce tabular embeddings, resulting in the highest study-level performance. Furthermore, we experiment with removing the cross-attention module, and instead apply the CLIP loss [6] directly between the original tabular embeddings and video embeddings, without any refinement of the tabular features based on the video content. In this setup, the tabular embeddings include information from all videos in an echo study, rather than being tailored to the specific video, which may introduce irrelevant features. Our results show that contrastive alignment alone is less effective than using cross-attention, which extracts video-relevant tabular features. This further highlights the advantage of the cross-attention module in addressing the challenge that each echo video corresponds to only a subset of information from study-level reports. Lastly, we show that using sample selection to avoid overfitting to samples with high label noise is effective for improving model performance. While label noise should ideally be consistent across both branches, the tabular branch, which is grounded in Doppler-derived markers that closely align with the final diagnosis, provides a more reliable signal for identifying mislabeled or ambiguous samples. By filtering out high-loss instances, our approach improves model robustness and reduces overfitting to spurious correlations. When we analyze the samples with high label noise, we find that they often contain discordant Doppler markers—instances where one out of three key measurements do not agree with each other—or borderline cases that lie near classification thresholds. For example, we identified a high-loss sample with an aortic valve area of 1.51 cm^2, which falls just above the normal threshold of 1.50 cm^2. In this case, the model struggles to classify it correctly, contributing to its high loss. Removing such samples prevents the model from overfitting to inconsistent or borderline data points, leading to greater overall performance improvements.

VII. Discussion

A. Performance of Multimodal versus Unimodal Models

Through a multimodal approach, we demonstrate that video analysis of AS can be improved by leveraging echo reports during training. Although MultiASNet outperforms all baselines and achieves SOTA results, this does not suggest that all multimodal architectures are better than unimodal ones. On our private dataset, every multimodal baseline performed worse than each of the video-based models [15], [17], [41]. This outcome may be attributed to a few factors. Firstly, echo data consists of misaligned video-report pairs. Since the multimodal baselines lack a mechanism to highlight features of the report that complement the paired videos, misalignment between modalities can introduce spurious features, ultimately degrading the quality of the learned video representations.

Additionally, multimodal models inherently have more parameters than unimodal ones, which can enhance performance but also require larger datasets and make training more challenging. Increased complexity may lead to overfitting on limited AS data or unstable optimization, reducing performance. Notably, MultiASNet has one of the fewest parameters among multimodal models and is comparable in size to some unimodal models (Table II), which may help mitigate these challenges and contribute to its strong results.

B. Performance on Minority Subgroups

Although bicuspid valve status was included as a binary input feature, MultiASNet showed poorer severity classification performance on cases involving bicuspid valves. This is likely partly caused by overfitting to tricuspid valve cases. This challenge was mitigated by DFR, which increased the model's study-level balanced accuracy on bicuspid valve cases by 7.9%. The effect was much larger on a study-level than cine-level due to the impact of DFR on model confidence. Since study-level metrics were derived from the mean prediction probabilities across all videos within a study, these predictions were sensitive to overly confident false predictions from any single video. Our analysis revealed that MultiASNet exhibited greater confidence in incorrect predictions than correct ones, a trend reversed following re-training with DFR. Consequently, with DFR, false video predictions were less likely to shift the study-level predictions toward an incorrect class. In contrast,

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

TABLE VI

ABLATION STUDY ON CINE- AND STUDY-LEVEL SEVERITY CLASSIFICATION FOR A PRIVATE DATASET USING MEAN BALANCED ACCURACY AND F1 SCORE METRICS AVERAGED ACROSS THREE RUNS, REPORTED AS "MEAN (STANDARD DEVIATION)".

