Artificial Intelligence-based Echocardiogram Video Classification by Aggregating Dynamic Information

Zi Ye[1,2], Yogan J. Kumar[2], Goh O. Sing[2], Fengyan Song[3], Xianda Ni[4], and Jin Wang[5*]
[1] Department of Information Technology, Wenzhou Polytechnic, Wenzhou 325035, China
[e-mail: yezi1022@gmail.com]
[2] Centre for Advanced Computing Technology, Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka 76100, Malaysia
[e-mail: yogan@utem.edu.my, goh@utem.edu.my]
[3] Shanghai Gen Cong Information Technology Co. Ltd, Shanghai 201200, China
[e-mail: songfy@ai-galaxy.com]
[4] Department of Ultrasonography, the First Affiliated Hospital of Wenzhou Medical University
Wenzhou 325003, China
[e-mail: xianda.ni@gmail.com]
[5] School of Computer & Communication Engineering, Changsha University of Science & Technology
Changsha 410004, China
[e-mail: jinwang@csust.edu.cn]
[*]Corresponding author: Jin Wang

Received December 28, 2020; revised January 23, 2021; accepted February 3, 2021; published February 28, 2021

Abstract

Echocardiography, an ultrasound scan of the heart, is regarded as the primary physiological test for heart disease diagnoses. How an echocardiogram is interpreted also relies intensively on the determination of the view. Some of such views are identified as standard views because of the presentation and ease of the evaluations of the major cardiac structures of them. However, finding valid cardiac views has traditionally been time-consuming, and a laborious process because medical imaging is interpreted manually by the specialist. Therefore, this study aims to speed up the diagnosis process and reduce diagnostic error by providing an automated identification of standard cardiac views based on deep learning technology. More importantly, based on a brand-new echocardiogram dataset of the Asian race, our research considers and assesses some new neural network architectures driven by action recognition in video. Finally, the research concludes and verifies that these methods aggregating dynamic information will receive a stronger classification effect.

Keywords: Classification, Deep Learning, Echocardiogram View, LSTM, Two-Stream Network

This work was supported by the Basic Research Project of Wenzhou, China (Grant No. G2020019), and also by National Natural Science Foundation of China (61772454, 62072056).

http://doi.org/10.3837/tiis.2021.02.007 **ISSN : 1976-7277**

1. Introduction

Medical diagnosis is usually a non-invasive approach. Studies of medical imaging have captured great attention in the deep learning and computer vision fields given that the results might benefit people significantly. Automatically identifying structures in medical images involve modeling their appearance over several subjects, but the natural variations of human anatomy may pose several challenges to this modeling process. This research focuses on the application and feasibility of neural networks for medical image recognition.

Echocardiography, an ultrasound scan of the heart, is the main physiological test for heart disease diagnoses. An echocardiographic exam often involves making visualized measurements of the anatomy. When using a 2D ultrasound probe to image the heart, several different views are obtained depending on the correct location and angulations of the probe before making any measurements [1]. Some of such views are identified as standard views because of the ease and presentation of the evaluations of the major cardiac structures in them. Therefore, the first imperative step in interpreting an echocardiogram is to decide the standard cardiac view, and then the automated classification will enhance the workflow and also enable non-stop scanning without pressing a single button.

Nevertheless, the recognition of ultrasound images is particularly difficult because the borders of the cardiac structures can be corroded by various kinds of noise. Besides, signal dropouts, speckle noise, and low contrast can be caused by the poor imaging quality of 2D echocardiogram videos. This would also lead to bias of the interpretation over the image, setting US data apart from other image modalities used in medicine [2].

Besides, the special anatomical structure of the heart also contributes to the difficulty of recognizing the standard cardiac view. Given substantial intra-view variability of different patients and much inter-view similarity of different classes – for example, parasternal short axis at the apex, papillary muscles, and mitral valve are all belonging to the short-axis view, and these three views share a similar structure in terms of the interventricular septum, left ventricle, and pericardium – the key criteria for distinguishing between them are the apex, papillary muscles, and mitral valve [3].

To assist echocardiographers in speeding up the diagnosis process by improving the use of echocardiography for precision medicine, the key objective of this research is to improve supervised deep learning to assess the capability of the state-of-the-art scene understanding methods proposed in computer vision to identify the standard cardiac views being visualized in echocardiograms.

In this paper, we first employed classic neural networks that were proven successful in quite a few different fields of medical image classification to classify individual video frames. However, these methods omitted the ability to use dynamic knowledge about how features such as ventricular walls or heart valves behave during the cardiac cycle. Therefore, we then evaluate some new deep learning mechanisms that the action recognition inspired, which was shown in the video, trying to capture the complementary information and features on continuous image sequences in video.

Novel computational models should no longer only achieve excessive precision, however, it also needs to have real-time output efficiency to have a translational effect in medicine. We extracted various sequence images from the video and put them into the neural networks to assess the impact of various frames over the classification accuracy. Timeliness and precision can both be taken into consideration when it comes to practical application, and it is proposed that the number of frames adopted can be minimized if the accuracy is sufficiently high.

Our main task in this research is to propose fully automatic and robust approaches to apply

the deep learning approaches for the classification of real-time cardiac views to improve their use in clinical practices. Furthermore, we explored if the incorporation of sequence information sustained by the moving heart's video images would achieve better success on the classification task. Compared to preceding research, this paper's contributions are as follows:

(1) Annotation and training on a brand-new echocardiogram dataset of Asian race, prepared by a private hospital in Malaysia. Current work of artificial intelligence in echocardiography are mostly based on European ethnicity, however, there was evidence of variations in the physiological and anatomical structures of the heart that could be related to racial and ethnic variations, and the standard reference values of echocardiographic measurements were provided by different countries to their citizens [4].

(2) Our study considers a selection of nine of the most common cardiac view: parasternal long axis (PLAX), parasternal short axis at aortic valve level (PSAX-AV), that at apical level (PSAX-AP), that at mitral valve level (PSAX-MV), that at papillary level (PSAX-MID), apical 2-chamber (A2C), apical 3-chamber (A3C), apical 4-chamber (A4C) and apical 5-chamber (A5C). Among them, A2C and A4C are the essential views for the measurement of Ejection Fraction (EF), which is widely used as an indicator of the severity of heart failure. We also consider a class of "OTHERS" for other cardiac aspects, because usually, a comprehensive study comprises more required views and measurements, such as Suprasternal (SSN) and Subcostal (SC). **Fig. 1 s**hows examples of related views.

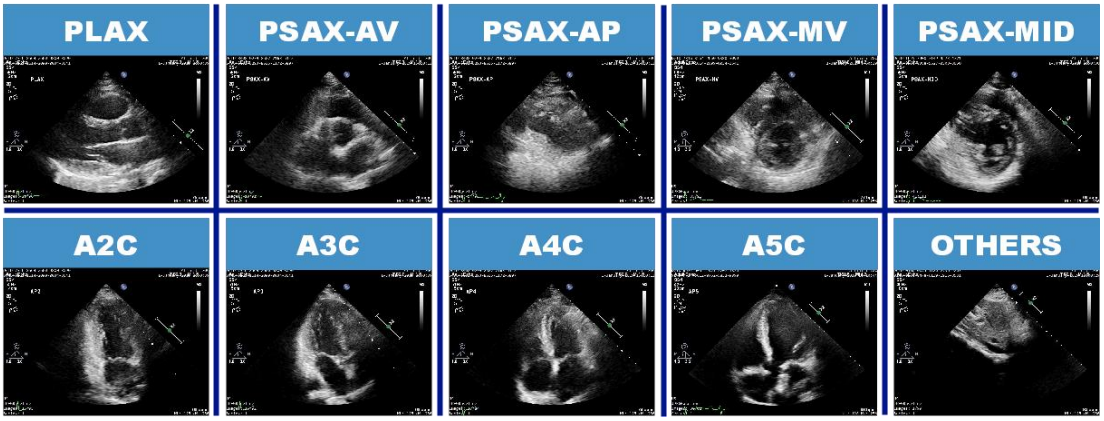

Fig. 1. In transthoracic echocardiography, nine cardiac views are acquired at arbitrary stages of the heart cycle. Examples of PLAX, PSAX-AV, PSAX-AP, PSAX-MV, PSAX-MID, A2C, A3C, A4C, and A5C, as well as non-assignable samples labeled 'OTHERS' are illustrated.

