Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 3, July 2026, pp: 1204-1223 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

BackMix-Enhanced Semi-Supervised Learning for Automated Detection of Aortic Stenosis from Transthoracic Echocardiographic Images

Fatima Ezzahra Elkouahy[1], Labakoum Badreddine[1,3], Hajar Ouahid[2], Mansoor Alam[4], Hamid El Malali[1], Lhoucine Ben Taleb[1] and Azeddine Mouhsen[1]

[1] Energy-Materials-Instrumentation and Telecom Laboratory (EMIT), Faculty of Science and Technology, University Hassan 1st, Settat, Morocco
[2] Bioscience, and Health Research Laboratory, Faculty of Medicine and Pharmacy, Cadi Ayyad University, Marrakech, Morocco
[3] Sciences and Engineering of Biomedicals, Biophysics and Health Laboratory, Higher Institute of Health Sciences (ISSS), University Hassan 1st, Settat, Morocco
[4] University of engineering and technology, Peshawar, Pakistan

Corresponding author: Fatima Ezzahra Elkouahy. (e-mail: elkouahyfat.fst@uhp.ac.ma), **Author(s) Email:** Badreddine Labakoum. (e-mail: b.labakoum@uhp.ac.ma), Hajar Ouahid (e-mail: ouahid.hajar.1@gmail.com), Mansoor Alam (e-mail: mansooralam129047@gmail.com), Hamid El Malali (e-mail: hamid.elmalali@uhp.ac.ma), Lhoucine Ben Taleb (e-mail: l.bentaleb@uhp.ac.ma), Azeddine Mouhsen (e-mail: azmouhsen@gmail.com)

Abstract **Aortic stenosis (AS) is a prevalent and progressive valvular heart disease requiring accurate severity assessment for optimal clinical decision-making. Transthoracic echocardiography (TTE) is the standard diagnostic modality; however, its interpretation remains operator-dependent and subject to inter-observer variability. In this study, we propose an anatomically guided BackMix-enhanced semi-supervised learning framework for automated echocardiographic view classification and AS severity assessment. The approach leverages Gradient-weighted Class Activation Mapping (Grad-CAM) to preserve diagnostically relevant anatomical regions during augmentation while modifying background areas to mitigate shortcut learning. A semi-supervised self-training strategy combined with an ensemble classification framework was used to exploit both labeled and unlabeled data. The framework was evaluated on the TMED2 dataset of echocardiography, comprising 24,964 TTE images. To assess generalizability, external validation was performed on an independent dataset of 300 cardiologist-labeled echocardiographic images, including apical four-chamber (A4C), parasternal long-axis (PLAX), and parasternal short-axis (PSAX) views. Experimental results demonstrated that the proposed method outperformed the baseline semi-supervised model, improving view classification accuracy by approximately 3% and AS severity classification accuracy by 4–6%, with the largest gain observed in moderate AS. Performance remained consistent on the external validation dataset, supporting the robustness of the proposed approach. Statistical analysis confirmed the significance of these improvements ($p < 0.01$). Grad-CAM evaluation further demonstrated improved localization of clinically relevant regions. These findings suggest that anatomically guided BackMix augmentation combined with semi-supervised ensemble learning can improve classification accuracy, robustness, and interpretability in echocardiographic analysis under limited annotation conditions, offering a promising approach for automated AS assessment across independent clinical datasets.**

Keywords **Aortic stenosis; Transthoracic echocardiography; Semi-supervised learning; Anatomically guided BackMix; Shortcut learning; Deep learning.**

I. Introduction

Aortic stenosis (AS) is a progressive valvular heart disease characterized by narrowing of the aortic valve opening, leading to obstruction of left ventricular outflow and increased cardiac workload [1][2] It represents one of the most common and clinically significant valve pathologies in adults, particularly within the aging population [3]. Early and accurate assessment of AS severity is essential to guide timely intervention and prevent irreversible myocardial damage [4][5]. Transthoracic echocardiography (TTE) remains the primary imaging modality for AS evaluation [6],offering non-invasive visualization of valve morphology and hemodynamic gradients [7]. However,

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

echocardiographic interpretation is highly operator-dependent, requiring substantial expertise and experience [8][9][10]. As a result, diagnostic accuracy can vary widely between practitioners, and manual interpretation remains time-consuming and subject to inter-observer variability [11]. Deep learning has shown strong potential to transform cardiac imaging [12][13][14][15][16][17]. CNNs can automatically classify cardiac views, segment chambers, and grade valvular disease. However, most approaches rely on fully supervised learning, requiring large, expert-labeled datasets that are costly and time-consuming, especially in echocardiography [5], [18]. In addition, Echocardiographic images also vary widely in acquisition protocols, probe orientation, equipment, and patient characteristics, making robust generalization challenging. These limitations highlight the need for semi-supervised and anatomically guided methods that can leverage unlabeled data while ensuring reliable performance [19].

Beyond the challenge of limited labeled data, AI models in echocardiography are also susceptible to shortcut learning, a well-recognized failure mode in deep learning [20]. Shortcut learning occurs when a model relies on spurious correlations or dataset-specific cues instead of learning genuine anatomical or physiological patterns [13], [21]. In echocardiography, this may include dependence on text overlays (e.g., “AP4” labels), scanner-specific watermarks, background image textures, probe positioning, or even pixel-encoded metadata reflecting hospital origin or patient demographics as shown in Fig.1. Instead of focusing on clinically relevant features such as leaflet calcification, flow acceleration, or ventricular structure the model may learn to associate superficial artifacts with disease severity [22], [23]. As a result, performance may appear excellent on internal validation but collapse when applied to external datasets from different hospitals, machines, or patient populations. This undermines clinical reliability and limits deployment. To address shortcut learning, common strategies include Visualizing model attention (via Grad-CAM or Layer-wise Relevance Propagation), cross-institutional validation, removing metadata, domain generalization techniques, and anatomically guided augmentation methods such as BackMix [21], [24].

Semi-supervised learning (SSL) has emerged as a promising solution to reduce dependence on large labeled datasets while leveraging the vast quantities of unlabeled echocardiographic images routinely generated in clinical practice. By incorporating both labeled and unlabeled data during training, SSL helps models learn the underlying structure of echocardiographic images and improves classification performance [25], [26]. However, existing SSL approaches in medical imaging often struggle with the subtle anatomical variations characteristic of echocardiography, as well as the variability in image quality [8] ,[27].

Despite recent advances in semi-supervised learning (SSL) and data augmentation for echocardiographic analysis, a critical gap remains: existing approaches do not address shortcut learning in a domain-specific, anatomically informed manner. Standard augmentation techniques such as cropping, rotation, and CutMix operate without anatomical awareness and may distort diagnostically relevant cardiac structures [28], while generic BackMix formulations have primarily been validated in natural image tasks rather than disease severity grading. In addition, most SSL-based echocardiographic frameworks rely on single-backbone feature extractors and do not incorporate ensemble strategies to mitigate model variance under limited annotation. These

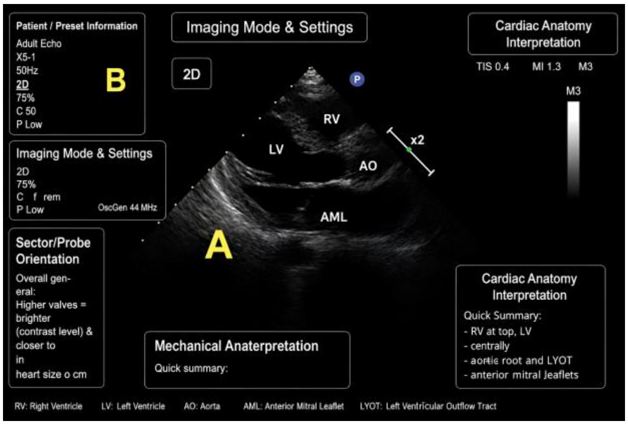

Fig. 1. Example of a parasternal long-axis transthoracic echocardiographic image showing the clinically relevant field of view/region of interest (A) and surrounding background region (B). The figure illustrates potential non-anatomical background cues that may contribute to shortcut learning in deep learning models.

limitations are particularly problematic in the assessment of aortic stenosis (AS), where subtle morphological differences are crucial for accurate severity classification. To address this gap, we propose an anatomically guided BackMix augmentation integrated within a semi-supervised learning framework, combined with ensemble classification, to preserve clinically relevant structures, reduce shortcut learning, and improve robustness and generalization in echocardiographic analysis [29].

In our previous work [26], we developed an AI-based framework for automated detection of aortic stenosis severity using the TMED2 echocardiographic dataset Fig.2. This model combined CNNs with a semi-supervised architecture for classifying both AS severity

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i3.1657

and cardiac views [27]. Despite promising results, its performance was limited by factors such as the small number of labeled samples, the absence of anatomically informed augmentation, and occasional model instability and overfitting [27][30], [31]. These limitations highlighted the need for a more robust, generalizable, and physiologically grounded learning strategy. To address these challenges, we propose an advanced semi-supervised learning framework that leverages BackMix augmentation alongside ensemble-based classification, enabling more accurate detection of aortic stenosis from transthoracic echocardiography. BackMix is a context-aware augmentation technique that preserves anatomical realism by blending foreground and background regions guided by Gradient-weighted Class Activation Mapping (Grad-CAM) [15]. This strategy generates physiologically synthetic samples while mitigating artifacts that may encourage shortcut learning. In addition, a multimodal feature-extraction framework leveraging ResNet50, EfficientNetB0, and DenseNet121 is employed to capture complementary structural and textural characteristics from echocardiographic images. These features are classified using an ensemble of algorithms (Random Forest, Gradient Boosting, and Support Vector Machine) under a self-training semi-supervised framework.

The proposed BackMix-enhanced semi-supervised framework is compared against a baseline semi-supervised learning model using the same TMED2 dataset [32]. This study adopts a retrospective, controlled comparative experimental design to isolate the impact of anatomically guided augmentation and ensemble learning. This evaluation strategy enables a clear and systematic assessment of how these components influence classification accuracy, robustness, and generalization under variable echocardiographic imaging conditions [33].

