Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Vision transformer-based multi-class classification of aortic calcification scores

Aort kalsifikasyon skorlarının görü dönüştürücü tabanlı çok sınıflı sınıflandırması

Mervenur Cakir[1,*], Murat Ekinci[2], Elif Baykal Kablan[3], Mursel Sahin[4]

[1,3] *Karadeniz Technical University, Department of Software Engineering, 61080, Trabzon, Türkiye*
[2] *Karadeniz Technical University, Department of Computer Engineering, 61080, Trabzon, Türkiye*
[4] *Karadeniz Technical University, Faculty of Medicine, Department of Cardiology, 61080, Trabzon, Türkiye*

Abstract

This study presents a fully ultrasound-based and cost-effective approach for the automatic grading of aortic valve calcification, which plays a critical role in the assessment of aortic stenosis. To eliminate radiation exposure associated with computed tomography, the proposed method relies exclusively on echocardiographic images and was trained on a dedicated dataset constructed for this study. A Vision Transformer (ViT) model operating on ROI frames extracted from the aortic valve region was developed to classify four calcification levels: mild, moderate, severe, and critical. During testing, the model achieved 85% accuracy and a macro-averaged precision of 0.74, demonstrating a reliable decision mechanism with a low false-positive tendency. To further evaluate the architectural choice, ImageNet pre¬trained ResNet50 and EfficientNet-B3 models were trained and tested under identical conditions. Although CNN-based architectures produced competitive results, the ViT model demonstrated more balanced performance, particularly in intermediate and advanced grades, and achieved the highest macro-averaged metrics among the evaluated models. By providing a radiation-free, reproducible, and economically accessible solution, the proposed framework offers a clinically applicable decision-support system for the evaluation of aortic stenosis.

Keywords: Aortic valve, Echocardiography, Calcium scoring, Classification, Vision transformer

Öz

Bu çalışma, aort darlığının değerlendirilmesinde kritik öneme sahip aort kapak kalsifikasyonunun otomatik derecelendirilmesi için tamamen ultrason tabanlı ve düşük maliyetli bir yaklaşım sunmaktadır. Bilgisayarlı tomografi yöntemlerinde maruz kalınan radyasyonu ortadan kaldırmak amacıyla yöntem yalnızca ekokardiyografi görsellerini kullanmakta ve çalışmaya özel oluşturulan bir veri seti ile eğitilmiştir. Aort kapak bölgesinden elde edilen ROI frameleri üzerinden çalışan Vision Transformer (ViT) modeli, düşük, orta, yüksek ve kritik olmak üzere dört kalsifikasyon seviyesini sınıflandırmak üzere geliştirilmiştir. Test aşamasında model %85 doğruluk ve 0.74 ortalama precision elde ederek güvenilir ve düşük yanlış pozitif oranına sahip bir karar mekanizması ortaya koymuştur. Mimarinin etkinliğini değerlendirmek amacıyla aynı veri seti üzerinde ImageNet ön-eğitimli ResNet50 ve EfficientNet-B3 modelleriyle karşılaştırmalı deneyler gerçekleştirilmiştir. CNN tabanlı modeller rekabetçi sonuçlar üretmekle birlikte, ViT modeli özellikle ara ve ileri derecelerde daha dengeli performans göstermiş ve makro ortalama metriklerde en yüksek başarıyı elde etmiştir. Ultrason temelli olması sayesinde radyasyona maruz bırakmayan, tekrarlanabilir ve ekonomik bir çözüm sunan bu yaklaşım, aort darlığının değerlendirilmesinde klinik olarak uygulanabilir bir karar destek sistemi sağlamaktadır.

Anahtar Kelimeler: Aort kapağı, Ekokardiyografi, Kalsiyum skoru, Sınıflandırma, Görü dönüştürücü

1 Introduction

Aortic stenosis is the most common valvular heart disease encountered in developed countries, particularly among the elderly population [1]. Although echocardiographic examination is the first-line diagnostic tool for aortic stenosis, it may be insufficient in some patients [2]. Especially in cases of low-gradient aortic valve disease, additional diagnostic methods are required. Aortic valve calcification is the most significant factor in the development of aortic stenosis, and the amount, extent, and density of calcification are closely related to the severity of the stenosis [3]. Clinical guidelines used in cardiology diagnosis and treatment recommend the assessment of the aortic valve calcium score [4]. The European Society of Cardiology guidelines suggest computed tomography (CT) imaging as the gold standard method for measuring the aortic valve calcium score.

The most commonly used method for quantifying calcium in CT imaging is the Agatston scoring [5]. This technique calculates the calcium score by multiplying the area of calcified plaque by the highest Hounsfield density value obtained. This scoring method interprets the lesions with CT densities greater than 130 Hounsfield units (HU) in at least two to three adjacent pixels, covering an area larger than 1 mm² as calcifications. The area and density of each segmented lesion are automatically measured by the software device under the supervision of an expert radiologist. This manual segmentation task for each slice is

* Sorumlu yazar / Corresponding author, e-posta / e-mail: mervecakir@ktu.edu.tr (M. Cakır)
Geliş / Received: 27.11.2025 Kabul / Accepted: 24.04.2026 Yayımlanma / Published: 20.07.2026
doi: 10.28948/ngumuh.1830650

time-consuming, and the accuracy of the results may vary depending on the experience and skill of the radiologist. Furthermore, several variations in the shape, size, and position of the aortic region of interest in each CT slice (open or closed valve phases) make manual segmentation even more challenging.

For each calcified lesion, the calcium score is calculated by multiplying the lesion area by a density score determined according to the lesion's density. The calcium score for each lesion is then determined, and the total calcium score for the patient is calculated semi-quantitatively. In this process, the radiologist's experience plays a significant role. Echocardiography-based approaches have traditionally relied on observer-dependent assessments, which inherently limit accuracy and reproducibility. To address this limitation, recent studies have increasingly focused on the development of deep learning–based fully automatic methods. Tang et al. [6] proposed DLFFNet, a novel network architecture designed for the detection of calcification in echocardiographic images. Similarly, Cakir et al. introduced an original dataset and method specifically tailored for aortic valve calcification segmentation [7, 8].

Elvas et al. [9] developed a two-stage fully automatic framework that combines the YOLOv5 model for valve localization with a CNN-based classifier for detecting the presence or absence of calcification. Their approach achieved 95% accuracy / 100% recall in valve localization and 92% accuracy / 100% recall in binary calcification classification. However, the study had two notable limitations: a relatively small dataset and the use of only a binary classification scheme (calcified vs. non-calcified). Elvas et al. [10] also applied transfer learning–based CNN models (e.g. VGG16) to classify aortic stenosis using magnetic resonance imaging (MRI) images and achieved 95% sensitivity and F1-score. While the results were promising, the lack of multi-center validation limited the generalizability of the findings and their potential for clinical adoption. Vaseli et al. [11] proposed an explainable AI–based model called ProtoASNet. This model classifies aortic stenosis (AS) severity into three levels (none, early, significant) and incorporates an uncertainty estimation component. The authors reported a balanced accuracy of 80%. This result indicates competitive performance. However, the classification remained restricted to three categories, and aortic valve calcification (AVC) staging was not addressed. This narrows the clinical applicability of the model.

