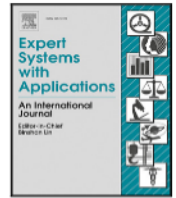




Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Expert Systems With Applications

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



Highlights

Automatic Cardiac Pathology Recognition in Echocardiography Images using Higher Order Dynamic Mode Decomposition and a Vision Transformer for Small Datasets

Andrés Bell-Navas^{*}, Nourelhouda Groun, María Villalba-Orero, Enrique Lara-Pezzi, Jesús Garicano-Mena, Soledad Le Clainche

- An automatic heart disease recognition system in real-time is proposed.
- A novel procedure creates a large annotated database from echocardiograms to train.
- Higher Order Dynamic Mode Decomposition (HODMD) used for first time in medicine.
- The HODMD algorithm improves accuracy by data augmentation and feature extraction.
- The designed Vision Transformer can be trained from scratch with small datasets.

Expert Systems With Applications xxx (xxxx) xxx

Graphical abstract and Research highlights will be displayed in online search result lists, the online contents list and the online article, but **will not appear in the article PDF file or print** unless it is mentioned in the journal specific style requirement. They are displayed in the proof pdf for review purpose only.

Automatic Cardiac Pathology Recognition in Echocardiography Images using Higher Order Dynamic Mode Decomposition and a Vision Transformer for Small Datasets

Andrés Bell-Navas^{a,*}, Nourelhouda Groun^{a,b}, María Villalba-Orero^{c,d}, Enrique Lara-Pezzi^c, Jesús Garicano-Mena^{a,e}, Soledad Le Clainche^{a,e}

^a ETSI Aeronáutica y del Espacio, Universidad Politécnica de Madrid, Pl. del Cardenal Cisneros, 3, 28040, Madrid, Spain

^b ETSI Telecomunicación, Universidad Politécnica de Madrid, Av. Complutense, 30, 28040, Madrid, Spain

^c Centro Nacional de Investigaciones Cardiovasculares (CNIC), C. de Melchor Fernández Almagro, 3, 28029, Madrid, Spain

^d Departamento de Medicina y Cirugía Animal, Facultad de Veterinaria - Universidad Complutense de Madrid, Av. Puerta de Hierro, 28040, Madrid, Spain

^e Center for Computational Simulation (CCS), 28660, Boadilla del Monte, Spain

ARTICLE INFO

Keywords:

Cardiac pathology recognition

Deep learning

Echocardiography imaging

Higher order dynamic mode decomposition

Vision transformers

ABSTRACT

Heart diseases are the main international cause of human defunction. According to the WHO, nearly 18 million people de cease each year because of heart diseases. Also considering the increase of medical data, much pressure is put on the health industry to develop systems for early and accurate heart disease recognition. In this work, an automatic cardiac pathology recognition system based on a novel deep learning framework is proposed, which analyses in real-time echocardiography video sequences. The system works in two stages. The first one transforms the data included in a database of echocardiography sequences into a machine learning-compatible collection of annotated images which can be used in the training phase of any kind of machine learning-based framework, including deep learning. This includes the use of the Higher Order Dynamic Mode Decomposition (HODMD) algorithm, for the first time to the authors' knowledge, for both data augmentation and feature extraction in the medical field. The second stage is focused on building and training a Vision Transformer (ViT), barely explored in the related literature. The ViT is adapted for an effective training from scratch, even with small datasets. The designed neural network analyses images from an echocardiography sequence to predict the heart state. The results obtained show the efficacy of the HODMD algorithm and the superiority of the proposed system, even outperforming pretrained Convolutional Neural Networks (CNNs), which are so far the method of choice in the literature.

1. Introduction

Heart diseases or cardiovascular diseases (CVDs) are the main cause of human defunction in the world (Arooj et al., 2022). According to the World Health Organization (WHO), a report conducted in 2019 (World Health Organization, 1999) shows that approximately 18 million people deceased that year precisely due to CVDs, supposing nearly a third of the total defunctions. Not only the number of patients is growing, but also medical data availability. Therefore, a lot of pressure is put on the health industry to develop systems for early and accurate detection of heart diseases, which usually takes much time, effort, and demanding resources. Researchers and medical experts mainly rely on images for the diagnosis, because of containing much information about the

heart state useful for the further clinical procedures to adopt. Echocardiography imaging is very widely used, in particular transthoracic echocardiography (TTE), as being rapid and non-ionizing for diagnosis, therapeutic planning, and preoperative management of patients. The widespread use of mobile (handheld) echocardiography devices has significantly increased the recognition rate of heart diseases. However, several challenges concerning the quality and characteristics of the imagery (e.g., poor contrast, noise) remain. These hamper the interpretations, which require the cardiologist's expertise.

On the other hand, as a tractable and repeatable modality, processing echocardiography images for diagnoses of cardiac pathologies through deep learning has become popular. This is because Artificial Intelligence (AI) has demonstrated to improve welfare and could make

* Corresponding author.

E-mail addresses: a.bell@upm.es (A. Bell-Navas), gr.nourelhouda@alumnos.upm.es (N. Groun), mvorero@ucm.es (M. Villalba-Orero), elara@cnic.es (E. Lara-Pezzi), jesus.garicano.mena@upm.es (J. Garicano-Mena), soledad.leclainche@upm.es (S.L. Clainche).

<https://doi.org/10.1016/j.eswa.2024.125849>

Received 22 July 2024; Received in revised form 23 September 2024; Accepted 17 November 2024

diagnoses automatic, quicker, more accurate and reproducible, reducing human mistakes and costs, and complementing decision-making processes. Several classification techniques have been proposed to help health-care experts in the diagnoses of heart diseases. For example, in [Wahlang et al. \(2021\)](#), a Long Short Term Memory (LSTM) and a Variational Autoencoder (VAE) were proposed for binary classification (normal or abnormal) and classification between six different types of regurgitation in 2D echo, 3D Doppler, and videographic images. Similarly, in [Madani, Ong, Tibrewal, and Mofrad \(2018\)](#), a supervised model and a semi-supervised generative adversarial network (GAN) were used to classify left ventricle hypertrophy (LVH) and 15-view echocardiography images. The work in [Vafaezadeh, Behnam, Hosseinsabet, and Gifani \(2023\)](#) proposes CarpNet, a novel structure which aggregates spatio-temporal information by integrating CNNs and a Transformer, for the first time, for diagnosis of Mitral Valve (MV) leaflet abnormalities in four common Carpentier functional classes in parasternal long-axis (PLA) echocardiographic videos. Although the classification of these four classes using few test samples is challenging, Inception and ResNet version 2 with transfer learning for the CNN part of CarpNet have demonstrated to be superior to image-based algorithms using only spatial information. In [Farhad et al. \(2023\)](#), novel zero-shot and few-shot learning based on a Siamese Neural Network were proposed for the first time for classification of LVH and healthy states using only 70 two-chamber (2CH) and four-chamber (4CH)-view echocardiography images. Contrarily to traditional zero-shot learning, this work does not require text vectors for image description. A shallow CNN was used as feature extractor, and the Euclidean distance was employed to compute the similarity between echo images. The proposed Siamese Network is superior to VGG16 and ResNet50 with transfer learning. The work from [Holste, Oikonomou, Mortazavi, Wang, and Khera \(2022\)](#) proposes self-supervised learning for effective representation learning for downstream fine-tuning on the task of automatic severe aortic stenosis (AS) diagnosis in echocardiogram videos with very little labelled data. To this end, positive pairs of different videos from the same PLAX view of the same patient are formed as a multi-instance contrastive learning. In this way, strong data augmentation which could potentially degrade signal is not required. A 3D ResNet18 is employed for the classification. In addition, the pretext task of frame re-ordering is used to incorporate temporal information, further aiding downstream fine-tuning, via a fully-connected classification head. The work in [Chen et al. \(2023\)](#) addresses hypertensive cardiomyopathy (HCM) detection in echocardiography videos from 4CH view as a weakly supervised learning problem with a multi-instance deep learning framework. It uses a CNN for effective incorporation of temporal correlated multi-instance learning, end-systole and end-diastole attributes. In addition, the work employs domain adversarial neural networks for systematic addressing of variable appearance due to noise, blur, contrast, etc. In this way, the snippet most representative of the heart state in the video becomes greatly discriminative. Using a non-local I3D model pretrained on Kinetics-400 data set as the 3D CNN feature extractor, the approach has demonstrated to outperform the vanilla I3D algorithm.

Among the aforementioned works, CNNs are mostly proposed and usually focus on either binary classification (i.e., one specific disease), or heart state categories related to the same disease (like in [Vafaezadeh et al., 2023](#)). Besides, works proposing alternative algorithms, e.g., Transformers, either use very simple networks or vanilla versions of the algorithms, without any further contributions or adaptations to better address the heart disease recognition task.

In this paper, a different solution is proposed, based on deep learning methods for recognition of different cardiac pathologies from echocardiography images. A novel procedure has been devised to create a large database from different sources of echocardiography images with heart diagnosis provided by human experts. One of the main purposes for this database creation process is to improve the discriminating capability of standardized neural network architectures in the usual scenario of having a limited number of samples. Precisely, in the

medicine field, the difficulty resides in obtaining a varied database of echocardiography images with high quality for an adequate training of deep learning algorithms. This is because very hard specialized work with specific knowledge about heart diseases is required, implying high costs. In addition, these sources of echocardiography images are heterogeneous regarding the image resolution and frame-rate, due to different acquisition conditions and sensors used. Therefore, it is not possible to use them simultaneously to create a larger database. The proposed procedure enables the homogenization of the different sources of echocardiography images to create a larger database required by the proposed deep learning algorithm, again, contributing to performance improvements. This also includes the use of a data-driven method, the Higher Order Dynamic Mode Decomposition (HODMD) ([Le Clainche & Vega, 2017](#)), for the first time to the authors' knowledge. Its purpose is to generate images with less noise and more discriminative features than the original data (with high noise by nature), associated to the different cardiac conditions, thus augmenting the training database and improving recognition accuracy. In fact, the HODMD modes contain temporal information, extracted from the sequences, which could be relevant to characterize different heart states. Regarding the proposed deep neural network, it incorporates modules to palliate the usual need for large databases, especially in Vision Transformers (ViT) ([Lee, Lee, & Song, 2021](#)), further addressing the difficulty to gather large high-quality databases in the medicine field. As a result, the proposed system can automatically estimate the heart state from echocardiography video sequences without the intervention of human experts during its operation. Moreover, that estimation can be performed in real-time. This work largely extends the preliminary study carried out in [Bell-Navas, Groun, Garicano-Mena, and Le Clainche \(2023\)](#) by proposing the new database creation procedure, the deep learning architecture, and giving experimental results about the cardiac pathology recognition performance. The source code used in this work will be incorporated into the next version release of the ModelFLOWS-app ([Hetherington et al., 2023](#)), which can be found in [ModelFLOWS research group \(2023\)](#).

