Computer Methods and Programs in Biomedicine 245 (2024) 108037

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

MV-MS-FETE: Multi-view multi-scale feature extractor and transformer encoder for stenosis recognition in echocardiograms

Danilo Avola [a], Irene Cannistraci [a], Marco Cascio [a], Luigi Cinque [a], Alessio Fagioli [a,*], Gian Luca Foresti [b], Emanuele Rodolà [a], Luciana Solito [a]

[a] Department of Computer Science, Sapienza University, Via Salaria 113, 00185, Rome, Italy
[b] Department of Mathematics, Computer Science and Physics, University of Udine, 33100 Udine, Italy

ARTICLE INFO

Keywords:
Aortic stenosis recognition
Echocardiograms
Multi-view
Feature extraction
Transformers

ABSTRACT

Background: aortic stenosis is a common heart valve disease that mainly affects older people in developed countries. Its early detection is crucial to prevent the irreversible disease progression and, eventually, death. A typical screening technique to detect stenosis uses echocardiograms; however, variations introduced by other tissues, camera movements, and uneven lighting can hamper the visual inspection, leading to misdiagnosis. To address these issues, effective solutions involve employing deep learning algorithms to assist clinicians in detecting and classifying stenosis by developing models that can predict this pathology from single heart views. Although promising, the visual information conveyed by a single image may not be sufficient for an accurate diagnosis, especially when using an automatic system; thus, this indicates that different solutions should be explored.

Methodology: following this rationale, this paper proposes a novel deep learning architecture, composed of a multi-view, multi-scale feature extractor, and a transformer encoder (MV-MS-FETE) to predict stenosis from parasternal long and short-axis views. In particular, starting from the latter, the designed model extracts relevant features at multiple scales along its feature extractor component and takes advantage of a transformer encoder to perform the final classification.

Results: experiments were performed on the recently released Tufts medical echocardiogram public dataset, which comprises 27,788 images split into training, validation, and test sets. Due to the recent release of this collection, tests were also conducted on several state-of-the-art models to create multi-view and single-view benchmarks. For all models, standard classification metrics were computed (e.g., precision, F1-score). The obtained results show that the proposed approach outperforms other multi-view methods in terms of accuracy and F1-score and has more stable performance throughout the training procedure. Furthermore, the experiments also highlight that multi-view methods generally perform better than their single-view counterparts.

Conclusion: this paper introduces a novel multi-view and multi-scale model for aortic stenosis recognition, as well as three benchmarks to evaluate it, effectively providing multi-view and single-view comparisons that fully highlight the model's effectiveness in aiding clinicians in performing diagnoses while also producing several baselines for the aortic stenosis recognition task.

1. Introduction

Aortic stenosis (AS) is one of the most prevalent valve diseases in developed countries [1]. It consists of the aortic valve narrowing, which restricts the blood flow from the left ventricle to the aorta [2]. Without this pathology, left ventricle contractions can easily move blood through the valve and into the aorta to eventually reach the rest of the body. During the left ventricle expansions, the aortic valve remains closed to prevent backward blood flow from the aorta. As the aortic valve becomes narrowed or constricted with AS, the left ventricle must generate a higher pressure with each contraction to move blood forward into the aorta [3]. In its early stages, the left ventricle compensates for this increased pressure by thickening its muscular walls (i.e., myocardial hypertrophy). In the later stages, the left ventricle dilates,

* Corresponding author.
E-mail address: fagioli@di.uniroma1.it (A. Fagioli).

https://doi.org/10.1016/j.cmpb.2024.108037
Received 13 February 2023; Received in revised form 27 December 2023; Accepted 15 January 2024
Available online 17 January 2024

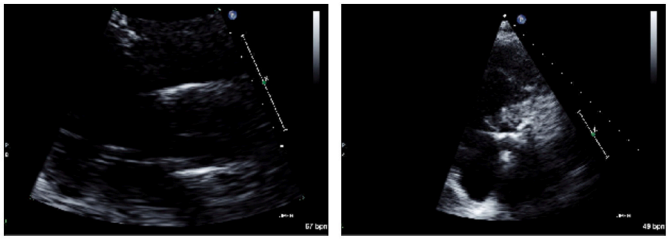


Fig. 1. Example of (left) PLAX and (right) PSAX views.

the wall thins, and the systolic function deteriorates, resulting in an impaired ability to pump blood forward and potentially causing backward blood leakages (i.e., regurgitation) [4]. While the effects of AS range from mild to severe, the long asymptomatic period experienced by many patients makes early detection challenging, which is aggravated by the most common cause of AS being age-related progressive calcification [5]. In fact, some individuals may not exhibit symptoms for many years, but once symptoms begin, mortality rates increase rapidly [1,6]. For this reason, screening is essential to avoid irreversible disease progression and otherwise preventable death [7]. To assess the stenosis, cardiologists can manually examine key frames acquired using different methods such as electrocardiogram (ECG), X-ray angiography (XRA), computerized tomography (CT), coronary CT angiography (CCTA) and echocardiograms (echos). The latter is the standard means for diagnosing and evaluating AS [8,9], being a non-invasive procedure with minimal risk for the patient that can be used to examine their heart. During the echo procedure, a transducer placed on the patient's chest produces sound waves that move through the body and bounce off structures in the heart, resulting in live videos of the heart walls and valves. Different views of the heart are generated depending on the angle and location of the transducer. The heart's four main cardiac ultrasound views are the parasternal long-axis (PLAX), parasternal short-axis (PSAX), Apical, and subxiphoid (subcostal). The two echocardiogram views in which the aortic valve is visible are the PLAX and the PSAX. The first one is obtained with the transducer image marker directed toward the patient's right ear and the sound beam directed to the spine. The second one is obtained by rotating the transducer 90 degrees clockwise with respect to PLAX. An example of these views is shown in Fig. 1. Although clinicians can diagnose stenosis by analyzing these images, examining medical results generally takes considerable time and increases the doctor's workload. In addition, intra- and inter-observer variations caused by other tissues, camera movements, and uneven lighting can significantly affect the visual inspection of aortic stenosis [1,10], indicating that different approaches are required to help clinicians make their diagnoses.