Overall, these studies demonstrate a growing interest in automating echocardiography-based valve assessment using deep learning. They also highlight several challenges including limited data availability and insufficient multi-class or clinically interpretable severity grading, that still constrain the clinical integration of such AI-assisted systems. In this study, a Vision Transformer (ViT)-based classification model was proposed for the automatic evaluation of AVC from transthoracic echocardiographic images. The developed framework first employed the AVD-YOLOv5 model to detect and localize the aortic valve region automatically. The extracted regions of interest (ROIs) were then divided into sequential frames and were classified using the ViT-based model to predict calcification severity on a four-level scale (0–3). This multi-class approach enables not only the detection of calcified valves but also the grading of calcification severity, thereby addressing one of the major limitations of previous binary classification studies. Unlike existing semi-automatic or single-center frameworks, the proposed method provides a fully automated and reproducible pipeline. This pipeline supports clinical decision-making and facilitates quantitative and objective AVC assessment in echocardiographic practice.

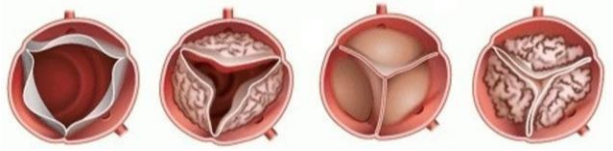

Figure 1. From left to right: Appearances of the aortic valve; normal in systole, calcified in systole, normal in diastole, and calcified in diastole.

2 Material ve methods

All deep learning experiments were conducted on a workstation equipped with 32 GB RAM and an NVIDIA GeForce RTX 3060 GPU with 12 GB of VRAM, using the Python programming language and the PyTorch framework. The training and validation phases of the model were conducted on this high-performance hardware infrastructure, which provided sufficient computational power and memory capacity to reduce training time and enhance computational efficiency.

2.1 Aortic valve dataset

In this study, a dataset consisting of 157 parasternal short-axis ultrasound images obtained during transthoracic echocardiography (TTE) examinations was used. The images were collected from 99 different patients and were originally stored in a standardized resolution of 1024 × 768 pixels. All images were acquired in the short-axis view, focusing on the anatomical plane of the aortic valve, which provides the most informative perspective for assessing calcification intensity and distribution.

For the AVD-YOLOv5 detection stage, the images were resized to 768 × 768 pixels, as the YOLO architecture requires square input dimensions. This resizing was performed while aiming to minimize information loss. During the testing phase, resizing was similarly applied for model compatibility; however, evaluations were conducted in accordance with the original clinical image format (1024 × 768), ensuring consistency with real-world clinical scenarios.

As illustrated in Figure 2, the dataset includes frames obtained from patients with varying degrees of aortic valve calcification. Each image was labeled according to aortic valve calcification scores (score 0–3) determined by experienced cardiologists. These scores reflect the severity and progression of calcific deposits: *score 0* represents normal valve tissue without visible calcification, *score 1*

corresponds to early-stage cases with mild focal calcific areas, *score 2* denotes moderate calcification with more distinct deposits, and *score 3* indicates advanced, heavily calcified valves. Thus, the dataset captures four progressive levels of aortic valve calcification.

To evaluate the generalization performance of the proposed model, the dataset was randomly divided into training, validation, and test subsets while preserving the proportional distribution of each calcification level. In the training subset, data augmentation techniques were applied to address class imbalance. These augmentation procedures were performed on the cropped aortic valve regions of interest obtained using the AVD-YOLOv5 model. The applied techniques included horizontal flipping, small rotations, brightness and contrast variations, slight blurring, and low-level Gaussian noise addition.

After the augmentation process, underrepresented classes were increased to match the size of the most populated class, resulting in a total of 305 images, up from the original 157. This balancing procedure effectively eliminated class imbalance, enabling the model to learn each calcification level more uniformly. Furthermore, this strategy reduced the risk of overfitting and enhanced the model's generalization capability on unseen echocardiographic data.

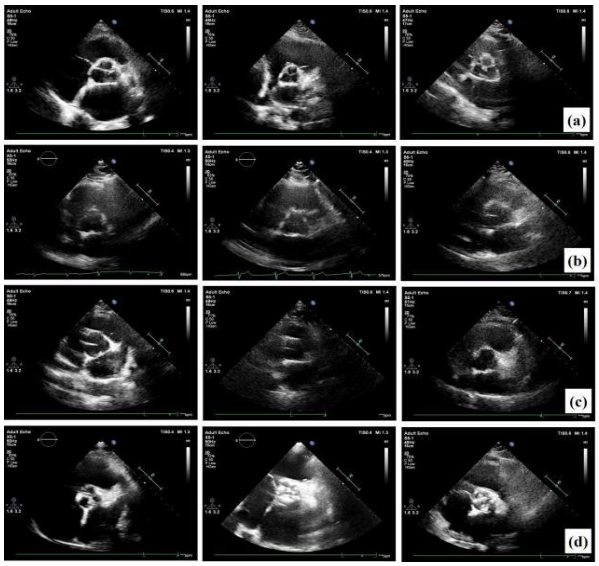

Figure 2. Some example echocardiography images showing different conditions: (a) diastole, (b) systole, (c) normal tissue, and (d) abnormal tissue

2.1.1 Data preprocessing

In raw echocardiographic images, the aortic valve represents a relatively small structure within a wide anatomical field; therefore, it is essential to identify and focus on this region before analysis. For this purpose, the images were annotated by expert cardiologists using the Image Labeler Toolbox in MATLAB 2022a. The aortic valve regions of interest (ROIs) were annotated using rectangular bounding boxes. Figure 3 illustrates an example of an echocardiographic frame where a cardiologist manually labeled the aortic valve region.

The annotated data were then processed using our previously developed AVD-YOLOv5 (Aortic Valve Detection-YOLOv5) model [7], which is a structurally modified version of the standard YOLOv5 architecture. This model automatically detected and localized the aortic valve ROI in each 1024×768 frame with high accuracy.

Using the detected ROI coordinates, only the aortic valve region was cropped from each original echocardiographic image. Using the detected ROI coordinates, the aortic valve region was cropped from each echocardiographic image. This step eliminated tissues and background patterns outside the ROI, enabling the subsequent classification model to focus on clinically meaningful anatomical information.

Following cropping, the images were labeled into four classes based on calcification scores to create labeled datasets. The classified data were then divided into three subsets: training, validation, and testing. The training subset was further expanded using various data augmentation techniques to improve the model's generalization ability.