The organization of the paper is as follows. Section 2 describes the proposed cardiac pathology recognition system. Section 3 introduces the database created to assess the proposed system. Section 4 summarizes the obtained results with the proposed system and makes a comparison with other algorithms. Finally, conclusions are drawn in Section 5.

2. Cardiac pathology recognition system

The cardiac pathology recognition system here proposed is aimed to analyse echocardiography images using an adapted neural network architecture to predict the heart states in real-time. The neural network is based on a ViT that introduces modules to palliate the need for large training databases. Therefore, the ViT can be effectively trained from scratch, even on small datasets. The system output is the prediction of the heart state from the input sequence of echocardiography images. The cardiac pathology recognition system has been carefully trained using a new and large cardiac pathology database, whose creation is a key contribution of this paper. The medical database creation consists of a multi-stage procedure applied to sequences of echocardiography images showing hearts with different states and pathologies. The designed procedure not only generates a machine learning-compatible medical database, but also deals with the fact that the echocardiography images have been acquired with different sensors, and therefore they have different characteristics (image resolution and frame-rate, among others). On the other hand, it alleviates the need to collect large databases required by ViTs ([Lee et al., 2021](#)), as this results very expensive in the medicine field, supposing much effort from experts.

[Fig. 1](#) depicts the block diagram of the proposed cardiac pathology recognition system, which is divided into two major stages: Medical Database Creation, and Pathology Recognition; those are described in detail in the next subsections.

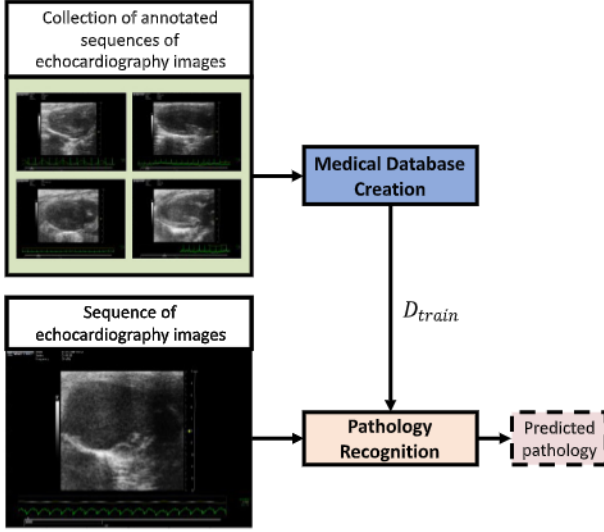


Fig. 1. Block diagram of the proposed cardiac pathology recognition system, with representations of the echocardiography images with non-heart regions (the electrocardiogram, medical information, and the black background).

2.1. Medical database creation

This stage creates a large annotated database from echocardiography images acquired with different sensors. The input consists of a heterogeneous collection of sequences of echocardiography images with additional medical information. Each sequence has the diagnosis of the heart state made by doctors. The output of this stage is a database composed of samples (as images) representing different heart states with homogenized spatial and temporal characteristics, adjusted to a target acquisition device. In this way, the resulting database can be used by machine learning algorithms. This Medical Database Creation stage can be further divided into two phases, as shown in Fig. 2: (1) Data Homogenization, and (2) Modal Decomposition-based Data Generation.

The first phase, Data Homogenization, extracts the images from the sequences, in the way that only the region representing the heart is retained, and all areas outside are discarded (containing medical information and the black background). This is achieved by using edge detection algorithms which locate the boundaries of the area corresponding to the echocardiography image (i.e., containing the heart). After that, the images containing text and graphics superposed on the heart area are filtered out. This is done to remove outliers, as containing artificial data which do not represent the heart. Therefore, these should not be features associated to different heart states to be learned by the proposed deep learning algorithm. For this purpose, Optical Character Recognition (OCR) has been adopted. This is because the artificial data in the studied databases always contain alphanumeric characters. Note that this makes sequences to have frames with different time intervals. However, in the second phase, Modal Decomposition-based Data Generation, the time interval between consecutive frames in a sequence must remain the same. This is accomplished by splitting a sequence with frames removed into sub-sequences, each one with the maximum possible number of consecutive frames with the same time interval. Each sub-sequence is therefore reinterpreted as an individual sequence of echocardiography images.

In the second and last phase, Modal Decomposition-based Data Generation, novel samples are generated from the homogenized sequences. These consist of images with more discriminative features than those of the original echocardiography data. This generation is done by sequentially applying the Singular Value Decomposition (SVD) (Sirovich, 1987) and the HODMD algorithms (Groun et al., 2022; Le Clainche

& Vega, 2017) on each homogenized sequence, and taking from the outputs of each algorithm. Therefore, this phase can be seen as a machine learning method based on the data physics, in which data augmentation with modal decomposition is performed. Conventional data augmentation techniques based on geometric transformations have been applied as well, as will be specified in Section 2.2, complementing this data augmentation based on modal decomposition. However, these conventional techniques are not based on physics and only explore slight variations of the actual samples. Therefore, these do not generate discriminative features which strengthen the differentiation between the heart states, so the conventional data augmentation techniques barely lead to performance improvements.

The steps performed in the Modal Decomposition-based Data Generation phase are as follows. First, the SVD algorithm is employed. This generates a series of modes and the reconstructions of the echocardiography images. Next, the HODMD algorithm is applied on these reconstructions obtained with the previous SVD algorithm. As a result, HODMD modes and reconstructions associated to the sequence are obtained. Section 2.1.1 describes in more detail the HODMD algorithm for feature extraction and data augmentation. Finally, the SVD algorithm is applied on the HODMD modes, obtaining SVD modes associated to the HODMD modes, and their reconstructions. From the data obtained in this latest use of the SVD algorithm, only the reconstructions of the HODMD modes are considered, and the modes are discarded. Note that a sequence must cover a number of heart cycles enough to be able to apply the HODMD algorithm and, therefore, the subsequent SVD algorithm. In this way, the temporal information is correctly characterized, and fair reconstructions can be obtained. This is accomplished by applying a threshold which indicates the minimum number of snapshots required in a sequence to apply the HODMD algorithm. Note that the reconstructions are leveraged due to having less noise than their associated original data, making the patterns of the different heart pathologies more distinct. In fact, original echocardiography imagery is typically afflicted by noise. Therefore, the SVD and HODMD algorithms further contribute to the extraction of discriminative features due to also effectively reducing noise as one of their main roles, improving the final recognition performance. In addition, the deep neural network can potentially learn better the temporal information of the sequences with the use of the obtained HODMD modes and their associated reconstructions (obtained with the second use of the SVD algorithm). It is due to that the HODMD modes describe the physical patterns associated to the dynamics of the data. These physical patterns can be related to heart diseases (Groun et al., 2022).

As a result, an annotated database with a large number of cardiac pathology samples is obtained, whose spatial and temporal characteristics are adjusted to the target sensor used in the proposed recognition system. This is used as a training database D_{train} to learn a deep neural model, which is the core of the Pathology Recognition stage described in Section 2.2. Table 2 summarizes the different combinations of the data generated by the SVD and HODMD algorithms which are taken to form D_{train} . Precisely, the combinations with HODMD modes and reconstructions of the echocardiography images lead to the best results, because of having more samples and also more discriminative features derived from modal decomposition.

2.1.1. Higher order dynamic mode decomposition

The Higher Order Dynamic Mode Decomposition algorithm (HODMD) (Le Clainche & Vega, 2017) is an extension of the Dynamic Mode Decomposition (DMD) (Schmid, 2010), extensively used in fluid dynamics and for diverse industrial applications (Groun et al., 2022; Vega & Le Clainche, 2021). HODMD decomposes spatio-temporal data (in this case, a video sequence of echocardiography images) into a number of DMD modes. Each one is associated to a frequency, a growth rate, and an amplitude. The most important modes represent the most discriminative patterns associated to the different heart states in echocardiography images. Specifically, according to Groun et al.

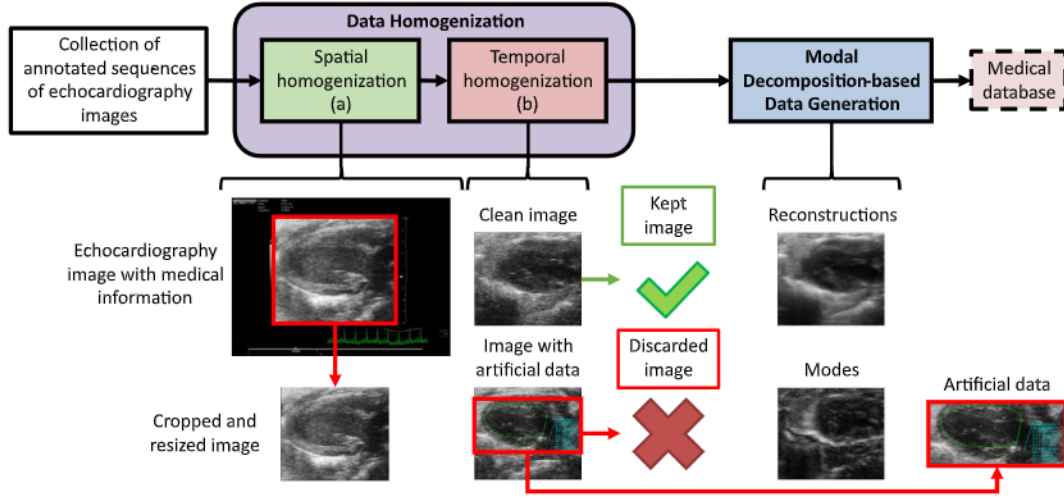


Fig. 2. Block diagram of the Medical Database Creation stage, composed of two phases: Data Homogenization, and Modal Decomposition-based Data Generation. The former phase is further divided into (a) Spatial homogenization and (b) Temporal homogenization.

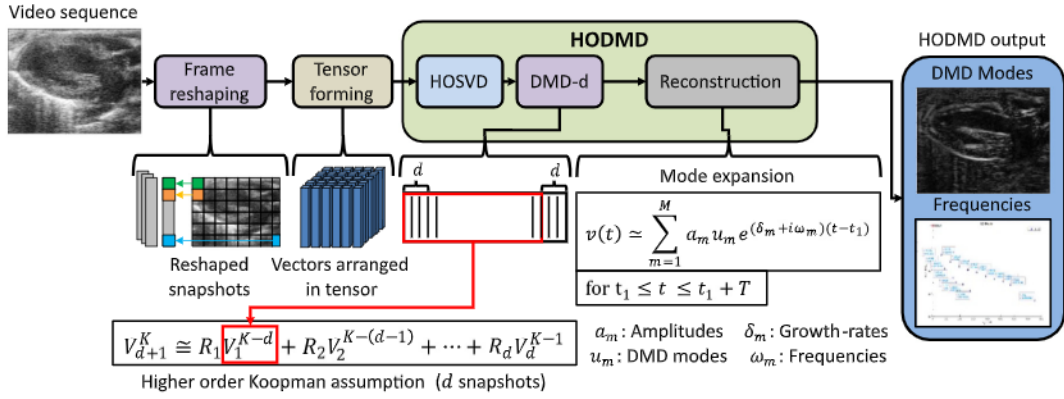


Fig. 3. Block diagram of the HODMD algorithm applied on a sequence of echocardiography images.