An immediate solution to the issues mentioned above lies in deep learning (DL) algorithms, which can provide a comprehensive and automated diagnosis of medical images. As a matter of fact, these methods are effectively being applied to diagnose different illnesses such as malignant thyroid nodules [11–13], COVID-19 [14–16], and others [17–19]. Indeed, DL algorithms are also used for AS prediction. For instance, the scheme in [20] proposes an automated ensemble approach leveraging self-supervised learning (SSL) of PLAX videos and convolutional neural networks (CNNs). Here, the ensemble model is capable of identifying severe AS from raw single-view 2D echos. Similarly, the work presented in [21] introduces a model based on a faster region based convolutional neural network (R-CNN) [22] to detect the aortic valve in video sequences acquired by the transducer. To classify the PLAX state, features are obtained from each frame of the sequence, concatenated, and subsequently processed by a temporal CNN model to predict whether or not there is stenosis. The authors of [23] describe a multi-task training procedure to predict AS severity and key parameters used in clinical AS assessment. This involves an architecture previously validated as the gold standard for video classification and regression tasks with echo [24], i.e., a residual network (ResNet) [25] with spatio-temporal convolutions. A different approach is reported in [26], presenting the development of a screening tool to identify patients affected by moderate to severe AS. This approach is based on an architecture inspired by DenseNet [27] and also takes into account the patient's age and sex for prediction. Another example of automatic AS recognition is proposed in [28], where the authors classify this condition in ECG images. Specifically, they first train a CNN on manually annotated data to extract relevant features from input images and learn to distinguish aortic stenosis. Then, they use the gradient-weighted class activation mapping (Grad-CAM) algorithm to analyze the output of the trained network and detect feature areas in the early time range of the one-beat ECG image. Finally, both schemes presented in [29] and [30] address the classification of four major cardiac diseases, including AS. In [29], the authors developed an embedded low-cost diagnostic tool for both medical professionals and personal use at home. This tool comprises two modalities; the first is based on a 1D-CNN for the classification of heart sounds, i.e., raw phonocardiogram (PCG) signals, and the second exploits a 2D-CNN to classify the spectrogram of the given/recorded heart sound signals. In [30], the authors designed a lightweight end-to-end convolutional recurrent neural network (CRNN) architecture that consists of two phases: representation learning and sequence residual learning. In the first phase, three parallel CNNs extract efficient time-invariant features from PCG, while in the second phase, a combination of bidirectional-long-short term memory (LSTM) and skip connections extract temporal features. A common aspect of these DL approaches is that they infer stenosis using a single image. However, the visual information conveyed by this view may not be sufficient to provide the clinician with an accurate diagnosis. In fact, even humans might have to examine more than a single image before correctly diagnosing an illness. Indeed, combining information from multi-view images has been crucial for improving the accuracy and robustness of automated methods in diagnosing several diseases [31–35]. While this modality is already being explored for other pathologies, existing methods addressing the AS recognition in echocardiograms are still based on the single-view, suggesting that further improvements might be achieved by investigating a multi-view approach.

Inspired by the results obtained by [36] using machine learning algorithms in AS diagnosis and by recent multi-view deep learning approaches analyzing other organs and pathologies [37,38], this study introduces a novel architecture to predict AS from multi-view echo images, i.e., the multi-view, multi-scale feature extractor and transformer encoder (MV-MS-FETE). In more detail, starting from PLAX and PSAX views, two parallel feature extractors derive feature maps at multiple scales. These maps are concatenated scale-wise and fed to a patch embedding module to generate a latent representation of the analyzed echos, which is then given as input to a transformer encoder predicting whether the patient is suffering from AS. Experiments evaluating the model were performed on a recently released public collection, i.e., the Tufts medical echocardiogram dataset (TMED) [36]. To the best of our knowledge, there are currently no other works addressing this dataset; therefore, an extensive analysis was conducted to produce multi-view and single-view benchmarks using well-known models such as SqueezeNet [39], ResNet101 [25], MobileNet V3 [40], EfficientNet B0 [41], and VGG19 [42]. The obtained results show that the proposed architecture outperforms the other models in both accuracy and F1-score, demonstrating its effectiveness in AS recognition.

Summarizing, the main contributions of this paper are:

- Exploring, for the first time in the literature to the best of our knowledge, a multi-view approach for diagnosing aortic stenosis in echocardiograms;
- Designing a novel architecture (MV-MS-FETE) that integrates multi-view images (PLAX and PSAX) and generates multi-scale features via parallel feature extractors, utilizing a transformer encoder to enhance performance in AS recognition;

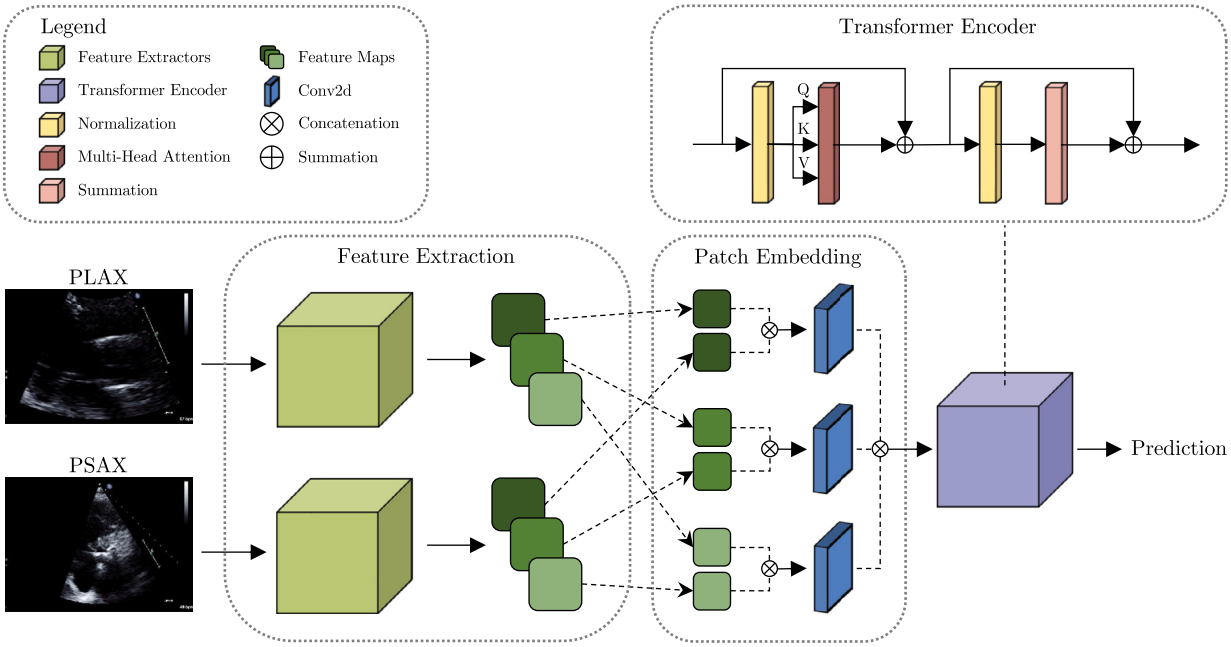


Fig. 2. MV-MS-FETE scheme overview. The model comprises two parallel feature extractors that generate multi-scale feature maps for both PLAX and PSAX views. A scale-wise map concatenation is then performed and patch embeddings are generated through 2D convolutions. A transformer encoder is finally tasked with the classification of these embeddings to recognize AS.