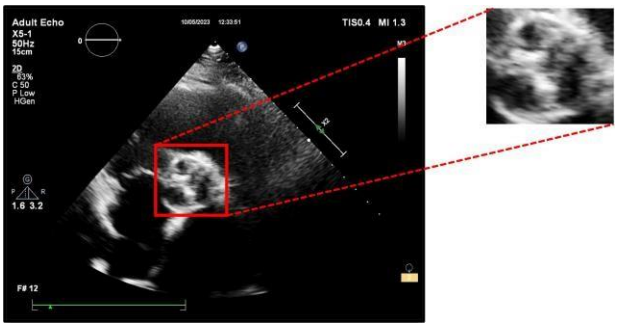

Figure 3. Image taken from the labeling step in the recommended aortic valve dataset. The red rectangular bounding boxes indicate the aortic valve regions of interest labeled by expert cardiologists.

2.1.2 Data augmentation

The augmentation procedures applied to the training dataset are designed to preserve the anatomical integrity of medical images. By following a gentle augmentation strategy, all transformations are kept within clinically acceptable limits so as not to complicate the visual interpretation of echocardiographic data. In this context, horizontal rotation ($p\approx0.5$), angular rotations ($\pm6°$), translation ($\pm2\%$), and shear transformations ($\pm2°$) were applied at a mild level to enable the model to learn directional variations, ensuring similarity to the diversity of imaging angles in real applications. Mild Gaussian blurring ($p\approx0.15$, $r\approx0.1–0.5$) was applied to increase robustness against focus errors, while brightness/contrast adjustments ($\alpha\approx0.97–1.05$, $\beta\approx\pm2$) were made to enhance the model's generalization ability across different imaging devices and patient conditions. Additionally, low-level Gaussian noise ($p\approx0.15$, $\sigma\approx2–6$) was added to increase robustness against random ultrasound-related artifacts

All augmentation procedures were applied only to the training set; the validation and test sets were kept in their original form to realistically reflect model performance in clinical applications. Before training, all images were resized

to 224×224 pixels and normalized using ImageNet statistics (mean: 0.485, 0.456, 0.406; standard deviation: 0.229, 0.224,0.225).

This normalization ensured compatibility with the input distribution of the pretrained Vision Transformer (ViT) model, facilitating more effective transfer learning.

After data augmentation was applied, the class distribution is summarized in Table 1. As seen, the "Moderate (1)", "Severe (2)", and "Critical (3)" classes initially contained fewer examples compared to the "Mild (0)" class and were therefore expanded through augmentation. After the augmentation process, which was performed to keep each class the same as the class with the highest data, the total number of images increased from 157 to 305. This process eliminated class imbalance, allowing the model to learn all classification severity levels more homogeneously. As a result, the model's generalization performance improved significantly, particularly for higher calcification grades (2-3 points).

Table 1. Number of images per class before and after data augmentation.

Classes (Labels)	Before Augmentation			After Augmentation		
	Train	Val	Test	Train	Val	Test
Mild (0)	63	9	19	63	9	19
Moderate (1)	6	0	3	63	0	3
Severe (2)	30	4	9	63	4	9
Critical (3)	9	1	4	63	1	4
Total sample	108	14	35	252	14	35

2.2 *Vision transformer (ViT)*

In this study, the Vision Transformer (ViT) architecture was employed as the primary model for the automatic classification of aortic valve calcification. ViT is an adaptation of the Transformer architecture - originally developed for natural language processing (NLP) - to the visual domain, and it has emerged as a powerful alternative to conventional convolutional neural networks (CNNs) due to its superior ability to capture global contextual dependencies in large-scale visual data (Dosovitskiy et al., 2020).

The core concept of ViT is to represent an image not as a two-dimensional matrix of pixels but as a sequence of patch embeddings, which can be processed similarly to word tokens in language models. Each patch is treated as an individual token, and the long-range dependencies between patches are modeled using the Multi-Head Self-Attention (MHSA) mechanism.

In the ViT-B/16 model used in this study, each echocardiographic region of interest (ROI) was resized to a fixed input resolution of 224×224 pixels and then divided into 16×16-pixel patches. This operation, computed as shown in Equation (1), produces 196 patches per image. Here, H and W denote the height and width of the image, and P represents the patch size. Each patch is then flattened and projected into a vector representation $X \in R^{P^2 \times C}$, where C= 3 corresponds to the number of image channels

$$N = \frac{H \times W}{P^2} = \frac{224 \times 224}{10^2} = 196 \qquad (1)$$

These patches are then transformed into D-dimensional vectors through a learnable embedding matrix $E \in R^{(P^2 \times C) \times D}$, where D = 768 for the ViT-B/16 model. As a result, the image is represented as a sequence of 196 embedded tokens. The obtained sequence is concatenated with a classification token ([CLS]), and a positional embedding is added to each token to preserve spatial ordering among the patches. This process enables the model to retain information about the relative positions of patches within the image. The final input sequence fed into the first Transformer layer is defined as shown in Equation (2).

$$Z_0 = [X_{CLS}; x_1^P E; x_2^P E; \ldots; x_N^P E] + E_{pos} \qquad (2)$$

Here, $Epos \in R(N+1) \times D$ represents the positional embedding matrix, and X_{CLS} denotes the special classification token used for image-level prediction. After passing through the Transformer layers, this [CLS] token encapsulates the global representation of the entire image and is subsequently fed into the classification head to produce the final output.

2.2.1 *Transformer layers*

The Transformer layers enable the deep representation of visual features and the modeling of global contextual relationships. Unlike conventional CNN architectures, ViT replaces convolutional operations with Multi-Head Self-Attention (MHSA) layers. The model consists of L identical encoder layers (for ViT-B/16, L=12), and each layer contains two primary components: the MHSA and the Feed-Forward Network (FFN) module

Multi-Head Self-Attention (MHSA): Each image patch (token) interacts with all other patches to learn contextual dependencies. This mechanism is particularly valuable in echocardiographic images, which often exhibit limited contrast and heterogeneous tissue textures. In such cases, calcification patterns are not always confined to localized regions; instead, the global context - such as reflections around the valve or brightness distribution near the aortic root - provides critical diagnostic information.

Mathematically, in each layer, the input sequence $X \in R^{N*D}$ is projected into three different linear transformations - Query (Q), Key (K), and Value (V) matrices - as expressed in Equation (3):

$$Q = \mathrm{XW_Q}, K = \mathrm{XW_K}, V = \mathrm{XW_V} \qquad (3)$$

Here $W_Q, W_K, W_V \in R^{D*dk}$ are learnable parameter matrices. The attention scores are computed as shown in Equation (4):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{dk}}\right)V \qquad (4)$$

This operation defines the relationship of each token with all other tokens, enabling the model to capture not only local

texture cues but also long-range structural dependencies across the image.

In the multi-head version, hhh parallel attention heads are employed to learn diverse contextual relationships simultaneously, and their outputs are concatenated and linearly transformed as expressed in Equation (5):

$$\mathrm{MHSA(X)} = [h\mathrm{ead}_1; \ldots\ldots\ldots; h\mathrm{ead}_h]\mathrm{W}_0 \qquad (5)$$

Here, $W0$ represents the output projection matrix. This mechanism enables the model to focus simultaneously on different tissue regions and clinically significant areas through multiple attention heads.

