IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 November 2024, accepted 1 December 2024, date of publication 4 December 2024, date of current version 12 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3510774

RESEARCH ARTICLE

Time-Distributed Vision Transformer Stacked With Transformer for Heart Failure Detection Based on Echocardiography Video

MGS M. LUTHFI RAMADHAN[1], ADYATMA W. A. NUGRAHA YUDHA[1], MUHAMMAD FEBRIAN RACHMADI[1,2], KEVIN MOSES HANKY JR TANDAYU[3,4], LIES DINA LIASTUTI[3,4], AND WISNU JATMIKO[1], (Senior Member, IEEE)

[1]Faculty of Computer Science, University of Indonesia, Depok City 16424, Indonesia
[2]Brain Image Analysis Unit, RIKEN Center for Brain Science, Wako 351-0106, Japan
[3]Department of Cardiology and Vascular Medicine, Faculty of Medicine, Universitas Indonesia, Depok City 16424, Indonesia
[4]National Cardiovascular Center, Harapan Kita Hospital, Jakarta 11420, Indonesia

Corresponding author: Mgs M. Luthfi Ramadhan (mgs.m01@ui.ac.id)

This work was supported by the Hibah Riset Internal Faculty of Computer Science, Universitas Indonesia, under Grant NKB-014/UN2.F11.D/HKP.05.00/2023.

ABSTRACT Heart failure is a disease many consider to be the number one global cause of death. Despite its mortality, heart failure is still underdiagnosed clinically, especially in a remote area that experiences cardiologists shortage. Existing studies have employed artificial intelligence to help with heart failure screening and diagnosis processes based on echocardiography videos. Specifically, most existing studies use a convolutional neural network that only captures the local context of an image hindering it from learning the global context of an image. Moreover, the frame sampling algorithms only sample certain consecutive frames which makes it questionable whether the dynamic of the left ventricle during a cardiac cycle is included. This study proposed a novel deep learning model consisting of a time-distributed vision transformer stacked with a transformer. The time-distributed vision transformer learns the spatial feature and then feeds the result to the transformer to learn the temporal feature and make the final prediction afterward. We also proposed a frame sampling algorithm by squeezing the video and sampling the frame after a certain interval. Consequently, the video still contains the sequential information up until the end of the video with some in-between frames removed by a certain interval. Thus, the dynamic of the left ventricle is preserved. Our proposed method achieved an F1 score of 95.81%, 96.19%, and 93.43% for the apical four chamber view, apical two chamber view, and parasternal long axis view respectively. The overall trustworthiness of our model is quantified using the NetTrustScore and achieved a score of 0.9712, 0.9767, and 0.9527 for the apical four chamber view, apical two chamber view, and parasternal long axis view respectively.

INDEX TERMS Deep learning, pattern recognition, heart failure, echocardiography, computer vision.

I. INTRODUCTION

Heart failure (HF) is a complication that reduces the heart's ability to pump blood to other organs of the body through the aorta [1]. It is mainly caused by structural or functional impairment of the ventricle that results in an abnormal ejection of blood [2]. The abnormal blood flow to the organs can cause some serious health problems such as kidney damage, liver damage, Pulmonary edema, Pulmonary hypertension, etc.

HF is a high-risk and high-probability disease with a prevalence of about 750,000 new cases each year and its probability to occur keeps increasing as a person ages [3], [4]. According to the research that has been done by [5], the risk of HF has increased to 24% which means that approximately 1 in 4 persons will experience HF in their lifetime. It remains

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Iliyasu.

the number one global cause of death and contributes significantly to worsening other health problems [6], [7], [8]. Research indicates that in 2020 the one-year mortality rate of HF is 30% for outpatients and 20% for inpatients [9].

With that being said, HF is still underdiagnosed clinically while having a high mortality and morbidity which is very terrible given its mortality rate [10]. Thus, diagnosing and treating HF properly is very important to prevent mortality and morbidity [11]. One way to diagnose HF is by using echocardiography which many consider to be the foundational examination for HF with reduced ejection fraction as well as preserved ejection fraction [10], [12], [13].