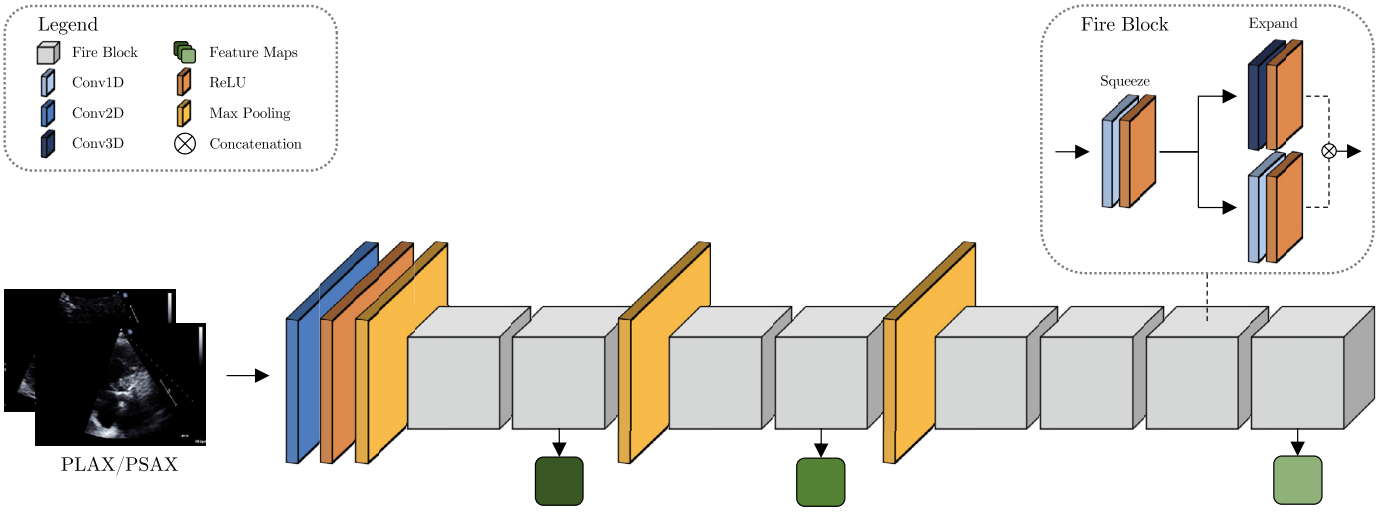


Fig. 3. Feature extractor scheme overview. The model generates multi-scale features through a series of fire blocks interleaved by max pooling operations that reduce the scale along the architecture.

- Establishing benchmarks for both single-view and multi-view aortic stenosis recognition on the recently released Tufts medical echocardiogram public dataset (TMED);
- Setting a new standard in accuracy and F1-score for multi-view AS recognition on the TMED public collection with the proposed MV-MS-FETE model.

2. Materials and methods

This section focuses on the introduction of the MV-MS-FETE architecture, which is inspired by the multi-view strategy presented in the work of Sun et al. [33], devised for mammographic image classification. Following this rationale, the model presented in this paper applies the multi-view paradigm to the classification of AS in echocardiograms and extends the idea described in [33] with multi-scale features and a transformer encoder. Multi-scale feature extraction is pivotal in capturing a comprehensive range of information from medical images, as it allows the model to recognize patterns and anomalies at various spatial resolutions. In medical imaging, different scales can reveal critical details, ranging from broad anatomical structures to minute pathological changes, and have been shown to enhance the accuracy of disease detection and classification [43]. Particularly, the extraction of features at multiple scales is crucial in echocardiography, where the variability in cardiac structures and the presence of pathological signs like valve calcification or ventricular hypertrophy demand analysis at different levels of granularity [44]. Regarding transformer models, they have recently gained traction in the field of medical image analysis due to their ability to model long-range dependencies and capture global contextual information [45,46]. The integration of a transformer encoder within the MV-MS-FETE architecture allows the encoder's self-attention mechanism to weigh the importance of different areas in an echocardiogram, thus enabling a more holistic interpretation of cardiac structures and

functions [47]. This characteristic is particularly beneficial for AS classification, where the assessment of cardiac morphology requires careful consideration of spatial relationships across the heart [48]. By leveraging multi-scale feature extractors and applying a transformer encoder, the MV-MS-FETE architecture can adaptively focus on relevant spatial contexts and effectively synthesize and analyze the complex patterns present in multi-view echocardiograms, leading to an improved diagnostic capability for AS. To enact these components, the MV-MS-FETE architecture takes as input two echo views (PLAX and PSAX) of a single patient and analyzes them through two parallel CNNs acting as feature extractors, tasked with generating multi-scale features, as reported in Section 2.1. The extracted features are then merged scale-wise, flattened, and concatenated by the patch embedding module described in Section 2.2. Finally, the model predicts the presence of aortic stenosis for the given patient from these embeddings, using the transformer encoder presented in Section 2.3. An overview of the proposed architecture is summarized in Fig. 2.

2.1. Feature extraction module

The first component of the proposed methodology extracts multi-scale feature maps through two parallel feature extraction networks that take as input a PLAX and a PSAX image, respectively. Both networks are uniformly structured, conforming to the architecture depicted in Fig. 3. In detail, the input heart view of either network is first fed to a 2D convolutional layer that outputs an activation map. A ReLU function is applied on this map to remove all negative values from the matrix while retaining the other ones. Then, a max pooling operation is used to reduce the spatial size of the map before generating the multi-scale features. These features are obtained via a series of eight fire blocks interleaved by two max pooling operations placed after the second and fourth modules to enable the multiple-scale analysis. Regarding the fire block, it is a building unit for CNNs, defined in the SqueezeNet [39] model, and has been found to be effective at generating features. It comprises a *squeeze* convolutional layer, which uses a 1x1 kernel, followed by an *expand* convolutional layer, composed of two paths using a 1x1 and a 3x3 kernel, respectively. With this structure, the model extracts three feature maps at different scales for a given input $view \in \{\text{PLAX, PSAX}\}$: X_1^{view} with shape $128 \times 15 \times 15$ at the second fire block, X_2^{view} with shape $256 \times 7 \times 7$ at the fourth fire block, and X_3^{view} with shape $512 \times 3 \times 3$ at the eight, and last, fire block. The resulting multi-scale feature maps derived by the parallel feature extractors, i.e., $X_{maps} = \{X_1^{\text{PLAX}}, X_2^{\text{PLAX}}, X_3^{\text{PLAX}}, X_1^{\text{PSAX}}, X_2^{\text{PSAX}}, X_3^{\text{PSAX}}\}$, are then given as input to the following component of the proposed model, i.e., the patch embedding module.