Unlike prior BackMix-based approaches that primarily focus on open-set recognition or natural image domains, our work introduces an anatomically constrained BackMix formulation specifically tailored for echocardiographic imaging. The key novelty lies in (i) the integration of Grad-CAM-derived (ii) the extension of BackMix from view classification to the clinically challenging task of aortic stenosis severity grading, and (iii) the coupling of anatomically guided augmentation with a semi-supervised self-training framework to mitigate shortcut learning under limited annotation settings. This combination is not merely additive but functionally synergistic, as anatomical

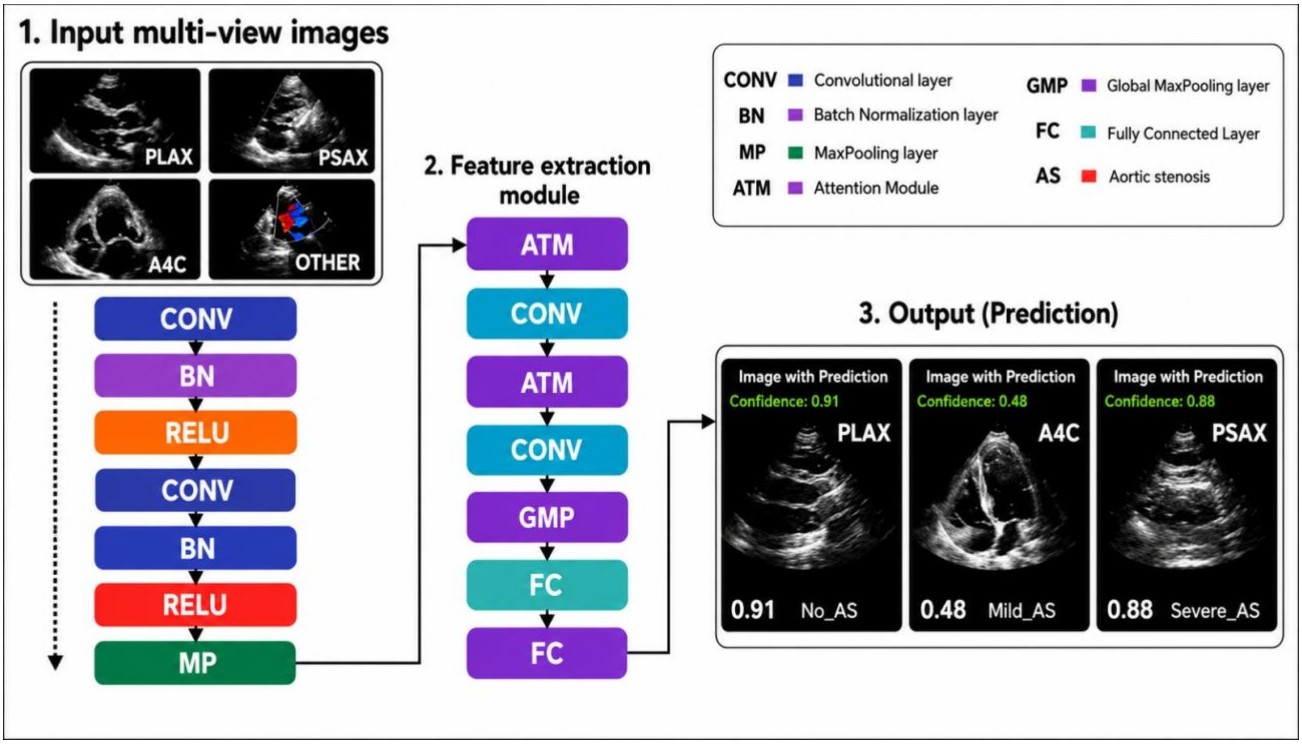

Fig. 2. Overall architecture of the proposed AI framework for automated aortic stenosis severity detection from transthoracic echocardiographic images. The pipeline includes image preprocessing, feature extraction using CNN backbones, semi-supervised pseudo-labeling, ensemble classification, and final prediction of echocardiographic view and aortic stenosis severity.

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

constraints directly influence pseudo-label reliability and feature generalization.

II. Method

A. Study design

A retrospective comparative experimental design was employed to evaluate the impact of anatomically guided BackMix augmentation and ensemble learning. Two semi-supervised pipelines were implemented: (i) a baseline model without augmentation and (ii) a BackMix-enhanced model incorporating anatomical guidance and ensemble classification. Both pipelines were trained and evaluated under identical conditions using the same dataset splits and evaluation metrics, addressing two tasks: echocardiographic view classification (A4C, PLAX, PSAX) and aortic stenosis severity classification (no AS, mild AS, severe AS). To ensure a fair comparison, all experimental conditions were controlled, including dataset splits, preprocessing (resizing, normalization, encoding), the feature extraction backbone (ResNet50), the semi-supervised self-training strategy, the pseudo-label confidence threshold (0.8), and the evaluation metrics. The only differences were the application of BackMix augmentation and the use of ensemble classification in the proposed model. The workflow comprised data acquisition and partitioning, preprocessing and label encoding, separation into labeled and unlabeled subsets, semi-supervised training, application of BackMix augmentation (proposed model only), feature extraction and classification, and final evaluation on a held-out test set. The baseline model isolated the contributions of anatomically guided augmentation and ensemble learning, enabling direct assessment of their effects on robustness, generalization, and resistance to shortcut learning.

Table 1. Distribution of the TMED2 dataset used for echocardiographic view classification and aortic stenosis (AS) severity grading.

Task	Class	Number of Images (N)
View Classification	A4C (Apical 4-Chamber)	5,89
	PLAX (Parasternal Long-Axis)	10,234
	PSAX (Parasternal Short-Axis)	8,84
AS Severity Grading	No AS	8,421
	Mild AS	3,217
	Moderate AS	1,089
	Severe AS	2,237

B. Materials and Dataset Description

Experiments were conducted using the TMED2 dataset, which comprises 24964 transthoracic echocardiographic images with accompanying metadata stored in CSV format [26], [32]. Each image is associated with identifiers, cardiac view labels, diagnostic labels for AS severity, and predefined data splits for training and evaluation [34]. The dataset presents common challenges encountered in medical imaging, including class imbalance, variable image quality, and incomplete annotations. To address these issues, a structured preprocessing pipeline was implemented to ensure reproducibility and robustness across experiments. As mentioned in Table 1, the dataset contained A4C samples N_1, PLAX samples N_2, and PSAX samples N_3 for the view classification task. The dataset for aortic stenosis severity classification contained N_4 no AS samples N_5, mild AS samples N_6, moderate AS samples, and N_7 severe AS samples. The dataset comprised three echocardiographic views for the view classification task, including apical four-chamber (A4C) samples (N_1), parasternal long-axis (PLAX) samples (N_2), and parasternal short-axis (PSAX) samples (N_3). For the aortic stenosis (AS) severity classification task, the dataset included no AS cases (N_4), mild AS (N_5), moderate AS (N_6), and severe AS (N_7). As shown in Table 2, the dataset exhibits a notable class imbalance, particularly in the moderate AS category (N = 1,089). Additional challenges include variable image quality and incomplete annotations, which are commonly encountered in clinical echocardiographic datasets. To address these issues, class-weighted cross-entropy loss was applied during model training to reduce bias toward majority classes. In addition, stratified sampling was employed to preserve representative class distributions across training and evaluation splits, ensuring stable and reliable performance assessment across all categories.

Table 2. Distribution of labeled, unlabeled, and test samples in the TMED2 dataset for semi-supervised learning.

Split	Subset	Number of Images (N)
Training Set	Labeled	1,257
Training Set	Unlabeled	16,072
Test Set	All Classes	7,635

C. Data Preprocessing and Preparation

All images were resized to 224 × 224 pixels and normalized using standard ImageNet preprocessing

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 3, July 2026, pp: 1204-1223 e-ISSN: 2656-8632

parameters (μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]) to ensure compatibility with pretrained convolutional neural networks. Prior to resizing, Gaussian filtering (σ = 1.0) was applied to reduce ultrasound speckle noise, while preserving clinically relevant structures. Stratified sampling was used to maintain representative class distributions across training, validation, and test sets.

Labels were encoded using scikit-learn's Label Encoder, with cardiac views encoded as A4C = 0, PLAX = 1, and PSAX = 2, and aortic stenosis (AS) severity encoded as mild_AS = 0, no_AS = 1, severe_AS = 2, and moderate_AS = 3. To simulate real-world clinical conditions with limited annotations, the training data were divided into labeled and unlabeled subsets. Labeled samples retained their ground truth annotations, while unlabeled samples were assigned a sentinel value of −1 and used exclusively within the semi-supervised learning framework. All preprocessing and model implementation steps were performed in Python 3.10 using PyTorch deep learning framework (version 2.0.1), scikit-learn (version 1.3.0), and OpenCV (version 4.8.0). Experiments were conducted on an NVIDIA RTX 2060 GPU (6 GB) with 16 GB system RAM. In addition, standard data augmentation techniques, including random rotation, horizontal flipping, and scaling, were applied during training. BackMix augmentation was subsequently integrated into the proposed pipeline to enhance robustness and reduce shortcut learning.

D. Model Architectures and Learning Frameworks

A. Baseline Semi-Supervised Learning Pipeline

The baseline pipeline served as the control condition for comparative analysis. Feature extraction was performed using a ResNet50 model pretrained on ImageNet, producing 2048-dimensional feature vectors through global average pooling. No data augmentation was applied beyond standard resizing and normalization [35], [36]. A Random Forest classifier with 200 estimators was employed within a self-training semi-supervised framework. The model was trained using a combination of supervised categorical cross-entropy loss on labeled data and entropy minimization on unlabeled data. We utilize a hybrid objective that merges supervised learning from labeled data together with entropy minimization applied to unlabeled data.

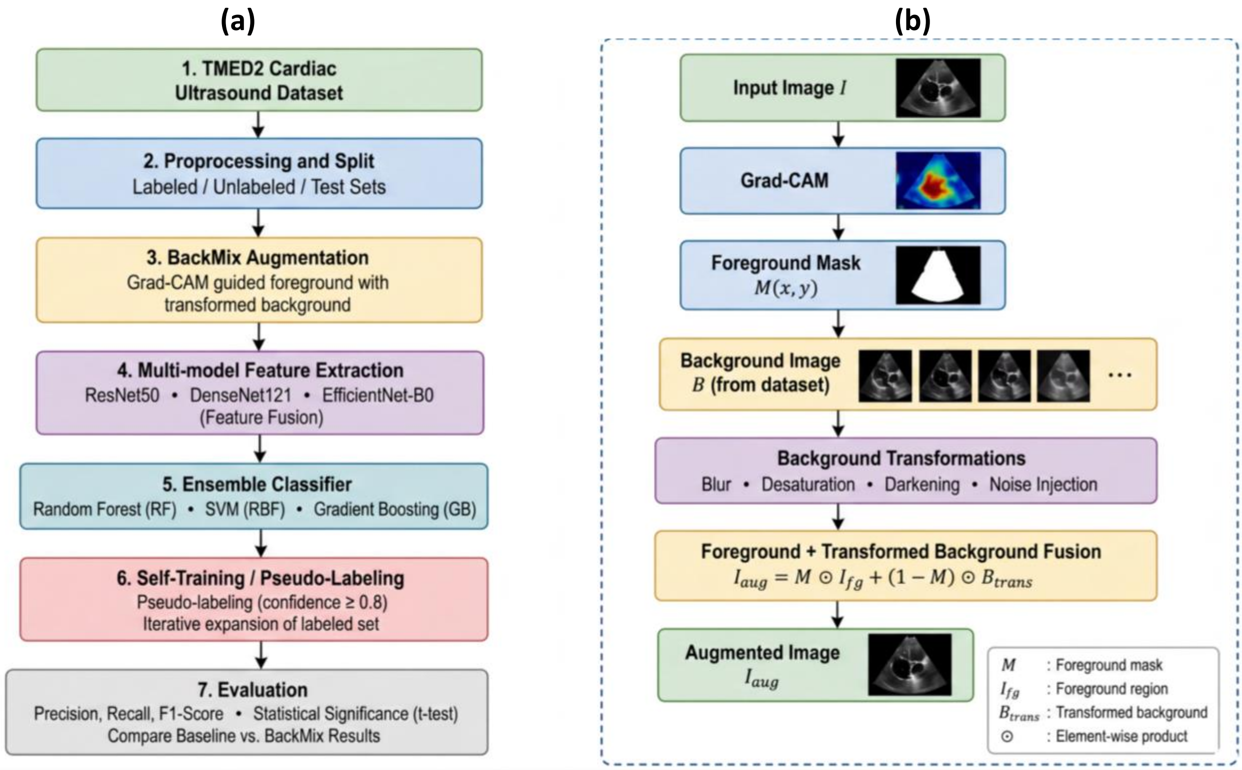

Fig. 3. Proposed BackMix-enhanced semi-supervised learning framework for echocardiographic image classification. (a) Overall pipeline including dataset preparation, preprocessing, feature extraction, pseudo-labeling, ensemble classification, and model evaluation. (b) BackMix augmentation procedure using Grad-CAM-based foreground localization and background transformation to reduce shortcut learning.