Echocardiography is the use of ultrasound to examine the heart's structure in a minimally invasive way which means that it doesn't involve cutting the skin open or entering body cavities. Thus, it has little to no risk or side effects [14]. The output of this technique is a medical imaging called an echocardiogram or echo. The personnel to operate echocardiography is mostly required to have at least an associate degree [15]. With that being said, the examination of echo has to be done by experts such as cardiologists and trained cardiovascular technicians. In addition to that, not all public health facility has cardiac ultrasounds due to their limited funding.

In Indonesia specifically, the medical sector faces many problems one of which is the specialist doctor shortage [16]. This problem is highly related to the geographical and demographic conditions of Indonesia. Indonesia is an archipelago consisting of around 17,000 islands and more than 280 million populations [17], [18]. Thus, it is necessary to think about detecting HF that can be operated by the medical workforce in areas that experience cardiologist shortages or accessibility problems. One way to do this is by employing artificial intelligence (AI) technology because it can examine medical imaging accurately with the ability to save time for the diagnosis process [19], [20].

The usage of AI for echocardiography in Indonesia is attempted by [21]. They proposed the Learning Intelligent for Effective Sonography (LIFES), a deep learning-based AI that is trained on echocardiography video of patients in Harapan Kita National Heart Center, Jakarta, Indonesia. Despite its astonishing success, there are several issues with what has been done by [21] and this study aims to fix that. The model selection in [21] is decided based on the highest F1 score. From our point of view, the F1 score is not a reliable metric given the condition of the dataset that has the positive class as the majority class. The F1 score relies heavily on the recall score and it is easy to obtain high recall in such conditions. Though a high recall is considered good for safety reasons, the false positive still comes with a cost such as unnecessary treatment and investigation [22]. Thus, in this study, we employed the Matthews correlation coefficient (MCC) score for our model selection purpose as it can measure the model performance better when the negative class is the minority class [23], [24].

LIFES consists of a convolutional neural network (CNN) namely VGG16 and long short-term memory (LSTM). The VGG16 is only used to extract spatial features and is frozen hindering it from learning the spatial information of echocardiography. It relies heavily on LSTM which only learns the temporal information. The VGG16 uses convolution operation which only takes $n \times n$ neighboring pixel into consideration where n is the size of the kernel. This hinders the VGG16 from understanding the global context of an image [25], [26]. For this reason, we proposed a vision transformer to learn the spatial information of echocardiography. The self-attention mechanism in the vision transformer allows it to learn the global context of an image resulting in better performance [25], [26].

The frame sampling in [21] is decided based on the shortest video in the dataset which turned out to be 41, 30, and 41 for apical four chamber (A4C), apical two chamber (A2C), and parasternal long axis (PLAX) respectively. Then, they sampled the video using only the first n frame of the video which doesn't guarantee a full cardiac cycle. To solve this, the number of frames to use in our study is decided by using the median sequence length of the dataset which turned out to be much longer than what [21] use. In addition to that, we also proposed a better sampling algorithm by squeezing the video and sampling the frame after a certain interval.

Lastly, in accordance with the emergence of a transformer in vision, we proposed a time-distributed vision transformer that is stacked with a transformer. The vision transformer learns the spatial feature and then feeds the result to the transformer to learn the temporal feature and make the final prediction afterward. Drawing on that, our research question is whether or not our proposed method performs better than [21]. To answer this, we compared our proposed algorithm against [21] on the same dataset, metrics, and environment as theirs.

The main contribution of this study lies in the following aspects:

- We introduced a novel architecture for HF detection;
- We introduced frame squeezing algorithms to downsample frame in each video;
- Our proposed method outperformed its comparison.

The rest of this paper is organized as follows. Section II provides related work. Section III gives a brief overview of the dataset used in this study. Section IV details our proposed method. Section V explains and discusses the result of our experiment. Section VI concludes this paper.

II. RELATED WORKS

Multiple studies have been done to diagnose HF using artificial intelligence. Akerman et al. [27] developed a 3D CNN to diagnose HF with preserved ejection fraction based on an A4C view of a transthoracic echocardiography video clip. The data is chunked frame-wise for each 30 frames. Thus, the input for the model is $30 \times 256 \times 256$. Lastly, the final prediction result is averaged throughout all the

chunked frames. They achieved a recall score of 87.80% and a specificity score of 81.90% on the testing set.