(3) The images acquired during the examinations will include some unnecessary information, for example, patient identifiers and electrocardiogram (ECG). Moreover, the ground truth mark is placed in the upper left corner of the image in this dataset. Therefore, we decided to adopt view segmentation prior to the view classification. The contours of the echocardiogram images can be discerned accurately by the segmentation pipeline. A map of related areas will be created.

(4) Unlike conventional machine learning techniques with hand-crafted characteristics [5, 6], a deep learning approach learns not only the feature classification but also the feature extraction directly from the training data [7]. Here we explored the efficacy of several classical convolutional neural network (CNN) architectures to acquire the cardiac anatomical characteristics. As a result, the best-performing network was Xception, which is a member of lightweight neural networks [8].

(5) Our research considers the movement of objects between frames, including long short term memory networks and two-stream networks [9]. After analyzing the consequences obtained from performed experiments, we concluded and verified that the methods including the features along the time dimension will receive a stronger classification effect.

(6) Theoretically, deep learning networks have better performance with more datasets. In our research, every single class has about 270 videos on average. Compared to the published work, this is not dramatically large. However, CNN had better performance compared to all the studied hand-crafted approaches. Particularly, our best-performing architecture outperforms other models built on the much larger dataset [10], and also achieved a result very close to the state-of-the-art.

Section 2 reviews the significant efforts which have been put into the automatic classification of echocardiographic views and also describes the approaches and techniques related to it briefly. In Section 3, the methodology used in this research is presented, here we constructed three separate sets of neural network architectures for the identification of cardiac views. Section 4 offers the results of the research. A comparison is made as well. Section 5 focuses on the discussion based on the results obtained from the previous section. Finally, Section 6 presents the conclusion also the suggestions for future work.

2. Related Work

The standard cardiac view classification is the foundation for intelligent analysis and interpretation of echocardiography. Massive studies have been made during recent years regarding the automatic view classification methods. Previous studies claim that the overall accuracy of image sequences, as stated by Østvik et al, is as high as 98.9 percent [11]. Generally, the inclusion of more views has greatly decreased accuracy, and with more data available, the outcome would have higher accuracy. Howard et al. recorded the largest data set collection to the best of our knowledge, comprising 6549 and 2183 videos for training and validation, respectively [12].

2.1 Traditional Machine Learning Methods

A majority of the previous studies have used traditional machine learning methods to recognize the standard cardiac view. Ebadollahi et al. for the first-time applied Markov Random Field to model to represent the constellation of the heart chambers partially, followed by inputting it to the Support Vector Machine classifier to look for the view label [13]. Balaji et al. proposed an automated classification algorithm based on histograms and statistical features to recognize the short parasternal axis, long parasternal axis, apical four-chamber, and apical two-chamber, which result in an average accuracy of 87.5 percent [14]. Khammis et al. suggested the use of Spatio-temporal feature extraction and supervised dictionary learning approach to classify three apical views with an average recognition rate of 95 percent [15]. These traditional machine learning methods can be summarized into two stages: first, the image can be represented by prior manual design features, and second different ML classification methods are applied to model and analyze these feature vectors [16, 17].

2.2 Convolutional Neural Networks Methods

The deep convolutional neural system has outperformed the traditional methods in many classification tasks during recent years. The ultrasound image analysis community has attached great attention to the CNNs [18]. Deep-learning methods can be understood as representation-learning methods with various representation levels. It is created by simple but

non-linear modules, each of which could change the representation to a higher level. Moreover, Yann LeCun et al. stated that the purpose of the convolutional neural networks is data processing. The data has various arrays. For instance, the architecture of a regular convolutional and the two-dimension images compose a series of stages. Besides, supervised deep learning is a type of advanced technology in the medical image analysis and computer vision fields. An objective function evaluating the error between the desired scores and the output scores is computed. The internal adjustable parameters are modified afterwards by the machine to decrease such an error. There might be massive adjustable weights in a regular deep-learning system. Meanwhile, it might also have massive labeled examples.

In the echocardiographic field, CNNs technology has already achieved reliable results. Using the VGG-16 network, Ali Madani et al. distinguished 12 video views, the accuracy of which was 97.8 percent [19]. Moreover, Zhang used a deep architecture with 13 layers, considered a large number of echocardiography view classes (23 views), applied to a large data set (14035 echocardiograms), and reported an 84 percent accuracy overall (including partially obscured views) [20]. However, in most of these works, the researchers only designed classic convolutional neural networks and the input is a single cardiac image, ignoring the dynamic information during the cardiac cycle.

2.3 Other Deep Learning Methods

Recently, some studies have extended the advanced convolutional neural systems to classify the echocardiographic video images by incorporating the dynamic information on how features shift during the cardiac cycle.

2.3.1 Recurrent Neural Networks

Traditional neural networks (including CNNs) assume that all inputs and outputs are independent of each other. This assumption is very limiting for many tasks, such as in natural language recognition, where contextual semantics are often required. A recurrent neural network performs the same operation on each element of the sequence, each operation relying on the results of previous calculations, the RNN remembers all the information that has been calculated for the current position. Thanks to its good temporal characteristics, RNNs are widely used in natural language processing, handwriting font synthesis, and action recognition.

2.3.2 Two Stream Networks

The classification method based on multiple frame integration inputs the original information of the video into the neural network for end-to-end learning and extracts the temporal information of the video through different stages of fusion in the neural network. However, the feature extraction engineering for data in the neural network is in a complete black box state, so we cannot observe whether the extracted features are effective enough to give timely correction. Simonyan [21] proposed a two-stream neural network structure. In this structure, the video is preprocessed and its temporal information is extracted manually. To maintain the spatial information of the video, two independent CNNs are adopted in the construction of the neural network structure, of which, one way is to process the spatial information of the video, using the original video frame as the input to represent the information related to the characters and scenes in the video, the other way is to process the temporal information of the video, using the temporal feature extracted manually as input to represent the motion-related information of the video. Finally, fuse the output by average to get the recognition result.