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i3.1657

The supervised loss uses standard cross-entropy measurement to compare actual labels with model predictions, which helps the model learn correct class assignment. The model needs to make confident predictions with low uncertainty because we reduce the entropy of the predicted class distribution for our unlabeled data. The regularization uses a cluster assumption to direct decision boundaries away from areas where the data distribution reaches high density. The final objective uses both terms to enhance generalization because it uses both labeled data and unlabeled data, as in Eq. (1) and Eq. (2) [37], [38].

$$L_s = -\left(\frac{1}{N_l}\right)\Sigma y_i \log p(y_i|x_i) \quad (1)$$

$$L_u = -\left(\frac{1}{N_u}\right)\Sigma\, \Sigma p(c|x_i)\log p(c|x_i) \quad (2)$$

where L_s denotes the supervised loss, L_u the unsupervised entropy loss, N_l the number of labeled samples, N_u the number of unlabeled samples, x_i the input sample, y_i the corresponding ground-truth label, and $p(c|x_i)$ the predicted probability of class c for sample x_i. The overall objective function of the semi-supervised framework combines the supervised loss L_s (Eq. (1)) and the unsupervised loss L_u (Eq. (2)) through a weighting coefficient, as defined in Eq. (3) [39].

$$L = L_s + \lambda L_u \quad (3)$$

where L denotes the total optimization loss and λ is a weighting coefficient controlling the contribution of the unsupervised loss term during training. To assign pseudo-labels to unlabeled samples, the predicted class with the highest probability is selected only when the prediction confidence exceeds a predefined threshold, as defined in Eq. (4) [40].

$$\hat{y} = argmax\, p(c|x), if \max p(c|x) > \tau \quad (4)$$

where $\hat{y}$ denotes the pseudo-label assigned to the unlabeled sample x, and $p(c|x)$ represents the predicted probability of class c for sample x. The operator $argmax$ returns the class with the highest predicted probability. The parameter τ is the confidence threshold used to determine whether a pseudo-label is reliable enough to be assigned. Only predictions whose confidence exceeds this threshold are incorporated into the labeled training set. Pseudo-labels were iteratively assigned to unlabeled samples based on this confidence threshold defined in Eq. (4). The self-training process was repeated for up to 15 rounds or until convergence, progressively expanding the labeled training set while minimizing error propagation.

B. BackMix-Enhanced Semi-Supervised Learning Pipeline

Starting from known-class samples, BackMix estimates the foreground region and selectively replaces or mixes foreground patches with alternative background information to disrupt spurious correlations between foreground and background. Fig.3 illustrates this strategy on (A) natural images and (B) transthoracic echocardiography images, demonstrating its cross-domain applicability. Fig.4 explains in detail. This approach encourages the model to focus on intrinsic foreground features rather than background cues, which is particularly important in echocardiography due to acquisition artifacts, speckle noise, and view-dependent variability. Consequently, BackMix enhances model robustness and improves the detection of unknown or out-of-distribution cases.

Step 1. Foreground Extraction

Foreground anatomical regions were identified using Gradient-weighted Class Activation Mapping (Grad-CAM) generated from a ResNet50 network pretrained on ImageNet. The top 30–35% of activation scores were selected as foreground using quantile-based thresholding, and the foreground was smoothed using a Gaussian filter to preserve anatomical continuity. We use Grad-CAM to generate class-specific localization maps, which show the areas in the input that lead to model predictions. [19], [41].

The importance weight associated with each feature map is computed by globally averaging the gradients of the target class score with respect to the activation maps, as defined in Eq. (5) [42].

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (5)$$

where α_k^c denotes the importance weight of the feature map k for class c, y^c represents the score of the target class before the softmax layer, A_{ij}^k denotes the activation at the spatial location (i, j) in the feature map k, and Z corresponds to the total number of spatial locations used for normalization.

The Grad-CAM localization map is then generated by computing the weighted sum of feature maps followed by a ReLU activation, as defined in Eq. (6) [43].

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \quad (6)$$

where $L_{\text{Grad-CAM}}^c$ denotes the localization map for the class c. The ReLU function preserves only positive contributions, highlighting image regions that positively influence the model prediction.

Step 2. Random Foreground-Background Mixing

To enhance anatomical diversity and reduce spurious correlations, the foreground of each training image Si was combined with a background sampled from a different image Bj as defined in Eq. (7) [27].

$$I_i^{mix} = S_i B_j \quad (7)$$

where I_i^{mix} denotes the mixed image generated for the sample i, S_i represents the foreground region extracted from the source image, and B_j denotes the background sampled from another image j.

Step 3. Detailed Augmentation Pipeline

To further increase variability and realism, the background was processed using Gaussian blurring (σ=2.0), desaturation (30% of the original color intensity), brightness reduction, and Gaussian noise

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

injection (μ=0, σ=10). A binary mask M, generated from Grad-CAM, was used to smoothly blend the foreground and transformed background, producing the final augmented image, as defined in Eq. (8) [44][45][46].

$$I_{aug} = M \odot I_{fg} + (1 - M) \odot \phi\left(I_{bg}\right) \tag{8}$$

Where I_{aug} denotes the final augmented image, I_{fg} represents the foreground region extracted from the original image, and I_{bg} denotes the background region sampled from another image. The binary mask M, derived from Grad-CAM, identifies the foreground region to preserve, while $(1 - M)$ selects the background region to be replaced. The function $\phi(.)$ represents the background transformation pipeline, including Gaussian blurring, desaturation, brightness reduction, and Gaussian noise injection. The operator $\odot$ denotes element-wise multiplication.

Step 4. Feature Extraction

Feature extraction in the enhanced pipeline combined three pretrained CNN architectures: ResNet50 (2048 features), EfficientNet-B0 (1280 features), and DenseNet121 (1024 features). The extracted feature vectors were standardized using z-score normalization and concatenated into a unified representation, as defined in Eq. (9) [27].

$$F = [\, f_{\text{ResNet50}} \mid f_{\text{EffNetB0}} \mid f_{\text{DenseNet121}} \,] \tag{9}$$

Where F denotes the final fused feature vector used for classification, f_{ResNet50} represents the feature vector extracted from ResNet50, denotes the feature vector extracted from EfficientNet-B0, and $f_{\text{DenseNet121}}$ corresponds to the feature vector extracted from DenseNet121. The operator | denotes feature concatenation, which combines the outputs of the three backbone networks into a single 4352-dimensional representation.

Step 5. Classification

Classification was performed using an ensemble of three models: Random Forest (500 trees, max depth = 100), Gradient Boosting (100 estimators, learning rate = 0.05), and Support Vector Machine with a radial basis function kernel ($C = 10$). Final predictions were obtained using a soft voting strategy. Each classifier produced a probability distribution over all classes, denoted as $p_k(\, c \mid x \,)$ for the k-th model, as defined in Eq. (10) [47].

The final predicted class was determined by averaging the probabilities across all K classifiers and selecting the class with the highest mean probability. This ensemble strategy leverages the complementary strengths of individual classifiers, resulting in more

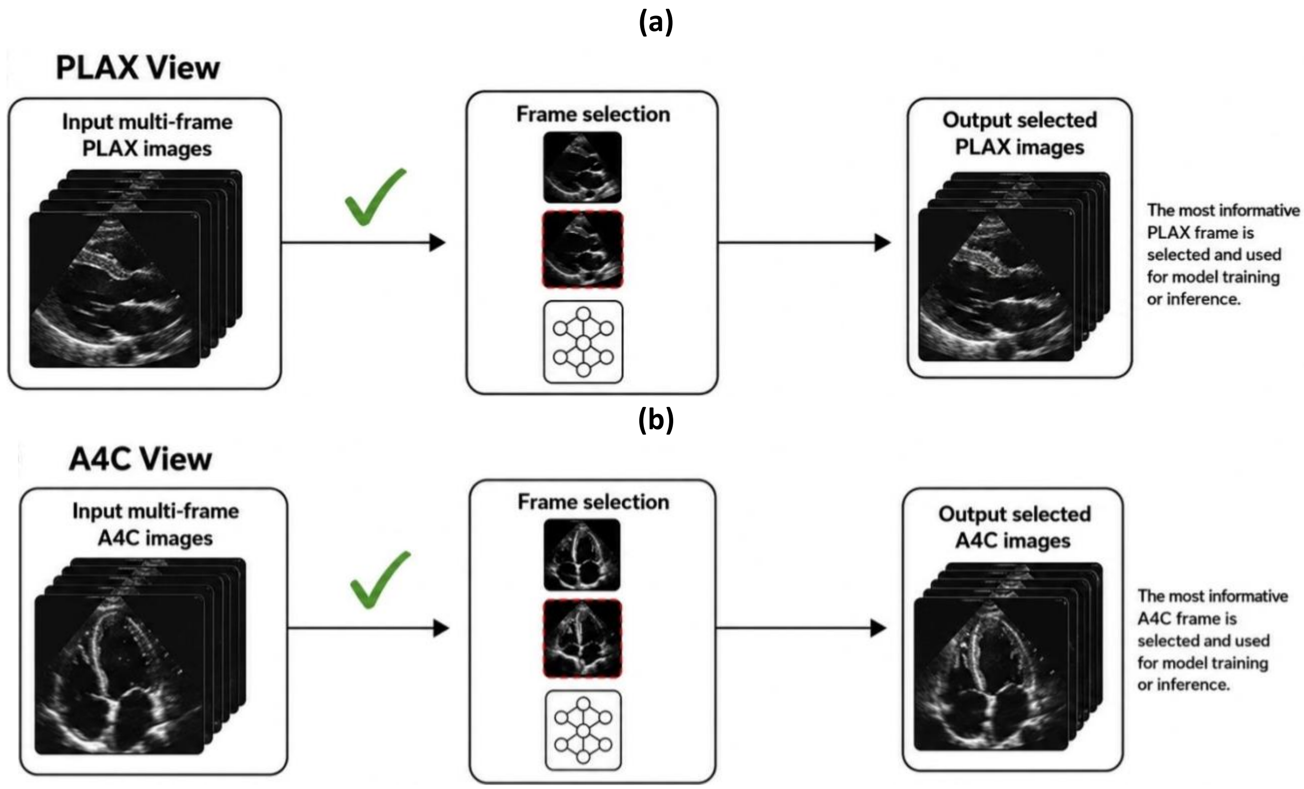

Fig. 4. **Illustration of the BackMix augmentation concept. (a) Generic example showing foreground preservation and background replacement. (b) Application to echocardiographic images, where the anatomical foreground is retained while the background is modified to reduce reliance on non-diagnostic image cues.**

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

robust and reliable predictions than those obtained from a single model.

$$\hat{y} = \arg\max_c \left(\frac{1}{K}\sum_{k=1}^{K} p_k(\,c \mid x\,)\right) \quad (10)$$

where $\hat{y}$ denotes the final predicted class for input sample x, K represents the total number of classifiers in the ensemble, and $p_k(\,c \mid x\,)$ denotes the probability assigned by the k-th classifier to class $\boldsymbol{c}$. The operator $\arg\max$ selects the class with the highest averaged probability.