2.2. Patch embedding module

The second component of the proposed architecture is entrusted with organizing the feature maps X_{maps} received by the feature extraction module so that the transformer encoder can use them to classify AS. This step is crucial to the correct implementation of the third component as the transformer architecture requires as input a sequence of patch embeddings with fixed dimension D. The first step to prepare this data is to merge the received maps scale-wise as follows:

$$X_i^{\otimes} = X_i^{PLAX} \otimes X_i^{PSAX}, \tag{1}$$

where $\otimes$ is the concatenation operation and $i \in [1, 2, 3]$. The resulting features have a shape of $X_1^{\otimes} = 256 \times 15 \times 15$, $X_2^{\otimes} = 512 \times 7 \times 7$, and $X_3^{\otimes} = 1024 \times 3 \times 3$, respectively.

Upon merging the multi-scale features, the patch embedding module must reshape them so that these representations can be concatenated together. The reasoning behind this procedure is twofold. Firstly, the standard transformer architecture requires a 1D sequence of token embeddings as input since it was devised for the natural language processing (NLP) task. Secondly, the extracted feature maps have different shapes and can only be concatenated when brought to a common size. To address this issue, the merged feature maps $X_i^{\otimes}$ are flattened in accordance with Dosovitskiy et al. [46]. However, differing from their work that uses a trainable linear projection, the proposed method implements a 2D convolutional layer for each concatenated feature map, with a kernel size k equal to the map dimension. For instance, map $X_1^{\otimes}$ has a shape of $256 \times 15 \times 15$, therefore, it will use a kernel $k = 15$ to flatten the map. Through this configuration, the model can automatically compute the mapping function used to extract the patch embeddings at training time. This ensures that each patch, i.e., one per feature map, has the same dimensionality D as required by the transformer, where D corresponds to the convolution output filters number. Note that the last step to prepare the embeddings for the transformer encoder entails their concatenation, resulting in an embedding P with shape $3 \times D$, where each patch represents a specific feature map extracted by the first module.

2.3. Transformer encoder module

The last module of the proposed model is a transformer encoder that predicts whether a patient has stenosis or not. The implementation follows the encoder design by [46] and [45], which contains two key components, namely, feed-forward layers in the form of a multi-layer perceptron (MLP), and a multi-headed attention (MHA) layer, that applies the attention mechanism in parallel. In more detail, the MHA divides its Query, Key, and Value parameters, i.e., Q, K, and V in Fig. 2, into N segments and passes each segment independently through separate heads. The results are then combined to produce a final attention score. Formally, this process is represented as:

$$MultiHead(Q, K, V) = \otimes(head_1, \dots, head_h)W^O, \tag{2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V), \tag{3}$$

where $\otimes$ is again the concatenation operation, while W_i^Q, W_i^K, W_i^V, and W^O are all learnable parameter matrices.

Regarding Equation (3), it represents a scaled attention computed via a dot-product [45] where the input consists of queries and keys of dimension d_k, and values of dimension d_v. Specifically, dot products between the query and all keys are first computed. These products are then divided by $\sqrt{d_k}$, and a $Softmax$ function is subsequently applied to derive the values' weights. Moreover, this attention function is computed simultaneously on the query set contained in matrix Q, using keys and values that are packed into matrices K and V. Formally, the outputs matrix W^O is computed via the following equation:

$$Attention(Q, K, V) = Softmax\left(\frac{Q_i^K}{\sqrt{d_k}}\right)V. \tag{4}$$

The resulting outputs are concatenated and linearly transformed through a MLP to perform the AS classification task. It is important to note that the MHA enables the model to focus on different parts of the sequence, effectively connecting the multi-scale features generated by the patch embedding module.

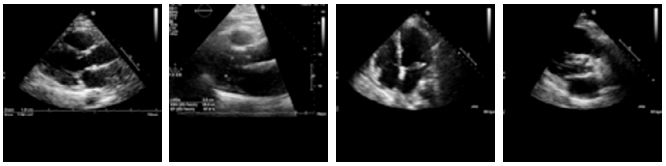
Finally, the entire model is trained to minimize the binary cross entropy (BCE) loss:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i), \tag{5}$$

where y and $\hat{y}$ represent the expected and predicted category, respectively.

3. Experimental results and discussion

This section assesses the proposed method's effectiveness in AS recognition. The public collection used to evaluate the model is first



(a) (b) (c) (d)

Fig. 4. Example of AS severity in PLAX views. In (a) no AS, in (b) mild AS, in (c) moderate AS, and in (d) severe AS.

Table 1
Dataset summary for the train, validation, and test sets.

	#Patients			#Images		
Split	AS	No AS	Total	AS	No AS	Total
Train	49	107	156	11680	4801	16481
Validation	16	36	52	4099	1518	5617
Test	16	36	52	4148	1542	5690

introduced in Section 3.1. Implementation details required to reproduce the experiments are described in Section 3.2. A comparison with multi-view approaches and the effectiveness of the transformer encoder are discussed in Section 3.3. Finally, two single-view benchmarks and the differences between single and multi-view strategies are analyzed in Section 3.4.

3.1. Dataset

The proposed model was tested on the public Tufts Medical echocardiogram dataset (TMED) [36], a recently published collection. It contains fully labeled data of 260 unique patients, capturing transthoracic echocardiogram (TTE) imagery acquired in routine care consistent with the American Society of Echocardiography (ASE) guidelines at Tufts Medical Center. All images have a dimension of 64×64 and are associated with an AS diagnosis label (i.e., none, mild, moderate, severe) assigned by a board-certified cardiologist. Furthermore, all images have a corresponding view label (i.e., PLAX, PSAX, other) provided by board-certified sonographers or cardiologists. Examples of AS severity are shown in Fig. 4.

To use this dataset as a binary classification benchmark for the stenosis recognition task, it was organized as follows: diagnoses labeled as mild, moderate, and severe, were merged into a single category, termed “AS”, while all images without this pathology were marked as “No AS”. Moreover, only the PLAX and PSAX views were retained, thus discarding all images labeled as “other” in order to maintain only consistent heart views. Finally, the dataset was split into train, validation, and test sets using a 3:1:1 ratio, as summarized in Table 1. In detail, the train set comprises 156 patients, accounting for a total of 16,481 images; the validation set contains 52 patients, amounting to a total of 5,617 images; while the test set includes the remaining 52 patients, totaling 5,690 images.

3.2. Implementation details

The proposed system was implemented using the Pytorch library, and all the experiments were executed on an Intel Core i7-7700HQ CPU @2.80 GHz with 16 GB RAM and a GeForce GTX 1050 graphics card. Each implemented model was trained for ten epochs using the Adam Optimizer [49], with an initial learning rate of 0.0025 and a batch size of eight. The highest-performing weights in relation to the validation set were used as the final configuration to evaluate each architecture on the test set. Furthermore, standard classification metrics were employed to evaluate all architectures, including accuracy, precision, recall, and F1-score.

Table 2
Multi-view SOTA benchmark. Results refer to the test set.