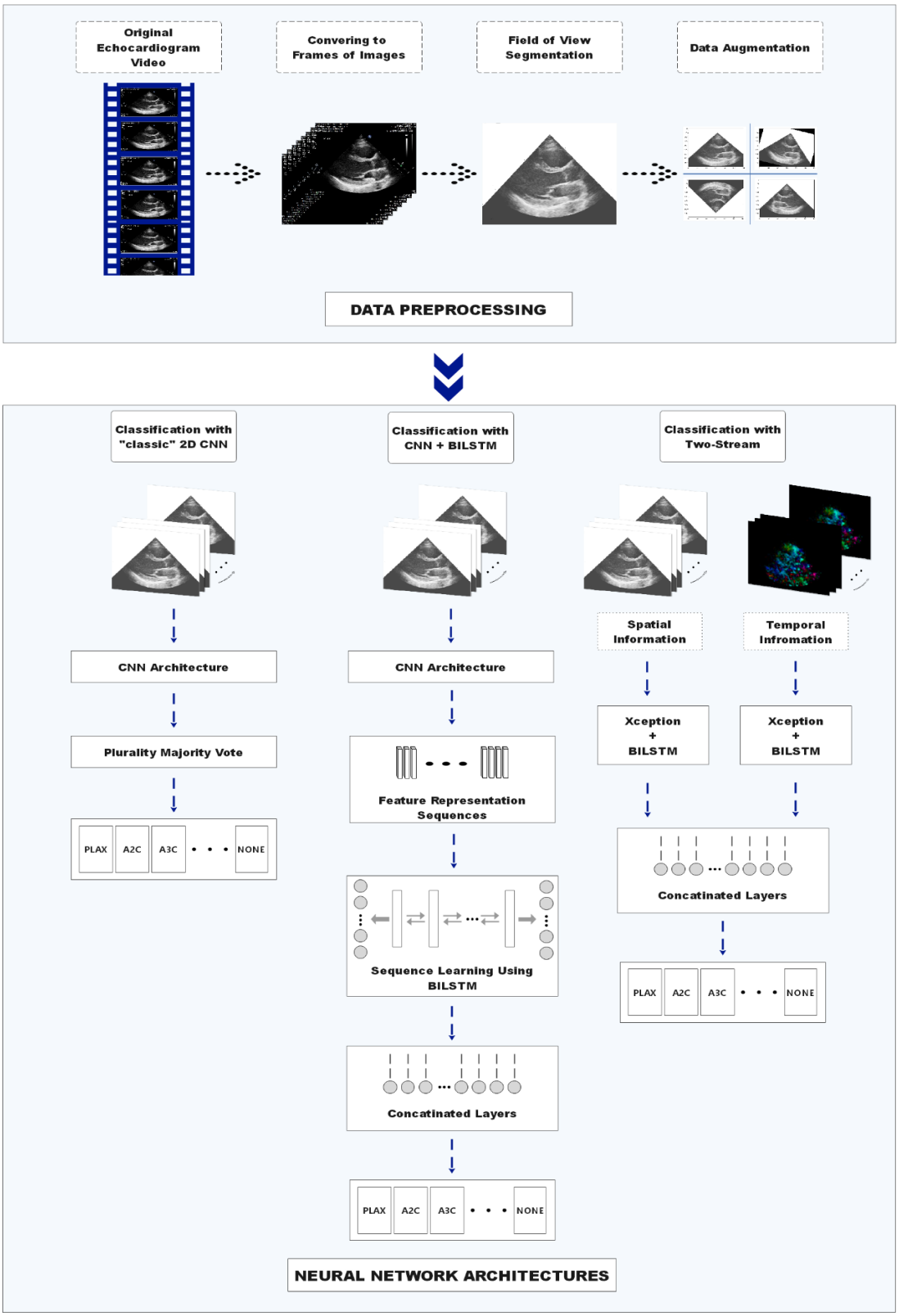

Fig. 2. The three classes of neural network architectures used in this study

The temporal motion information of video can be represented by optical flow [22]. The concept of optical flow was first proposed by Gibson. Optical flow is the flow of light, which refers to the trajectory of pixel change generated by the instantaneous motion of a space object in the video. Such instantaneous motion can be represented by the relative position of the same object in two adjacent frames of the video. The optical flow is generally generated by the movement of objects, backgrounds, and cameras. Gao et al. firstly developed a two-strand CNN architecture, integrating both the temporal and spatial information that is sustained through the video images of the moving heart. This classifies the results of 8 viewpoint categories in the best way, the accuracy of which is 92.1 percent [3]. Nonetheless, most researches currently rely only on spatial information. We assume that the classification precision will be further improved with the consideration of sequence information.

3. Deep Learning Architectures for Echo Views Classification

Howard et al. discussed four groups of CNN architectures for the classification task, which serves as the main inspiration for this project [12]. Here we evaluated three separate sets of new architectures, which are shown in **Fig. 2**.

3.1 Classic 2D CNNs Architectures

We assessed three distinct CNN architectures, i.e., VGG16 [23], Inception-Resnet-V2 [24], and Xception, Each of them is an advanced network design for image recognition. The tail layers of these classical CNN architectures are replaced and improved. The last fully-connected layer must be ten nodes, corresponding to ten probabilities for each processed image. It is worth mentioning that during the testing procedure, the final predicted label is collated by plurality voting on multiple frames of the video, the complete steps are described in **Fig. 3**.

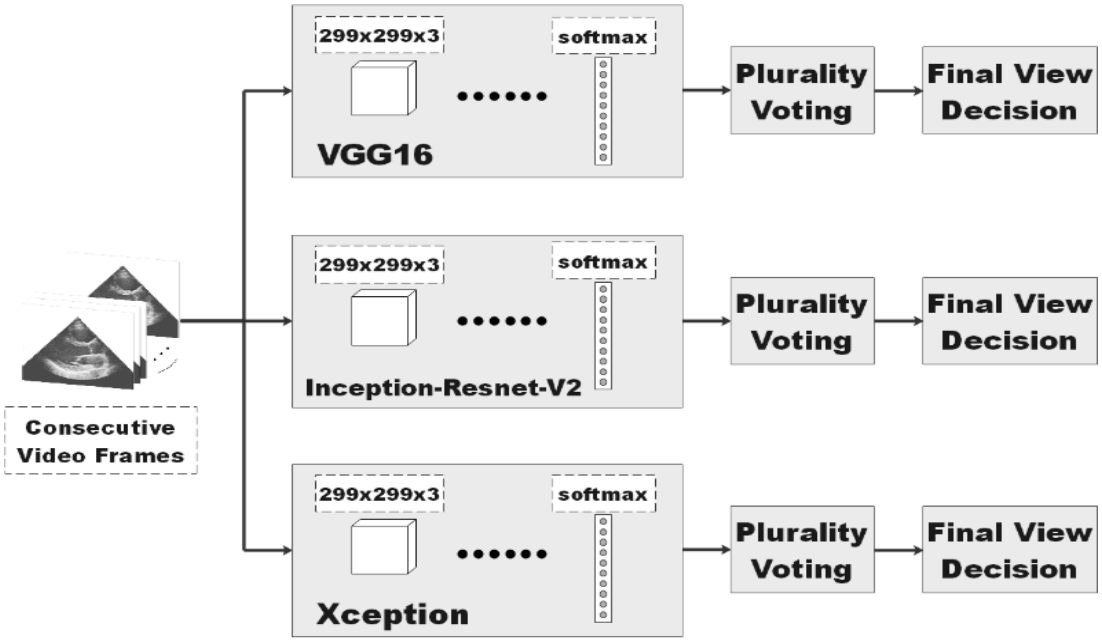

Fig. 3. The procedure of video classification for classic 2D CNNs

3.2 CNN+BiLSTM Architectures

In this study, an entire video is regarded as a sequence of two-dimensional images, passing through several classical CNN frame by frame. After eliminating the extreme fully-connected layer of the CNN model, the output from each frame is a 512-dimensional vector, which is placed sequentially into a Bi-directional Long Short-Term Memory network (Bi-LSTM). It is then followed by a hidden layer with 128 nodes, to extract the dynamic features. The LSTM network introduces additional neurons to "archive" the previous sequence frames information, and the output of the current moment is determined by the past information and present input data.

The structure of CNN+BiLSTM architecture is illustrated in **Fig. 4**, and here we also recover a series of "gate" state variables inside the LSTM cells, which are shown in the following formula [25]:

$$i_t = \sigma(W_i fc^t + U_i h_{t-1} + b_i) \tag{1}$$

$$o_t = \sigma(W_o fc^t + U_o h_{t-1} + b_o) \tag{2}$$

$$f_t = \sigma\left(W_f fc^t + U_f h_{t-1} + b_f\right) \tag{3}$$

$$\hat{C}_t = tanh(W_c fc^t + U_c h_{t-1} + b_c) \tag{4}$$

$$C_t = i_t \odot \hat{C}_t + f_t \odot C_{t-1} \tag{5}$$

$$h_t = o_t \odot \tanh(C_t) \tag{6}$$

where $\odot$ represents the element-wise product and $\sigma(\cdot)$ is the sigmoid activation function; fc^t represents the completely-connected layer input at time step t to the LSTM cell; W_i, W_o, W_f, W_c, U_i, U_o, U_f and U_c refer to the weight matrices corresponding to different state parameters; b_i, b_f, b_c and b_o are bias vectors; i_t, o_t, f_t, C_t and h_t refer to the input gate, output gate, forget gate, cell state, and hidden state, respectively. $\hat{C}_t$ is the candidate cell state before the combination of the forget gate and the prior cell state (C_{t-1}).

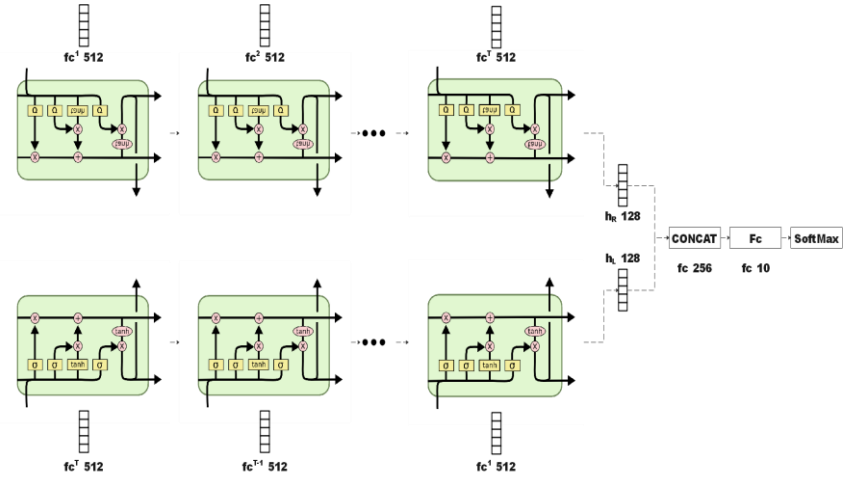

Fig. 4. The structure of CNN+BiLSTM architectures for video classification

3.3 Spatiotemporal-BiLSTM Architecture

The proposed Spatiotemporal-BiLSTM architecture in this paper is shown in **Fig. 5** and it mainly contains three modules, which are illustrated in the subsections below.