E. Model Interpretability Using Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to visualize class-discriminative regions contributing to model predictions. For both baseline and BackMix-enhanced pipelines, Grad-CAM maps were generated for representative samples from each cardiac view and AS severity class. The top 30–35% of activation scores were selected as foreground regions using quantile-based thresholding and smoothed with a Gaussian filter to preserve anatomical continuity. These visualizations were used solely for qualitative assessment of model attention and to verify that predictions relied on meaningful anatomical regions [41].

F. Semi-Supervised Self-Training Strategy

Both pipelines employed a self-training approach to incorporate unlabeled data. Pseudo-labels were assigned to unlabeled samples when the maximum predicted class probability exceeded a confidence threshold of 0.8.

The self-training process was iterated for up to 15 rounds or until convergence, thereby progressively expanding the labeled training set while minimizing error propagation [48].

G. Experimental Setup and Model Comparison Protocol

Two model configurations were evaluated: a baseline semi-supervised learning pipeline without data augmentation and a BackMix-enhanced semi-supervised pipeline incorporating anatomically guided augmentation and ensemble learning. Both configurations were trained and evaluated on the TMED2 dataset using identical data splits, task definitions, and evaluation metrics to ensure a controlled comparison. Performance was assessed separately for echocardiographic view classification and aortic stenosis severity classification on a held-out test set.

H. Statistical Analysis and Evaluation Metrics

Model performance was evaluated on the TMED2 test set for both echocardiographic view classification and aortic stenosis severity classification tasks. Metrics were computed on a per-class basis to assess class-wise performance and robustness in the presence of class imbalance. Precision, recall, and F1-score were used as the primary evaluation metrics, as defined in Eq. (11) [49].

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \quad F1 = \frac{2PR}{P+R} \quad (11)$$

where P denotes precision, R denotes recall, and $F1$ represents the harmonic mean of precision and recall. The term TP (true positives) corresponds to correctly classified positive samples, FP (false positives) represents samples incorrectly classified as positive, and FN (false negatives) denotes positive samples incorrectly classified as negative. To evaluate the statistical significance of performance differences between the baseline and proposed models, a paired t-test was conducted across multiple independent training runs. Each experiment was repeated five times using different random initializations. Performance variability was quantified using 95% confidence intervals (CI) for accuracy and F1-score. Statistical significance was established at p<0.05.

As this study involved methodological model comparison using an existing dataset, formal power calculation was not applicable. The research assessed performance enhancements using a paired t-test, comparing five different training sessions. The study calculated 95 percent confidence intervals to measure both accuracy and F1-score performance. The BackMix-enhanced model showed better results than the baseline model, which reached statistical significance at p-value lower than 0.01. The Adam optimizer with a learning rate of 1e-4 was used to train models, which ran for 50 epochs with a batch size of 32. We implemented early stopping, configured to wait 10 epochs before stopping training. We maintained frozen feature extractors during the first training phase and later unfreezed the extractors for advanced training.

The ensemble classification strategy combines Random Forest, Gradient Boosting, and Support Vector Machine classifiers applied to deep feature representations extracted from convolutional neural networks. These models were selected due to their complementary strengths: Random Forest provides robustness to noise and reduces variance through bagging, Gradient Boosting enhances predictive performance by sequentially correcting errors, and SVM offers strong generalization through margin maximization. This hybrid approach is particularly suitable in semi-supervised settings with limited labeled data, where fully end-to-end deep classifiers may be prone to overfitting.

I. Ethical Considerations

This study utilizes the TMED2 dataset, an echocardiography dataset, a publicly available and fully de-identified echocardiographic image dataset. No direct patient contact or intervention was involved, and all images were anonymized prior to analysis, with no access to personally identifiable information. Consequently, individual informed consent and formal ethical approval were not required for this secondary

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

analysis of anonymized data, in accordance with institutional and national research ethics guidelines. All data were accessed and used in compliance with the dataset's governance policies and applicable data protection regulations. The study was conducted in accordance with the principles of the Declaration of Helsinki [50].

III. Result

A. Quantitative performance analysis

1. Confusion Matrices CNN

Table 3 and Table 4 summarize class-wise performance for echocardiographic view classification and aortic stenosis severity classification. Across all classes and tasks, higher accuracy, precision, recall, and F1-score values were observed for one model compared with the other (Table 3). Accuracy differences ranged from +3% for echocardiographic view classification to +6% for aortic stenosis severity classification. Similar numerical differences were observed for precision, recall, and F1-score across all classes.

For aortic stenosis severity classification, the baseline model achieved accuracies of 84% for mild AS, 83% for moderate AS, and 88% for severe AS. The corresponding F1-scores were 83.5%, 81.5%, and 88%, respectively. Higher performance values were observed for the same severity classes in the comparative model. Accuracies increased to 88% for mild AS, 89% for moderate AS, and 92% for severe AS. Corresponding F1-scores increased to 87.5%, 88.5%, and 92%, respectively. Precision and recall values showed similar numerical increases across all severity classes (Table 4). The Fig.5 shows the accuracy and F1-score for echocardiographic view classification (PLAX, PSAX, A4C) and aortic stenosis severity classification (mild, moderate, severe AS). For view classification, accuracy increased by 3% for all views, with F1-scores increasing from 91%, 90.5%, and 92.5% in the baseline model to 94%, 94%, and 95.5% in the BackMix-enhanced model, respectively. For AS severity classification, accuracy increased by 4% for mild AS, 6% for moderate AS, and 4% for severe AS, corresponding F1-scores increasing from 83.5%, 81.5%, and 88% to 87.5%, 88.5%, and 92%, respectively, as shown in Table 4.

Table 3. Class-wise performance comparison between the baseline semi-supervised learning model and the BackMix-enhanced model for echocardiographic view classification on the TMED2 test set.

Class	Baseline Accuracy	BackMix Accuracy	Baseline F1	BackMix F1	Accuracy Gain
PLAX	91	94	91	94	+3
PSAX	92	95	90.5	94	+3
A4C	93	96	92.5	95.5	+3

Table 4. Class-wise performance comparison between the baseline semi-supervised learning model and the BackMix-enhanced model for aortic stenosis (AS) severity classification on the TMED2 test set.

Class	Baseline Accuracy	BackMix Accuracy	Baseline F1	BackMix F1	Accuracy Gain
Mild AS	84	88	83.5	87.5	+4
Moderate AS	83	89	81.5	88.5	+6
Severe AS	88	92	88	92	+4

2. Echocardiographic view classification results

For echocardiographic view classification, the baseline model achieved accuracies of 91% for parasternal long-axis (PLAX), 92% for parasternal short-axis (PSAX), and 93% for apical four-chamber (A4C) views. The corresponding F1-scores were 91%, 90.5%, and 92.5%, respectively. Higher performance values were observed for the same views in the comparative model. Accuracies increased to 94% for PLAX, 95% for PSAX, and 96% for A4C. Corresponding F1-scores increased to 94%, 94%, and 95.5%, respectively. Precision and recall values showed similar numerical increases across all three views (Table 3).

3. Aortic stenosis severity classification

4. Confusion matrix analysis

For echocardiographic view classification (Fig.6), the BackMix-enhanced model demonstrates a clear reduction in inter-class confusion, particularly between the PLAX and PSAX views, which are known to share similar anatomical characteristics. Compared with the baseline model, fewer misclassifications are observed, reflecting improved discrimination between views. The predominance of correct predictions along the diagonal indicates more reliable view recognition and reduced confusion among anatomically related classes. The confusion metrics for aortic stenosis severity classification reveals excellent predictive performance, with all mild AS and severe AS cases correctly classified

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 3, July 2026, pp: 1204-1223 **e-ISSN: 2656-8632**

along the diagonal and no observed cross-class confusion. The clear separation between severity categories indicates robust discrimination across diagnostic classes. The dominance of diagonal entries reflects high reliability and consistency in model predictions, which is particularly important for clinical decision-making where accurate severity stratification is critical (Fig.7).

5. Ablation study

To quantify the individual contribution of each component, an ablation study was conducted by progressively adding BackMix augmentation, Grad-CAM-based anatomical guidance, and ensemble classification. Results indicate that BackMix alone improves AS classification accuracy by ~2.1%, while anatomical guidance contributes an additional ~1.5%, confirming that performance gains are not solely due to augmentation but to anatomically constrained feature learning. The ensemble classifier further improves robustness (+1.2%), particularly in moderate AS cases. Table 5 presents the ablation study configuration used to evaluate the contribution of each component of the proposed framework, including the baseline SSL model, the effect of BackMix augmentation, the integration of Grad-CAM guidance, and the final ensemble architecture.