Model	Accuracy	Precision	Recall	F1-score
MVMDCNN [33]	87.35%	**94.70%**	89.81%	91.00%
SqueezeNet [39]	88.14%	91.16%	96.38%	93.17%
ResNet 101 [25]	88.66%	93.80%	**97.45%**	94.05%
MobileNet V3 [40]	87.97%	93.39%	95.95%	93.05%
EfficientNet B0 [41]	88.59%	93.57%	93.39%	93.21%
VGG 19 [42]	81.66%	89.01%	89.48%	89.12%
Proposed Model	**90.31%**	93.47%	97.41%	**94.36%**

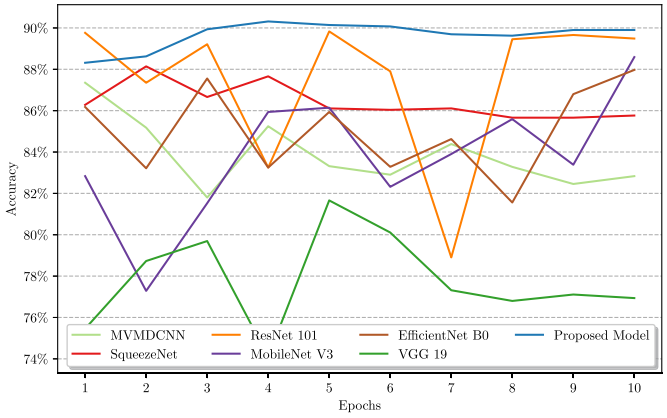


Fig. 5. Multi-view accuracy comparison on the validation set.

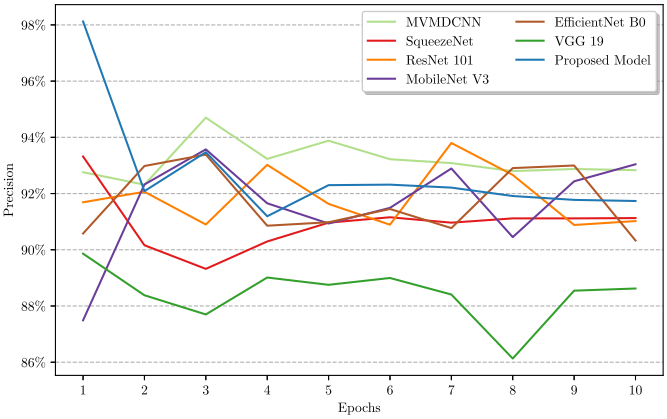


Fig. 6. Multi-view precision comparison on the validation set.

3.3. MS-MV-FETE performance evaluation

The TMED dataset is a recently published collection and, to the best of our knowledge, there currently are no works addressing AS recognition on this dataset. Thus, to report a state-of-the-art comparison, experiments were performed using the multi-view mammography method described in [33], which inspired the proposed approach, as well as several well-known models such as SqueezeNet [39], ResNet101 [25], MobileNet V3 [40], EfficientNet B0 [41], and VGG19 [42]. All these models were pre-trained on the ImageNet dataset [50] and fine-tuned on TMED, while the model described in [33] was trained directly on TMED. Note that to present a fair comparison, all pre-trained models were used as backbone feature extractors in the architecture proposed in [33] so that all networks leveraged the same multi-view paradigm. Moreover, all experiments were performed on the dataset splits mentioned in Section 3.1 according to the protocol described in Section 3.2.

A benchmark summarizing the experimental results of multi-view approaches is reported in Table 2. As can be observed, all models achieve high performances across all metrics, with the proposed method

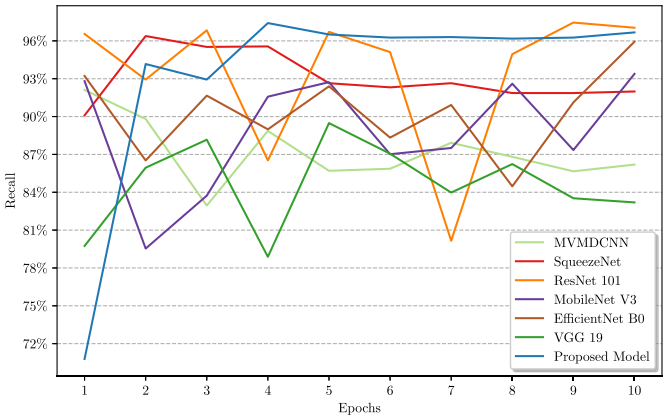


Fig. 7. Multi-view recall comparison on the validation set.

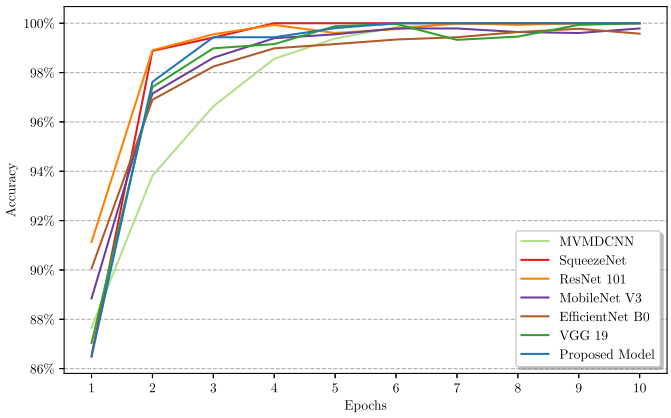


Fig. 9. Multi-view accuracy convergence on the training set.

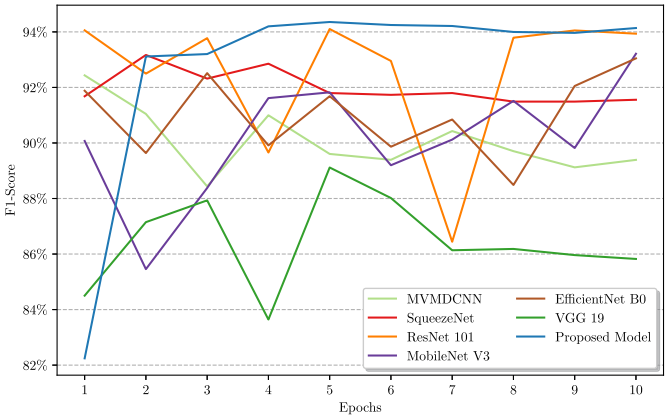


Fig. 8. Multi-view F1-Score comparison on the validation set.