Somewhat similar research also has been done by [28]. They ensembled DenseNet Transformer and 3D CNN to diagnose HF based on A4C and PLAX views. Each video is downsampled to 16 frames resulting in a tensor with a shape of $16 \times 64 \times 64 \times 3$ for sequence length, width, height, and channel respectively. They achieved an F1 score of 87.00%.

Naser et al. [29] developed a CNN-based model to detect views from echocardiography video. They compared the performance of 2D CNN and 3D CNN. Each video is downsampled by using only 10 consecutive middle frames of the video. For the 2D CNN experiment, the sampled 10 frames are treated as still images and counted as 10 individual instances. On the other hand, for the 3D CNN experiment, the 10 frames are stacked and treated as a single instance. Their finding is that the 2D CNN achieved an accuracy score of 96.80% while the 3D CNN achieved an accuracy score of 96.30%.

Even though the 3D CNN is very popular in video processing, it does come with limitations. One limitation of 3D CNN is that they are computationally expensive and have excessive memory usage due to the use of a 3D convolution kernel that has relatively more parameters compared to 2D convolution kernel [30], [31], [32], [33]. A straightforward fix to this problem is to simply keep the 2D CNN as the spatial feature extractor and share the weight in the temporal dimension. This can be achieved by wrapping the 2D CNN using a time-distributed layer. However, this approach still requires a sequential model to be stacked on top of the time-distributed layer to learn the temporal features such as recurrent neural network (RNN), LSTM, Transformer, etc.

Several researchers have used a time-distributed layer and some sequential models for video classification. Athira et al. [34] proposed time-distributed VGG16 stacked with LSTM for precise action recognition. They tested their model on the UCF50 [35] dataset and achieved an accuracy score of 98%. This result is significantly better than the Long Term Recurrent Convolutional Neural Network [36] and ConvLSTM [37].

In echocardiography specifically, a similar approach also has been done by [38]. They conducted experiments to classify echocardiography views. They employed a two-stream CNN that consisted of a time-distributed CNN and a temporal CNN and achieved an accuracy score of 96.10%. They also conducted a standalone time-distributed CNN which achieved an accuracy score of 95.30%. These results are relatively better than the 3D CNN which achieved an accuracy score of 89.60%. The same approach has also been done by [21]. They proposed a VGG16 stacked with LSTM for HF detection. The VGG16 is weight-shared in the temporal dimension. Even though it is not directly stated by [21], what they did with the VGG16 is essentially wrapping it using a time-distributed layer. They achieved an F1 score of 92.94%, 94.54%, and 91.01% for A4C, A2C, and PLAX respectively. Liu et al. [39] also proposed a similar approach. They utilize ResNet18 to sample 16 frames out of a video and then feed the sampled frame to a pre-trained ViT L to extract its spatial feature. Lastly, the extracted feature is then fed to BiLSTM stacked with an attention mechanism to detect left ventricular hypertrophy. They achieved an F1 score of 92.15%.

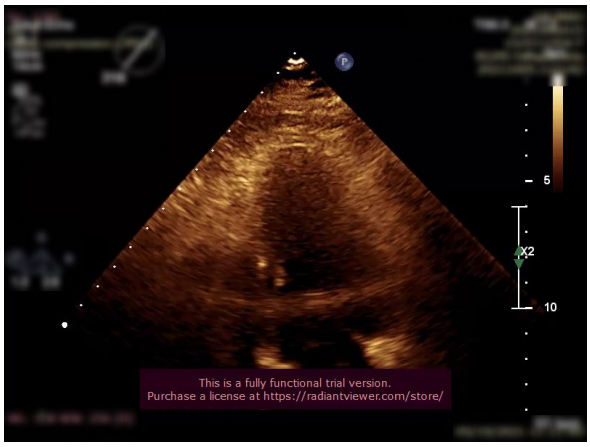


FIGURE 1. Example of non-grayscale video.