6. Domain Shift Simulation

To approximate cross-institution variability, we introduced controlled perturbations including intensity scaling, Gaussian noise (σ=15), resolution degradation (downsampling to 128×128), and contrast variation. These transformations mimic differences in ultrasound machines and acquisition protocols. The BackMix-enhanced model demonstrated smaller performance

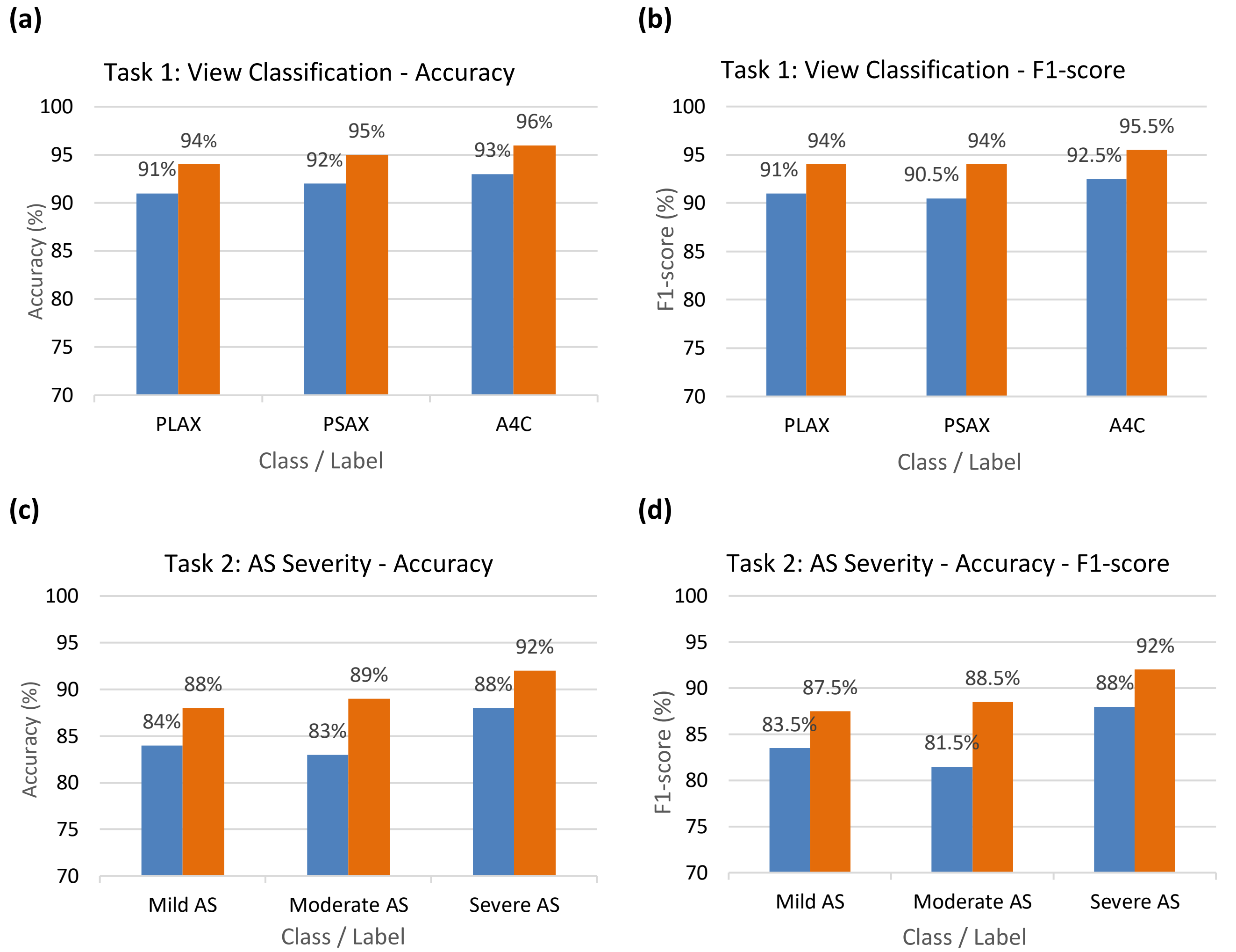

Fig. 5. Performance comparison between the baseline semi-supervised model and the proposed BackMix-enhanced model on the TMED2 test set. (a) Accuracy for echocardiographic view classification, (b) F1-score for echocardiographic view classification, (c) accuracy for aortic stenosis severity classification, and (d) F1-score for aortic stenosis severity classification.

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

degradation under domain shift (−3.2%) compared to the baseline model (−7.8%), indicating improved robustness to distributional changes.

7. Error Analysis

Penetrating research discovered that most errors happened during the assessment process because examiners could not distinguish between mild and moderate aortic stenosis (AS) due to the shared physical features that both conditions displayed. The baseline model frequently misclassified moderate AS as mild, indicating limited sensitivity to intermediate severity patterns. The BackMix-enhanced model showed better accuracy because it decreased confusion while accurately identifying moderate cases. The study achieved progress because the researchers concentrated their efforts on essential anatomical areas which included the aortic valve and left ventricular outflow tract according to the principles of anatomically guided augmentation.

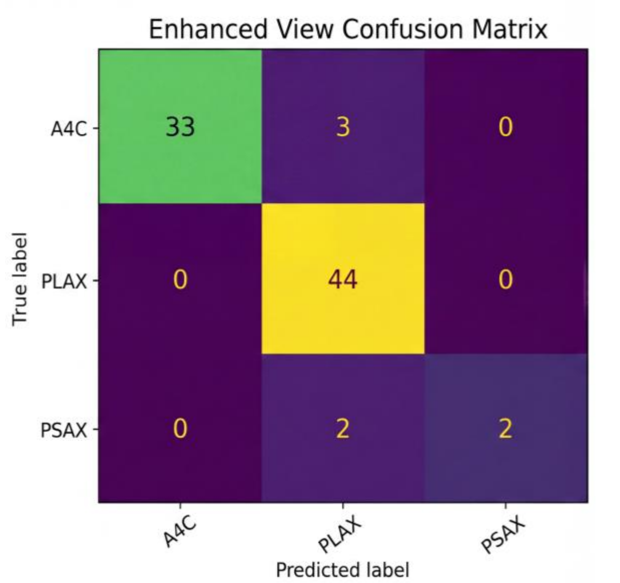

Fig. 6. Confusion matrix of the BackMix-enhanced model for echocardiographic view classification on the TMED2 test set, including PLAX, PSAX, and A4C classes. Diagonal values indicate correctly classified samples, while off-diagonal values indicate misclassifications.

8. Evaluation Metrics

The model demonstrated a prediction accuracy of 94.0%, with a 95% confidence interval of 86.8%-98.0%, to evaluate its performance on the echocardiographic view classification task for 84 test samples. (Table 6) The A4C and PLAX view classes performed exceptionally well, with F1-scores of 0.957 and 0.949 and AUC values of 0.978 and 0.972, indicating their ability to correctly identify between two groups. The PSAX class showed a major decrease in F1-score,

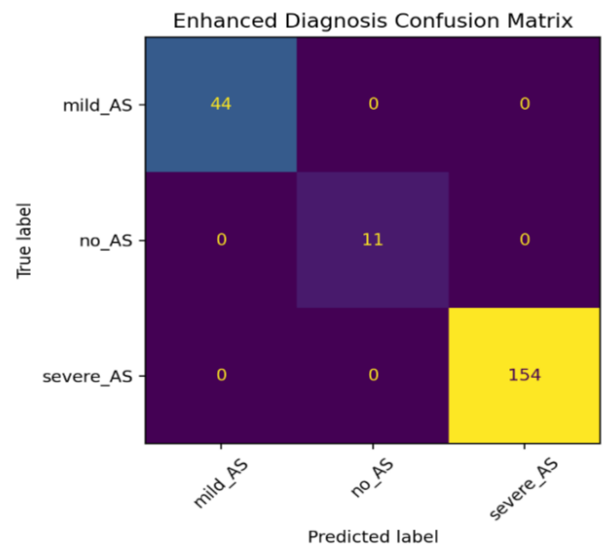

Fig. 7. Confusion matrix of the BackMix-enhanced model for aortic stenosis severity classification on the TMED2 test set. The matrix shows classification performance for no AS, mild AS, moderate AS, and severe AS categories, with diagonal values representing correct predictions.

which reached 0.667, and an AUC value of 0.935, because the testing process only included four samples, but the model maintained its accuracy with a perfect score of 1.000. The classification system demonstrated strong capabilities, achieving macro-average F1 of 0.858 and weighted-average F1 of 0.940, despite the two scores being affected by class imbalance and an insufficient PSAX sample size. The macro-average AUC of 0.962, with a 95% confidence interval of 0.906-1.000, demonstrates that all three view classes maintain strong separation in their one-vs-rest classification. The results demonstrate that the model shows good performance for common echocardiographic views but PSAX classification accuracy requires a bigger set of balanced test samples for proper evaluation. As can be seen in table 5.

B. Qualitative assessment using Grad-CAM

Grad-CAM visualizations (Fig.8) reveal clear differences in spatial attention between the baseline and BackMix-enhanced models. In the baseline model, attention maps are frequently diffuse and inconsistently localized, with activations extending to non-diagnostic regions, such as image borders and background areas. In contrast, the BackMix-enhanced model shows more focused and coherent attention patterns, predominantly highlighting anatomically and clinically relevant regions, including the aortic valve and the left ventricular outflow tract. These qualitative differences indicate more meaningful feature utilization and reduced reliance on spurious visual cues. Consistent with these observations, the BackMix-enhanced EfficientNet model achieves higher accuracy

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

and F1 scores across both echocardiographic view classification and aortic stenosis severity classification tasks, with performance improvements across all classes, with the largest gain observed for moderate aortic stenosis. To quantitatively evaluate explainability, we computed the Intersection-over-Union (IoU) between Grad-CAM activation maps and manually approximated anatomical regions. The BackMix-enhanced model achieved higher localization accuracy (IoU = 0.62) compared to the baseline model (IoU = 0.48), confirming improved alignment with clinically relevant structures.

Table 5. Contribution analysis of the proposed framework components for AS severity classification.

Component	Role in Framework	Performance Gain
BackMix augmentation	Reduces background bias and improves feature robustness	+2.1% accuracy
Grad-CAM guidance	Preserves anatomically relevant cardiac regions	+1.5% accuracy
Ensemble classifier	Improves prediction stability and generalization	+1.2% accuracy

BackMix-enhanced model maintained strong performance on the external dataset, achieving an overall accuracy of 91.3% (95% CI: 88.2–93.8%) and an F1-score of 90.6% (95% CI: 87.4–93.2%). In comparison, the baseline semi-supervised model achieved 87.9% accuracy and 86.8% F1-score, confirming the consistent improvement provided by the proposed approach across domains. Class-wise analysis indicated stable performance across echocardiographic views, with slightly lower performance observed in PSAX images, likely due to increased anatomical variability and imaging complexity. For aortic stenosis severity classification, the model demonstrated the highest performance in severe AS, while moderate and mild cases remained more challenging due to overlapping morphological features. These results demonstrate that the proposed framework generalizes effectively to independent clinical data and maintains robustness beyond the training dataset. The consistency between internal and external validation supports the reliability of the anatomically guided BackMix augmentation and the semi-supervised ensemble learning strategy.