Such an ability is particularly valuable in echocardiographic imaging, where anatomical structures often exhibit limited contrast and heterogeneous texture patterns. In these cases, calcification signals are not always confined to localized regions; instead, they gain diagnostic meaning through the global context, such as reflections around the aortic valve or the brightness distribution in the aortic root region.

Feed-Forward Network (FFN): Following the MHSA layer, the internal representation of each token is transformed through fully connected layers and a ReLU activation function. The FFN structure is defined as shown in Equation (6):

$$FFN(x) = \mathrm{W2}_{\sigma}(\mathrm{W}_{1x} + \mathrm{b}_1) + \mathrm{b}_2 \qquad (6)$$

Here, σ represents the ReLU activation function. This structure enhances the representational capacity of the model by nonlinearly remapping each token's features.

In each Transformer layer, Layer Normalization (LN) and Residual Connections are applied to prevent information loss and to improve training stability. The output of each layer is updated as shown in Equations (7–8):

$$Z_l' = MHSA\big(LN(Z_{l-1})\big) + Z_{l-1} \qquad (7)$$

$$Z_l = FFN\big(LN(Z_l')\big) + Z_l' \qquad (8)$$

These operations are iteratively applied across all layers $l = 1,2,\ \ ,L$, where the output of each layer serves as the input to the next one. This hierarchical and recursive structure enables the model to gradually capture both low-level textural features and high-level structural relationships within the image. As a result, the Vision Transformer effectively learns a rich, multi-scale representation of the echocardiographic data, integrating fine-grained local patterns with global anatomical context relevant to aortic valve calcification.

2.2.2 Classification layer and model customization

In the final stage of the model, the [CLS] token is used to represent the global contextual information of the entire image and is passed through a fully connected (FC) layer for classification. This layer was reconfigured to align with the **four-class (score 0–3)** aortic valve calcification task designed in this study. In other words, instead of the original ImageNet classification head, a new output layer was defined as:

$$Head: R765 \rightarrow R5 \qquad (9)$$

The remaining layers of the model were initialized with pre-trained weights from ImageNet. This initialization enabled the model, unlike conventional CNN architectures, to effectively capture long-range correlations between spatially distant regions. Since the model had already learned low-level visual patterns such as edges, textures, and contrast variations from large-scale natural images, it could focus on learning high-level, fine-grained distinctions such as variations in calcification density and tissue rigidity specific to the medical domain. This property provides a significant advantage for tasks such as aortic valve calcification classification, where the global anatomical context carries crucial clinical meaning and local pixel-level variations alone may not be sufficient for an accurate diagnosis

2.2.3 Training strategy and regularization

In this study, the term *"Original ViT"* refers to the standard ImageNet pre-trained ViT-B/16 architecture trained with conventional cross-entropy loss and default optimization settings. The *"Proposed ViT"* retains the same transformer backbone without any structural modification. Instead, it incorporates task-specific training strategies designed for multi-class aortic calcification grading. These include label smoothing to improve probability calibration in borderline classes, cosine annealing with warm-up scheduling for stable convergence, gradient clipping for optimization stability, and controlled augmentation to address class imbalance. Therefore, the novelty of the proposed model lies in its optimized training framework rather than in architectural redesign.

The model was trained using the Adam optimization algorithm, which retains the adaptive learning capabilities of Adam while correctly applying the weight decay term for improved generalization. In this study, the learning rate was set to 2×10^{-4}, the weight decay coefficient to 5×10^{-5}, the batch size to 32, and the input image size to 224×224 pixels. To enhance the stability and convergence of the training process, several regularization and scheduling strategies were employed:

- **Label Smoothing (0.05):** This technique prevents the model from becoming overconfident in its predictions by distributing a small portion of the probability mass to non-target classes. It encourages smoother decision boundaries and enhances generalization, particularly in borderline cases where calcification grades (e.g., score 1↔2) exhibit gradual transitions rather than discrete separations.
- **Warmup + Cosine Annealing:** The learning rate was linearly increased during the first five epochs (warmup phase) and then gradually reduced following a cosine decay schedule for the remaining epochs. This two-stage adjustment mitigates the characteristic learning rate sensitivity observed in Transformer-based architectures and helps the model achieve stable convergence during early training stages.

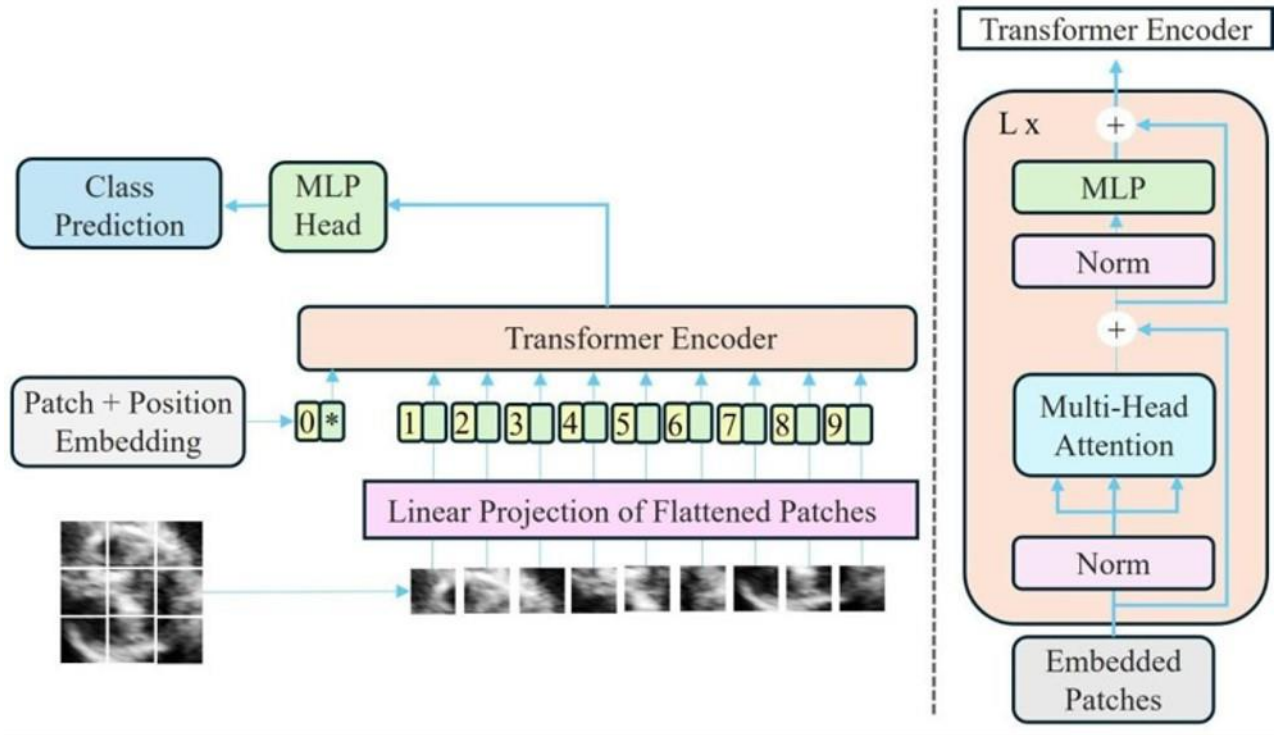

Figure 4. The general architecture of the Vision Transformer (ViT) model

- **Gradient Clipping ($\|g\|_2 \leq 1.0$):** Applied to prevent numerical instability caused by sudden gradient explosions, ensuring smoother and more stable parameter updates.
- **Dropout (0.5):** Introduced in the classification head to reduce overfitting by randomly deactivating neurons during training, thereby improving the model's ability to generalize.