III. DATASET

The dataset used in this study is echocardiography videos from the accredited echocardiography laboratory in Harapan Kita National Heart Center, Jakarta, Indonesia. The population in this dataset is the patients who underwent the transthoracic echocardiography test from January 2020 until October 2021. The instance in the dataset that doesn't satisfy the inclusion criteria is then excluded from the study. The inclusion criteria are as follows:

- Patient with clinical shortness of breath;
- Patient doesn't have arrhythmia atrial fibrillation;
- Patient has all the A4C, A2C, and PLAX views;
- Patient has a complete medical record and clinical data.

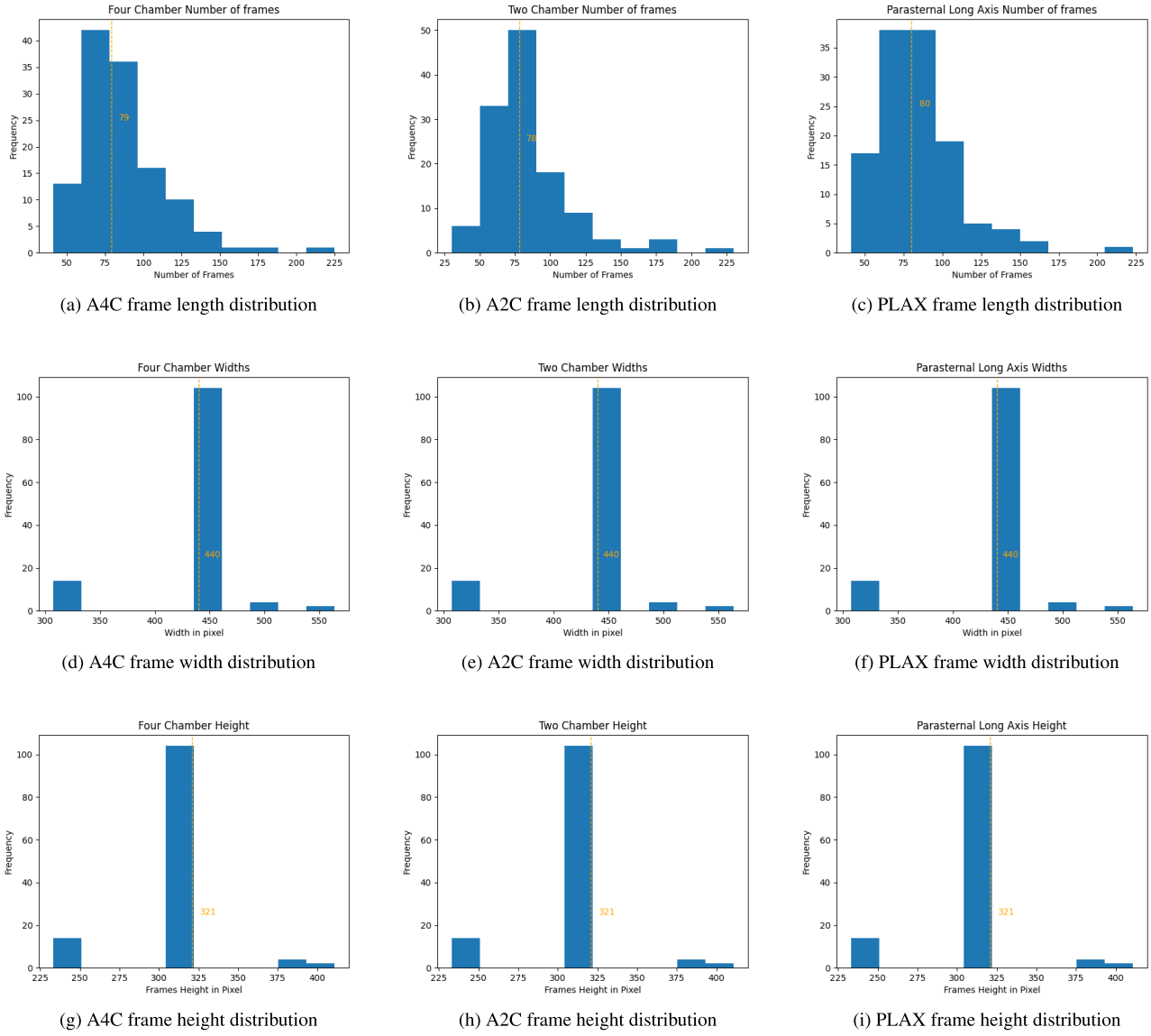
It resulted in 124 instances for all three views. It was then examined by cardiologists and trained cardiovascular technicians using GE Vivid E9 echocardiography machines and M5Sc transducers. Multiple features are extracted to help examine the echocardiography videos such as velocity, left ventricle end-systolic volume, left ventricle end-diastolic volume, ejection fraction, mitral inflow velocity and mitral annular early diastolic velocity ratio, left atrial volume, and left atrial volume index. The dataset was then annotated into two categories namely HF and normal with a class ratio of 63% and 37% for HF and normal respectively.