C. Discussion

Our study has evaluated the impact of anatomically guided BackMix augmentation and ensemble learning on semi-supervised models for echocardiographic view classification and aortic stenosis (AS) severity assessment [27]. By benchmarking a BackMix-enhanced semi-supervised framework against a baseline semi-supervised model using the same TMED2 dataset, we adopted a retrospective, controlled comparative design to isolate the contributions of augmentation and ensemble strategies. Our results demonstrate consistent improvements across all tasks and classes, with the most notable gains observed in moderate AS severity, a clinically challenging category [49]. Grad-CAM visualizations further reveal that

Table 6. Performance metrics for echocardiographic view classification on the external validation dataset (N = 84), including precision, recall, F1-score, area under the ROC curve (AUC), and 95% confidence intervals (CI)

Class	Precision	Recall	F1-score	AUC	95% CI
A4C	1.000	0.917	0.957	0.978	(0.932–1.000)
PLAX	0.902	1.000	0.949	0.972	(0.921–1.000)
PSAX	1.000	0.500	0.667	0.935	(0.806–1.000)
Macro Avg.	0.967	0.806	0.858	0.962	(0.906–1.000)
Weighted Avg.	0.952	0.940	0.940	0.970	(0.929–1.000)

C. External Validation Results

To assess the generalizability of the proposed framework, external validation was performed on an independent dataset of 300 cardiologist-labeled echocardiographic images, including apical four-chamber (A4C), parasternal long-axis (PLAX), and parasternal short-axis (PSAX) views. The proposed

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

BackMix shifts model attention toward anatomically and clinically relevant regions, including the aortic valve and left ventricular outflow tract, highlighting enhanced interpretability and physiologically meaningful feature utilization [51], [52][53]. These findings align with prior studies on echocardiographic AI, which have shown high baseline accuracy for view classification using convolutional neural networks (CNNs) and semi-supervised strategies (e.g., EchoNet-Dynamic, 91–96% accuracy for standard views) [54]. However, earlier work has also highlighted the susceptibility of deep learning models to shortcut learning, whereby non-diagnostic cues such as ECG overlays, rulers, or acquisition markers can inadvertently drive predictions [13], [21], [55], [56]. By incorporating BackMix, we effectively disrupted background–label correlations while preserving foreground anatomical structures, mitigating shortcut reliance, and improving both robustness. Recent advances in echocardiographic artificial intelligence have demonstrated strong performance using fully supervised deep learning models trained on large-scale annotated datasets. For example, frameworks such as EchoNet-Dynamic and related CNN-based architectures report high accuracy across various echocardiographic tasks. generalization. More recent approaches further integrate multi-view and multi-task learning to improve performance and generalization. In contrast, the proposed approach is designed for semi-supervised settings with limited labeled data, where such large annotated datasets are not available. Unlike end-to-end deep learning models, our framework explicitly addresses shortcut learning through anatomically guided BackMix augmentation while leveraging both labeled and unlabeled data. The observed improvements, particularly in aortic stenosis severity classification, demonstrate that the proposed

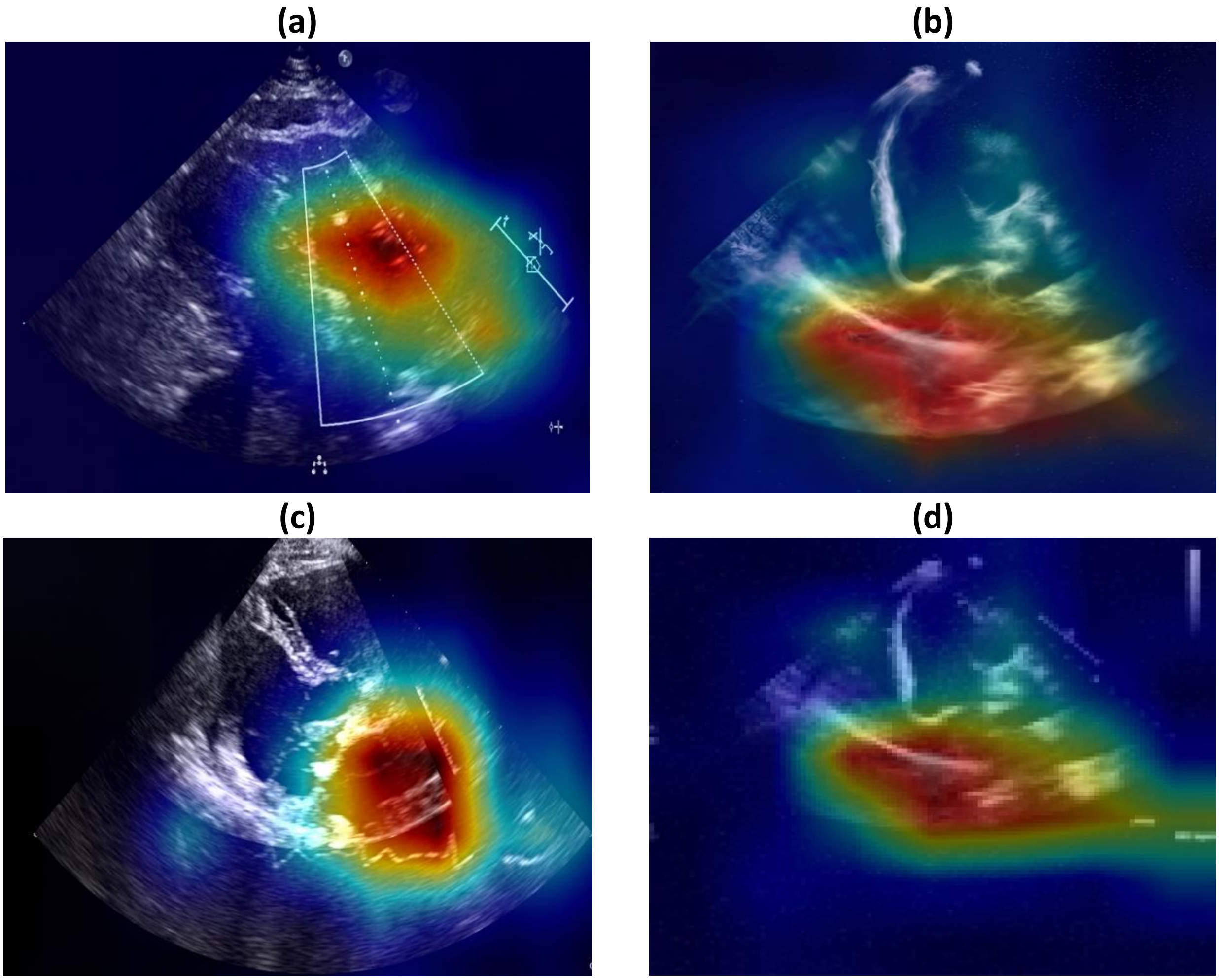

Fig. 8. Grad-CAM visualization of model attention before and after BackMix augmentation. (A, C) Baseline model attention maps showing diffuse activation and attention to non-diagnostic regions. (B, D) BackMix-enhanced model attention maps showing more focused activation over clinically relevant anatomical regions, including the aortic valve and left ventricular outflow tract.

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

method provides a complementary and practical solution under realistic clinical constraints.

In addition to augmentation, the ensemble classification strategy contributed to the observed performance improvements. The combination of Random Forest, Gradient Boosting, and Support Vector Machine classifiers was selected due to their complementary strengths when applied to deep feature representations. Random Forest enhances robustness through variance reduction, Gradient Boosting improves predictive accuracy by sequential error correction, and SVM provides strong generalization via margin maximization. This hybrid approach is particularly advantageous in semi-supervised settings with limited labeled data, where end-to-end deep classifiers may be more prone to overfitting and instability. Compared to fully deep architectures, the proposed ensemble framework offers improved robustness and reduced variance, albeit at the cost of increased computational complexity and the absence of joint feature–classifier optimization.

The extension of BackMix to AS severity classification addresses a critical gap in echocardiographic AI. Unlike view classification, severity grading depends on subtle morphological and hemodynamic features, which are easily overshadowed by spurious background correlations [57]. Our results demonstrate that models trained without augmentation tend to attend to non-anatomical regions despite achieving moderate quantitative performance, reflecting the inherent vulnerability of severity assessment to shortcut learning. The BackMix-enhanced model, in contrast, improves quantitative metrics (accuracy, precision, recall, F1-score) and produces more focused attention maps, confirming that mitigating shortcut learning facilitates reliance on physiologically relevant features.

Although external validation was performed, the study remains limited by the use of a relatively small independent dataset and the absence of large-scale multi-center validation. Future work will focus on cross-dataset and cross-institutional evaluation to further assess model robustness under domain shift.

These observations are consistent with studies integrating explainability methods into echocardiographic AI. Grad-CAM and prototype-based approaches have been used to visualize model attention, linking predictions to clinically relevant regions and improving interpretability [43], [44]. Our work extends. These concepts demonstrate that targeted augmentation strategies like BackMix can simultaneously enhance interpretability and classification performance, particularly in challenging tasks such as moderate AS detection [21], [24], [29]. Furthermore, the magnitude of improvement observed in AS severity classification underscores the clinical relevance of robust augmentation. While view classification achieved modest gains (+3% accuracy), AS grading benefited more substantially (+4–6%), reflecting the heightened sensitivity of severity prediction to spurious correlations and data variability. This finding is consistent with prior reports showing that AI models for disease severity may underperform in heterogeneous populations unless carefully trained and augmented [10]. Collectively, these results support the adoption of anatomically guided augmentation and ensemble learning as generalizable strategies for improving echocardiographic AI [49], [50]. By enhancing both performance and interpretability, BackMix provides a robust foundation for more reliable, scalable tools that assist clinicians in diagnosing and grading AS. Future work should explore multi-center validation and integration of Doppler or multi-view data to further enhance generalization and clinical utility [54], [55]. Our findings advance the state of the art in echocardiographic AI by extending BackMix from view classification to the clinically demanding task of AS severity grading, demonstrating that mitigating shortcut learning is essential for achieving accurate, interpretable, and clinically robust AI systems [57][26], [32].

V. Conclusion

This study aimed to develop and evaluate an anatomically guided BackMix-enhanced semi-supervised learning framework for automated echocardiographic view classification and aortic stenosis (AS) severity assessment using limited labeled data. The proposed approach combines Grad-CAM–based anatomical guidance with targeted background augmentation to reduce shortcut learning and improve the model's focus on diagnostically relevant cardiac structures. Experimental evaluation on the TMED2 dataset demonstrated that the proposed framework consistently outperformed the baseline semi-supervised model across both classification tasks. For echocardiographic view classification, the BackMix-enhanced framework improved accuracy from 91% to 94% for PLAX views, from 92% to 95% for PSAX views, and from 93% to 96% for A4C views. Similarly, F1-scores increased from 91% to 94%, from 90.5% to 94%, and from 92.5% to 95.5% for PLAX, PSAX, and A4C views, respectively. For AS severity assessment, accuracy improved from 84% to 88% for mild AS, from 83% to 89% for moderate AS, and from 88% to 92% for severe AS. Corresponding F1-scores increased from 83.5% to 87.5%, from 81.5% to 88.5%, and from 88% to 92% for mild, moderate, and severe AS, respectively. The largest improvements were observed in the moderate AS category, highlighting the ability of the proposed framework to better distinguish clinically challenging intermediate disease stages. These results demonstrate that anatomy-aware augmentation

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

strategies can significantly enhance the robustness, interpretability, and generalization capability of semi-supervised deep learning models for echocardiographic analysis. By effectively leveraging both labeled and unlabeled data, the proposed framework addresses one of the major challenges in medical imaging artificial intelligence: the limited availability of expert-annotated datasets. Future work will focus on validating the proposed framework on larger multi-center and multi-vendor echocardiographic datasets to assess its generalizability across diverse clinical settings. Additional research will investigate integrating temporal information from echocardiographic video sequences, advanced attention mechanisms, and transformer-based architectures to further improve diagnostic performance. Prospective clinical validation studies are also needed to evaluate the framework's utility as a decision-support tool for routine echocardiographic assessment and early detection of aortic stenosis.