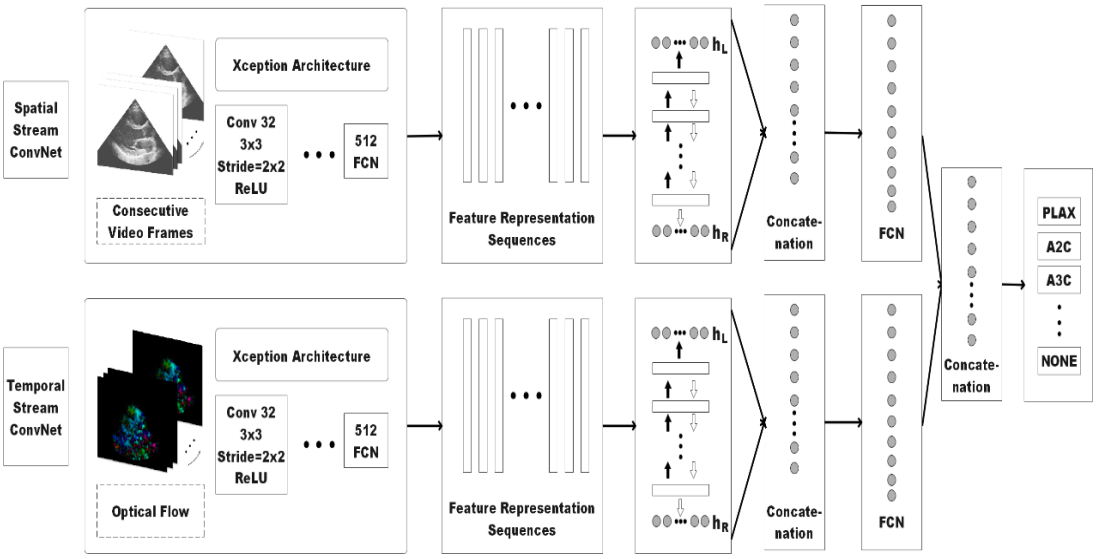

Fig. 5. The structure of Spatiotemporal-BiLSTM architecture for video classification

3.3.1 Convolutional-to-Fully Connected

The first module generates two independent streams of data: a temporal and a spatial stream would guarantee that the model would be able to identify the information from both temporal motion and visual appearance [26, 27]. The dense optical flow technique is used to calculate the acceleration along the time direction of every point. The sequential video frames are processed by both streams. A time-distributed Xception network is created. Two feature sequences represent the input video after feature extraction. There are 512 dimensions of every feature vector.

Similar to the CNN+BiLSTM Architectures in Section 2, let the last fully-connected layer output of i^{th} video frame for both streams be fc_{sp}^{i} and fc_{te}^{i} respectively, then the feature representation sequences for two streams of this sample video can be defined as,

$$fc_{sp} = \left[fc_{sp}^{1}, fc_{sp}^{2}, fc_{sp}^{3}, \ldots, fc_{sp}^{T}\right] \tag{7}$$

$$fc_{te} = \left[fc_{te}^{1}, fc_{te}^{2}, fc_{te}^{3}, \ldots, fc_{te}^{T}\right] \tag{8}$$

where T refers to the total number of the sample video frames.

3.3.2 Fully Connected-to-BiLSTM

In the second module, the extracted feature representation sequences are then fed to the BiLSTM network, the outputs can be written as follows,

$$h_{sp} = BiLSTM\left(fc_{sp}\right) \tag{9}$$

$$h_{te} = BiLSTM(fc_{te}) \tag{10}$$

3.3.3 BiLSTM-to-Classification

In our third module, the outputs from the above BiLSTM network are individually connected to a fully connected layer with 10 nodes, then the outputs of (11) and (12) are merged. Afterwards, they are passed via a soft-max layer for the assessment of the final classification, which could be represented as,

$$fcn_{sp} = FC(h_{sp}) \tag{11}$$

$$fcn_{te} = FC(h_{te}) \tag{12}$$

$$h = concat(fcn_{sp}, fcn_{te}) \tag{13}$$

$$y = softmax\{FC(h)\} \tag{14}$$

4. Experimental Results and Analysis

4.1 Data Acquisition

All datasets were collected and de-identified at a private hospital in Malaysia, with waived consent under the Institutional Review Board (IRB). Methods were performed following relevant regulations and guidelines. There were echocardiogram studies from 267 patients being chosen randomly. Besides, 7994 were extracted from the hospital's echocardiogram database in DICOM format. Out of the sample, the videos with the following classes are selected and annotated manually by a board-certified echocardiographer, finally categorized into ten different folders: PLAX, PSAX-AV, PSAX-MV, PSAX-AP, PSAX-MID, A4C, A5C, A3C, and A2C. The folder containing "OTHERS" does not belong to any of the nine standard views. **Table 1** summarizes the data indicating the class balance.

Table 1. The overview of the dataset

	PLAX	**PSAX-AV**	**PSAX-MV**	**PSAX-AP**	**PSAX-MID**
Patients	258	248	247	253	254
Videos	308	264	263	268	269
	A5C	**A4C**	**A3C**	**A2C**	**OTHERS**
Patients	254	256	242	248	121
Videos	273	275	259	264	250

It is worth noting that some patients have multiple examinations on a regular basis, in order to avoid patient overlapping between Training and Test Sets, this subset of 2693 videos from 10 views was randomly split using Python into Training, validation, and Test datasets strictly according to different patients with a ratio of approximately 2.5:1:1. Therefore, each dataset affirmatively included videos from the echocardiographic records of different patients to

ensure each sample is independent of the other.

The number of patients in Training, Validation, and Testing datasets is shown particularly in **Fig. 6**, and the video files generated from these patients for each standard cardiac view are described in **Fig. 7**. The test dataset was applied for the evaluation of the performance of the final trained model.

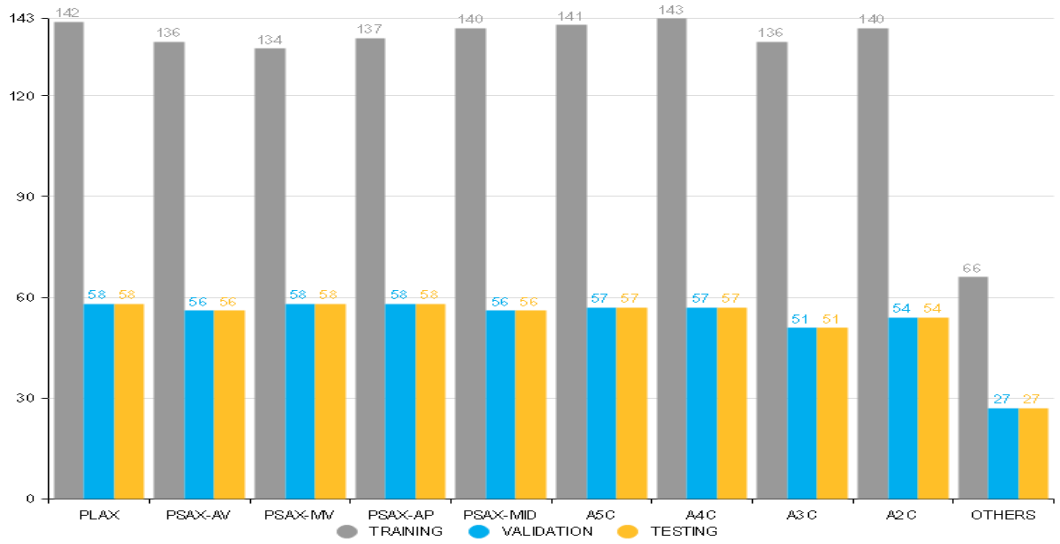

Fig. 6. The numbers of patients for every ten viewpoints in the database and applied for training, validation, and testing correspondingly