Collectively, these strategies effectively minimized the tendency toward overfitting, which is a common issue in medical imaging tasks with limited data, and contributed to maintaining a balanced overall accuracy across all calcification classes.

2.3 Evaluation metrics

To evaluate the model's performance, several widely used classification metrics were employed. In this study, the classification performance of the model was quantitatively analyzed using accuracy, recall (sensitivity), precision, F1-score, specificity, negative predictive value (NPV), and ROC-AUC (Receiver Operating Characteristic – Area Under the Curve) values.

For each test sample, the model's predictions were compared with the corresponding ground truth labels, and the confusion matrix was constructed based on these matches. The confusion matrix contains the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Through these values, both the model's ability to correctly classify samples and the nature of its misclassifications were thoroughly examined.

Accuracy represents the proportion of correctly classified samples among all predictions and is defined as in Equation (10):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

However, in imbalanced datasets, accuracy alone is not a sufficient indicator of performance. Therefore, additional metrics were used to evaluate the model's ability to correctly identify positive samples and the reliability of its positive predictions. Recall (sensitivity) measures how well the model captures actual positive cases, while precision indicates how many of the model's positive predictions are truly correct. These metrics are calculated as shown in Equation (11):

$$Recall = \frac{TP}{TP + FN}, \qquad Precision = \frac{TP}{TP + FP} \tag{11}$$

Precision and recall are particularly critical in medical decision-support systems, as a high recall value ensures that diseased or high-risk cases are not overlooked, while a high precision value minimizes false positives, thereby reducing unnecessary alerts or misdiagnoses. To balance these two aspects, the F1-score, which represents their harmonic mean, was employed and is defined as in Equation (11):

$$Recall = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recal}} \tag{12}$$

The F1-score provides a fairer representation of the model's overall performance, especially when the class distribution is imbalanced.

Finally, the model's overall discriminative ability across different decision thresholds was evaluated using the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) value. The AUC quantifies the model's capacity to distinguish between positive and negative classes, where values approaching 1.0 indicate stronger separability and more reliable classification performance.

In the multi-class aortic valve calcification classification task, all these metrics were computed individually for each class and then averaged using the macro-average strategy. This approach treats each class equally, mitigating the impact of class imbalance on the overall evaluation. Through this comprehensive analysis, not only the global accuracy but also the model's capability to correctly discriminate among calcification grades was thoroughly assessed.

In the context of medical image analysis, high sensitivity (recall) ensures that pathological cases are not overlooked, while high specificity prevents healthy individuals from being falsely identified as diseased. Therefore, the joint interpretation of these metrics substantiates the clinical reliability and validity of the proposed Vision Transformer–based classification framework, confirming its potential applicability in real-world echocardiographic assessment of aortic valve calcification.

3 Results and discussion

The experimental results demonstrated that the proposed Vision Transformer (ViT)–based classification model exhibited strong performance in the automatic grading of aortic valve calcification. The model was trained on 224×224 pixel ROI images cropped from echocardiographic frames, where the aortic valve regions were automatically detected using the AVD-YOLOv5 architecture. The training process was conducted for 200 epochs, with an initial learning rate of 2×10^{-4}, which was gradually reduced using a cosine annealing strategy to ensure stable convergence.

The application of controlled augmentation techniques, including horizontal flipping, small-angle rotation, mild blurring, brightness and contrast adjustments, and low-level Gaussian noise, improved dataset variability without affecting clinical relevance. As a result, overfitting was mitigated and model generalization was enhanced.

Additionally, the integration of a weighted oversampling strategy after data equalization further reinforced the balance in class representation during training, ensuring that the model learned more uniformly across all calcification grades. Collectively, these methodological refinements contributed to a robust and reliable classification performance in distinguishing varying degrees of aortic valve calcification from echocardiographic images.

As presented in Table 2, the performance metrics obtained during the test phase indicate that the model achieved a high overall accuracy (0.84), with recall and precision values showing a balanced distribution across all classes. Notably, the F1-scores obtained for the moderate and severe calcification groups (score 2–3) demonstrate the model's strong ability to distinguish between different levels of calcification severity accurately.

Additionally, the ROC curve analysis revealed AUC values ranging between 0.77 and 0.98 for all classes, confirming that the proposed Vision Transformer–based model effectively differentiates between positive and negative samples with high precision. These results collectively emphasize the robustness and clinical potential of the developed framework for reliable, automated grading of aortic valve calcification from echocardiographic images.

3.1 Class-based performance analysis

The class-based performance evaluation of the model was analyzed through the confusion matrix and per-class performance metrics. For Class 0 (low calcification), the results yielded Precision = 0.75, Recall = 0.94, and F1-score = 0.83, indicating that this class achieved the highest overall performance among all categories. The model demonstrated particularly strong sensitivity in detecting low-level calcific regions, successfully distinguishing subtle textural and brightness variations associated with early-stage calcification.

Table 2. Class-wise and average evaluation metrics of the ViT-based aortic valve calcification classification model.

Class (Label)	PRE	REC	ACC	FM
Mild (0)	0.73	1.00	0.80	0.84
Moderate (1)	0.99	0.33	0.94	0.49
Severe (2)	0.66	0.22	0.77	0.33
Critical (3)	0.60	0.75	0.91	0.66
Overall Avg.	0.749	0.576	0.857	0.586

For Class 0 (mild calcification), the model achieved Precision = 0.73, Recall = 1.0, and F1 = 0.84, indicating that it performed best in identifying low-grade calcification. The high recall value shows that the model can reliably detect mild tissue calcifications with minimal false negatives, which is clinically valuable for early-stage identification.

For Class 1 (moderate calcification), the results were Precision = 0.99, Recall = 0.33, F1 = 0.49, and Acc = 0.94. Despite the high AUC value, the relatively low recall suggests that some moderately calcified samples were misclassified into adjacent classes. This behavior is likely due to the subtle and overlapping echogenic patterns observed between mild and severe calcification levels.

For Class 2 (severe calcification), the model obtained Precision = 0.66, Recall = 0.22, F1 = 0.33, and Acc = 0.77, showing a moderate level of discrimination capability. The relatively lower precision and recall values indicate that this class often overlaps with both moderate and critical levels, reflecting the transitional nature of echogenicity in mid-to-high calcification stages.