(2022), these most relevant modes are typically associated with the highest amplitudes. This decomposition also allows to identify and filter out noisy modes, which, in turn, tend to be those associated to small amplitudes (i.e., a_m in Eq. (2) below). In particular, these consist in high frequencies related to the high noise inherent to the echocardiography imagery, which, therefore, do not contribute with relevant features of the data and of the heart states. In this way, reconstructions of each frame from the input sequence could be obtained, with less noise and with more discriminative features associated to the different heart states. The HODMD algorithm used in this work is specifically the multidimensional iterative HODMD algorithm (Groun et al., 2022; Le Clainche, Vega, & Soria, 2017). It is based on the Higher Order Singular Value Decomposition (HOSVD) (Tucker, 1966), which extends the SVD algorithm by applying it along each spatial dimension to better clean the data.

Fig. 3 depicts the steps performed by the HODMD algorithm on a sequence of echocardiography images, which are as follows. Assuming that the sequence has been homogenized as described in Section 2.1, each frame (or snapshot) is reshaped into a vector of dimensions $N_p = N_x \times N_y$, where N_x and N_y denote the spatial resolution of the snapshots, and N_p the number of pixels of the snapshot. Once all vectors have been computed, a tensor is formed. The dimensions of the tensor are $N_p \times K$, where K is the number of snapshots of the sequence. Then, a dimensionality reduction is carried out by applying the HOSVD algorithm. After that, a decomposition in eigenvalues and eigenvectors

is conducted. In this case, a number of d subsequent snapshots is considered as a contribution given by the HODMD algorithm to improve the precision.

For this purpose, the higher order Koopman assumption is applied, which relates an echocardiography image with their d previous snapshots with the expression as follows:

$$V_{d+1}^K \approx R_1 V_1^{K-d} + R_2 V_2^{K-(d-1)} + \dots + R_d V_d^{K-1}, \quad (1)$$

where V_{d+1}^K represents the measurement of the image one time step into the future, related with their d previous K equispaced snapshots by the R_i Koopman operators. Next, the DMD modes u_m are computed. Finally, a mode expansion process is performed, obtaining the frequencies, the growth-rates, and the amplitudes, using the following expression:

$$v(t) \approx \sum_{m=1}^M a_m u_m e^{\delta_m + i\omega_m(t-t_1)} \quad \text{for } t_1 \leq t \leq t_1 + T, \quad (2)$$

where $v(t)$ represents the spatio-temporal data (i.e., the sequence of echocardiography images) as an expansion of M DMD modes; t is the time, T the sampled timespan, a_m the (real) amplitudes, u_m the normalized spatial modes, δ_m the growth rates, and ω_m the frequencies. The Koopman operator (Le Clainche and Vega (2017)) allows to relate measurements from a non-linear system (in this case, echocardiography images) in consecutive time steps with a linear operator of infinite dimension. This means that the HODMD algorithm, although formulated with linear operators, as an approximation of the Koopman operator, can deal with the non-linearity which could be present in the

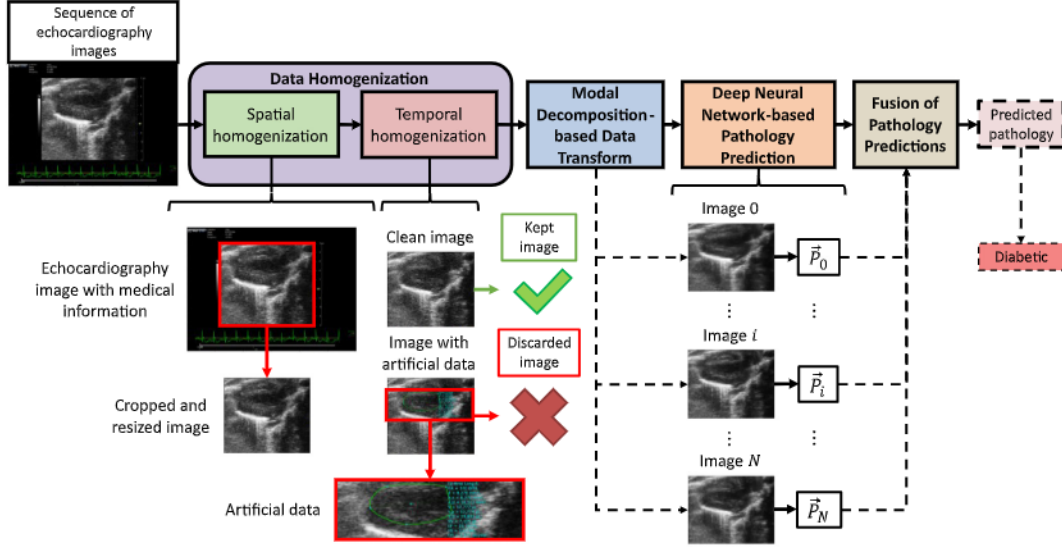


Fig. 4. Block diagram of the Pathology Recognition stage.

echocardiography imagery. All these steps are applied iteratively until the number of HOSVD modes converges, i.e., remains the same. The DMD modes allow to eventually reconstruct each snapshot (each frame) from the input sequence.

Among the obtained DMD modes, only the most representative ones are selected for the feature extraction and data augmentation. These representative modes are those containing characteristic patterns associated to the different heart states. The selection of modes is performed by first comparing their associated frequencies with those which are characteristic of the different heart states, previously identified with the HODMD algorithm in Groun et al. (2022). The frequencies related to the different heart states are those associated to higher amplitudes. Therefore, their corresponding DMD modes describe the physical patterns associated to the dynamics of the echocardiography imagery. In this way, the modes with higher amplitudes, and with frequencies more similar to those of the heart states (identified in Groun et al., 2022) are taken. Regarding the other modes, which tend to have higher frequencies and smaller amplitudes, are associated to the noise of the echocardiography imagery, and so discarded. The most representative DMD modes, together with the reconstructions, can be leveraged by the deep neural network to improve the recognition performance.

2.2. Pathology recognition

The Pathology Recognition stage predicts heart states from the corresponding input echocardiography images in real-time. Specifically, the input is a sequence of echocardiography images obtained by medical devices. It is adapted to the target acquisition device used in production in the previous Medical Creation Stage. The output of the Pathology Recognition stage is a cardiac pathology prediction which indicates the most probable heart state.

The Pathology Recognition stage can be divided into four phases, as can be seen in Fig. 4: (1) Data Homogenization, (2) Modal Decomposition-based Data Transform, (3) Deep Neural Network-based Pathology Prediction, and (4) Fusion of Pathology Predictions. In the first phase, Data Homogenization, the input sequence is homogenized as already described in Section 2.1.

In the second phase, Modal Decomposition-based Data Transform, the same process as in the Modal Decomposition-based Data Generation phase described in Section 2.1 is performed. This is, images with less noise and more discriminative features than the original echocardiography data are obtained by sequentially applying the SVD and the HODMD algorithms, increasing the differentiation between the heart

states and improving the performance. Note that, in this case, the resulting images from this phase are exclusively taken as input of the deep neural network for the prediction of the heart state of the corresponding sequence. Therefore, this phase only acts as a feature extractor. However, the Modal Decomposition-based Data Generation phase leverages the output of each applied algorithm to enlarge the training database D_{train} , directly fed to the deep neural network for training, containing more discriminative samples for the differentiation between the heart states, and addressing the limited availability of high-quality samples usual in the medicine field. In Table 3, Table 4, and in Table 6, different types of data resulting from the Modal Decomposition-based Data Transform phase have been tested as input of the deep neural network for the prediction of heart states. In particular, the use of the reconstructions of the original echocardiography images resulting from the modal decomposition techniques, especially those from the HODMD algorithm, leads to better performances. This fact demonstrates the efficacy of these techniques to extract more discriminative features associated to the different heart states from echocardiography images and to filter out the high noise inherent, leading to more reliable heart state predictions.

The third phase, Deep Neural Network-based Pathology Prediction, computes the cardiac pathology prediction for each image generated in the previous phase, Modal Decomposition-based Data Transform. For this purpose, a deep neural network architecture based on the ViT is used. It incorporates the Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA) modules (Lee et al., 2021), making the deep neural network specifically adapted to limited medical data, and to effectively learn to differentiate between heart states. This is different to the general trend in the related literature about heart disease recognition in echocardiography images, because standard architectures are proposed with no specific modifications to deal with this typical scenario in the medicine field of having a low number of high-quality samples. The two modules, SPT, and LSA, jointly deal with the problem of lack of locality inductive bias inherent to ViTs, improving the performance in the usual scenario of limited samples. This allows to further reduce the dependence on very large training datasets with high quality, whose elaboration is very costly and requires very hard specialized work. Moreover, the modules further help the proposed deep neural network to perform better than pretrained Convolutional Neural Networks even from scratch, as will be seen in Section 4.2, reducing training computational resources. These modules are complementary (Lee et al., 2021), in the way that the absence of any of them would make the ViT to have lower locality inductive bias, and so a worse performance.

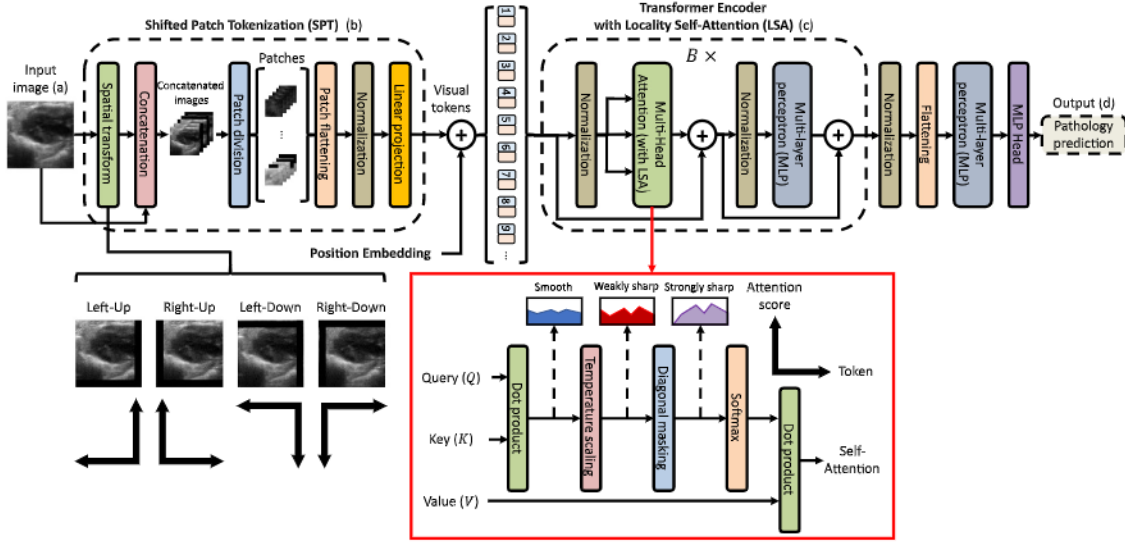


Fig. 5. Architecture of the proposed deep neural network. (a) The input: an image from an echocardiography video sequence, in the form of an original echocardiography sample, a mode, or a reconstruction obtained with the SVD or the HODMD algorithms. (b) The SPT module for more embedding of spatial information in each visual token. (c) The Transformer Encoder with multi-head attention based on LSA for more locality inductive bias. (d) The predicted heart state of the input image.