Method					Cine-level		Study-level	
Tab data	Tab encoder	CLIP	Cross-attn	Sam Selec	Bacc	F1	Bacc	F1
					73.5(0.7)	0.74(0.01)	76.9(1.7)	0.77(0.02)
✓	MLP	✓	✓	✓	73.1(2.1)	0.73(0.02)	74.8(1.5)	0.75(0.02)
✓	FT-trans	✓		✓	73.0(1.2)	0.73(0.01)	76.9(0.9)	0.77(0.01)
✓	FT-trans	✓	✓		72.6(1.1)	0.73(0.01)	77.3(2.1)	0.77(0.02)
✓	FT-trans	✓	✓	✓	**75.5(0.3)**	**0.75(0.01)**	**80.4(0.5)**	**0.80(0.01)**

TABLE VII

ABLATION STUDY ON CINE- AND STUDY-LEVEL SEVERITY CLASSIFICATION FOR A PRIVATE DATASET USING MEAN BALANCED ACCURACY AND F1 SCORE METRICS AVERAGED ACROSS THREE RUNS WITHOUT BICUSPID VALVE CASES, REPORTED AS "MEAN (STANDARD DEVIATION)".

Method					Cine-level		Study-level	
Tab data	Tab encoder	CLIP	Cross-attn	Sam Selec	Bacc	F1	Bacc	F1
					77.4(0.2)	0.77(0.01)	80.1(1.3)	0.81(0.01)
✓	MLP	✓	✓	✓	76.9(2.3)	0.77(0.02)	78.3(1.3)	0.79(0.02)
✓	FT-trans	✓		✓	76.2(2.0)	0.76(0.02)	80.3(1.3)	0.80(0.02)
✓	FT-trans	✓	✓		78.3(2.5)	0.78(0.02)	81.6(3.4)	0.81(0.03)
✓	FT-trans	✓	✓	✓	**79.2(0.2)**	**0.79(0.01)**	**84.1(0.3)**	**0.84(0.00)**

TABLE VIII

COMPARISON OF BALANCED ACCURACY AND F1 SCORE FOR SEVERITY CLASSIFICATION ON CINE- AND STUDY-LEVELS FOR ZERO-SHOT, PRETRAINED AND FINETUNED MODELS [26] ON THE PRIVATE DATASET

Model	Cine-level		Study-level	
	bACC	F1	bACC	F1
Xu et al. [26] (Zero-shot)	30.2	0.29	33.0	0.19
Xu et al. [26] (Finetune)	43.4	0.38	45.7	0.39
Xu et al. [26] (Pretrain + Finetune)	70.5	0.70	76.5	0.76
Ours	**75.5**	**0.75**	**80.4**	**0.80**

TABLE IX

THE EFFECTIVENESS OF DEEP FEATURE REWEIGHTING ON SEVERITY CLASSIFICATION FOR IMPROVING THE CINE- AND STUDY-LEVEL BALANCED ACCURACY AND F1 SCORE OF MULTIASNET WITH MINORITY (BICUSPID VALVE) SUBGROUPS

Model	Cine-level		Study-level	
	bACC	F1	bACC	F1
MultiASNet	59.3%(0.03)	0.47(0.02)	59.8%(0.01)	0.49(0.05)
MultiASNet with Classifier Retraining	**60.9%(0.04)**	**0.48(0.03)**	**67.7%(0.03)**	**0.54(0.04)**

DFR was far less impactful at the cine-level because only the correctness of the prediction was considered; all false predictions were treated equally, regardless of the model's confidence.

MultiASNet also exhibited an especially high false positive rate for cases with a normal AS classification, which contributed to an F1 score that was much lower than the balanced accuracy. This likely occurred because stenosis appears dissimilar in bicuspid valves than in tricuspid valves. In patients with bicuspid valves, AS tends to develop at a much earlier age and is characterized by less valve calcification, making it challenging for machine learning models to differentiate diseased bicuspid valves from normal tricuspid valves (see Fig. 4). DFR did not substantially lower the high false positive rate for normal cases, possibly due to the limited representation of normal bicuspid valve cases in the training dataset (5 echo studies), which may have hindered the ability of the feature extractor to learn essential characteristics for this subgroup.