IV. PROPOSED METHOD

A. EXPLORATORY DATA ANALYSIS

Exploratory data analysis is performed to analyze the condition of our dataset. Our preprocessing method is justified according to the result of exploratory data analysis.

The echocardiography videos in our dataset have already been extracted and compressed using audio video interleave



(a) A4C frame length distribution (b) A2C frame length distribution (c) PLAX frame length distribution

(d) A4C frame width distribution (e) A2C frame width distribution (f) PLAX frame width distribution

(g) A4C frame height distribution (h) A2C frame height distribution (i) PLAX frame height distribution

FIGURE 2. The result of exploratory data analysis.

(AVI) format. Each frame contains textual information such as subject name, hospital name, date time, etc. These videos vary in terms of sequence, width, and height. Most of the videos are in grayscale except for the textual information. Fig. 1 shows an example of a non-grayscale video.

Fig. 2a, 2b, and 2c show the distribution of sequence length in our dataset. The dashed yellow line indicates its median value. All three views have positive skewness with a value of 1.8622, 1.8616, and 1.7663 for A4C, A2C, and PLAX respectively.

Fig. 2d, 2e, and 2f show the distribution of width in pixels, while Fig. 2g, 2h, and 2i show its height. The width and height are centered around 440 and 321 with a very little outlier. Note that, these width and height histograms are the result after we implemented our cropping algorithm which we explained in Section IV-B1.

B. PREPROCESSING

1) FRAME CROPPING

The textual information is irrelevant and not intended to be used in our study. For that reason, each frame is refocused by center cropping it in a certain coordinate. The relative location of the left ventricle in each video is consistent, thus, we constructed a cropping algorithm that applies to all videos in our dataset. The cropping algorithm is provided in algorithm 1.

Note that, the divisors to compute x, y, w, and h are found manually by trial and error. The x, y, w, and h also depend on the original frame width and height, making it adaptable

Algorithm 1 Frame Cropping Algorithm

Input: *frame*
Output: *croppedFrame*
  for each *frame* **do**
    $frameWidth \leftarrow getFrameWidth(frame)$
    $frameHeight \leftarrow getFrameHeight(frame)$
    $x \leftarrow int(\frac{frameWidth}{6.666})$
    $y \leftarrow int(\frac{frameHeight}{5})$
    $w \leftarrow int(\frac{frameWidth}{1.4545})$
    $h \leftarrow int(\frac{frameHeight}{1.5})$
    $croppedFrame \leftarrow frame[y : y + h, x : x + w]$
  end for

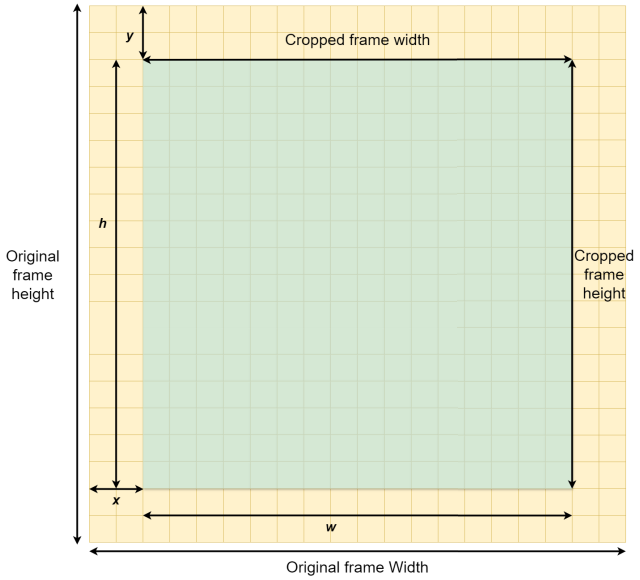


FIGURE 3. Cropping algorithm illustrated.

to various frame heights and widths. Fig. 3 illustrates this cropping algorithm while Fig. 5 shows a real example.