The architecture of the ViT is represented in Fig. 5. It is composed of the SPT module, with the aim to exploit the spatial relations between neighbouring pixels in the tokenization process, and a Transformer Encoder adapted with LSA. The input is an image representing a heart in a determined state. As a result of the previous phase, Modal Decomposition-based Data Transform, this input image can be a reconstruction, or a mode obtained with the SVD or the HODMD algorithms. This fact makes the deep neural network to have the advantage of being able to directly use as input images, in the form of original echocardiography data (in case of not using the Modal Decomposition-based Data Transform phase), reconstructions, or modes obtained with the modal decomposition techniques. However, as said previously, using the reconstructions of the original echocardiography images obtained with the HODMD algorithm as input for the deep neural network has led to the best performance. The output is a vector of score predictions, i.e., the confidences which represent each heart state. The ViT, as a deep neural network, is the one which deals with the non-linear relationships which could be present between the input and the output, in this case, between echocardiography images and heart states, regardless of whether the SVD and the HODMD algorithms have been employed or not. However, as will be seen in Section 4, the use of the SVD and HODMD algorithms for both feature extraction and data augmentation, and the fact that the proposed ViT is specifically adapted to limited databases of echocardiography images, allow this proposed deep neural network to effectively model the relationship existing between echocardiography images and heart states. In addition, as already seen in Section 2.1.1, the HODMD algorithm, formulated with linear operators, is an approximation of the Koopman operator, allowing to deal with the non-linearities existing in the echocardiography imagery.

The processing of the input starts with a Shifted Patch Tokenization (SPT) module, in which four diagonally shifted versions are first generated. Then, these images are concatenated with the input, and a division in non-overlapping patches is performed. Next, the spatial dimension of the patches is flattened, followed by a normalization layer and a linear projection. Before processing the patches with the Transformer Encoder, a patch encoding layer is used to add positional information to them. By embedding more spatial information in each visual token, the locality inductive bias is increased, so the receptive field to visual tokens, improving standard tokenization, and thus the final performance of the ViT on limited medical data.

The backbone which processes the visual tokens is based on a Transformer Encoder, whose multi-head attention incorporates LSA to further increase the locality inductive bias. This has been done with eight stacked Transformer blocks (so $B = 8$), which are lower than the typical $B = 12$ to avoid overfitting, especially considering this scenario of limited samples. Each Transformer block includes an adapted multi-head attention layer, with four heads, projection dimension 64, and dropout factor 10%. LSA is aimed to increase the locality inductive bias, making the attention of the ViT locally focused. This is done by sharpening the distribution of attention scores in two ways: one, by learning the temperature parameters of the Softmax function, with scaling to avoid a very small gradient for this Softmax function, so to precisely sharpen the attention scores. And two, by removing the self-token relation with the use of diagonal masking. This forces to suppress the diagonal components of the similarity matrix computed by the query and key, increasing the attention scores between different tokens.

The adapted multi-head attention layer is followed by an empty branch (skipped connection) and a Multi-layer Perceptron (MLP) with two units, each one composed of a fully connected layer with the Gaussian Error Linear Unit (GELU) activation function. The use of skip connections addresses the vanishing gradient problem, inherent to deep neural networks, to considerably improve training convergence. The dropout factor is also here 10%, and the number of neurons of the dense layers has been set to 128 and 64, respectively. The Transformer block finally introduces a skip connection, connecting the output of the previous one to the output of this MLP. After all the B Transformer blocks, features are flattened and introduced into a Dropout layer and two head units, each one composed of a fully connected layer with the GELU activation function. The dropout factor of these layers is 50%, and the number of elements of the dense layers is 1024 and 512, respectively. Finally, a fully connected layer is used, whose output is a vector of score predictions containing the confidences of representing each heart state.

The deep neural network uses the annotated cardiac pathology database of medical images, D_{train} , obtained in the Medical Database Creation stage described in Section 2.1 for training. Concerning the optimization technique, the AdamW algorithm has been used. The employed learning rate strategy is based on the warm-up cosine policy (Lee et al., 2021) that computes the learning rate at iteration i by the expression $\lambda_i = 0.5 \times \lambda_t \times (1 + \cos(\pi \times (i - N_w) / (N_{iter} - N_w)))$. The target learning rate is fixed to $\lambda_t = 0.001$, and the warmup steps N_w

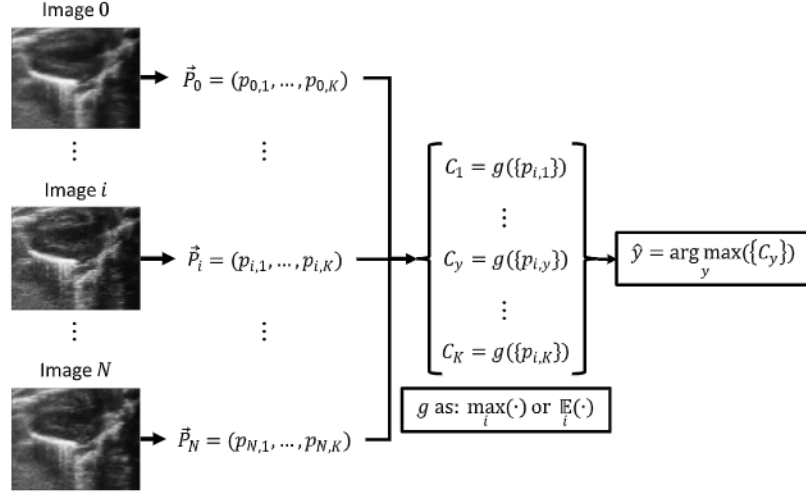


Fig. 6. Illustration of the Fusion of Pathology Predictions phase. The predictions corresponding to each heart state class from every image are fused, and the highest value determines the predicted heart state class of a test sequence.

to $10\% \times N_{iter}$, where N_{iter} is the maximum number of iterations. The momentum and weight decay are set to 0.9 and 0.0001, respectively. The batch size has been set to 64 images due to the limitations of the physical memory of the GPUs available (Nvidia A100). About the database division, a splitting scheme of 60%–20% for training and validation, respectively, has been adopted, balancing the cardiac categories to ensure that every set contains approximately the same number of samples for each heart state. The batch generation scheme minimizes the adverse effects of data imbalance by forcing a minimum proportion of samples of every class. In this way, the skewed predictions towards the heart state classes with more samples are prevented. The batch generation also includes data augmentation techniques based on geometric transformations (namely: resizing, random horizontal flips, random rotations, and random zooms), applied to the echocardiography data. In this way, the model becomes more robust to different perspectives acquired of the heart. In addition, these data augmentation techniques complement to the Modal Decomposition-based Data Generation phase already explained in Section 2.1, marginally improving the performance compared to the modal decomposition.

In the fourth and last phase, Fusion of Pathology Predictions, the most probable heart state of the input test sequence is determined by combining multiple predictions. This process is illustrated in Fig. 6. Each image from the sequence used as input of the previous phase, Deep Neural Network-based Pathology Prediction, has associated a vector of score predictions $\vec{P}_i = (p_{i,1}, \dots, p_{i,K})$, that is, the confidences of belonging to each class. The set of scores $\{p_{i,y}\}$ is then used to compute a unique prediction score of the heart state y . For this purpose, the average and the maximum scores have been tested. Finally, the most probable heart state class is selected by $\hat{y} = \arg \max_y(\{C_y\})$, as long as $\max_y(\{C_y\})$ reaches a minimum threshold. Otherwise, the test sequence is classified as undetermined. Note that relying on a sequence of echocardiography images instead of a single image to predict pathologies is more robust, as using more data can help doctors to make more confident heart diagnoses. In addition, certain heart pathologies might lead to arrhythmias or patterns in the heart cycles which, therefore, constitute temporal information which can be obtained from sequences.

3. Database

The cardiac pathology recognition system here proposed has been tested and validated using a database of echocardiography data. It is composed of video sequences taken from several mice in diverse heart

conditions based on mouse models, whose diagnoses were provided by experts: diabetic cardiomyopathy (DBC), healthy (HL), myocardial infarction (MI), and cardiac hypertrophy (TAC hypertrophy from now on) (TACHYP). These models follow the protocols approved by the Centro Nacional de Investigaciones Cardiovasculares (CNIC) Institutional Animal Care and Research Advisory Committee of the Ethics Committee of the Regional Government of Madrid (PROEX177/17) (Groun et al., 2022). Also, the performed experiments comply with the ARRIVE guidelines. Standard bidimensional (2D) PLAX images of the left ventricle of the heart were obtained and used in this work. The acquisition was performed based on TTE using a high-frequency ultrasound system (Vevo 2100, Visualsonics Inc, Canada) with a 40-MHz linear probe (Groun et al., 2022). A sample image from each heart state is shown in Fig. 7. As can be seen, the echocardiography imagery inherently has much noise, as expected in this modality of data. In addition, different perspectives of the LAX view are taken, boosting the complexity of the cardiac pathology recognition task. Only the area of the heart, i.e., the region of interest (ROI), is taken for the recognition system, as specified in the Data Homogenization phase described in Section 2.1. The obtained heart areas have different resolutions, showing the heterogeneity of the echocardiography imagery: in mean and standard deviation, the spatial resolution is $659.10 \pm 54.80 \times 581.04 \pm 18.18$ pixels. However, the aspect factor among the images barely changes and it is close to 1, so these are practically square: 1.13 ± 0.09 . Therefore, resizing to square images for the deep neural networks barely deforms the heart areas, and does not suppose a degradation of the heart disease recognition performance, as preserving the spatial characteristics characterizing the different heart states.

A summary of the main characteristics of the cardiac pathology database is presented in Table 1. It is composed of the original echocardiography images, and the associated SVD-based data and the HODMD-based data. These data have been generated as a result of the Modal Decomposition-based Data Generation phase described in Section 2.1. The aim is to address the difficulty in the medicine field to elaborate a varied database with high quality enough to properly train deep learning algorithms, as implying high costs and very hard specialized work. Note that the HODMD-based data were obtained after applying the HODMD algorithm on the SVD reconstructions of the original images. Also note that the total number of original snapshots equals the number of associated reconstructed ones after applying the SVD algorithm, but the amount of reconstructions obtained after using the HODMD algorithm is less or equal.