C. Analysis of Cross-Attention

The cross-attention mechanism in our model plays a crucial role in aligning video and tabular data, enhancing the model's ability to focus on features most relevant to detecting AS. While the tabular data initially contains 35 features selected by cardiologists based on clinical relevance to AS, the attention mechanism allows the model to prioritize those that carry more

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

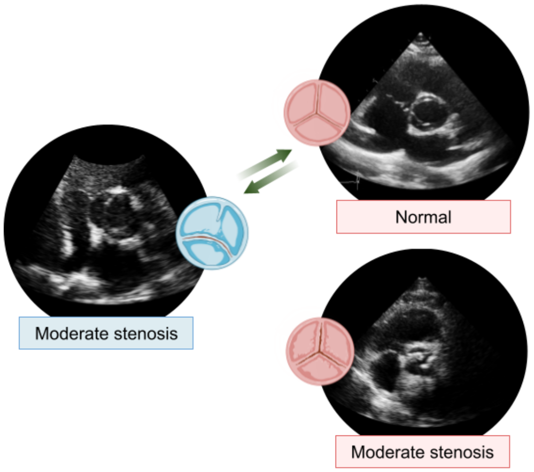

Fig. 4. Bicuspid valve (BV) with moderate stenosis can appear more similar to a tricuspid valve (TV) that is normal than moderately stenotic (shown by green arrows). Subtle differences between diseased BV and normal TV likely contributed to a high false positive rate for patients without AS (normal). This failure mode was resolved with class reweighting using DFR.

discriminative power in the context of the specific input video for the task. Rather than excluding features outright, the model adjusts the importance given to each feature in relation to the visual cues from the video data, improving the overall prediction accuracy.

Averaging the attention values across all frames reveals which features were consistently emphasized by the model during training, indicating their higher relevance for AS detection. The most important features identified by the model include:

- ROOT: Aortic root dimension, critical for assessing AS severity.
- AVA: Aortic valve area, fundamental for AS assessment.
- Age: Known risk factor for AS progression.
- Body Surface Area: An important clinical variable that helps normalize various cardiovascular measurements.
- AoPG: Aortic Peak Gradient, another vital parameter that directly reflects the severity of AS.

Conversely, the least important features include:

- MV Prosthetic: Whether the patient has a prosthetic mitral valve.
- LVEF: Left ventricle ejection fraction (EF) range.
- Mitral Regurgitation: Improper closure of the mitral valve leading to backward blood flow into the left atrium.
- MV Sam: Abnormal anterior mitral valve motion during systole.
- Bicuspid: Presence of a bicuspid aortic valve, an anatomical variant relevant to AS.

Using cross-attention, the model reduces the impact of features like MV Prosthetic or Mitral Regurgitation, which align less with the video data, while emphasizing those features more tightly correlated with AS from B-mode echo video. Although having a bicuspid valve is clinically important for AS diagnosis, the model considers it less directly relevant, potentially due to the relatively small number of bicuspid cases in the dataset. More broadly, the least important features tend to be less anatomically aligned with the aortic valve—focusing instead on the mitral valve or left ventricular function.

D. Clinical Importance of Misclassifications

For screening and diagnosis, the most harmful mistake occurs when clinically significant AS is misclassified as normal. Among untreated patients with clinically significant AS, it is estimated that 56-67% of them would die within just 5 years [51]. However, with a timely intervention that involved valve replacement, patients with clinically significant AS would achieve long-term survival rates similar to those of the general population [52]. On our private dataset, we found that only 0.4% of studies with clinically significant AS were misclassified by MultiASNet as normal. For the detection of clinically significant AS, MultiASNet achieved 84% sensitivity and 87% specificity. Gulic et al. [2] showed that cardiologists who used POCUS without Doppler achieved similar levels of sensitivity (83%) and specificity (84%) for the detection of clinically significant AS in patients with newly discovered heart murmur. The ability for MultiASNet to achieve high sensitivity for detecting clinically significant AS demonstrates its potential as a screening tool for POCUS users. Nevertheless, it is still unknown how the same DL model would perform after finetuning on POCUS datasets or on datasets that include cases with discordant Doppler measurements. We leave these investigations for future work.