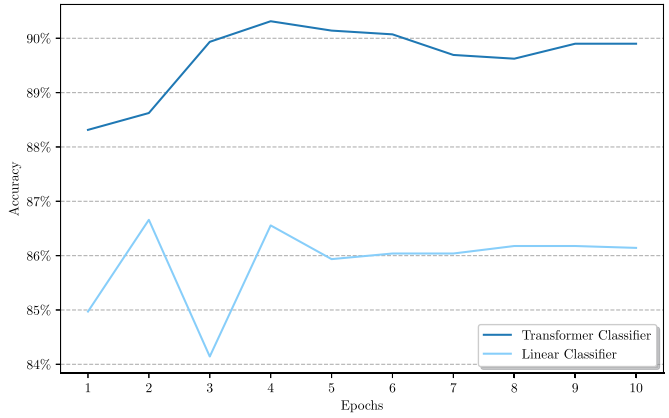


Fig. 10. Classifier ablation study performed on the validation set.

achieving the highest accuracy and F1-score. This outcome highlights, on the one hand, the effectiveness of multi-view methods as all models perform well on the AS recognition task and, on the other hand, the advantage of the multi-scale features and transformer encoder of the devised method as they enable higher performances. What is more, the proposed model has more stable performances during the ten training epochs, as can be observed in Figs. 5 to 8 that report accuracy, precision, recall, and F1-score plots of the selected models on the validation set. These metrics provide further insights on the models as they indicate their capability to correctly classify AS (accuracy), produce few false positives (precision), miss few true positives (recall), and report a holistic view of the models' overall performance in AS recognition (F1-score). By achieving high and stable performance across all metrics at training time, these metrics indicate that the proposed approach can grasp more relevant and consistent information from the PLAX and PSAX views compared to other models and can assist clinicians in making more informed diagnoses, given the small number of false negatives and false positives. Indeed, although all architectures converge early on the training set, as shown in Fig. 9, the only model able to retain consistent performances across all epochs is the proposed MV-MS-FETE. This indicates that the devised solution can generalize better on the shown data compared to the other models and does not suffer from overfitting. Such an outcome has a twofold explanation. First, the multi-scale features allow the model to retain more information and analyze different aspects of the input echo, which is crucial given the relatively small image size, i.e., 64×64. Second, the transformer encoder extracts meaningful details from the patch embeddings generated through the CNNs described in Section 2.2, thus helping to reach higher metrics. These aspects can also be appreciated by substituting the transformer encoder with a standard linear classifier performing the AS recognition. In fact, after this modification, the model can still retain its stability across all training epochs. However, performances drop considerably compared to the entire MV-MS-FETE, i.e., by ≈4% as illustrated in Fig. 10, thus demonstrating the proposed strategy's effectiveness.

3.4. Single-view benchmarks

Multi-view approaches achieve considerable performances; however, single-view experiments must also be performed to complete the assessment of the chosen pre-trained models on the TMED dataset. To this end, tests were conducted using either a PLAX or PSAX view as input for the various architectures. The obtained results are reported in Table 3 and Table 4 for the PLAX and PSAX view, respectively. As can be observed, all models achieve significant performances across all metrics with either view, indicating that they can extract relevant characteristics from the input image, an expected behavior as these images are both useful to recognize AS. More interestingly, all pre-trained architectures tend to reach higher scores when using a PSAX view with the exception of MobileNet V3, that shows a preference for PLAX views. This outcome suggests that PSAX images might contain more information for a neural network, which is also usually the case when they are analyzed by clinicians [51].

Independently of the input view, apart from ResNet 101 that performs particularly well on PSAX and can even best multi-view models on various metrics, all architectures have lower performances compared to the proposed MV-MS-FETE, as well as their respective multi-view counterparts. Indeed, by comparing the single-view models with their corresponding multi-view version, i.e., Table 3 and Table 4 against Table 2, there is a performance increase of up to ≈5/8% depending on the metric, e.g., accuracy and recall for the EfficientNet B0. This out-

Table 3
PLAX single-view benchmark. Results refer to the test set.

Model	Accuracy	Precision	Recall	F1-score
SqueezeNet [39]	88.43%	89.04%	**97.98%**	93.30%
ResNet 101 [25]	**88.86%**	91.85%	94.87%	**93.34%**
MobileNet V3 [40]	83.53%	92.14%	87.42%	89.72%
EfficientNet B0 [41]	80.57%	**93.99%**	81.59%	87.35%
VGG 19 [42]	82.59%	87.74%	91.63%	89.64%
Proposed Model*	90.31%	93.47%	97.41%	94.36%

*Multi-view approach shown as reference.

Table 4
PSAX single-view benchmark. Results refer to the test set.

Model	Accuracy	Precision	Recall	F1-score
SqueezeNet [39]	89.51%	91.08%	96.71%	93.81%
ResNet 101 [25]	**91.74%**	92.75%	**97.59%**	**95.11%**
MobileNet V3 [40]	81.47%	92.91%	83.87%	88.16%
EfficientNet B0 [41]	83.24%	**93.70%**	85.36%	89.33%
VGG 19 [42]	84.61%	85.90%	97.24%	91.22%
Proposed Model*	90.31%	93.47%	97.41%	94.36%

*Multi-view approach shown as reference.

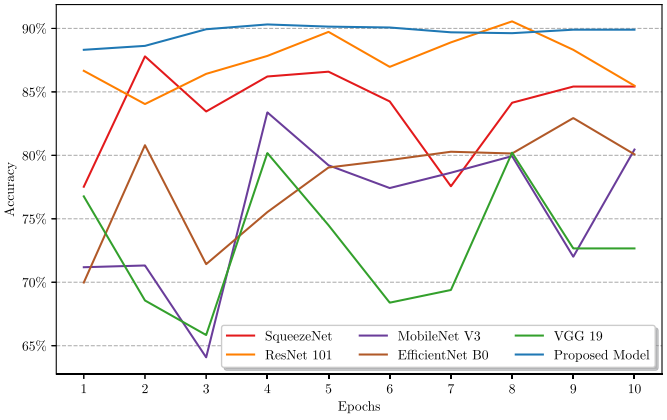


Fig. 11. PLAX single-view comparison on the validation set.

come corroborates the effectiveness of multi-view approaches that can leverage information derived from diverse views to obtain significant results on the AS recognition task. Moreover, similarly to the multi-view scenario, when analyzing performances epoch-wise, it is clear that single-view approaches suffer from the same issue of their multi-view implementation. Specifically, these models tend to overfit on the training data and generate irregular performance on the validation set, as can be observed in Fig. 11 and Fig. 12. This behavior differs from the proposed multi-view method that remains consistent throughout the various epochs, fully highlighting the MV-MS-FETE effectiveness in the aortic stenosis recognition task.