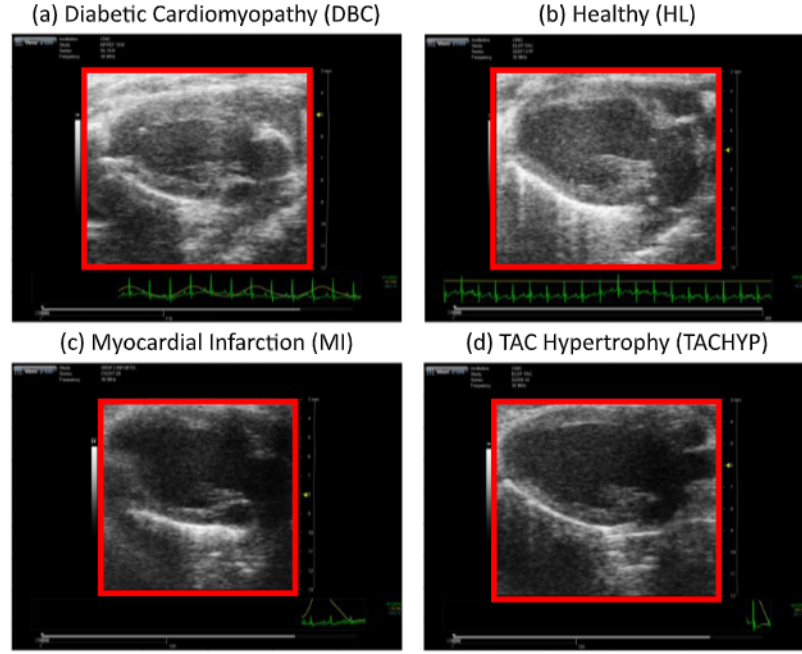


Fig. 7. Sample images from the different heart states of the echocardiography data. The regions of interest are delimited by the red boundaries.

Table 1
Summary of the main characteristics of the cardiac pathology database.

Heart state	Set	# Sequences	# Snapshots	# SVD modes	HODMD-based data	
					# Modes	# Reconstructed
Diabetic Cardiomyopathy (DBC)	Training	20	2276	100	1137	2214
	Validation	9	796	45	408	735
	Test	10	785	50	351	715
Healthy (HL)	Training	21	2382	105	1176	2341
	Validation	10	893	50	480	832
	Test	8	871	40	405	837
Myocardial Infarction (MI)	Training	19	2425	95	1092	2425
	Validation	6	900	30	360	900
	Test	6	900	30	345	900
TAC Hypertrophy (TACHYP)	Training	19	2243	95	960	2443
	Validation	6	720	30	333	720
	Test	6	720	30	345	720
Total		140	15 911	700	7392	15 582

Table 2
Summary of the combinations for D_{train} used in the Medical Database Creation stage.

Case	Original	SVD-based data		HODMD-based data		SVD (HODMD modes)
		Reconstructed	Modes	Reconstructed	Modes	Reconstructed
1	×					
2		×				
3	×	×	×			
4	×	×				
5		×	×			
6				×		
7				×	×	
8		×		×		
9	×	×		×		
10				×	×	×
11		×	×	×	×	
12		×	×	×	×	×

This is due to the minimum number of frames required in a sequence in order to apply the HODMD algorithm, which is not achieved in some of the sequences. Note that the training, validation and test sets comprise different sequences, which are the ones that both achieve the splitting scheme proposed in this work and a reasonable balance in the number of samples in each class.

Table 2 shows the experimented cases of the training database D_{train} , each one indicating which types of data are used to form that. Different types of test data, resulting from the Modal Decomposition-based Data Transform phase described in Section 2.2, have been assessed as well. In this way, the impact of each type of data on the final performance could be evaluated, including the robustness to noise,

as the original echocardiography imagery inherently has much noise. The final column, i.e., SVD (HODMD Modes), represents the use of the reconstructions of the HODMD modes after applying the SVD algorithm on them.

4. Results

The algorithm here considered has been tested with the cardiac pathology database using the different cases for the training database D_{train} . Additionally, it has been compared with CNN-based algorithms, including Inception-version 3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), ResNet50-version 2 (He, Zhang, Ren, & Sun, 2016), and VGG16 (Simonyan & Zisserman, 2014). The following metrics have been utilized to measure the performance of each algorithm in terms of classification accuracy and computational cost: precision (P), recall (R), F1-Score (F), accuracy, and average processing time per image $\bar{t}$. The precision (P) is the percentage of correct detections, that is, true positives (TP), with respect to all samples detected as the heart state (either images or videos, depending on the case, explicitly indicated): true positives (TP) and false positives, (FP , samples wrongly detected as the heart state). This metric is expressed as follows:

$$P = \frac{TP}{TP + FP} \quad (3)$$

The recall (R) represents the percentage of correct detections (TP) with respect to the total number of existing samples from the heart state: true positives (TP) and false negatives, (FN , actual samples from the heart state wrongly classified), expressed by:

$$R = \frac{TP}{TP + FN} \quad (4)$$

The F1-Score (F) is the harmonic mean of the precision and the recall, given by:

$$F = 2 \cdot \frac{P \cdot R}{P + R} \quad (5)$$

The accuracy is the percentage of the correctly predicted samples with respect to the total number of samples:

$$Accuracy = \frac{\# \text{ Correct predictions}}{\# \text{ Samples}} \quad (6)$$

To represent global values of the metrics, three types have been used (exemplifying with the accuracy): the per-image accuracy (w/o), the per-image accuracy after the Fusion of Pathology Predictions phase (w/), described in Section 2.2, and the per-sequence accuracy. The per-image accuracy is the one which considers each sample as an image. The one after the Fusion of Pathology Predictions phase measures the accuracy when the samples (as images) are classified as the predicted heart state in its associated sequence. Finally, the per-sequence accuracy considers each sample as a sequence.

The average processing time per sample $\bar{t}$ is the average time taken by the proposed system to process a sample (either image or sequence) and give the heart state classification. It is broken down into the SVD algorithm ($\bar{t}_{SVD}$), the HODMD algorithm, comprising the HOSVD dimensionality reduction ($\bar{t}_{HOSVD}$) prior to the HODMD itself ($\bar{t}_{HODMD}$) (Bell-Navas et al., 2023), and the Deep Neural Network-based Pathology Prediction phase ($\bar{t}_{pred}$). Unless the contrary is explicitly specified, the times are measured in milliseconds.

The results have been obtained using a cluster with Intel Xeon Gold 6240R and three Tesla A100 GPU working in parallel for training, and Intel Xeon Gold 6230 and one Tesla V100 GPU for testing. In this way, not only the high amount of data conforming the created medical database is managed, but also the training convergence is sped up and eased.

4.1. Analysis of types of data

First, a study to evaluate the performance in terms of accuracy of the training cases on different types of data used for test, i.e., test data obtained with the Modal Decomposition-based Data Transform phase described in Section 2.2, is presented. To that end, the configuration of the SVD and HODMD algorithms, the deep neural network, and of the prediction is initially introduced. Later on, the influence of the types of data used for both training and test on the overall performance of the system is studied. Lastly, the performance in each heart state for the best configuration is shown.

1. Configuration: Regarding the Modal Decomposition-based Data Generation and Transform phases, the configuration is as follows. For the SVD algorithm, the number of modes provided has been set to 5, considered adequate for the reconstruction according to the contribution of the modes in the frames. For the HODMD algorithm, the number of snapshots used on each sequence K has been set to the total number of frames of the corresponding sequence. In this way, the maximum number of cardiac cycles possible can be captured and the cardiac and respiratory frequencies can be more accurately obtained. The time interval between snapshots $\Delta t = 4$ ms and the tolerances $\epsilon_{SVD} = \epsilon_{DMD} = 5e-4$ are based on Groun et al. (2022). This is, Δt is based on the used ultrasound scanner, and its value has been configured by a specialist, in the way that the images acquired properly give the cardiac and respiratory frequencies. Regarding the tolerances for the dimensionality reduction and for the amplitude truncation steps, respectively, these are larger than the noise level, and lead to a reasonable number of frequencies associated to the studied heart states. The index d has been configured for each sequence in the way that fair reconstructions of the snapshots are obtained. In particular, for the diabetic cardiomyopathy and healthy states, depending on the sequence, $d = \lfloor K/3 \rfloor$ or $d = \lfloor K/5 \rfloor$ works better. Regarding the myocardial infarction and TAC hypertrophy sequences, $d = 50$ is considered adequate, which corresponds nearly to $d = \lfloor K/3 \rfloor$ in the majority of these sequences. These values of d follow the recommendations for the calibration process described in Le Clainche and Vega (2017). In fact, other values of d make the reconstructions to present severe distortions and artifacts which do not describe the dynamics of the data and do not reproduce adequately the heart cycles. Therefore, discriminative features associated to the heart states are not extracted in these cases.

Concerning the Deep Neural Network-based Pathology Prediction phase, the input images have been resized to 256×256 and the patch size has been set to 32, due to the memory limitations of the GPU devices available. In fact, deeper models of the proposed ViT could not be tested for this reason. In addition, considering the actual scenario of limited samples, deeper models would be too complex and more prone to overfitting, worsening the performance. On the other hand, reducing the complexity of the Transformer Encoder of the proposed ViT described in Section 2.2 did not lead to significant changes in the performance, even tending to a degradation. This is because the ViT becomes too shallow, and so not capable to extract features complex enough from the medical data.

Finally, regarding the Fusion of Pathology Predictions phase, maximum and average scores have been used to calculate the global prediction score per class. This is, to calculate a global prediction score per class in a test sequence, the maximum of the predictions associated to a class is calculated to obtain the global prediction score of that class, or instead the average of the predictions.

Table 3

Comparison results with different types of data for test and training cases using the proposed framework and the average score value from the predictions per class. Refer to Table 2 for the specifications of each training case.