For Class 3 (critical calcification), the results were Precision = 0.60, Recall = 0.75, F1 = 0.66, and Acc = 0.91, These values reveal that the model exhibits high specificity and confidence when correctly identifying critical cases but tends to misclassify

some highly calcified samples into less severe categories, possibly due to signal saturation and acoustic shadowing effects in ultrasound imaging.

As shown in Figure 5, the confusion matrix demonstrates that the model maintains high performance in classifying samples with low calcification, while most misclassifications occur between neighboring severity levels. This pattern is expected, as transitions between mild, moderate, and severe calcification in echocardiographic images exhibit gradual rather than abrupt changes in echogenicity and texture.

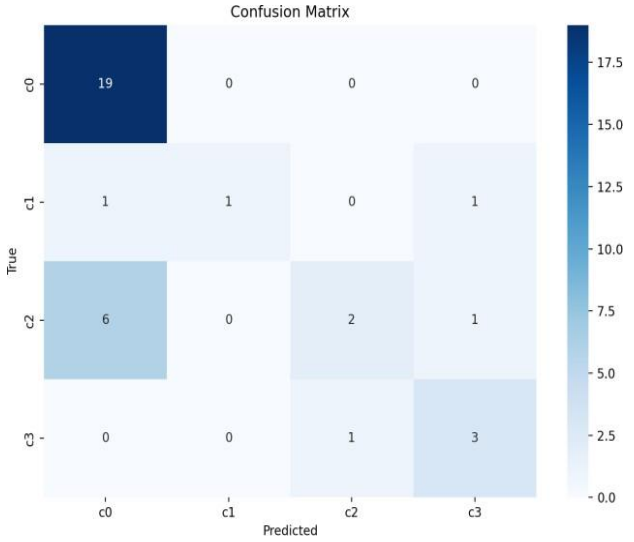

Figure 5. Confusion matrix of the ViT model

For the lowest calcification level (score 0), the model shows excellent discriminative ability, correctly identifying all 19 samples without misclassifying any of them. This result confirms that the model robustly detects tissues with minimal or absent calcific deposits, where structural boundaries and grayscale characteristics are well preserved. For score 1, only one sample was correctly classified, while another was misclassified as score 0, and one as score 3. The limited number of samples in this class likely contributes to this variability, but the errors also reflect the clinical ambiguity of early-stage calcification, where subtle brightening or faint hyperechoic spots may resemble either normal tissue or more advanced stages, depending on the imaging angle and acoustic attenuation.

The score 2 (moderate calcification) class exhibits a clear pattern of misclassification into lower and higher categories. Specifically, 6 samples were assigned to score 0, while 2 were predicted correctly and 1 was labeled as score 3. This distribution highlights the transitional nature of moderate calcification: although its echogenicity is elevated, its spatial homogeneity does not yet match that of heavily calcified valves. As a result, the model's uncertainty in this middle region can be considered clinically meaningful and consistent with known inter-observer variability.

For the highest calcification level (score 3), the model correctly classified 3 samples, while 1 was misclassified as score 2. This confusion can be attributed to shadowing artifacts and signal saturation commonly present in severely calcified valves, which blur boundaries and flatten intensity gradients. Such acoustic effects may cause moderate and severe calcification to appear visually similar, even to expert observers.

Overall, the clustering of misclassifications around adjacent classes instead of being randomly distributed suggests that the model captures the clinical progression of calcification severity. This behavior is consistent with the uncertainty commonly observed among human experts and indicates that the model's predictions follow physiologically meaningful patterns.

Figure 6 presents the class-wise ROC curves. All ROC curves remain well above the diagonal line, confirming that the model possesses strong discriminative capability across all calcification stages. The AUC values were obtained as follows: Class 0 = 0.84, Class 1 = 0.70, Class 2 = 0.75, and Class 3 = 0.97.

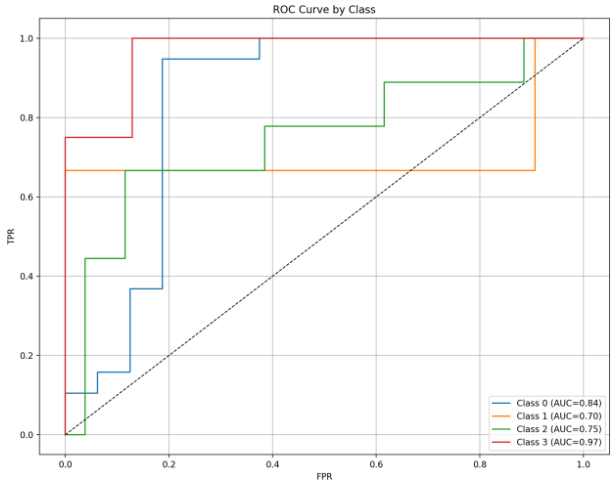

Figure 6. Roc curve

These findings show that the model performs exceptionally well in distinguishing samples from the lowest and the highest calcification classes. The AUC value of 0.84 for score 0 demonstrates strong sensitivity in detecting non-calcified valve textures, while the remarkably high AUC of 0.97 for score 3 indicates highly reliable recognition of severe calcification patterns. The moderate AUC values for score 1 and score 2 reflect known challenges in differentiating early and intermediate calcification, where intensity and texture characteristics overlap with neighboring categories.

Overall, the ROC curves confirm that the ViT-based classifier successfully captures global context and structural patterns even under challenging imaging conditions characterized by acoustic shadowing and contrast variability. These results reinforce the robustness of the proposed method and highlight its potential for reliable clinical use in echocardiography-based calcification assessment.

Figure 7 shows attention maps for correctly and incorrectly classified examples for each calcification class. The top row shows correctly classified examples, while the bottom row shows incorrectly classified examples.

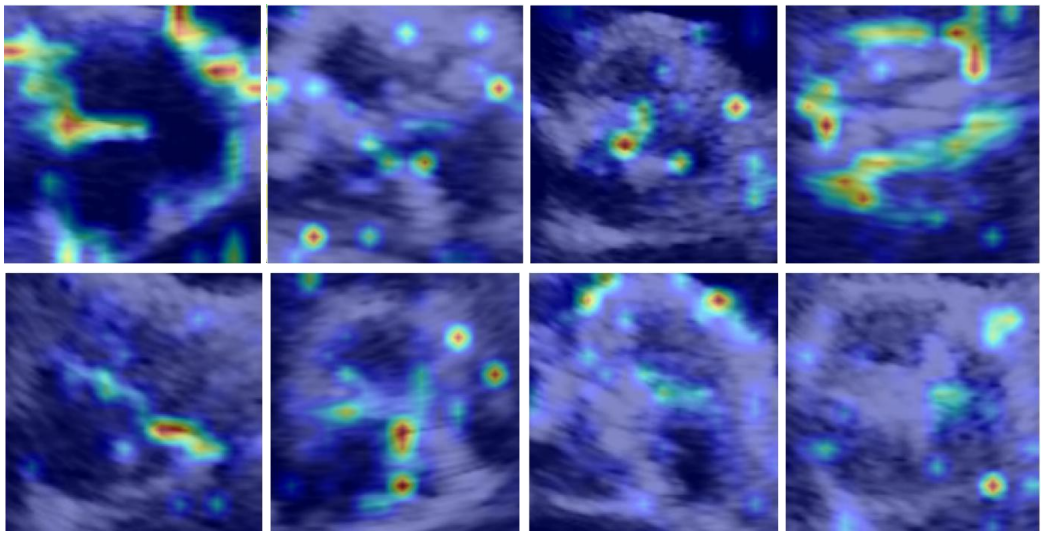

Figure 7. Attention map comparison between correctly and misclassified cases across calcification grades. The upper row illustrates representative correctly classified examples, while the lower row displays misclassified cases. From left to right, the columns correspond to mild, moderate, severe, and critical calcification grades, respectively.