4. Conclusion

This paper presented a novel architecture for aortic stenosis recognition from echocardiogram views, i.e., the multi-view, multi-scale feature extractor and transformer encoder (MV-MS-FETE) model. The proposed strategy capitalizes on information derived from PLAX and PSAX views through two parallel feature extractors that generate multi-scale feature maps to analyze multiple characteristics of the input, effectively performing, for the first time in literature, AS recognition from echos in a multi-view setting. The feature maps are concatenated scale-wise and converted into patch embeddings so that a transformer encoder can use them to predict whether the patient suffers from AS. Extensive experiments were performed on a recently published public collection, i.e., the Tufts medical echocardiogram dataset (TMED), where the proposed MV-MS-FETE model set new state-of-the-art performances for multi-view approaches in terms of accuracy and F1-score metrics. In detail, several well-known models were used as the backbone of an effective multi-view mammography architecture [33] to provide a fair multi-view benchmark. Furthermore, single-view experiments were conducted to evaluate the chosen literature architectures, resulting in two additional benchmarks associated with the PLAX and PSAX views. The obtained results showed that multi-view architectures outperform their single-view counterparts with the exception of ResNet 101, which exhibited a preference for PSAX views. Regarding the proposed model, it also reported more stable performances throughout the training epochs, indicating that the multi-scale strategy extracts robust features from the input images. Finally, additional tests confirmed the advantages of the transformer encoder over a standard linear one, demonstrating the effectiveness of the proposed approach on the AS recognition task.

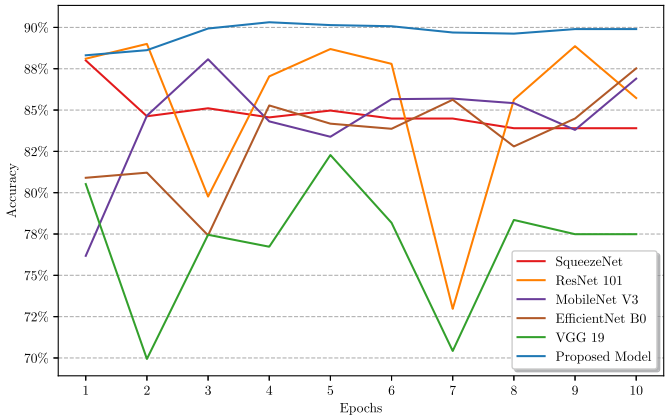


Fig. 12. PSAX single-view comparison on the validation set.

CRediT authorship contribution statement

Danilo Avola: Conceptualization, Methodology, Supervision. **Irene Cannistraci:** Software, Writing – original draft, Visualization. **Marco Cascio:** Software, Supervision, Writing – review & editing. **Luigi Cinque:** Resources, Supervision, Conceptualization. **Alessio Fagioli:** Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Gian Luca Foresti:** Resources, Supervision, Conceptualization. **Emanuele Rodolà:** Resources, Supervision, Conceptualization. **Luciana Solito:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the MIUR under grant “Departments of Excellence 2018-2022” of the Department of Computer Science of Sapienza University; the “Smart unmannEd AeRial vehiCles for Human likE monitoRing (SEARCHER)” project of the Italian Ministry of Defence (CIG: Z84333EA0D); the Departmental Strategic Plan (DSP) of the University of Udine - Interdepartmental Project on Artificial Intelligence (2020-25); and the ERC Starting Grant no. 802554 (SPECGEO).

References

[1] B.A. Carabello, W.J. Paulus, Aortic stenosis, Lancet 373 (9667) (2009) 956–966.

[2] J. Ross Jr, E. Braunwald, Aortic stenosis, Circulation 38 (1s5) (1968), V–61.

[3] F.J. Rogers, Aortic stenosis: new thoughts on a cardiac disease of older people, Journal of Osteopathic Medicine 113 (11) (2013) 820–828.

[4] W.J. Manning, Asymptomatic aortic stenosis in the elderly: a clinical review, JAMA 310 (14) (2013) 1490–1497.

[5] P. Lancellotti, J. Magne, R. Dulgheru, M.-A. Clavel, E. Donal, M.A. Vannan, J. Chambers, R. Rosenhek, G. Habib, G. Lloyd, et al., Outcomes of patients with asymptomatic aortic stenosis followed up in heart valve clinics, JAMA Cardiol. 3 (11) (2018) 1060–1068.

[6] J.J. Thaden, V.T. Nkomo, M. Enriquez-Sarano, The global burden of aortic stenosis, Prog. Cardiovasc. Dis. 56 (6) (2014) 565–571.

[7] J.-M. Kwon, S.Y. Lee, K.-H. Jeon, Y. Lee, K.-H. Kim, J. Park, B.-H. Oh, M.-M. Lee, Deep learning–based algorithm for detecting aortic stenosis using electrocardiography, J. Am. Heart Assoc. 9 (7) (2020) e014717.

[8] H. Baumgartner, J. Hung, J. Bermejo, J.B. Chambers, A. Evangelista, B.P. Griffin, B. Iung, C.M. Otto, P.A. Pellikka, M. Quiñones, Echocardiographic assessment of valve stenosis: eae/ase recommendations for clinical practice, European Journal of Echocardiography 10 (1) (2009) 1–25.

[9] H. Baumgartner, J. Hung, J. Bermejo, J.B. Chambers, T. Edvardsen, S. Goldstein, P. Lancellotti, M. LeFevre, F. Miller Jr, C.M. Otto, et al., Recommendations on the echocardiographic assessment of aortic valve stenosis: a focused update from the European association of cardiovascular imaging and the American society of echocardiography, European Heart Journal-Cardiovascular Imaging 18 (3) (2017) 254–275.

[10] V. Kamperidis, V. Delgado, N.M. van Mieghem, A.-P. Kappetein, M.B. Leon, J.J. Bax, Diagnosis and management of aortic valve stenosis in patients with heart failure, European Journal of Heart Failure 18 (5) (2016) 469–481.

[11] X. Zhang, V.C. Lee, J. Rong, J.C. Lee, F. Liu, Deep convolutional neural networks in thyroid disease detection: a multi-classification comparison by ultrasonography and computed tomography, Comput. Methods Programs Biomed. 220 (2022) 106823.

[12] D. Avola, L. Cinque, A. Fagioli, S. Filetti, G. Grani, E. Rodolà, Multimodal feature fusion and knowledge-driven learning via experts consult for thyroid nodule classification, IEEE Trans. Circuits Syst. Video Technol. 32 (5) (2021) 2527–2534.

[13] S. Sorrenti, V. Dolcetti, M. Radzina, M.I. Bellini, F. Frezza, K. Munir, G. Grani, C. Durante, V. D'Andrea, E. David, et al., Artificial intelligence for thyroid nodule characterization: where are we standing?, Cancers 14 (14) (2022) 3357.

[14] Z. Liao, Y. Song, S. Ren, X. Song, X. Fan, Z. Liao, Voc-dl: deep learning prediction model for Covid-19 based on voc virus variants, Comput. Methods Programs Biomed. 224 (2022) 106981.

[15] D. Avola, A. Bacciu, L. Cinque, A. Fagioli, M.R. Marini, R. Taiello, Study on transfer learning capabilities for pneumonia classification in chest-X-rays images, Comput. Methods Programs Biomed. 221 (2022) 106833.