		Per-image accuracy (w/o)											
		Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12
SVD	Original	0.49	0.56	0.54	0.56	0.57	0.48	0.55	0.59	0.52	0.54	0.55	0.62
	Reconstruction	0.52	0.57	0.53	0.59	0.56	0.50	0.55	0.58	0.52	0.54	0.55	0.63
	Modes	0.24	0.17	0.38	0.18	0.35	0.23	0.38	0.22	0.14	0.42	0.40	0.47
HODMD	Reconstruction	0.53	0.56	0.52	0.58	0.54	0.50	0.57	0.58	0.51	0.56	0.56	0.64
	Modes (abs)	0.39	0.38	0.45	0.41	0.48	0.40	0.59	0.40	0.37	0.53	0.57	0.61
		Per-image accuracy (w/): Average											
SVD	Original	0.52	0.60	0.57	0.60	0.64	0.51	0.57	0.64	0.60	0.57	0.64	0.64
	Reconstruction	0.56	0.57	0.57	0.57	0.60	0.51	0.57	0.64	0.57	0.57	0.60	0.68
	Modes	0.27	0.17	0.53	0.13	0.53	0.23	0.60	0.17	0.10	0.47	0.50	0.57
HODMD	Reconstruction	0.60	0.56	0.49	0.56	0.56	0.49	0.60	0.64	0.52	0.60	0.56	0.68
	Modes (abs)	0.48	0.50	0.50	0.50	0.52	0.43	0.64	0.54	0.42	0.61	0.60	0.64
		Per-sequence accuracy: Average											
SVD	Original	0.60	0.57	0.53	0.57	0.73	0.57	0.67	0.73	0.70	0.63	0.73	0.73
	Reconstruction	0.63	0.67	0.67	0.67	0.70	0.57	0.67	0.73	0.67	0.63	0.70	0.77
	Modes	0.27	0.17	0.53	0.13	0.53	0.23	0.60	0.17	0.10	0.47	0.50	0.57
HODMD	Reconstruction	0.69	0.65	0.58	0.65	0.65	0.58	0.69	0.73	0.62	0.65	0.65	0.77
	Modes (abs)	0.50	0.54	0.54	0.54	0.58	0.46	0.69	0.58	0.46	0.65	0.65	0.69
# Samples (train)		9326	9326	19 047	18 652	9721	9223	13 588	18 549	27 875	17 953	23 309	27 674

Table 4

Comparison results with different types of data for test and training cases using the proposed framework and the maximum score value from the predictions per class. Refer to Table 2 for the specifications of each training case.

		Per-image accuracy (w/o)											
		Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12
SVD	Original	0.49	0.56	0.54	0.56	0.57	0.48	0.55	0.59	0.52	0.54	0.55	0.62
	Reconstruction	0.52	0.57	0.53	0.59	0.56	0.50	0.55	0.58	0.52	0.54	0.55	0.63
	Modes	0.24	0.17	0.38	0.18	0.35	0.23	0.38	0.22	0.14	0.42	0.40	0.47
HODMD	Reconstruction	0.53	0.56	0.52	0.58	0.54	0.50	0.57	0.58	0.51	0.56	0.56	0.64
	Modes (abs)	0.39	0.38	0.45	0.41	0.48	0.40	0.59	0.40	0.37	0.53	0.57	0.61
		Per-image accuracy (w/): Maximum											
SVD	Original	0.48	0.64	0.57	0.57	0.63	0.51	0.57	0.64	0.57	0.64	0.55	0.64
	Reconstruction	0.52	0.60	0.53	0.57	0.59	0.51	0.57	0.64	0.53	0.61	0.60	0.64
	Modes	0.20	0.13	0.57	0.17	0.67	0.20	0.53	0.20	0.13	0.50	0.60	0.53
HODMD	Reconstruction	0.56	0.60	0.49	0.56	0.56	0.52	0.56	0.64	0.52	0.60	0.56	0.68
	Modes (abs)	0.22	0.36	0.51	0.36	0.63	0.45	0.57	0.49	0.36	0.61	0.60	0.60
		Per-sequence accuracy: Maximum											
SVD	Original	0.57	0.60	0.53	0.53	0.70	0.57	0.67	0.73	0.67	0.70	0.63	0.73
	Reconstruction	0.60	0.70	0.63	0.67	0.67	0.57	0.67	0.73	0.63	0.67	0.70	0.73
	Modes	0.20	0.13	0.57	0.17	0.67	0.20	0.53	0.20	0.13	0.50	0.60	0.53
HODMD	Reconstruction	0.65	0.69	0.58	0.65	0.65	0.62	0.65	0.73	0.62	0.65	0.65	0.77
	Modes (abs)	0.23	0.38	0.54	0.38	0.69	0.50	0.62	0.54	0.38	0.65	0.65	0.65
# Samples (train)		9326	9326	19 047	18 652	9721	9223	13 588	18 549	27 875	17 953	23 309	27 674

2. Performance: Tables 3 and 4 reveal the impact of the training cases presented in Table 2 using the different types of data for testing on the performance of the proposed system, with and without the fusion of predictions, via the global prediction score using the average and the maximum, respectively. The results show a general improvement in the training cases with respect to the case 1 (with only original images), considering that, in general, more training samples are available, as expected. However, the variation of the performance is also affected by the specific types of training data: for example, although having a similar number of training samples, the case 12 leads to much better performances than those with the case 9. The former uses SVD, HODMD modes and the SVD reconstructions of the HODMD modes, but the latter uses the original images instead. Also note that there is a consistent improvement in the training cases with HODMD modes, and also with their corresponding SVD reconstructions, i.e., cases 7, 10, 11, and 12. Therefore, these two facts show the importance of using HODMD modes for training. This is because the HODMD modes describe the dynamics of the data and contain the temporal information,

leveraged by the neural network, which could be relevant to distinguish the different heart states. On the other hand, among the types of test data, a general improvement is produced with the reconstructed images, demonstrating the contribution of the Modal Decomposition-based Data Transform phase to a better heart state recognition performance. This is because these have less noise than the original data (inherently with much noise) and, therefore, contain more discriminative features associated to the different heart states. In addition, there are more reconstructed images than modes in both training and testing, which implies that the neural network can learn better from image reconstructions, and also that there are more test data in each sequence in order to more reliably determine its heart state. So, this shows the effectiveness of the SVD and HODMD algorithms for data augmentation and extraction of more discriminative features associated to the different heart states, improving the discriminating capability of the proposed ViT in the usual scenario of having a limited number of samples. Moreover, this also demonstrates the usefulness of the SVD and HODMD algorithms to address the difficulty to elaborate varied databases with high

Table 5Cardiac pathology recognition performance using P , R , F and accuracy obtained with the best configuration of the proposed framework.

Heart state	Per-image results									Per-sequence results					
	(w/o)			(w/): Average			(w/): Maximum			Average			Maximum		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Diabetic Cardiomyopathy (DBC)	0.58	0.70	0.64	0.56	0.73	0.63	0.56	0.73	0.63	0.71	0.71	0.71	0.71	0.71	0.71
Healthy (HL)	0.70	0.57	0.64	0.82	0.64	0.72	0.82	0.64	0.72	0.86	0.86	0.86	0.86	0.86	0.86
Myocardial Infarction (MI)	0.68	0.65	0.66	0.73	0.67	0.70	0.73	0.67	0.70	0.71	0.83	0.77	0.71	0.83	0.77
TAC Hypertrophy (TACHYP)	0.50	0.65	0.56	0.62	0.67	0.64	0.62	0.67	0.64	0.80	0.67	0.73	0.80	0.67	0.73
Total	0.62	0.64	0.63	0.68	0.68	0.67	0.68	0.68	0.67	0.77	0.77	0.77	0.77	0.77	0.77
Accuracy	0.64			0.68			0.68			0.77			0.77		

Per-image confusion matrix (w/o)							Per-image confusion matrix (w/): Average						
	DBC	HL	MI	TACHYP	Undetermined	# Total samples		DBC	HL	MI	TACHYP	Undetermined	# Total samples
DBC	503	0	0	62	150	715	DBC	525	0	0	0	190	715
HL	12	473	0	334	18	837	HL	0	537	0	300	0	837
MI	229	0	585	71	15	900	MI	300	0	600	0	0	900
TACHYP	22	0	168	468	62	720	TACHYP	0	0	120	480	120	720
# Total detections	766	473	753	935	245	3172	# Total detections	825	537	720	780	310	3172

Per-sequence confusion matrix: Average						
	DBC	HL	MI	TACHYP	Undetermined	# Total samples
DBC	5	0	0	0	2	7
HL	0	6	0	1	0	7
MI	1	0	5	0	0	6
TACHYP	0	0	1	4	1	6
# Total detections	6	6	6	5	3	26

Fig. 8. Confusion matrices obtained with the best configuration of the proposed framework (absolute values).

Per-image confusion matrix (w/o)							Per-image confusion matrix (w/): Average						
	DBC	HL	MI	TACHYP	Undetermined	# Total samples		DBC	HL	MI	TACHYP	Undetermined	# Total samples
DBC	0.70	0.00	0.00	0.09	0.21	715	DBC	0.73	0.00	0.00	0.00	0.27	715
HL	0.01	0.57	0.00	0.40	0.02	837	HL	0.00	0.64	0.00	0.36	0.00	837
MI	0.25	0.00	0.65	0.08	0.02	900	MI	0.33	0.00	0.67	0.00	0.00	900
TACHYP	0.03	0.00	0.23	0.65	0.09	720	TACHYP	0.00	0.00	0.17	0.66	0.17	720
# Total detections	766	473	753	935	245	3172	# Total detections	825	537	720	780	310	3172

Per-sequence confusion matrix: Average						
	DBC	HL	MI	TACHYP	Undetermined	# Total samples
DBC	0.71	0.00	0.00	0.00	0.29	7
HL	0.00	0.86	0.00	0.14	0.00	7
MI	0.17	0.00	0.83	0.00	0.00	6
TACHYP	0.00	0.00	0.17	0.66	0.17	6
# Total detections	6	6	6	5	3	26

Fig. 9. Confusion matrices obtained with the best configuration of the proposed framework (normalized values).

quality enough in the medicine field in terms of costs and very hard specialized work. On the other hand, these results also show the effectiveness of the SVD and of the HODMD algorithms to reduce the high amount of noise inherent to the original echocardiography images, contributing to performance improvements, and making the proposed system more robust to noise. If another kind of noise were added to the echocardiography images, the SVD and the HODMD algorithms would also effectively clean the data, leading to more discriminative features associated to the different heart diseases, and so to performance improvements as well.

Table 5 presents the recognition results for each heart state using the best configuration inferred from Tables 3 and 4: training case 12 and HODMD reconstructions as test data, with and without fusion of predictions, using the average or the maximum. Similarly, Figs. 8 and 9 depict the confusion matrices obtained using this best configuration, with and without fusion of predictions, using the average, representing absolute and normalized values, respectively. Note that the column ‘undetermined’ is referred to the samples which have been misclassified as a different heart state of the database. The performance is balanced among the four considered heart states, like the number of test samples in terms of both images and sequences. A 77% average per-sequence

accuracy is reached, which is high enough for medical applications considering the low number of original training samples (images) from each class, much lower than the samples required by typical deep learning algorithms (hundreds of thousands or even millions of samples). Therefore, if a much higher number of samples were available (which is unfeasible to achieve in the medicine field for the high difficulty in terms of costs and specialized hard work), the different heart states would be better recognized and the associated features better learned. This result demonstrates that the proposed deep neural network can effectively model the non-linear relationship which could be present between the input and the output, in this case, between echocardiography images and heart states. In addition, this proves the capability of the HODMD algorithm to deal with the non-linearities which could be present in the echocardiography imagery. On the other hand, note that the fusion of predictions consistently improves the performance, increasing the number of true positives (*TP*) and decreasing the number of false negatives (*FN*), so the recall (*R*). This is because sequences which have been correctly classified do not have all their images classified as the same heart state. However, the fusion of predictions considers the whole set of images from the sequence to determine the final prediction. In the situation of having a predominant number of correctly classified images in a sequence, the misclassified images are corrected, reducing the number of false negatives (*FN*) and increasing at the same time the number of true positives (*TP*). An increase of the false positives (*FP*) could be produced with the fusion of predictions, and so a reduction in the precision (*P*), which is the case of the diabetic cardiomyopathy state as a result of a misclassified sequence. However, the false positives (*FP*) do not increase as much as the true positives (*TP*), resulting in a global improvement in the recognition.