2) FRAME SAMPLING

After refocusing, the video's frame is sampled to a fixed number of frames. Our idea is for the minority to adjust in accordance with its majority. Thus, the length of the sequence to be used is decided using central tendency. It can be inferred from Fig. 2a, 2b, and 2c that the sequence length distributions are positively skewed. Because of this skewness, the median score is used to measure the central tendency as it is not skewed by an extremely long or short sequence.

The median length of the A4C, A2C, and PLAX is 79, 78, and 80 respectively. However, we decided to round it up and use 80 as the length of the sequence to be used. If a video's sequence length is less than 80, the video is pre-padded using a zero frame repetitively until its length becomes 80. On the other hand, if a video's sequence length is more than 80, the video is sampled by squeezing the sequence down to 80. This sampling algorithm is provided in algorithm 2 and illustrated in Fig. 4.

Though the video is sampled in the length dimension, it still contains the sequential information up until the end of the

Algorithm 2 Frame Sampling Algorithm

Input: *video*
Output: *sampledVideo*
Require: *frame2Use*
  $nFrame \leftarrow getFrame(video)$
  if $frame2Use > nFrame$ **then**
    while $frame2Use$!= $nFrame$ **do**
      $zeroFrame \leftarrow createZeroMatrix()$
      Insert *zeroFrame* to *video*
      $nFrame \leftarrow getFrame(video)$
    end while
    $sampledVideo = video$
  else
    $incr \leftarrow \frac{nFrame}{frame2Use}$
    $sampledVideo = []$
    for $i = 0$ to $frame2use - 1$ **do**
      $index \leftarrow int(i \times incr)$
      Append $video[index]$ to *sampledVideo*
    end for
  end if

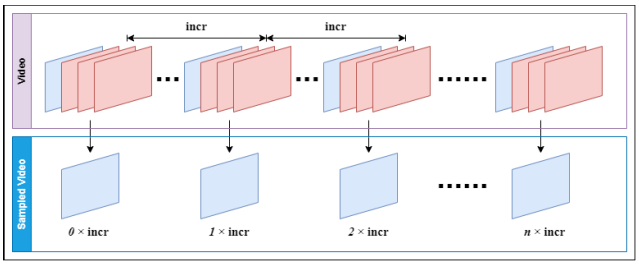


FIGURE 4. Frame sampling algorithm illustrated.