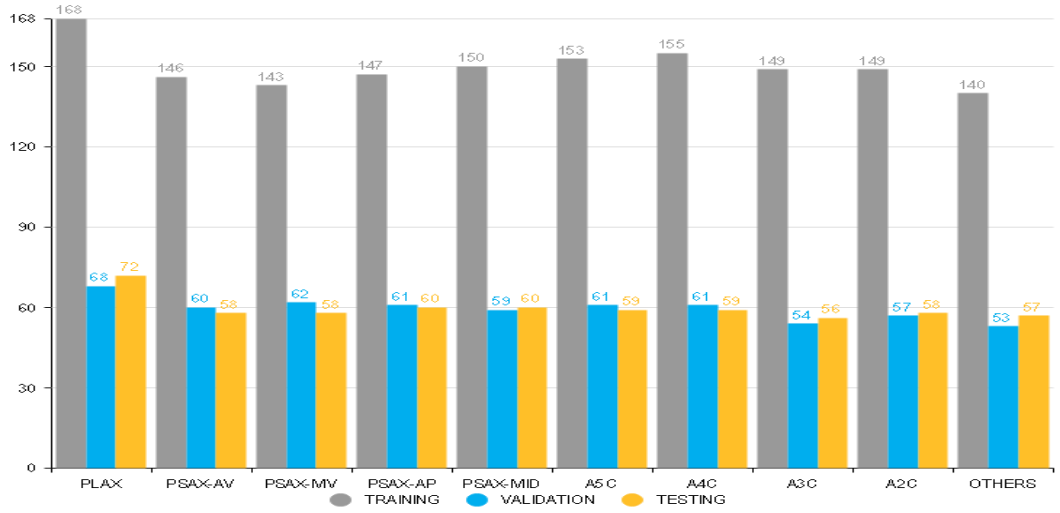

Fig. 7. The numbers of videos for every ten viewpoints in the database and applied for training, validation, and testing correspondingly

4.2 Data Preprocessing

Each DICOM-formatted echocardiogram video comes with a set of visual features that hold less relevance with the video identification, such as class labels, patient digital identifiers, electrocardiogram loops, study duration, etc. Therefore, some related structure segmentations

are used to explore the predictive models with particular visual features and also simplifying the classification task. Our segmentation pipeline is shown in **Fig. 8**.

First of all, each echocardiogram video is divided into constituent frames and changed into images in YBR format, because the YBR format has a better FoV segmentation performance than other color spaces. Morphological Opening and Closing Operation was then applied to remove irrelevant details in every single image and in the meanwhile obtained a rough contour about the main Field of View (FoV). Next, the left and right boundary lines of FoV are determined according to the minimum slope, followed by calculating the angle degree formed by two straight lines on two-sides. Finally, the masks were drawn on the Field of View outline containing the clinically appropriate area with pixels set to 1 inside the mask and pixels set to 0 outside the mask. Using this approach offers an effective way to predict a segmentation map over the key FoV, which is applied before the classification of view to the input image.

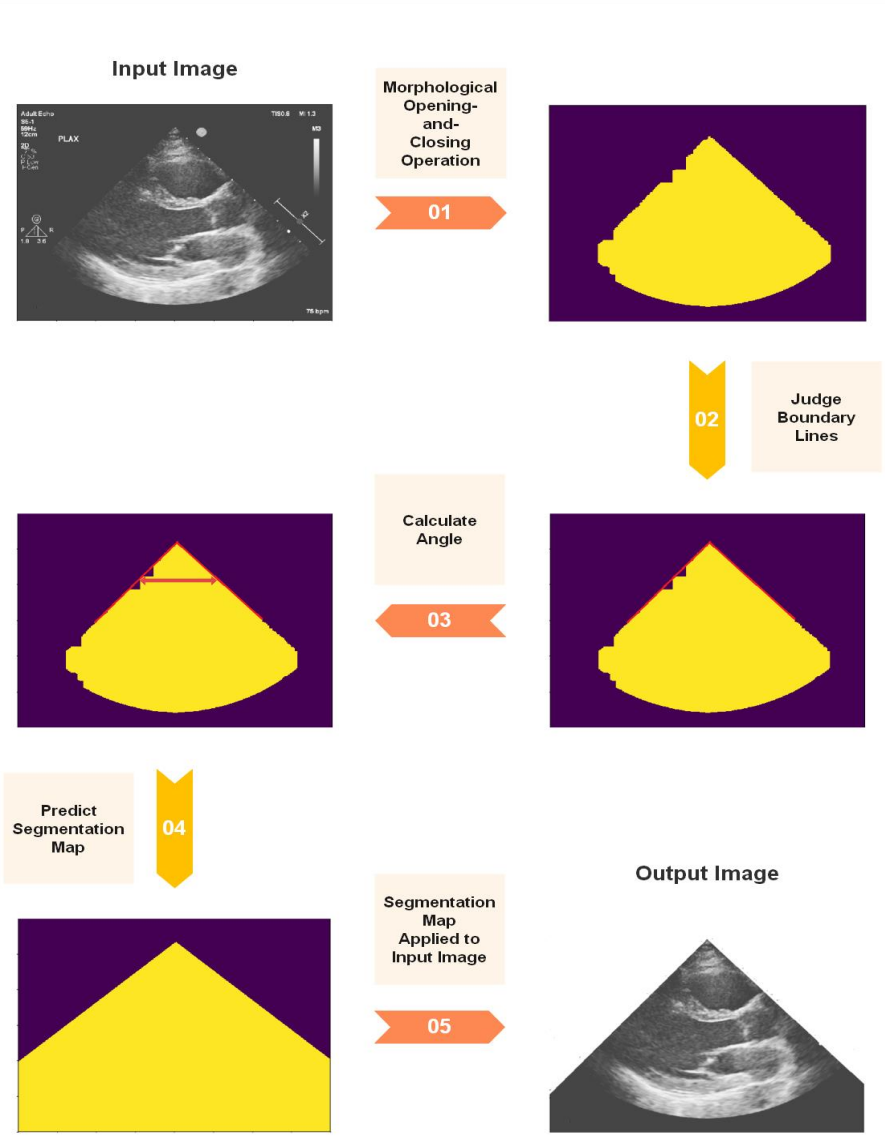

Fig. 8. Procedures used for removing the class label and segmenting relevant visual structures

Also as per standard practice, to enhance the generalization of the unseen data and also the robustness of the model, data were augmented at run-time by random crop with a scale of 0.75 to 1, rotations of up to 20 degrees, and horizontal/vertical flips. The true label for each data is a one-hot vector that corresponds to the view of that sample.

4.3 Performance of Different Architectures

4.3.1 Model Training

Python 3.8.3 was used to design the neural network architectures and to establish the learning environment, in which the Pytorch framework was utilized. Experiments were implemented through a workstation. An Ubuntu 18.04 operating mechanism was used to install the workstation. The hardware consisted of an Intel(R) Xeon(R) Gold 5220 CPU with a clock speed of 2.20 GHz, 64GB RAM, and an NVIDIA Quadro RTX 8000 GPU with 48GB of memory.

All echocardiogram videos formatted by DICOM were converted to YBR (three channels) and then divided into constituent frames, followed by converting into 299x299-pixel standardized images. In addition, 256-pixel values were scaled from [0.255] to [0.1] for each red, green and blue light. Afterwards, each dataset would be subtracted with the mean of the training data based on the standards for image recognition tasks.

However, the method is different when training models incorporate dynamic information. By cutting the entire video into several 30-frame videos firstly, these separate short videos were then put into the architectures for training. We researched the performance of three different neural network architectures for the classical 2D CNNs: VGG16, Inception-Resnet-V2, and Xception. Using mini-batch gradient descent with a batch size of 128, the training was conducted for 10 epochs. One epoch is described in machine learning as a complete passage with training data, among which the independent data set would be used merely for testing purposes. Weights derived from training on ImageNet would be used for the initialization of each network [28]. It is a large image database for object recognition.

The CNN+BiLSTM structure comprised the trained classic 2D network which placed multiple frames of each echocardiogram. Before training, the CNN part used for feature extraction and representation received model weights from their trained Classic 2D CNNs models, whereas the Uniform Initialization was used on the Bi-LSTM network part.

Finally, the Spatiotemporal-BiLSTM Architecture included two distinct CNN+BilSTM "streams" (one "spatial" stream and one "temporal" stream), which separately process the spatial and temporal features of a video until the data is merged and the view is finally determined. Firstly, the Temporal-BiLSTM network was individually trained for video identification through the optical flow data only. This emphasizes how the structures of a video move between two sequential frames. Those weights learned by the individual Spatio-BiLSTM network and Temporal-BiLSTM network were saved and then applied as the initial weights for the final Spatiotemporal-BiLSTM network. We found this will avoid unnecessary training time and lead to significantly faster convergence with improved accuracy. For both CNN+BiLSTM Architectures and Spatiotemporal-BiLSTM Architecture, the batch size used was 6, and the training was set for 15 epochs with early stopping to prevent overfitting.