4.2. Comparison with alternative algorithms

In this second part of the experiments, the performance of the proposed system has been compared against other alternative approaches, but still using the same framework, where the ViT has been replaced by other deep neural networks. In particular, the comparison includes a CNN trained from scratch, and the pretrained models Inception-v3, ResNet50-v2, and VGG16. First, the configuration of the proposed system is introduced. Next, the configuration and the necessary adaptations of the alternative algorithms are described. Finally, an overall performance comparison between the proposed ViT and other deep neural networks is presented, using the most representative training cases (1, 7, 10, 11, 12) and the main types of test data (Original and HODMD-based reconstructions).

1. Configuration and adaptation of the alternative algorithms: Regarding the configuration of the SVD and HODMD algorithms, the ViT, and of the prediction, it is the same as in Section 4.1. However, about the Fusion of Pathology Predictions phase, only the average scores are used to calculate the global prediction score per class in a test sequence. Before presenting the parameters selected for the CNN, Inception-v3, ResNet50-v2, and VGG16, some considerations must be taken into account. The input of these deep neural networks has three channels, but the cardiac pathology database used in this work comprises one-channel data. In addition, their architectures were designed to classify into the classes from the ImageNet challenge (Russakovsky et al., 2015). Therefore, adaptations were required to enable them to not only work with one-channel data, but also to classify into the heart states from the cardiac pathology database. To obtain three-channel data, the input has been replicated three times and concatenated in the channel dimension before being introduced into the deep neural networks. For classification, the final layer of these models is replaced by a series of dense layers as follows. In the Inception-v3 and VGG16 models, features are flattened, and two dense

layers are introduced: the first one has 128 elements and the second one the number of heart states. In the case of ResNet50-v2, a single dense layer, whose amount of elements is the number of heart states, is used. In addition, global average pooling is applied to the output of the last convolutional layer. The activation function after every dense layer is a rectifier linear unit (ReLU), except from the final one, which is a Softmax. For the initialization of Inception-v3, ResNet50-v2, and VGG16, pretrained weights from the ImageNet challenge (Russakovsky et al., 2015) have been used, instead of random ones. These pretrained weights give a good initialization, and therefore improve the training convergence and the quality of the models obtained after fine-tuning using the cardiac pathology database. This means that these algorithms use an auxiliary dataset in addition to the cardiac pathology database. As the proposed ViT and the CNN do not use that auxiliary dataset, it supposes an advantage for the algorithms.

The CNN is composed of three convolutional layers, each one followed by a max pooling layer. The activation function after every convolutional layer is a ReLU, and the number of filters have been set to 16, 32, and 64, respectively. The sizes are 3×3 and 2×2 for the convolutional and max pooling layers, respectively. After that, features are flattened, and two dense layers are introduced: the first one has 128 elements and the second one the number of heart states. The activation function of the dense layers are a ReLU and a Softmax, respectively. Similarly to the proposed ViT, the input images have been resized to 256×256 . This image resizing is also mandatory due to the memory limitations of the GPU devices for training.

To reduce randomness in the results because of the training convergence of the algorithms in each experiment, each algorithm is trained four times using the configuration explained above. Then, the results have been averaged, and the standard deviation of the results has been calculated.

2. Performance: Table 6 presents a comparison of the cardiac pathology recognition performance, using the medical database and the proposed system with either the ViT, the CNN, Inception-v3, ResNet50-v2, or VGG16. Only the training cases 1, 7, 10, 11, and 12 have been considered, as being the most representative of the contribution of the SVD and HODMD algorithms with respect to the use of only original images for training. Regarding the types of test data, original images and HODMD-based reconstructions have been compared, as mainly representing the influence of the SVD and HODMD algorithms as feature extractors. As can be observed in the table, the two best results in terms of recognition performance have been obtained using the proposed ViT and the ResNet50-v2, according to the mean and standard deviations of the accuracies. Moreover, these even outperform the CNN without pretraining, showing the efficacy of using more complex models, with techniques to better train them (e.g., residual networks, or the SPT and LSA modules in the proposed ViT). As deep neural networks, both show to be capable of effectively modelling the non-linear relationships which could be present between echocardiography imagery and heart states. In addition, there is a general improvement in the training cases with respect to the case 1 and in the HODMD-based reconstructions for test data. Therefore, the alternative algorithms also benefit from the use of the SVD and HODMD algorithms for both data augmentation and feature extraction, also incorporating temporal information thanks to the HODMD modes. This improvement also demonstrates the effectiveness to deal with the high noise inherent to the original echocardiography images and, therefore, the robustness of the proposed system with the alternative algorithms to noise. Again, other kinds of noise additionally incorporated into the echocardiography imagery would be effectively reduced by the SVD and

Table 6

Comparison of the recognition performance among different algorithms with the main types of data and training cases using the average score value from the predictions per class. Refer to Table 2 for the specifications of each training case.

		Per-image accuracy (w/o)				
Algorithm	Test data	Case 1	Case 7	Case 10	Case 11	Case 12
CNN	Original	0.40 ± 0.11	0.35 ± 0.17	0.44 ± 0.10	0.29 ± 0.05	0.45 ± 0.12
	HODMD reconstruction	0.36 ± 0.11	0.37 ± 0.20	0.50 ± 0.16	0.34 ± 0.14	0.51 ± 0.19
Inception-v3	Original	0.49 ± 0.01	0.49 ± 0.03	0.45 ± 0.06	0.42 ± 0.09	0.45 ± 0.08
	HODMD reconstruction	0.44 ± 0.02	0.51 ± 0.04	0.45 ± 0.09	0.42 ± 0.12	0.46 ± 0.09
ResNet50-v2	Original	0.47 ± 0.01	0.46 ± 0.07	0.48 ± 0.05	0.50 ± 0.05	0.46 ± 0.06
	HODMD reconstruction	0.45 ± 0.03	0.54 ± 0.03	0.56 ± 0.03	0.58 ± 0.01	0.54 ± 0.02
VGG16	Original	0.45 ± 0.10	0.39 ± 0.03	0.30 ± 0.09	0.38 ± 0.07	0.45 ± 0.05
	HODMD reconstruction	0.33 ± 0.06	0.47 ± 0.05	0.32 ± 0.12	0.44 ± 0.13	0.54 ± 0.06
ViT	Original	0.52 ± 0.02	0.56 ± 0.03	0.56 ± 0.03	0.57 ± 0.03	0.56 ± 0.04
	HODMD reconstruction	0.54 ± 0.02	0.58 ± 0.04	0.57 ± 0.03	0.58 ± 0.03	0.57 ± 0.05
		Per-image accuracy (w/): Average				
CNN	Original	0.38 ± 0.15	0.36 ± 0.24	0.48 ± 0.11	0.28 ± 0.08	0.49 ± 0.16
	HODMD reconstruction	0.39 ± 0.12	0.41 ± 0.16	0.52 ± 0.21	0.30 ± 0.11	0.55 ± 0.21
Inception-v3	Original	0.52 ± 0.05	0.53 ± 0.05	0.54 ± 0.10	0.46 ± 0.11	0.59 ± 0.09
	HODMD reconstruction	0.47 ± 0.06	0.58 ± 0.08	0.46 ± 0.12	0.46 ± 0.09	0.52 ± 0.09
ResNet50-v2	Original	0.48 ± 0.03	0.53 ± 0.14	0.53 ± 0.10	0.57 ± 0.03	0.50 ± 0.12
	HODMD reconstruction	0.47 ± 0.06	0.63 ± 0.06	0.63 ± 0.05	0.65 ± 0.03	0.63 ± 0.05
VGG16	Original	0.51 ± 0.12	0.39 ± 0.08	0.30 ± 0.08	0.40 ± 0.09	0.46 ± 0.07
	HODMD reconstruction	0.33 ± 0.07	0.47 ± 0.08	0.33 ± 0.11	0.49 ± 0.17	0.57 ± 0.06
ViT	Original	0.55 ± 0.03	0.58 ± 0.02	0.57 ± 0.06	0.63 ± 0.02	0.61 ± 0.04
	HODMD reconstruction	0.57 ± 0.03	0.61 ± 0.02	0.59 ± 0.07	0.61 ± 0.04	0.61 ± 0.06
		Per-sequence accuracy (w/): Average				
CNN	Original	0.49 ± 0.14	0.38 ± 0.30	0.57 ± 0.14	0.31 ± 0.07	0.57 ± 0.20
	HODMD reconstruction	0.42 ± 0.13	0.38 ± 0.25	0.59 ± 0.24	0.31 ± 0.11	0.63 ± 0.22
Inception-v3	Original	0.59 ± 0.07	0.61 ± 0.07	0.59 ± 0.10	0.55 ± 0.12	0.68 ± 0.10
	HODMD reconstruction	0.50 ± 0.07	0.62 ± 0.11	0.51 ± 0.15	0.48 ± 0.12	0.57 ± 0.12
ResNet50-v2	Original	0.56 ± 0.03	0.60 ± 0.14	0.63 ± 0.11	0.65 ± 0.06	0.58 ± 0.12
	HODMD reconstruction	0.50 ± 0.03	0.67 ± 0.05	0.67 ± 0.05	0.66 ± 0.02	0.69 ± 0.07
VGG16	Original	0.54 ± 0.15	0.42 ± 0.08	0.34 ± 0.12	0.43 ± 0.09	0.50 ± 0.09
	HODMD reconstruction	0.30 ± 0.09	0.51 ± 0.10	0.38 ± 0.12	0.51 ± 0.21	0.62 ± 0.07
ViT	Original	0.68 ± 0.05	0.68 ± 0.02	0.65 ± 0.07	0.72 ± 0.02	0.70 ± 0.05
	HODMD reconstruction	0.65 ± 0.03	0.70 ± 0.02	0.66 ± 0.08	0.70 ± 0.04	0.70 ± 0.07
# Samples (train)		9326	13 588	17 953	23 309	27 674

HODMD algorithms, leading to performance improvements with the alternative algorithms. On the other hand, note that, in the training case 1, a better performance is generally achieved when using original test images. This can be explained by the fact that the models learn to recognize different heart states in noisy images instead of in filtered ones, obtaining more confident predictions in the original images.