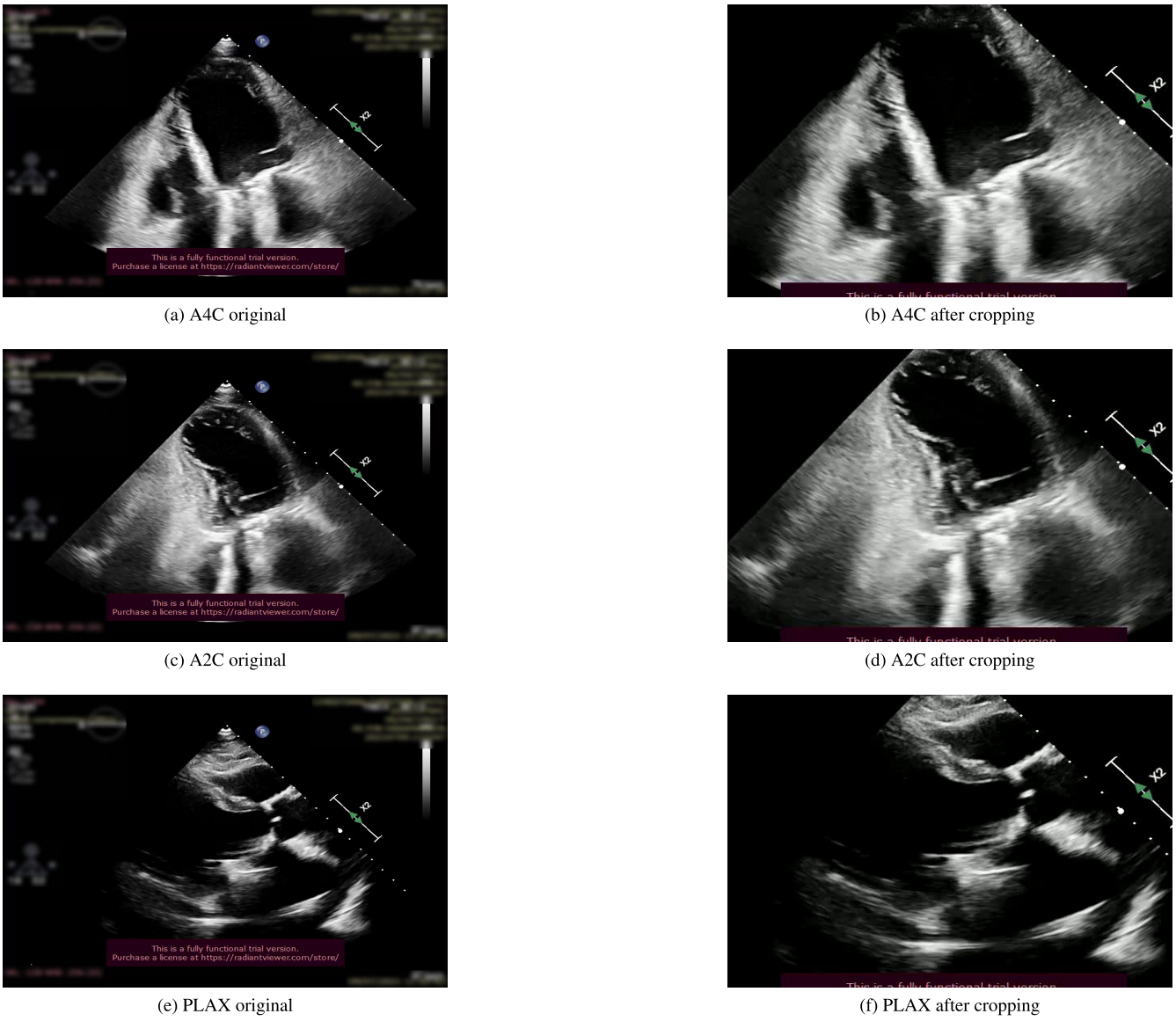
video with some in-between frames removed by a certain interval. Thus, the dynamic of the left ventricle is preserved.

3) RESIZE AND NORMALIZATION

The median width and height are centered around 440 × 321 for all three views. However, due to a limit in our computational resources. We decided to resize each frame to 96×96 so that we don't exhaust our computational resources. Moreover, the ejection fraction is mainly recognized based on the dynamic of the left ventricle [40], [41]. Thus we trade the spatial dimension for a longer sequence.

The red, green, and blue (RGB) channels are also irrelevant and a waste of computational resources. Most of the left ventricles in our video are in the grayscale setting with only a small amount in the RGB setting. This RGB channel has the potential to deceive our algorithm and bring it into the overfitting problem. Because of this reason, we decided to get rid of the RGB channel and convert all the videos in our dataset into grayscale.

Lastly, we normalize our data by dividing it by the maximum value of a pixel which is 255 so that each pixel ranges from $0 - 1$. With the preprocessing stage being done, our final tensor shape is $80 \times 96 \times 96 \times 1$ for sequence length, width, height, and channel respectively.



(a) A4C original (b) A4C after cropping

(c) A2C original (d) A2C after cropping

(e) PLAX original (f) PLAX after cropping

FIGURE 5. Cropping results.

C. CLASSIFICATION METHOD

In this study, we proposed a vision transformer that is wrapped using a time-distributed layer and then stacked a transformer on top of it. Our idea is for the vision transformer to learn the spatial information of each frame and represent it in an encoded sequence of spatial feature vectors. These sequences of spatial feature vectors are then fed into the transformer layer to learn the temporal information of the whole video. This model is end-to-end connected allowing the gradient of the loss function to flow backward throughout the whole network and gives it the capability to learn both the spatial information and temporal information of the video simultaneously. The overall diagram of the time-distributed vision transformer stacked with transformer is illustrated in Fig. 6.