For all the models mentioned above, categorical cross-entropy loss between predictions and true labels was back-propagated through the network to compute gradient descent and the weights were updated using the ADAM (Adaptive Moment Estimation) [29] optimizer with ReduceLROnPlateau scheduler, which allowed dynamic learning rate reduction based on some validation measurements. There was training for every network until the validation loss

plateaued. The models were saved after every epoch. The test used the model with the highest validation accuracy during 5 epochs for the final assessment.

4.3.2 Model Evaluation

For performance assessment, multiple metrics will be used over the test dataset. The general accuracy would be calculated as the percentage of correctly classified samples. To visualize the output of multi-view classifiers and their related errors, confusion matrices will be measured and plotted as heat maps.

For classical CNN model evaluations, the single testing images will be classified by referring to the view having the most possibility. The plurality voting of multiple images of a specific video will be used to classify the test videos. Similarly, for the other two sets of architectures, the test video is determined also by plurality voting on multiple 30-frame videos generated from the entire video.

Comparisons of the overall accuracy for different architectures are shown in **Table 2**. Both Xception and Inception-Resnet-V2 won the best performance model for classical 2D CNN design, with an accuracy of 93.13%. However, Xception shows stronger feature extraction when combined with BiLSTM to learn the sequential problem (94.30% accuracy versus 93.80% accuracy), indicating that Xception is the best feature extraction architecture for cardiac ultrasound images.

Table 2. Table demonstrating the overall accuracy for each of the networks on the test set

Classical 2D CNNs Architecture	**Overall Accuracy**
Xception	93.13%
VGG16	92.96%
Inception-Resnet-V2	93.13%
CNN+BiLSTM Architecture	**Overall Accuracy**
Xception	94.30%
VGG16	92.46%
Inception-Resnet-V2	93.80%
Spatiotemporal-BiLSTM Architecture	**Overall Accuracy**
Xception	93.80%

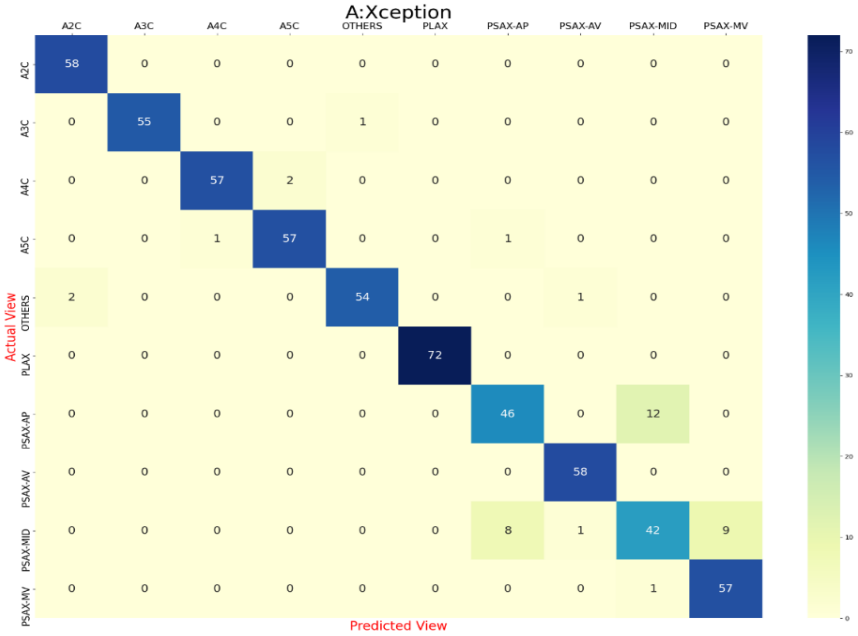

Fig. 9. Echocardiogram view classification through Xception. Confusional matrix demonstrating actual view labels on the y-axis, and neural network-predicted view labels on the x-axis by view category for video classification

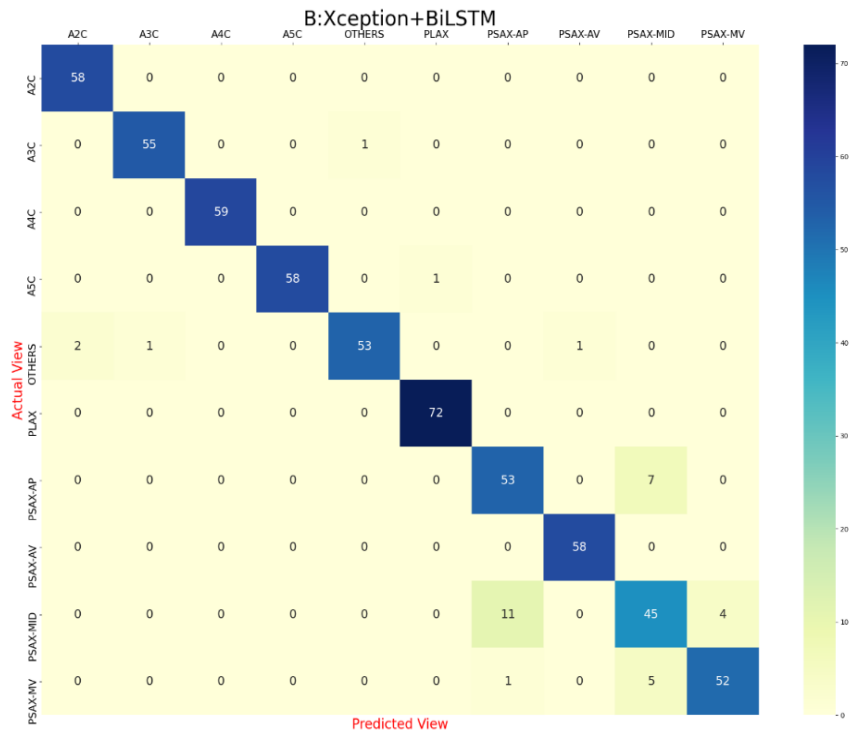

Fig. 10. Echocardiogram view classification by Xception+BiLSTM

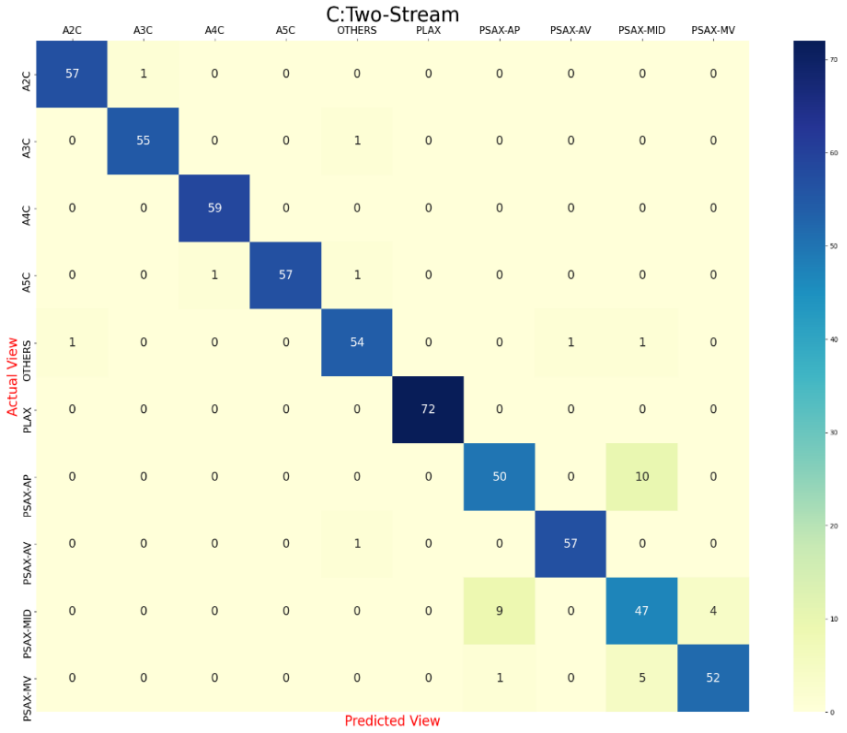

Fig. 11. Echocardiogram view classification by Two-Stream Model

Surprisingly, the two-stream architecture with the addition of optical flow analysis reduces the video classification accuracy (93.80% accuracy) and the best-performing architecture design is Xception+BiLSTM in this study. The confusion matrix is shown in **Fig. 10**. Most matrix entries are diagonal, suggesting that the model was making mostly accurate predictions. There are misclassifications of 10-15% between PSAX-MID and PSAX-AP, so some improvements are needed in the correct classification of these two views. **Fig. 9-11** also show the rates of disagreement with only a few percent in the best-performing systems. These errors look predominantly clustered among parasternal short-axis views representing anatomically adjacent imaging planes.