VIII. Conclusion

We introduce MultiASNet, a framework for robust training with tabular and video data modalities that outperforms all state-of-the-art models in detecting AS and classifying its severity at the study level. A central innovation of MultiASNet lies in its use of structured tabular report data to guide training—improving representation learning from video—while remaining video-only at inference. This allows the model to benefit from rich contextual information without introducing additional input requirements or workflow changes at deployment. Additionally, unlike prior work, which primarily focuses on static imaging modalities such as X-ray or MRI, MultiASNet is tailored for ultrasound video, a fundamentally different domain. Our use of cross-attention enables the model to prioritize clinically relevant tabular features while discounting misleading or weakly aligned inputs. Notably, MultiASNet achieves high balanced accuracy for AS severity detection, with 93.0% on the private dataset and 83.9% on the public dataset. In the context of POCUS, accurate AS detection is more critical than severity classification since detection prompts referral for Doppler imaging, the gold standard for detailed assessment. We further show the addition of sample selection reduces overfitting to noisy labels, resulting in better performance. As for clinical impact, we show that MultiASNet has a low risk of making critical errors, where it fails to detect AS in patients with severe AS. Future research will aim to address current limitations in DL models for AS diagnosis, particularly biases against underrepresented patients with bicuspid valves, and explore how to deal with subpopulation

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

imbalance in these patients when there is a large amount of unlabeled data, which is often the case with medical imaging.

References

[1] M. Thoenes, P. Bramlage, P. Zamorano, *et al.*, "Patient screening for early detection of aortic stenosis (as)—review of current practice and future perspectives," *Journal of Thoracic Disease*, vol. 10, no. 9, p. 5584, 2018.

[2] T. G. Gulič, J. Makuc, G. Prosen, and D. Dinevski, "Pocket-size imaging device as a screening tool for aortic stenosis," *Wiener Klinische Wochenschrift*, vol. 128, pp. 348–353, 2016.

[3] C. M. Otto, R. A. Nishimura, R. O. Bonow, *et al.*, "2020 ACC/AHA guideline for the management of patients with valvular heart disease," *Journal of the American College of Cardiology*, vol. 77, no. 4, e25–e197, 2021.

[4] J. Minners, M. Allgeier, C. Gohlke-Baerwolf, R.-P. Kienzle, F.-J. Neumann, and N. Jander, "Inconsistencies of echocardiographic criteria for the grading of aortic valve stenosis," *European Heart Journal*, vol. 29, no. 8, pp. 1043–1048, 2008.

[5] J. Minners, M. Allgeier, C. Gohlke-Baerwolf, R.-P. Kienzle, F.-J. Neumann, and N. Jander, "Inconsistent grading of aortic valve stenosis by current guidelines: Haemodynamic studies in patients with apparently normal left ventricular function," *Heart*, vol. 96, no. 18, pp. 1463–1468, 2010.

[6] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[7] M. K. Grzeszczyk, S. Płotka, B. Rebizant, *et al.*, "TabAttention: Learning attention conditionally on tabular data," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Springer Nature Switzerland, 2023, pp. 347–357, ISBN: 9783031439902.

[8] P. Hager, M. J. Menten, and D. Rueckert, "Best of both worlds: Multimodal contrastive learning with tabular and imaging data," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2023, pp. 23 924–23 935.

[9] W. Huang, "Multimodal contrastive learning and tabular attention for automated alzheimer's disease prediction," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023, pp. 2465–2474.

[10] S. Pölsterl, T. N. Wolf, and C. Wachinger, "Combining 3D image and tabular data via the dynamic affine feature map transform," in *Lecture Notes in Computer Science*. Springer International Publishing, 2021, pp. 688–698, ISBN: 9783030872403.

[11] W. Kim, S. Chun, T. Kim, D. Han, and S. Yun, "HYPE: Hyperbolic entailment filtering for underspecified images and text," *arXiv (Cornell University)*, 2024.