3.2 Overall performance and clinical ınterpretation

The comparative evaluation presented in Table 3 demonstrates a significant performance improvement when transitioning from the original ViT architecture to the proposed ViT-based method. While the baseline ViT struggles with class imbalance and produces imbalanced recall and F1 scores in certain classification categories, the proposed architecture achieves clinically meaningful results with overall Accuracy = 0.8571, Sensitivity = 0.7493, Recall = 0.5763, and F1 score = 0.5861. These values indicate a significant improvement in both discrimination ability and robustness.

Another important point is that the proposed model exhibits high sensitivity but moderate recall behavior. This shows that the model effectively minimizes false positives. In clinical applications, reducing false alarms is critical to preventing unnecessary interventions. The increase in average sensitivity from 0.2309 (Original ViT) to 0.7493 (Proposed ViT) demonstrates that the new approach provides much more reliable predictions and a lower risk of misclassification across all calcification stages.

The results obtained for each class further highlight this improvement. For the mild (0) class, the proposed method achieved a recall of 1.00 and an F1 score of 0.8444. The original ViT, despite its high sensitivity, faces the issue of low recall (0.33). This demonstrates that the proposed model more effectively captures subtle brightness and tissue changes associated with early-stage calcification, enabling more reliable identification of low-intensity cases.

The Moderate (1) class has been historically challenging due to overlapping visual characteristics. The proposed approach improves performance in this category. While the Original ViT demonstrates high recall (≈1.0), it exhibits extremely low precision (0.0666), suggesting a strong tendency toward overprediction. The proposed model achieves a more balanced trade-off between precision and recall (Precision = 0.9999, Recall = 0.3333), indicating improved stability and more clinically consistent decision-making.

For Severe (2) and Critical (3) calcification levels, improvements are particularly important from a clinical perspective. The original ViT failed to detect these categories (Sensitivity = Recall = 0.00), limiting its applicability in advanced disease grading. In contrast, the proposed method achieved F1-scores of 0.3333 for Severe and 0.6666 for Critical calcification, indicating a clear improvement in identifying advanced pathological changes. The improved performance in the Critical class suggests better recognition of bright calcium deposits, irregular tissue appearance, and signal loss regions that commonly occur in advanced calcification on echocardiographic images.

In addition to classification performance, the computational efficiency of the proposed model was evaluated. The average inference time per image was measured as 7.45 ms, indicating that the model is suitable for real-time or near real-time clinical deployment.

Collectively, these improvements can be attributed to ROI-based preprocessing steps and the representational advantages of the Vision Transformer architecture in a clinical context. The self-attention mechanism enables the model to capture broader spatial dependencies and holistic contrast variations. These elements are often difficult for traditional CNNs to model effectively due to their localized kernel structures. Furthermore, leveraging pre-trained ImageNet weights increases feature richness, enabling the model to generalize effectively even with limited medical imaging data.

Table 4 presents a comparative evaluation of ResNet50, EfficientNet-B3, and the proposed Vision Transformer (ViT)-based model for multi-class aortic calcification score classification. While all architectures demonstrate strong performance in specific categories, the proposed ViT model exhibits a relatively more balanced performance profile across severity levels.

Table 3. Comparison of Class-Wise performance metrics between original ViT and proposed ViT model

Method	**ViT (Original)**				**ViT (Proposed Method)**			
Class (Label)	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Mild (0)	0.5185	0.8571	0.3333	0.480	0.8000	0.7307	1.0000	0.8444
Moderate (1)	0.4814	0.0666	0.9999	0.125	0.9428	0.9999	0.3333	0.4999
Severe (2)	0.7407	0.0	0.0	0.000	0.7714	0.6666	0.2222	0.3333
Critical (3)	0.7777	0.0	0.000	0.000	0.9142	0.6000	0.7500	0.6666
Overall Avg.	0.6296	0.230952	0.3333	0.1512	0.8571	0.7493	0.5763	0.5861

Table 4. Performance comparison of the task-optimized ViT and ImageNet pre-trained ResNet50 and EfficientNet-B3 models under the same experimental setup

Method	**ResNet50**				**EfficientNet-B3**				**Proposed-ViT**			
Class (Label)	Precision	Accuracy	Recall	F1 M.	Precision	Accuracy	Recall	F1 M.	Precision	Accuracy	Recall	F1 M.
Mild (0)	0.86	0.91	1.00	0.92	0.82	0.88	1.00	0.90	0.73	0.80	1.00	0.84
Moderate (1)	0.28	0.82	0.66	0.40	0.50	0.91	0.33	0.39	0.99	0.94	0.33	0.49
Severe (2)	0.00	0.74	0.00	0.00	0.55	0.77	0.55	0.55	0.66	0.77	0.22	0.33
Critical (3)	0.50	0.88	0.75	0.59	0.99	0.91	0.25	0.40	0.60	0.91	0.75	0.66
Overall Avg.	0.412	0.84	0.60	0.48	0.72	0.87	0.53	0.56	0.74	0.85	0.57	0.58

In correctly classified examples (top row), the model's attention is concentrated on clinically significant areas of the aortic valve. In Class 0 (True: 0, Pred: 0) examples, attention is more widespread and evenly distributed across the valve structure, consistent with the absence of a distinct calcified area. In Class 1 examples, the model is observed to focus on small and limited bright areas seen on the valve leaflets. This indicates that the model can distinguish early and mild calcification.

In Class 2 examples, attention maps are concentrated on more distinct and clustered bright bands. These bands correspond to areas of moderate calcification. In Class 3 examples, warm colors are seen to be concentrated on large and intense bright areas covering a large part of the cap structure. This indicates that advanced calcification is perceived by the model as a widespread and intense increase in brightness.

In contrast, in misclassified samples (bottom row), the attention distribution partially deviates from the expected main calcification areas, shifting to scattered bright spots, image artifacts, or unclear transition zones between calcified and non-calcified tissues. In some cases, the model focuses on a limited bright area rather than the overall spread of calcification, leading to an underestimation or overestimation of disease severity.

Overall, in correctly classified samples, the model's attention is concentrated on anatomically and clinically meaningful calcification areas, while in misclassifications, attention shifts to more ambiguous or secondary areas. These findings indicate that the model's decision-making process is largely based on actual calcification areas and that attention maps enhance the model's interpretability.