[16] N. Subramanian, O. Elharrouss, S. Al-Maadeed, M. Chowdhury, A review of deep learning-based detection methods for Covid-19, Comput. Biol. Med. (2022) 105233.

[17] G. Wang, X. Luo, R. Gu, S. Yang, Y. Qu, S. Zhai, Q. Zhao, K. Li, S. Zhang, Pymic: a deep learning toolkit for annotation-efficient medical image segmentation, Comput. Methods Programs Biomed. (2023) 107398.

[18] G. Placidi, D. Avola, M. Ferrari, D. Iacoviello, A. Petracca, V. Quaresima, M. Spezialetti, A low-cost real time virtual system for postural stability assessment at home, Comput. Methods Programs Biomed. 117 (2) (2014) 322–333.

[19] D. Avola, L. Cinque, A. Fagioli, G. Foresti, A. Mecca, Ultrasound medical imaging techniques: a survey, ACM Comput. Surv. 54 (3) (2021) 1–38.

[20] G. Holste, E. Oikonomou, B. Mortazavi, K. Faridi, E. Miller, J. Forrest, R. McNamara, H. Krumholz, Z. Wang, R. Khera, Automated detection of severe aortic stenosis using single-view echocardiography: A self-supervised ensemble learning approach, medRxiv.

[21] C.-A. Hatfaludi, C.F. Ciusdel, A. Toma, L.M. Itu, Deep learning based aortic valve detection and state classification on echocardiographies, in: Proceedings of the International Power Electronics and Motion Control Conference (PEMC), 2022, pp. 275–280.

[22] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, in: Proceedings of the Advances in Neural Information Processing Systems, NIPS, 2015, pp. 1–9.

[23] T. Ginsberg, R.-e. Tal, M. Tsang, C. Macdonald, F.T. Dezaki, J. van der Kuur, C. Luong, P. Abolmaesumi, T. Tsang, Deep video networks for automatic assessment of aortic stenosis in echocardiography, in: Proceedings of the International Workshop on Advances in Simplifying Medical Ultrasound (ASMUS), 2021, pp. 202–210.

[24] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C.P. Langlotz, P.A. Heidenreich, R.A. Harrington, D.H. Liang, E.A. Ashley, et al., Video-based ai for beat-to-beat assessment of cardiac function, Nature 580 (7802) (2020) 252–256.

[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[26] M. Cohen-Shelly, Z.I. Attia, P.A. Friedman, S. Ito, B.A. Essayagh, W.-Y. Ko, D.H. Murphree, H.I. Michelena, M. Enriquez-Sarano, R.E. Carter, et al., Electrocardiogram screening for aortic valve stenosis using artificial intelligence, Eur. Heart J. 42 (30) (2021) 2885–2896.

[27] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.

[28] E. Hata, C. Seo, M. Nakayama, K. Iwasaki, T. Ohkawauchi, J. Ohya, Classification of aortic stenosis using ecg by deep learning and its analysis using grad-cam, in: Proceedings of the International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020, pp. 1548–1551.

[29] R.C. Joshi, J.S. Khan, V.K. Pathak, M.K. Dutta, Ai-cardiocare: artificial intelligence based device for cardiac health monitoring, IEEE Trans. Human-Mach. Syst. 52 (6) (2022) 1292–1302.

[30] S.B. Shuvo, S.N. Ali, S.I. Swapnil, M.S. Al-Rakhami, A. Gumaei, Cardioxnet: a novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings, IEEE Access 9 (2021) 36955–36967.

[31] J. Wang, X. Liu, F. Wang, L. Zheng, F. Gao, H. Zhang, X. Zhang, W. Xie, B. Wang, Automated interpretation of congenital heart disease from multi-view echocardiograms, Med. Image Anal. 69 (2021) 101942.

[32] S. Wang, M. Zhou, O. Gevaert, Z. Tang, D. Dong, Z. Liu, T. Jie, A multi-view deep convolutional neural networks for lung nodule segmentation, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017, pp. 1752–1755.

[33] L. Sun, J. Wang, Z. Hu, Y. Xu, Z. Cui, Multi-view convolutional neural networks for mammographic image classification, IEEE Access 7 (2019) 126273–126282.

[34] D.M. Vigneault, W. Xie, C.Y. Ho, D.A. Bluemke, J.A. Noble, ω-net (omega-net): fully automatic, multi-view cardiac mr detection, orientation, and segmentation with deep neural networks, Med. Image Anal. 48 (2018) 95–106.

[35] Y. Zhao, B. Ma, T. Che, Q. Li, D. Zeng, X. Wang, S. Li, Multi-view prediction of Alzheimer's disease progression with end-to-end integrated framework, J. Biomed. Inform. 125 (2022) 103978.

[36] Z. Huang, G. Long, B. Wessler, M.C. Hughes, A new semi-supervised learning benchmark for classifying view and diagnosing aortic stenosis from echocardiograms, in: Proceedings of the Machine Learning for Healthcare Conference (MLHC), 2021, pp. 614–647.

[37] J. Sheng, S.-K. Lam, Z. Li, J. Zhang, X. Teng, Y. Zhang, J. Cai, Multi-view contrastive learning with additive margin for adaptive nasopharyngeal carcinoma radiotherapy prediction, in: Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR), 2023, pp. 555–559.

[38] L. Xia, J. An, C. Ma, H. Hou, Y. Hou, L. Cui, X. Jiang, W. Li, Z. Gao, Neural network model based on global and local features for multi-view mammogram classification, Neurocomputing 536 (2023) 21–29.

[39] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, K. Keutzer, Squeezenet: alexnet-level accuracy with 50x fewer parameters and <0.5 mb model size, preprint, arXiv:1602.07360.

[40] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: efficient convolutional neural networks for mobile vision applications, preprint, arXiv:1704.04861.

[41] M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, in: Proceedinngs of the International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.

[42] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.

[43] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: full training or fine tuning?, IEEE Trans. Med. Imaging 35 (5) (2016) 1299–1312.

[44] L. Li, W. Ding, L. Huang, X. Zhuang, V. Grau, Multi-modality cardiac image computing: a survey, Med. Image Anal. (2023) 102869.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 1–11.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint, arXiv:2010.11929.

[47] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: a survey, ACM Comput. Surv. 54 (10s) (2022) 1–41.

[48] N. Ahmadi, M. Tsang, A. Gu, T. Tsang, P. Abolmaesumi, Transformer-based spatio-temporal analysis for classification of aortic stenosis severity from echocardiography cine series, IEEE Trans. Med. Imaging (2023) 1.

[49] D.P. Kingma, J. Ba Adam, A method for stochastic optimization, preprint, arXiv: 1412.6980.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.

[51] P. Pibarot, H. Baumgartner, M.-A. Clavel, N. Côté, S. Orwat, Aortic valve stenosis, in: The ESC Textbook of Cardiovascular Imaging, 2021, p. 161.