[12] B. J. J. Velders, R. H. H. Groenwold, A. P. Kappetein, J. Braun, R. J. M. Klautz, and M. D. Vriesendorp, "Measurement error in echocardiographic assessment of aortic stenosis: an epidemiological consideration of research methodology and clinical practice," *European Heart Journal*, vol. 43, no. Supplement_2, ehac544.2863, Oct. 2022, ISSN: 0195-668X.

[13] L. Jiang, Z. Zhou, T. Leung, L. Li, and L. Fei-Fei, "Mentornet: Regularizing very deep neural networks on corrupted labels," *CoRR*, vol. abs/1712.05055, 2017. arXiv: 1712.05055.

[14] F. Fooladgar, M. N. N. To, P. Mousavi, and P. Abolmaesumi, "Manifold dividemix: A semi-supervised contrastive learning framework for severe label noise," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2024, pp. 4012–4021.

[15] N. Ahmadi, M. Y. Tsang, A. N. Gu, T. S. M. Tsang, and P. Abolmaesumi, "Transformer-based spatio-temporal analysis for classification of aortic stenosis severity from echocardiography cine series," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 366–376, 2024.

[16] M. Mokhtari, N. Ahmadi, T. S. M. Tsang, P. Abolmaesumi, and R. Liao, "GEMTrans: A general, echocardiography-based, multi-level transformer framework for cardiovascular diagnosis," in *Machine Learning in Medical Imaging*, X. Cao, X. Xu, I. Rekik, Z. Cui, and X. Ouyang, Eds., Cham: Springer Nature Switzerland, 2024, pp. 1–10, ISBN: 978-3-031-45676-3.

[17] H. Vaseli, A. N. Gu, S. N. Ahmadi Amiri, *et al.*, "ProtoASNet: Dynamic prototypes for inherently interpretable and uncertainty-aware aortic stenosis classification in echocardiography," in Springer Nature Switzerland, 2023, pp. 368–378.

[18] Z. Huang, X. Yu, B. S. Wessler, and M. C. Hughes, "Semi-supervised multimodal multi-instance learning for aortic stenosis diagnosis," *arXiv preprint arXiv:2403.06024*, 2024.

[19] A. N. Gu, M. Tsang, H. Vaseli, T. Tsang, and P. Abolmaesumi, "Reliable multi-view learning with conformal prediction for aortic stenosis classification in echocardiography," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2024, pp. 327–337.

[20] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 932–18 943, 2021.

[21] M. Christensen, M. Vukadinovic, N. Yuan, and D. Ouyang, "Multimodal foundation models for echocardiogram interpretation," *arXiv (Cornell University)*, 2023.

[22] S. Ebrahimi, S. O. Arik, Y. Dong, and T. Pfister, "Lanistr: Multimodal learning from structured and unstructured data," *arXiv (Cornell University)*, 2023.

[23] A. Singh, R. Hu, V. Goswami, *et al.*, "Flava: A foundational language and vision alignment model," in *2022*

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15 617–15 629.

[24] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," *arXiv preprint arXiv:2210.10163*, 2022.

[25] S. Zhang, Y. Xu, N. Usuyama, *et al.*, "BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[26] H. Xu, G. Ghosh, P.-Y. Huang, *et al.*, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," *arXiv preprint arXiv:2109.14084*, 2021.

[27] B. Ni, H. Peng, M. Chen, *et al.*, "Expanding language-image pretrained models for general video recognition," in *European Conference on Computer Vision*, Springer, 2022, pp. 1–18.

[28] M. Vukadinovic, X. Tang, N. Yuan, *et al.*, "Echoprime: A multi-video view-informed vision-language model for comprehensive echocardiography interpretation," *arXiv preprint arXiv:arXiv:2410.09704v1*, 2024.

[29] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[30] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" *Advances in Neural Information Processing Systems*, vol. 35, pp. 507–520, 2022.