5. Discussion

5.1 Dynamic Neural Networks Perform Significantly Better

Two approaches were successful in terms of integrating dynamic information: CNN+BiLSTM and two-stream network. The error rates were lower for both two-stream network and the best CNN+BiLSTM network than that for the best classical 2D CNN.

Most of these benefits tend to be improved by the discrimination between specific pairs of views that are hard for regular CNNs. For instance, as shown in **Fig. 12**, the short-axis view of the papillary muscle, and that of the apical level are obtained by positing the probes in the second and third intercostals of the left sternum, followed by making a cross-section of the papillary muscle and the apical level during the measurement. Therefore, the structures of these two cardiac views have great similarities, making it difficult to distinguish. Furthermore, it should be noted that the misclassifications made by such advanced networks are close to the origins of differences of opinion between human experts.

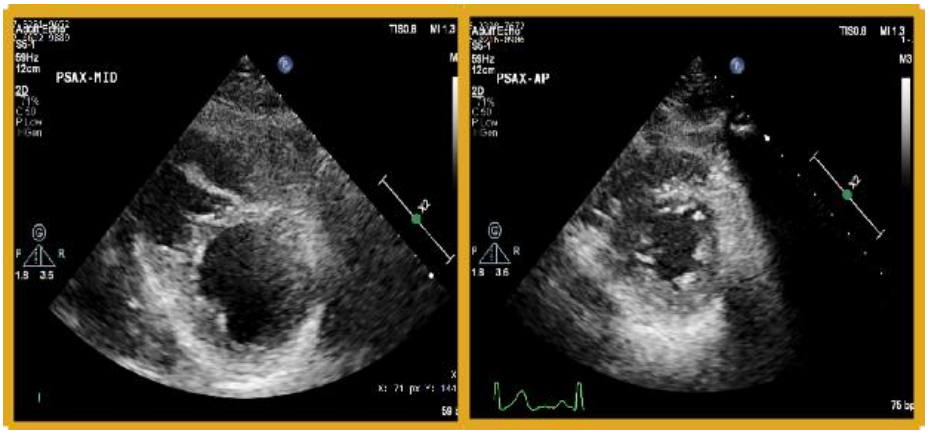

Fig. 12. Examples of the short-axis view of the papillary muscle, and that of the apical level

The long and short-term memory network introduces additional neuron traversal to "record" the previous input sequence information, and the output at the current moment is determined by the state variables and input variables. For echocardiogram videos, the distinctive papillary muscle and mitral valve only appear during diastole. Consequently, the main feature knowledge can be learned through recurrent neural networks.

5.2 No Rule on the Accuracy of Different Frames

We further examine the effect of the number of frames on classification accuracy performance to see if an optimal point exists. Such optimal point needs to make sure that the video frames are efficiently small but still can keep the excessive structural information. The features of video with various frames (15 and 30) are extracted through different CNN architectures and input into BiLSTM for classification training, and the models are also evaluated on the testing clips of different frames (15,30,60, and entire). The obtained results are presented in **Table 3**.

Table 3. Classification accuracy on various video frames

	Trained On 30-Frame Clips			**Trained On 15-Frame Clips**			
CNN+BiLSTM Architecture	Entire Video	30 Frames	60 Frames	Entire Video	15 Frames	30 Frames	60 Frames
Xception	94.30%	92.69%	92.89%	94.14%	92.65%	93.10%	93.03%
VGG16	92.46%	91.95%	92.58%	92.46%	90.65%	91.21%	91.74%
Inception-Resnet-V2	93.80%	92.50%	93.13%	93.63%	91.93%	92.12%	92.54%

We may conclude that the models trained on the basis of 15- and 30-frame videos respectively have little difference in classification accuracy on the overall echo videos. However, it illustrates a strong pattern of growing accuracy as the number of frames increases when it comes to testing on videos with different frames. Therefore, we can consider extracting relatively smaller frames to achieve accurate classification, thereby reducing the time of feature extraction, but also apply to videos with appropriate frames in realistic applications, such as a full cardiac cycle.

5.3 Optical Flow Provides a Slight Improvement

From **Table 4**, we can see that the integration of optical flow analysis has provided evident improvements to the 30- and 60-frames test sets but reduces the accuracy of the entire video test sets. This is because our testing methodology is to split the entire video into several 30-frame clips with an interval of 5 frames, and the optical flow information on some of these short clips does not work very well and therefore can be perceived as noise, leading to the final identification mistake.

Table 4. Comparison of CNN+BiLSTM structure and Spatiotemporal-BiLSTM structure

CNN+BiLSTM Architecture	**Entire Video**	**30 Frames**	**60 Frames**
Xception	94.30%	92.69%	92.89%
VGG16	92.46%	91.95%	92.58%
Inception-Resnet-V2	93.80%	92.50%	93.13%
Spatiotemporal-BiLSTM Architecture	**Entire Video**	**30 Frames**	**60 Frames**
Xception	93.80%	93.19%	93.27%

5.4. Additional Factors Leading to Misclassification

There are other key reasons for the misclassification shown as follow:
(1) The image quality is poor, and the image acquisition fails through the probe during the doctors' scanning process.
(2) The features are not obvious, and the characteristics of an abnormal cardiac view are somewhat different from the standard one.
(3) Multiple views are appearing in one single ultrasound video.
(4) Great similarities between different classes. The key explanation for the similarities between classes is that the apical two-chamber, apical three-chamber, and apical four-chamber are all adjacent apical views. For the apical four-chamber view, the probe is positioned at the apical impulse point, with the beam pointing at the right sternoclavicular joint, while for the apical three-chamber, the probe is rotated 120°counterclockwise based on the position of the apical four-chamber. For the apical two-chamber, it is rotated 60° counterclockwise. Therefore, a small jitter during the sonographer's measurement will cause enlarged similarity of the three views and eventually result in model prediction errors. This similarity illustrated the occasional misclassification of echocardiogram videos, most of which included views that may look similar to human eyes.

6. Conclusion and Future Work

Heart disease is the most common circulatory system disease, and echocardiography has the characteristics of being non-invasive and non-radiation, which has become the preferred method for evaluation of cardiac structure and function. Although some semi-automatic analysis software has been utilized [30, 31], the complex heart structure and low-quality images have brought great difficulty to the classification task of standard echocardiographic views. In this study, a series of different neural network architectures are proposed, including classical 2-dimensional CNN structures, CNN+BiLSTM Structures, and two-stream structure. In this investigation, the result showed that these architectures incorporating information described how the structures moved during the cardiac cycle, which has a better performance compared to all the traditional CNNs. The Xception+BiLSTM network has the best performance. The accuracy of the results of the classification was up to 94.30%. This indicates that the Xception network has the best performance in terms of automatically identifying the

discriminative features for echo video images. We believe that further development towards fully cardiac cycle recognition from echo images will increase the standard view classification accuracy.

Once the echo views are identified, the next task is to extract useful information from the relevant views. Such information includes diagnostics for a particular heart valve disease. For instance, aortic stenosis means to narrow down the aortic valve opening to constrain the blood outflow to the aorta. One common diagnostic is the aortic valve opening area (AVA) when it is open at its widest. Since the aortic valves are visible in PLAX, PSAX-AV, A5C, and A3C views, it is worth investigating how to obtain this measurement from these views [10]. Another significant evaluation method of cardiac function is the segmentation of the left ventricle, estimating ejection fraction, and assessing cardiomyopathy, which can be performed based on the apical-2-chamber and apical-4-chamber at end-systole and end-diastole [32].

Acknowledgments

This work is supported by Pantai Hospital Ayer Keroh, Malaysia. The authors would also like to thank Universiti Teknikal Malaysia Melaka for supporting this research. Professor Jin Wang is the corresponding author.