Regarding the per-image accuracy (w/o), the proposed ViT achieves the best results. Nonetheless, ResNet50-v2 performs very closely to it. This is mainly because ResNet50-v2 has been pretrained with the external ImageNet database, which contributes to compensate for the fact that the cardiac pathology database has relatively a lower number of samples in comparison with other databases used by deep learning techniques. In this sense, a larger cardiac pathology database could significantly improve the proposed framework with the alternative models, trained from scratch or not. On the other hand, the adoption of natural images for transfer learning in the proposed ViT would not be precisely as effective as using this larger cardiac pathology database. This is because medical images, like echocardiograms, have different characteristics to those of natural images like in the ImageNet database. Specifically, echocardiography images are based on standardized views and features with subtle fine texture differences, also with usually different pixel intensity distributions due to the different acquisition conditions. On the contrary, natural images focus on broader, more easily distinguishable patterns, and contain high-level semantic features which could not be useful for medical imaging tasks like heart disease recognition, explaining poor performances of pretrained models like in Farhad et al. (2023).

Regarding the results obtained, the training case 7 has led to the best performance in the ViT, showing the improvement given by the HODMD-based data, but with practically no impact when incorporating SVD-based data, as possibly barely giving new training data (case 11). On the contrary, the per-sequence accuracy is the best one in the proposed ViT in the training case 11 and with more difference, also with respect to ResNet50-v2, showing that the images which have been correctly classified lead to more correctly classified sequences. In addition, a higher performance is obtained in the case of using original images in test. Once more, this is due to that the distribution and confidences of the images correctly classified among the test sequences are more favourable than in the case of HODMD test data, obtaining more correctly classified test sequences.

Computational cost of the different phases of the prediction process is summarized in Table 7. It is important to note that the SVD and the HODMD algorithms are applied on sequences rather than on single images, but the time measures represented in the table are averages per image. As can be seen, the HOSVD dimensionality reduction process is the hardest part by far (Bell-Navas et al., 2023). Although all the compared deep learning algorithms have very low prediction times, the proposed ViT is the fastest one, which is desirable considering the lower number of parameters and the better recognition performance. This result also encourages the deployment of the proposed ViT using less resources. In every case, all time values are inside of real-time requirements, which is appealing for several medical applications.

For short, the best overall performance is achieved with the proposed ViT in terms of both recognition per-image and per-sequence accuracy and operation time. This is achieved even

Table 7

Computational cost of different phases of the cardiac pathology recognition system using the algorithms compared.

Model	# Parameters	$\bar{t}_{SVD}$	$\bar{t}_{HOSVD}$	$\bar{t}_{HODMD}$	$\bar{t}_{pred}$
CNN	7 396 872				2
Inception-v3	31 240 744				3
ResNet50-v2	23 575 048	5.1	591	0.648	3.2
VGG16	18 909 768				4.1
ViT	5 730 512				1.5

training the proposed ViT from scratch, unlike Inception-v3, ResNet50-v2, and VGG16, supposing an advantage, because of reducing training computational resources. Lastly, recognition accuracy is expected to improve with larger medical image databases based on heart diseases.

5. Conclusions

Heart diseases are the main cause of human mortality in the world. This makes their recognition a task of great importance. In this work, a real-time cardiac pathology recognition system for echocardiography video sequences has been presented to face this challenging and demanding task. The two major contributions of this paper are the creation of a large annotated medical database from echocardiography sequences with heterogeneous acquisition features, and a deep neural network, based on a ViT, designed for an effective training from scratch, even with small datasets. These contributions proposed jointly deal with the usual situation of having a limited number of samples in the medicine field, reducing the dependence on very large training datasets. Precisely, the elaboration of varied databases with high quality enough is very difficult, implying high costs and very hard specialized work from experts in heart diseases. In addition, the general trend in the related literature about heart disease recognition in echocardiography images does not propose architectures adapted to this typical situation of limited databases. Thus, the necessity to adopt techniques specifically designed to address this problem. During the creation of the database, the HODMD algorithm has been employed for the first time (to the best of the authors' knowledge) as a feature extractor and as a data augmentation technique in the medical field. This algorithm has demonstrated to improve the cardiac pathology recognition performance, as also incorporating temporal information from the HODMD modes, useful to better characterize the different heart states. The results obtained have proved that the proposed system outperforms CNNs, even with pretraining. We anticipate that, if longer image datasets about heart diseases were available, overfitting and robustness would be better.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Andrés Bell-Navas reports financial support was provided by Spanish Ministry of Science and Innovation. Nourelhouda Groun reports financial support was provided by Spanish Ministry of Science and Innovation. Enrique Lara-Pezzi reports financial support was provided by Spanish Ministry of Science and Innovation. Soledad Le Clainche reports financial support was provided by the Ministerio de Ciencia, Innovación y Universidades. Enrique Lara-Pezzi reports financial support was provided by Comunidad de Madrid. María Villalba-Orero reports financial support was provided by Comunidad de Madrid. María Villalba-Orero reports financial support was provided by Juan de la Cierva Incorporación. Enrique Lara-Pezzi reports a relationship with Spanish National Cardiovascular Research Center that includes: employment and funding grants. María Villalba-Orero reports a relationship with

Spanish National Cardiovascular Research Center that includes: employment and funding grants. If there are other authors, they 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

The authors acknowledge the Grants TED2021-129774B-C21, TED2021-129774B-C22, and PLEC2022-009235, funded by the Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033) and by the European Union "NextGenerationEU"/PRTR, and the Grant PID2023-147790OB-I00, funded by the Ministerio de Ciencia, Innovación y Universidades (MICIU/AEI/10.13039/501100011033/FEDER, UE); the first one to A.B-N and N.G, the next two to E.L-P, and the last one to S.L. The authors also acknowledge the Grant PEJ-2019-TL/BMD-12831 from Comunidad de Madrid to E.L-P and to M.V-O, and a Juan de la Cierva Incorporación Grant (IJCI-2016-27698) to M.V-O. The CNIC is supported by the Instituto de Salud Carlos III (ISCIII), the Ministerio de Ciencia, Innovación y Universidades (MICIU), and the Pro CNIC Foundation, and is a Severo Ochoa Center of Excellence (Grant CEX2020-001041-S funded by MICIU/AEI/10.13039/501100011033).

Data availability

The data that have been used are confidential.

References

- Arooj, S., Rehman, S. u., Imran, A., Almuhaimeed, A., Alzahrani, A. K., & Alzahrani, A. (2022). A deep convolutional neural network for the early detection of heart disease. *Biomedicine*, 10(11), 2796. <https://doi.org/10.3390/biomedicine10112796>.
- Bell-Navas, A., Groun, N., Garicano-Mena, J., & Le Clainche, S. (2023). Optimized higher order dynamic mode decomposition analysis of electrocardiography datasets. In *25th conf. of ILAS* (pp. 91–92).
- Chen, J., Zhang, X., Yuan, J., Shao, R., Gan, C., Ji, Q., et al. (2023). Weakly supervised video-based cardiac detection for hypertensive cardiomyopathy. *BMC Medical Imaging*, 23(1), 163. <https://doi.org/10.1186/s12880-023-01035-0>.
- Farhad, M., Masud, M. M., Beg, A., Ahmad, A., Ahmed, L. A., & Memon, S. (2023). A data-efficient zero-shot and few-shot siamese approach for automated diagnosis of left ventricular hypertrophy. *Computers in Biology and Medicine*, 163, Article 107129. <https://doi.org/10.1016/j.combiomed.2023.107129>.
- Groun, N., Villalba-Orero, M., Lara-Pezzi, E., Valero, E., Garicano-Mena, J., & Le Clainche, S. (2022). Higher order dynamic mode decomposition: From fluid dynamics to heart disease analysis. *Computers in Biology and Medicine*, 144, Article 105384. <https://doi.org/10.1016/j.combiomed.2022.105384>.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In *Computer vision—ECCV 2016: 14th European conference, Amsterdam, the Netherlands, October 11–14, 2016, proceedings, part IV 14* (pp. 630–645). Springer, https://doi.org/10.1007/978-3-319-46493-0_38.
- Hetherington, A., Corrochano, A., Abadía-Heredia, R., Lazpita, E., Muñoz, E., Díaz, P., et al. (2023). ModelFLOWS-app: data-driven post-processing and reduced order modelling tools. *arXiv preprint arXiv:2305.17150*.
- Holste, G., Oikonomou, E. K., Mortazavi, B., Wang, Z., & Khera, R. (2022). Self-supervised learning of echocardiogram videos enables data-efficient clinical diagnosis. *arXiv preprint arXiv:2207.11581*.
- Le Clainche, S., & Vega, J. M. (2017). Higher order dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, 16(2), 882–925.
- Le Clainche, S., Vega, J. M., & Soria, J. (2017). Higher order dynamic mode decomposition of noisy experimental data: The flow structure of a zero-net-mass-flux jet. *Experimental Thermal and Fluid Science*, 88, 336–353. <https://doi.org/10.1016/j.expthermflusci.2017.06.011>, URL <https://www.sciencedirect.com/science/article/pii/S089417771730184X>.
- Lee, S. H., Lee, S., & Song, B. C. (2021). Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*.
- Madani, A., Ong, J. R., Tibrewal, A., & Mofrad, M. R. (2018). Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *npj Digital Medicine*, 1(1), 59. <https://doi.org/10.1038/s41746-018-0065-x>.
- ModelFLOWS research group (2023). ModelFLOWS-app. URL <https://github.com/modelflows/ModelFLOWS-app>.

- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 211–252. <http://dx.doi.org/10.1007/s11263-015-0816-y>.
- Schmid, P. J. (2010). Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656, 5–28.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- Sirovich, L. (1987). Turbulence and the dynamics of coherent structures. I. Coherent structures. *Quarterly of Applied Mathematics*, 45(3), 561–571.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conf. comput. vis. pattern recognit.* (pp. 2818–2826).
- Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3), 279–311.
- Vafaezadeh, M., Behnam, H., Hosseinsabet, A., & Gifani, P. (2023). CarpNet: Transformer for mitral valve disease classification in echocardiographic videos. *International Journal of Imaging Systems and Technology*, 33(5), 1505–1514. <http://dx.doi.org/10.1002/ima.22885>.
- Vega, J. M., & Le Clainche, S. (2021). Higher order dynamic mode decomposition. In J. M. Vega, & S. Le Clainche (Eds.), *Higher order dynamic mode decomposition and its applications* (pp. 29–83). Academic Press, <http://dx.doi.org/10.1016/B978-0-12-819743-1.00009-4>.
- Wahlang, I., Maji, A. K., Saha, G., Chakrabarti, P., Jasinski, M., Leonowicz, Z., et al. (2021). Deep learning methods for classification of certain abnormalities in echocardiography. *Electronics*, 10(4), 495. <http://dx.doi.org/10.1186/s12880-023-01035-0>.
- World Health Organization (1999). Cardiovascular diseases (CVDs). URL [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).