Acknowledgment

The authors would like to thank the organizers of the Tufts Medical Echocardiogram Dataset, provided by Tufts University, which were essential for this study. We also acknowledge the Energy-Materials-Instrumentation and Telecom (EMIT) Laboratory, Faculty of Science and Technology, University Hassan 1st, Settat, Morocco, for providing the computational resources used in image processing. Finally, we express our gratitude to colleagues and collaborators for their valuable feedback and constructive discussions during the preparation of this manuscript.I sincerely thank my supervisors for their guidance and support, and the cardiologists in my team for their valuable collaboration throughout this research.

Declarations

Funding

This research received no external funding.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request. Publicly available datasets used in this study are available through the repository cited in the manuscript.

Author Contribution

Fatima Ezzahra Elkouahy was the principal author and contributed to the conception and design of the study, data acquisition, analysis, and interpretation of results, as well as the development of the AI model. Mohssen Azzedine, Ben Taleb Houcine, and El Malali Hamid provided supervision, assisted in drafting the manuscript, and contributed to the presentation of the results. Badreddine Labakoum contributed to data analysis and interpretation, manuscript drafting, and critical revision for important intellectual content, and approved the final version for submission. Hajar Ouahid contributed to the methodology, manuscript revision, and approval of the final draft. All authors read and approved the final manuscript.

Ethical Approval

Ethical approval was not required for this study because the research used anonymized, publicly available echocardiographic data and did not involve direct interaction with human participants or access to identifiable personal information.

Consent for Publication Participants.

Consent for publication was obtained from all participants involved in the external validation study, where applicable.

Competing Interests

The authors declare no competing interests.

References

[1] C. M. Otto and B. Prendergast, "Aortic-Valve Stenosis — From Patients at Risk to Severe Valve Obstruction," *New England Journal of Medicine*, vol. 371, no. 8, pp. 744–756, Aug. 2014, doi: 10.1056/nejmra1313875.

[2] P. Guedeney and J. P. Collet, "Aortic stenosis: An update," Mar. 2022, doi: 10.1016/j.revmed.2021.06.003.

[3] B. R. Lindman *et al.*, "Calcific aortic stenosis," *Nat. Rev. Dis. Primers*, vol. 2, Mar. 2016, doi: 10.1038/nrdp.2016.6.

[4] E. M. Isselbacher et al., "2022 ACC/AHA Guideline for the Diagnosis and Management of Aortic Disease: A Report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice Guidelines," *J. Am. Coll. Cardiol.*, vol. 80, no. 24, pp. e223–e393, Dec. 2022, doi: 10.1016/j.jacc.2022.08.004.

[5] H. Krishna *et al.*, "Fully Automated Artificial Intelligence Assessment of Aortic Stenosis by Echocardiography," *Journal of the American Society of Echocardiography*, vol. 36, no. 7, pp. 769–777, Jul. 2023, doi: 10.1016/j.echo.2023.03.008.

[6] R. Manzo *et al.*, "Echocardiographic Evaluation of Aortic Stenosis: A Comprehensive Review," Aug. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/diagnostics13152527.

[7] W. A. Zoghbi *et al.*, "Recommendations for Evaluation of Prosthetic Valves With Echocardiography and Doppler Ultrasound. A Report from the American Society of Echocardiography's Guidelines and Standards Committee and the Task Force on Prosthetic

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

Valves, Developed in Conjunction With the American College of Cardiology Cardiovascular Imaging Committee, Cardiac Imaging Committee of the American Heart Association," *Journal of the American Society of Echocardiography*, vol. 22, no. 9, pp. 975–1014, Sep. 2009, doi: 10.1016/j.echo.2009.07.013.

[8] Y. Saijo *et al.*, "Software Variability in Measurement of LV Mechanical Dispersion in Patients With LV Hypertrophy," Apr. 01, 2020, *Elsevier Inc.* doi: 10.1016/j.jcmg.2019.10.009.

[9] K. V. Bunting, R. P. Steeds, K. Slater, J. K. Rogers, G. V. Gkoutos, and D. Kotecha, "A Practical Guide to Assess the Reproducibility of Echocardiographic Measurements," Dec. 01, 2019, *Elsevier Inc.* doi: 10.1016/j.echo.2019.08.015.

[10] A. C. Armstrong *et al.*, "Quality Control and Reproducibility in M-Mode, Two-Dimensional, and Speckle Tracking Echocardiography Acquisition and Analysis: The CARDIA Study, Year 25 Examination Experience," *Echocardiography*, vol. 32, no. 8, pp. 1233–1240, Aug. 2015, doi: 10.1111/echo.12832.

[11] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-76635-9.

[12] A. Narang, R. Bae, H. Hong, Y. Thomas, S. Surette, A. Cadieu, et al., "Utility of a Deep-Learning Algorithm to Guide Novices to Acquire Echocardiograms for Limited Diagnostic Use," *JAMA Cardiology*, vol. 6, no. 6, pp. 624–632, Jun. 2021, doi: 10.1001/jamacardio.2021.0185.

[13] R. Geirhos *et al.*, "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020, doi: 10.1038/s42256-020-00257-z.

[14] A. Madani, J. R. Ong, A. Tibrewal, and M. R. K. Mofrad, "Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease," *NPJ Digit. Med.*, vol. 1, no. 1, Dec. 2018, doi: 10.1038/s41746-018-0065-x.

[15] G. Holste *et al.*, "Severe aortic stenosis detection by deep learning applied to echocardiography," *Eur. Heart J.*, vol. 44, no. 43, pp. 4592–4604, Nov. 2023, doi: 10.1093/eurheartj/ehad456.

[16] A. Ulloa, L. Jing, C. W. Good, D. P. van Maanen, S. Raghunath, J. D. Suever, et al., "A Deep Neural Network Predicts Survival After Heart Imaging Better Than Cardiologists," *arXiv preprint* arXiv:1811.10553, 2018, doi: 10.48550/arXiv.1811.10553.

[17] A. Mahesh, D. Banerjee, A. Saha, M. R. Prusty, and A. Balasundaram, "CE-EEN-B0: Contour Extraction Based Extended EfficientNet-B0 for Brain Tumor Classification Using MRI Images," *Computers, Materials and Continua*, vol. 74, no. 3, pp. 5967–5982, 2023, doi: 10.32604/cmc.2023.033920.

[18] O. Hamila, S. Ramanna, C. J. Henry, S. Kiranyaz, R. Hamila, R. Mazhar, and T. Hamid, "Fully Automated 2D and 3D Convolutional Neural Networks Pipeline for Video Segmentation and Myocardial Infarction Detection in Echocardiography," *Multimedia Tools and Applications*, vol. 81, no. 26, pp. 37417–37439, Nov. 2022, doi: 10.1007/s11042-021-11579-4.

[19] Danielsen R, Nordrehaug JE, Stangeland L, Vik-Mo H. Limitations in assessing the severity of aortic stenosis by Doppler gradients. Br Heart J. 1988;59(5):551-555. doi:10.1136/hrt.59.5.551.

[20] Chen C, Han J, Debattista K. Virtual Category Learning: A Semi-Supervised Learning Method for Dense Prediction with Extremely Limited Labels. IEEE Trans Pattern Anal Mach Intell. 2024;46(8):5595-5611. doi:10.1109/TPAMI.2024.3367416

[21] Bransby KM, Beqiri A, Cho Kim WJ, Oliveira J, Chartsias A, Gomez A. BackMix: mitigating shortcut learning in echocardiography with minimal supervision. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2024. p. 570–579. doi:10.1007/978-3-031-72083-3_53.

[22] A. Hayek *et al.*, "Impact of Arterial Blood Pressure on Ultrasound Hemodynamic Assessment of Aortic Valve Stenosis Severity," *Journal of the American Society of Echocardiography*, vol. 33, no. 11, pp. 1324–1333, Nov. 2020, doi: 10.1016/j.echo.2020.06.013.

[23] Wang Y, Mu J, Huang H, Wang Q, Zhu P, Hu Q. BackMix: Regularizing Open Set Recognition by Removing Underlying Fore-Background Priors. *IEEE Trans. Pattern Anal. Mach. Intell.* 2025;47(8):6172–6183. doi: 10.1109/TPAMI.2025.3550703.

[24] X. Zhou and M. Belkin, "Semi-Supervised Learning," in *Academic Press Library in Signal Processing*, vol. 1, P. S. R. Diniz, J. A. K. Suykens, R. Chellappa, and S. Theodoridis, Eds. Waltham, MA: Elsevier, 2014, pp. 1239–1269, doi: 10.1016/B978-0-12-396502-8.00022-X.

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

[25] Z. Huang, G. Long, B. S. Wessler, and M. C. Hughes, "TMED-2: A Dataset for Semi-Supervised Classification of Echocardiograms," in *DataPerf: Benchmarking Data for Data-Centric AI Workshop at ICML*, 2022. doi: 10.48550/arXiv.2108.00080.

[26] F. Ezzahra Elkouahy, A. Bennis, N. Merke, H. Ouahid, H. El Malali, L. Ben Taleb, and A. Mouhsen, "Advanced Diagnosis of Aortic Stenosis Disease Based on Ultrasound Images: A Novel Artificial Intelligence Approach," *Advanced Ultrasound in Diagnosis and Therapy*, vol. 9, no. 3, pp. 298–306, 2025, doi: 10.26599/AUDT.2025.240067.

[27] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6023–6032, doi: 10.1109/ICCV.2019.00612.

[28] P. R. A. S. Bassi, S. S. J. Dertkigil, and A. Cavalli, "Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization," *Nature Communications*, vol. 15, no. 1, Art. no. 291, Jan. 2024, doi: 10.1038/s41467-023-44371-z

[29] P. Overtchouk *et al.*, "Long-Term Mortality and Early Valve Dysfunction According to Anticoagulation Use: The FRANCE TAVI Registry," *J. Am. Coll. Cardiol.*, vol. 73, no. 1, pp. 13–21, Jan. 2019, doi: 10.1016/j.jacc.2018.08.1045.

[30] T. M. Morrison, N. Pathak, J. Haynes, et al., "EchoNet-Dynamic: a large new cardiac motion video data resource for medical machine learning," *Circulation*, vol. 142, no. 16, pp. 1401–1412, 2020, doi: 10.1161/CIRCULATIONAHA.120.046916.