At the class level, all models perform well in the Mild category. In the Moderate and Severe classes, CNN-based models show noticeable fluctuations between precision and recall, whereas the ViT model demonstrates a more conservative behavior, particularly in reducing false positives. In the Critical category, the ViT achieves a more balanced trade-off between precision and recall, resulting in a more stable F1-score. Overall, in classes where adjacent severity levels are difficult to distinguish, the ViT model appears to provide more controlled and consistent decision behavior.

When macro-averaged metrics are considered, the proposed ViT model achieves the highest precision and F1-score among the evaluated architectures. Although performance differences vary across classes, the ViT demonstrates relatively more consistent behavior across severity grades. This stability may be advantageous in multi-grade medical classification tasks where transitions between adjacent classes are clinically relevant.

Overall, these findings suggest that the transformer-based architecture may offer advantages in modeling global contextual relationships in certain categories. Nevertheless, CNN-based architectures also exhibit competitive performance, and model selection should be guided by dataset characteristics and clinical requirements.

From a clinical perspective, the proposed ViT-based framework demonstrates reliable capacity to distinguish clinically relevant calcification stages using a radiation-free method. Improvements in accuracy, inter-class stability, and the ability to detect advanced calcification severity indicate the system's suitability for real-world diagnostic support. These findings highlight the potential of Transformer-based architectures to provide a more holistic and interpretable understanding of heart valve morphology compared to traditional convolutional methods.

4 Conclusions

In this study, a Vision Transformer (ViT)-based classification model was proposed for the automatic evaluation of aortic valve calcification from transthoracic echocardiographic images. The developed system first utilized the aortic valve regions detected by the AVD-YOLOv5 model and subsequently classified the frames extracted from these ROIs to predict calcification scores (ranging from 0 to 3).

During the training process, gentle data augmentation techniques were applied to mitigate the effects of class imbalance and enhance the model's generalization ability. In addition, regularization strategies such as cosine annealing, label smoothing, and gradient clipping were employed to improve the stability of the training process. The experimental results demonstrated that the ViT-based architecture achieved high discriminative performance across both low and high calcification levels. The high ROC-AUC values indicate that the model can serve as a clinically reliable decision support tool for aortic valve calcification assessment.

In conclusion, the proposed ViT-based approach provides a radiation-free, fast, and fully automated framework for evaluating aortic valve calcification from echocardiographic images. This study demonstrates the feasibility of applying an artificial intelligence-based approach for the early detection and progression monitoring of calcification using echocardiography, a widely accessible and non-invasive imaging modality.

The model's high accuracy and discriminative capability suggest that, when integrated into clinical decision support systems, it has the potential to standardize diagnostic evaluations among cardiologists. This would help reduce observer-dependent variability and enable objective and repeatable quantification of calcification severity.

In the future, validating the proposed system on larger and multi-center datasets will be critical for achieving robust and generalizable performance across data acquired from different ultrasound devices and imaging conditions. Furthermore, extending the model to video-based (temporal) ViT-LSTM hybrid architectures could enable tracking of calcification dynamics over time and incorporate valve motion analysis into the diagnostic framework. These advancements will contribute significantly to the development of a clinically reliable, AI-assisted diagnostic tool for real-world healthcare applications.

CRediT authorship contribution statement

Mervenur Cakir: Data organization, Methodology, Software, Visualization, Writing—first draft; **Murat Ekinci:** Consulting, Conceptualization, Methodology; **Elif Baykal Kablan:** Method; Verification, Writing, Review, and Editing; **Mursel Sahin:** Data editing, Sourcing, Project management.

Acknowledgment

This study was fully supported by the TUBITAK Research Project 222S110.

Conflict of interest

The authors declare that there is no conflict of interest.

Ethics declaration

The study was approved by the ethical review boards on Farabi Hospital (2022/190).

Similarity rate (iThenticate): 16%

Declaration regarding the use of generative artificial intelligence (AI) and AI-powered technologies in the writing process

While preparing this paper, the authors used the ChatGPT tool to improve the language and readability. After using this tool, the authors reviewed and edited the content as necessary and are fully responsible for the published content.

References

[1] R. L. J. Osnabrugge et al., "Aortic stenosis in the elderly: disease prevalence and number of candidates for transcatheter aortic valve replacement: a meta-analysis and modeling study," *Journal of the American College of Cardiology*, vol. 62, no. 11, pp. 1002-1012, 2013. https://doi.org/10.1016/j.jacc.2013.05.015

[2] L. Tastet et al., "Impact of aortic valve calcification and sex on hemodynamic progression and clinical outcomes in AS," *Journal of the American College of Cardiology*, vol. 69, no. 16, pp. 2096-2098, 2017. https://doi.org/10.1016/j.jacc.2017.02.037

[3] T. A. Pawade, D. E. Newby, and M. R. Dweck, "Calcification in aortic stenosis: the skeleton key," *Journal of the American College of Cardiology*, vol. 66, no. 5, pp. 561-577, 2015. https://doi.org/10.1016/j.jacc.2015.05.066

[4] V. Falk et al., "2017 ESC/EACTS Guidelines for the management of valvular heart disease," *European Journal of Cardio-Thoracic Surgery*, vol. 52, no. 4, pp. 616-664, 2017. https://doi.org/10.1093/ejcts/ezx324

[5] A. S. Agatston et al., "Quantification of coronary artery calcium using ultrafast computed tomography," Journal of the American College of Cardiology, vol. 15, no. 4, pp. 827-832, 1990. https://doi.org/10.1093/ejcts/ezx324

[6] L. Tang et al., "DLFFNet: A new dynamical local feature fusion network for automatic aortic valve calcification recognition using echocardiography," *Computer Methods and Programs in Biomedicine*, vol. 243, p. 107882, 2024. https://doi.org/10.1016/j.cmpb.2023.107882

[7] M. Çakır et al., "AVD-YOLOv5: a new lightweight network architecture for high-speed aortic valve detection from a new and large echocardiography dataset," *Medical & Biological Engineering & Computing*, vol. 62, no. 8, pp. 2511-2528, 2024. https://doi.org/10.1007/s11517-024-03090-3

[8] M. Cakir et al., "A New Aortic Valve Calcium Scoring Framework for Automatic Calcification Detection in Echocardiography," *Journal of Imaging Informatics in Medicine*, pp. 1-19, 2025. https://doi.org/10.1007/s10278-025-01576-6

[9] L. B. Elvas et al., "Deep learning for automatic calcium detection in echocardiography," *BioData Mining,* vol. 17, no. 1, p. 27, 2024. https://doi.org/10.1186/s13040-024-00381-1

[10] L. B. Elvas et al., "AI-based aortic stenosis classification in MRI scans," *Electronics*, vol. 12, no. 23, p. 4835, 2023. https://doi.org/10.3390/electronics12234835

[11] H. Vaseli et al., "Protoasnet: Dynamic prototypes for inherently interpretable and uncertainty-aware aortic stenosis classification in echocardiography," in Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2023. https://doi.org/10.48550/arXiv.2307.14433