References

[1] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, "Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks," *IEEE Journal Biomedical Health Informatics*, vol. 19, no. 5, pp. 1627-1636, Sep. 2015. Article (CrossRef Link)

[2] R. G. Dantas, E. T. Costa, and S. Leeman, "Ultrasound speckle and equivalent scatterers," *Ultrasonics*, vol. 43, no. 6, pp. 405-420, May 2005. Article (CrossRef Link)

[3] X. Gao, W. Li, M. Loomes, and L. Wang, "A fused deep learning architecture for viewpoint classification of echocardiography," *Inform Fusion*, vol. 36, pp. 103-113, July 2017. Article (CrossRef Link)

[4] A. M. El Missiri, K. A. L. El Meniawy, S. A. S. Sakr, and A. S. E. D. Mohamed, "Normal reference values of echocardiographic measurements in young Egyptian adults," *Egypt Heart Journal*, vol. 68, no. 4, pp. 209-215, Dec. 2016. Article (CrossRef Link)

[5] A. A. M. Jamel and B. Akay, "A Survey and systematic categorization of parallel K-means and Fuzzy-c-Means algorithms," *Computer Systems Science and Engineering*, vol. 34, no. 5, pp. 259-281, Sep. 2019. Article (CrossRef Link)

[6] L. Aguilar, S. W. Nava-Diaz, and G. Chavira, "Implementation of decision trees as an alternative for the support in the decision-making within an intelligent system in order to automatize the regulation of the VOCs in non-industrial inside environments," *Computer Systems Science Engineering*, vol. 34, no. 5, pp. 297-303, 2019. Article (CrossRef Link)

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, May 2015. Article (CrossRef Link)

[8] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 1251-1258, 2017. Article (CrossRef Link)

[9] X. Hong, X. Zheng, J. Xia, L. Wei, and W. Xue, "Cross-Lingual Non-Ferrous Metals Related News Recognition Method Based on CNN with A Limited Bi-Lingual Dictionary," *Computers, Materials, and Contiuna*, vol. 58, no. 2, pp. 379-389, 2019. Article (CrossRef Link)

[10] S. Y. Tan, "Automated Interpretation of Echocardiograms Technical Milestone Report," to be published. Article (CrossRef Link)

[11] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen, and L. Lovstakken, "Real-time standard view classification in transthoracic echocardiography using convolutional neural networks," *Ultrasound in Medicine and Biology*, vol. 45, no. 2, pp. 374-384, Feb. 2019. Article (CrossRef Link)
[12] J. P. Howard, J. Tan, M. J. Shun-Shin, D. Mahdi, A. N. Nowbar, A. D. Arnold, Y. Ahmad, P. McCartney, M. Zolgharni, N. W. F. Linton, N. Sutaria, B. Rana, J. Mayet, D. Rueckert, G. D. Cole, and D. P. Francis, "Improving ultrasound video classification: an evaluation of novel deep learning methods in echocardiography," *Journal of Medical Artificial Intelligence*, vol. 3, no. 4, Mar. 2020. Article (CrossRef Link)
[13] S. Ebadollahi, S. F. Chang, and H. Wu, "Automatic view recognition in echocardiogram videos using parts-based representation," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 2-9, 2004. Article (CrossRef Link)
[14] G. N. Balaji, T. S. Subashini, and N. Chidambaram, "Automatic classification of cardiac views in echocardiogram using histogram and statistical features," *Procedia Computer Science*, vol. 46, pp.1569-1576, 2015. Article (CrossRef Link)
[15] H. Khamis, G. Zurakhov, V. Azar, A. Raz, Z. Friedman, and D. Adam, "Automatic apical view classification of echocardiograms using a discriminative learning dictionary," *Medical Image Analysis*, vol. 36, pp. 15-21, Feb. 2017. Article (CrossRef Link)
[16] Z. Liu, B. Xiang, Y. Song, H. Lu, and Q. Liu, "An Improved Unsupervised Image Segmentation Method Based on Multi-Objective Particle, Swarm Optimization Clustering Algorithm," *Computers, Materials and Continua*, vol. 58, no. 2, pp. 451-461, 2019. Article (CrossRef Link)
[17] M. Long and Y. Zeng, "Detecting Iris Liveness with Batch Normalized ConVolutional Neural Network," *Computers, Materials and Continua*, vol. 58, no. 2, pp. 493-504, 2019. Article (CrossRef Link)
[18] W. Lu, L. Quan, and L. Ping, "Image Classification using Optimized MKL for SSPM," *Intelligent Automation and Soft Computing*, vol. 25, no. 2, pp. 249-257, 2019. Article (CrossRef Link)
[19] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate view classification of echocardiograms using deep learning," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1-8, Mar. 2018. Article (CrossRef Link)
[20] J. Zhang, S. Gajjala, P. Agrawal, G. H. Tison, L. A. Hallock, L. Beussink-Nelson, M. H. Lassen, E. Fan, M. A. Aras, C. Jordan, K. E. Fleischmann, M. Melisko, A. Qasim, S. J. Shah, R. Bajcsy, and R. C. Deo, "Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy," *Circulation*, vol. 138, no. 16, pp. 1623-1635, Sep. 2018. Article (CrossRef Link)
[21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, pp. 568-576, 2014. Article (CrossRef Link)
[22] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. of Eupopean Conference on Computer Vision*, pp. 25-36, 2004. Article (CrossRef Link)
[23] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: Vgg and residual architectures," *Fronters in Neuroscienc*, vol. 13, Mar. 2019. Article (CrossRef Link)
[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, Aug. 2016. Article (CrossRef Link)
[25] C. Zhang and Y. Tian, "Automatic video description generation via lstm with joint two-stream encoding," in *Proc. of the 23rd International Conference on Pattern Recognition(ICPR)*, pp. 2924-2929, 2016. Article (CrossRef Link)
[26] H. Wu, Q. Liu, and X. Liu, "A review on deep learning approaches to Image classification And object segmentation," *Computers, Maerial and Continua*, vol. 60, no. 2, pp. 575-597, 2019. Article (CrossRef Link)

[27] S. Zhou, L. Chen, and V. Sugumaran, "Hidden two-stream collaborative learning network for action recognition," *Computers, Maerial and Continua*, vol. 63, no. 3, pp. 1545-1561, 2020. Article (CrossRef Link)
[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, May 2017. Article (CrossRef Link)
[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of the 3rd International Conference for Learning Representations*, 2015. Article (CrossRef Link)
[30] Z. Xu, Q. Zhou, and Z. Yan, "Special Section on Recent Advances in Artificial Intelligence for Smart Manufacturing - Part I," *Intellignet Automaiton and Soft Computing*, vol. 25, no. 4, pp. 693-694, 2019. Article (CrossRef Link)
[31] F. Duran and M. Teke, "Design and Implementation of an Intelligent Ultrasonic Cleaning Device," *Intellignet Automaiton and Soft Computing*, vol. 25, no. 3, pp. 441-449, 2019. Article (CrossRef Link)
[32] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, pp. 252-256, Mar. 2020. Article (CrossRef Link)

Zi Ye received bachelor degree in Mathematics & Statistical Science from University College London, UK in 2009, Master's degree in Applied Statistics from University of Oxford, UK in 2010. She is now pursuing her PhD in Universiti Teknikal Malaysia Melaka. Her research interests involve Artificial Intelligence & Machine Learning.

Yogan Jaya Kumar is a Senior Lecture in Universiti Teknikal Malaysia Melaka. He earned his bachelor degree and master degree from Universiti Sains Malaysia. He completed his Ph.D. in 2014 in the field of Computer Science. His research involves in the field of Text Mining, Information Extraction and AI applications.

Goh Ong Sing is an Assistant Vice Chancellor at Office of Industry and Community Network. His main research interest is in the development of intelligent agent, machine learning and speech technology, conversational robot and mobile services. He has led research grants funded by Malaysian Government's Intensified Research.

Fengyan Song received the M.S. degree from Physics Department of Peking University, he is proficient in quantum field theory. He now switches to the field of artificial intelligence.

Xianda Ni is now a Deputy Chief Physician in the Department of Ultrasonography from The First Affiliated Hospital of Wenzhou Medical University, P.R CHINA. He was a visiting scholar in Azienda Ospedaliera "Carlo Poma", Italy in the year of 2007. His main research is three-dimensional echocardiography.

Jin Wang received the M.S. degree from Nanjing University of Posts and Telecommunications, China in 2005. He received Ph.D. degree from Kyung Hee University Korea in 2010. Now, he is a professor at Changsha University of Science and Technology. He has published more than 400 international journal and conference papers. His research interests mainly include wireless sensor network, network performance analysis and optimization etc. He is an IET Fellow, and IEEE Senior member.