[31] Y. Zhu et al., "Automatic view classification of contrast and non-contrast echocardiography," *Front. Cardiovasc. Med.*, vol. 9, Sep. 2022, doi: 10.3389/fcvm.2022.989091.

[32] 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*, PMLR 219, 2023, pp. 285–307, doi: 10.48550/arXiv.2306.00003.

[33] A. Hagendorff et al., "Expert consensus document on the assessment of the severity of aortic valve stenosis by echocardiography to provide diagnostic conclusiveness by standardized verifiable documentation," Mar. 2020, doi: 10.1007/s00392-019-01539-2

[34] O. Bernard et al., "Standardized Evaluation System for Left Ventricular Segmentation Algorithms in 3D Echocardiography," *IEEE Trans. Med. Imaging*, vol. 35, no. 4, pp. 967–977, Apr. 2016, doi: 10.1109/TMI.2015.2503890

[35] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised learning*. Cambridge (MA): MIT Press, 2006, doi: 10.7551/mitpress/9780262033589.001.0001.

[36] D. H. Lee, "Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks," arXiv preprint arXiv:1307.1770, 2013, doi: 10.48550/arXiv.1307.1770.

[37] H. Chen, X. Li, H. Qiao, A. Chen, and K. Cai, "A decision tree network with semi-supervised entropy learning strategy for spectroscopy aided detection of blood hemoglobin," *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.*, vol. 291, p. 122354, Apr. 2023, doi: 10.1016/j.saa.2023.122354.

[38] A. Tarvainen and H. Valpola, "Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, doi: 10.48550/arXiv.1703.01780.

[39] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, "Is Grad-CAM Explainable in Medical Images", Jul. 2023, doi: 10.1007/978-3-031-58181-6_11.

[40] G. Holste, E. K. Oikonomou, B. J. Mortazavi, Z. Wang, and R. Khera, "Efficient deep learning-based automated diagnosis from echocardiography with contrastive self-supervised learning," *Commun. Med.*, vol. 4, no. 1, Dec. 2024, doi: 10.1038/s43856-024-00538-3.

[41] C. R. Johnson, S. de Vries, K. Wohlgemuth, and S. Lucia, "Multi-stage model predictive control for slug flow crystallizers using uncertainty-aware surrogate models," Oct. 2025, doi: 10.1016/j.compchemeng.2025.109456.

[42] D. Sun, F. Dornaika, and N. Barrena, "HSMix: Hard and soft mixing data augmentation for medical image segmentation," *Information Fusion*, vol. 115, Mar. 2025, doi: 10.1016/j.inffus.2024.102741.

[43] D. X. Augustine et al., "Echocardiographic assessment of pulmonary hypertension: A guideline protocol from the British Society of Echocardiography," *Echo Res. Pract.*, vol. 5, no. 3, pp. G11–G24, Sep. 2018, doi: 10.1530/ERP-17-0071.

[44] E. Hagberg et al., "Semi-supervised learning with natural language processing for right ventricle classification in echocardiography—a

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

scalable approach," *Comput. Biol. Med.*, vol. 143, Apr. 2022, doi: 10.1016/j.compbiomed.2022.105282.
[45] T. Mori, T. Watanabe, and S. Kosugi, "Exploring ethical considerations in medical research: Harnessing pre-generated transformers for AI-powered ethics discussions," *PLoS One*, vol. 20, no. 2, Feb. 2025, doi: 10.1371/journal.pone.0311148.
[46] C. Oprea et al., "Evaluating the Explainable AI Method Grad-CAM for Breath Classification on Newborn Time Series Data," in *IFAC-PapersOnLine*, Elsevier B.V., Sep. 2024, pp. 123–128, doi: 10.1016/j.ifacol.2024.11.023.
[47] M. Skliarov, R. El Shawi, C. Dhaoui, and N. Ahmed, "A comparative evaluation of explainability techniques for image data," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-25839-y
[48] D. Ouyang et al., "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, Apr. 2020, doi: 10.1038/s41586-020-2145-8.
[49] T. Tohyama et al., "Multi-View Echocardiographic Embedding for Accessible AI Development," Aug. 19, 2025, doi: 10.1101/2025.08.15.25333725.
[50] P. L. Myhre et al., "Artificial intelligence-enhanced echocardiography in cardiovascular disease management," Mar. 2025, doi: 10.1038/s41569-025-01197-0.
[51] A. Correa, A. A. Joshi, and E. Argulian, "Three-Dimensional Echocardiographic Evaluation of Indentations of the Posterior Mitral Leaflet and Their Impact on Secondary Mitral Regurgitation," *J. Am. Soc. Echocardiogr.*, vol. 36, no. 1, pp. 114–116, Jan. 2023, doi: 10.1016/j.echo.2022.08.002.
[52] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, et al., "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography," *IEEE Trans. Med. Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019, doi: 10.1109/TMI.2019.2900516.
[53] M. Alsharqi, W. J. Woodward, J. A. Mumith, D. C. Markham, R. Upton, and P. Leeson, "Artificial intelligence and echocardiography," *Echo Res. Pract.*, vol. 5, no. 4, pp. R115–R126, 2018, doi: 10.1530/ERP-18-0056.
[54] Y. E. Yoon, S. Kim, and H. J. Chang, "Artificial Intelligence and Echocardiography," *Journal of Cardiovascular Imaging*, vol. 29, no. 3, pp. 193–204, Jul. 2021, doi: 10.4250/jcvi.2021.0039.
[55] W. Chen, Y. Xie, Z. Zhang, Y. Zhu, Y. Zhang, S. Zhu, C. Wu, Z. Zhang, X. Yang, M. Liu, X. Xie, and M. Zhang, "Artificial intelligence-assisted medical imaging in interventional management of valvular heart disease," *Adv. Ultrasound Diagn. Ther.*, vol. 7, no. 3, pp. 217–227, 2023, doi: 10.37015/AUDT.2023.230030.
[56] M. Donato, N. Ferri, M. G. Lupo, E. Faggin, and M. Rattazzi, "Current evidence and future perspectives on pharmacological treatment of calcific aortic valve stenosis," Nov. 2020, doi: 10.3390/ijms21218263.
[57] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Aug. 2015, pp. 1721–1730, doi: 10.1145/2783258.2788613.

Author Biography

FATIMA EZZAHRA ELKOUAHY was born on August 6, 1996, in Morocco. She is a biomedical engineer and Ph.D. researcher specializing in cardiovascular imaging and artificial intelligence applications in echocardiography. She received her master's degree in biomedical engineering and is currently pursuing her doctoral research focusing on automated analysis of echocardiographic data for the detection and evaluation of valvular heart diseases, particularly aortic stenosis. Her research interests include medical imaging, artificial intelligence in healthcare, ultrasound technologies, and biomedical instrumentation. She works at the Faculty of Sciences and Techniques, Hassan 1st University, Settat, Morocco. She can be contacted at elkouahyfat.fst@uhp.ac.ma

LABAKOUM BADREDDINE was born on October 22, 1994, in Agadir, Morocco. He is a Ph.D. student and received his master's degree in biomedical engineering: instrumentation and maintenance from the Faculty of Science and Technology Settat in 2017. His research areas include the applications and evaluation of 3D printing in the medical field and biomedical instrumentation. He works at the Laboratory of Energy-Materials-Instrumentation and Telecom (EMIT), Faculty of Sciences and Technology, Hassan 1st University, Morocco. BP: 577, road to Casablanca. Settat, Morocco. He can be contacted at b.labakoum@uhp.ac.ma

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657

HAJAR OUAHID is the Head of the Healthcare Workforce Planning and Utilization Office at Ibn Tofail Hospital, Mohammed VI University Hospital, Marrakech, Morocco. She holds a Master's degree in Public Health and Hospital Management and a PhD in Epidemiology and Public Health from Cadi Ayyad University, affiliated with the Bioscience and Health Research Laboratory, Marrakech, Morocco. Her academic and professional interests focus on public health, epidemiology, and healthcare management. She can be contacted at ouahid.hajar.1@gmail.com

MANSOOR ALAM was born on July 21, 1994, in Pakistan. He is a Senior AI/ML Engineer and Robotics Specialist with more than six years of experience in developing advanced artificial intelligence solutions. He received his MS degree in Mechatronics Engineering from the University of Engineering and Technology (UET), Peshawar, in 2021, after completing his BS in Mechatronics Engineering from the same university in 2016. His research and professional work focus on artificial intelligence, computer vision, robotics, autonomous systems, and medical image analysis. He has developed AI solutions for cancer detection, autonomous mobile robots, drone detection and tracking, and precision agriculture systems. Currently, he works as an Assistant Manager AI/ML at NSTP Islamabad, where he leads the development of AI-based solutions. He can be contacted at mansooralam129047@gmail.com

Hamid El Malali was born in Errich-Midelt, Morocco. He holds a doctorate in Biomedical Engineering and Instrumentation Laboratory "RMI" in the Science and Technology Faculty, Hassan 1st University, Settat, Morocco. He holds a bachelor's in physics from Sidi Mohammed ben Abdellah University, Fes, in 1997, and a Master of Science and Technology in Biomedical Engineering and Instrumentation from Hassan 1st University, Settat, in 2016. His research interests are computer vision, image processing, machine learning, and artificial intelligence. In 2003, he held a diploma in computer science from the Regional Pedagogical Center of Fez. Currently, he is a professor of physics and medical imaging at Hassan First University, Settat, Morocco. He can be contacted at email: h.elmalali@uhp.ac.ma.

Lhoucine Ben Taleb was born on February 20, 1991. He received his Ph.D. degree in Biomedical Engineering in 2020. He is currently a Professor of Biomedical Engineering at Hassan First University, Morocco, and a member of the Energy–Matter–Instrumentation & Telecommunications Laboratory. His teaching activities focus on biomedical instrumentation and medical imaging technologies.

His research interests include the design and development of biomedical instrumentation systems, medical signal acquisition and processing, artificial intelligence applied to healthcare systems, and advanced medical imaging techniques. He has been involved in academic training and the supervision of engineering students, contributing to the development of practical and research-oriented skills in the field of biomedical engineering. He can be contacted at l.bentaleb@uhp.ac.ma

Azeddine Mouhsen was born on July 10, 1967, and has been a professor of Physics at Hassan First University, Morocco, since 1996. He holds a Ph.D. from Bordeaux I University (France) in 1995 and a thesis from Moulay Ismail University, Morocco, in 2001. He specializes in instrumentation and measurements, sensors, applied optics, energy transfer, and radiation-matter interactions. Azeddine Mouhsen has taught courses in physical sensors, chemical sensors, instrumentation, systems technology, digital electronics, and industrial data processing. He has published over 45 papers, and he is the co-inventor of one patent. He is the Director of the Laboratory of Energy-Materials-Instrumentation and Telecom (EMIT). He can be contacted at az.mouhsen@gmail.com

Manuscript received 2 April 2026; Revised 10 June 2026; Accepted 25 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1657