[31] Y. Gorishniy, I. Rubachev, and A. Babenko, "On embeddings for numerical features in tabular deep learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 991–25 004, 2022.

[32] S. Ö. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 6679–6687.

[33] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 932–18 943, 2021.

[34] H. Iida, D. Thai, V. Manjunatha, and M. Iyyer, "Tabbie: Pretrained representations of tabular data," *arXiv preprint arXiv:2105.02584*, 2021.

[35] J. Gardner, J. C. Perdomo, and L. Schmidt, "Large scale transfer learning for tabular data via language modeling," *arXiv preprint arXiv:2406.12031*, 2024.

[36] C. Xue, L. Yu, P. Chen, Q. Dou, and P.-A. Heng, "Robust medical image classification from noisy labeled data with global and local representation guided co-training," *IEEE Transactions on Medical Imaging*, vol. 41, no. 6, pp. 1371–1382, 2022.

[37] Z. Huang, G. Long, B. Wessler, and M. C. Hughes, "A new semi-supervised learning benchmark for classifying view and diagnosing aortic stenosis from echocardiograms," in *Machine Learning for Healthcare Conference*, PMLR, 2021, pp. 614–647.

[38] B. S. Wessler, Z. Huang, G. M. Long Jr, *et al.*, "Automated detection of aortic stenosis using machine learning," *Journal of the American Society of Echocardiography*, vol. 36, no. 4, pp. 411–420, 2023.

[39] Z. Huang, G. Long, B. Wessler, and M. C. Hughes, "TMED 2: A dataset for semi-supervised classification of echocardiograms," in *DataPerf: Benchmarking Data for Data-Centric AI Workshop*, 2022.

[40] Z. Huang, B. S. Wessler, and M. C. Hughes, "Detecting heart disease from multi-view ultrasound images via supervised attention multiple instance learning," in *Proceedings of the 8th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, vol. 219, PMLR, 2023, pp. 285–307.

[41] T. Ginsberg, R.-e. Tal, M. Tsang, *et al.*, "Deep video networks for automatic assessment of aortic stenosis in echocardiography," in *Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021*, Springer, 2021, pp. 202–210.

[42] M. Christensen, M. Vukadinovic, N. Yuan, *et al.*, "Vision-language foundation model for echocardiogram interpretation," *Nature Medicine*, vol. 30, pp. 1481–1488, 2024.

[43] X. Liu, Q. Wang, Y. Hu, *et al.*, "End-to-end temporal action detection with transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5427–5441, 2022, ISSN: 1941-0042.

[44] Z. Liao, H. Girgis, A. Abdi, *et al.*, "On modelling label uncertainty in deep neural networks: Automatic estimation of intra-observer variability in 2D echocardiography quality assessment," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1868–1883, 2019.

[45] T. Zhang, X. Yue, Y. Li, and H. Sun, "Tablellama: Towards open large generalist models for tables," *arXiv preprint arXiv:2311.09206*, 2023.

[46] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[47] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "Tabpfn: A transformer that solves small tabular classification problems in a second," *arXiv preprint arXiv:2207.01848*, 2022.

[48] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," *arXiv (Cornell University)*, 2022.

[49] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi, "Change is hard: A closer look at subpopulation shift," *arXiv (Cornell University)*, 2023.

[50] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[51] G. Strange, S. Stewart, D. Celermajer, *et al.*, "Poor long-term survival in patients with moderate aortic stenosis," *Journal of the American College of Cardiology*, vol. 74, no. 15, pp. 1851–1863, 2019.

[52] P. Varadarajan, N. Kapoor, R. C. Bansal, and R. G. Pai, "Survival in elderly patients with severe aortic stenosis is dramatically improved by aortic valve replacement: Results from a cohort of 277 patients aged 80 years,"

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3609319

European Journal of Cardio-Thoracic Surgery, vol. 30, no. 5, pp. 722–727, 2006.

Authorized licensed use limited to: SRM Institute of Science and Technology Kattankulathur. Downloaded on September 15,2025 at 19:22:32 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.