Fig. 7 details the proposed vision transformer architecture. The patch size used in this study is 10×10 which results in 81 patches that contain 100 pixels each. However, for simplicity, the illustration in Fig. 7 is simplified to only 9 patches. The projection dimension used is 128, thus after the positional embedding, the tensor shape that is received by the transformer encoder is 81×128 for the number of patches and projection dimension respectively. The transformer encoder consists of six stacks each having four multi-head attentions. The resulting feature of the transformer encoder is then flattened and fed into the multilayer perceptron (MLP) head which consists of two stacks of MLP each having 1024 and 512 units with rectified linear unit (ReLU) activation function followed by a dropout layer with a dropout rate of 0.1.

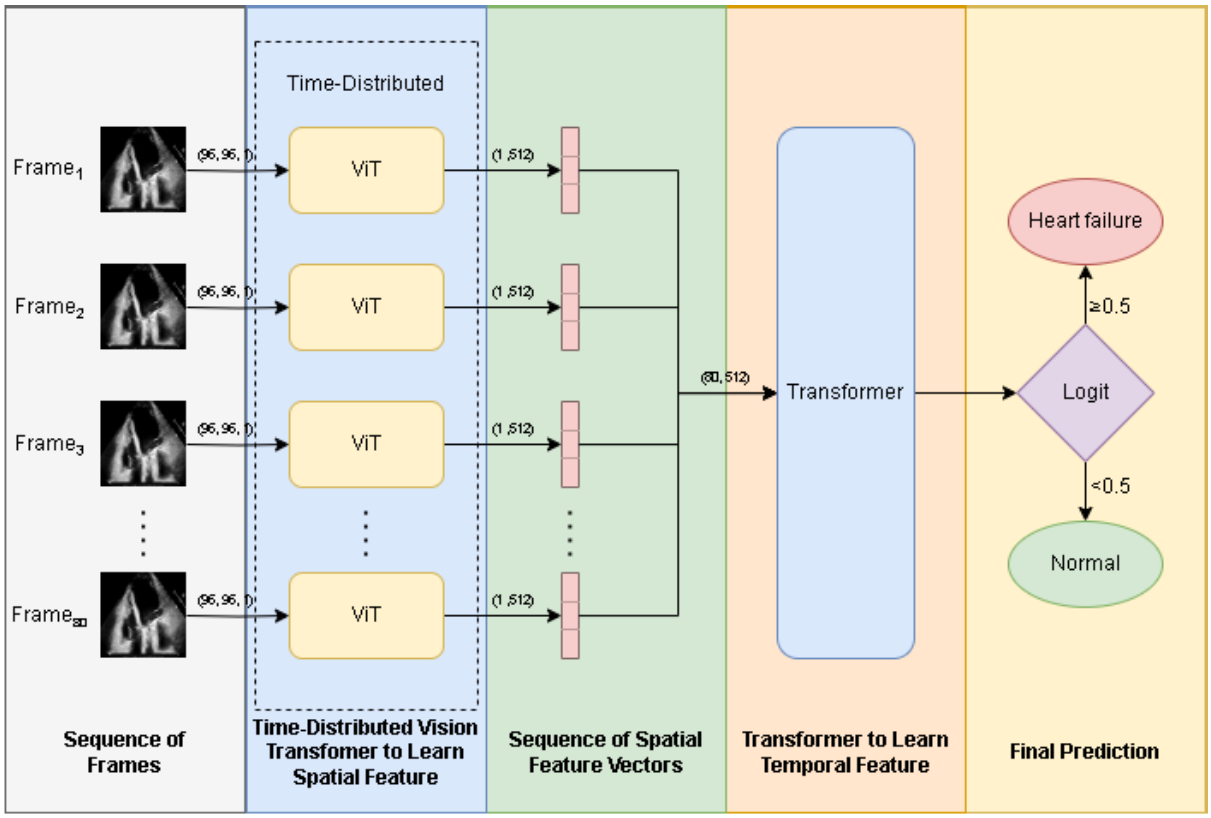


FIGURE 6. Architecture of time-distributed vision transformer-transformer.

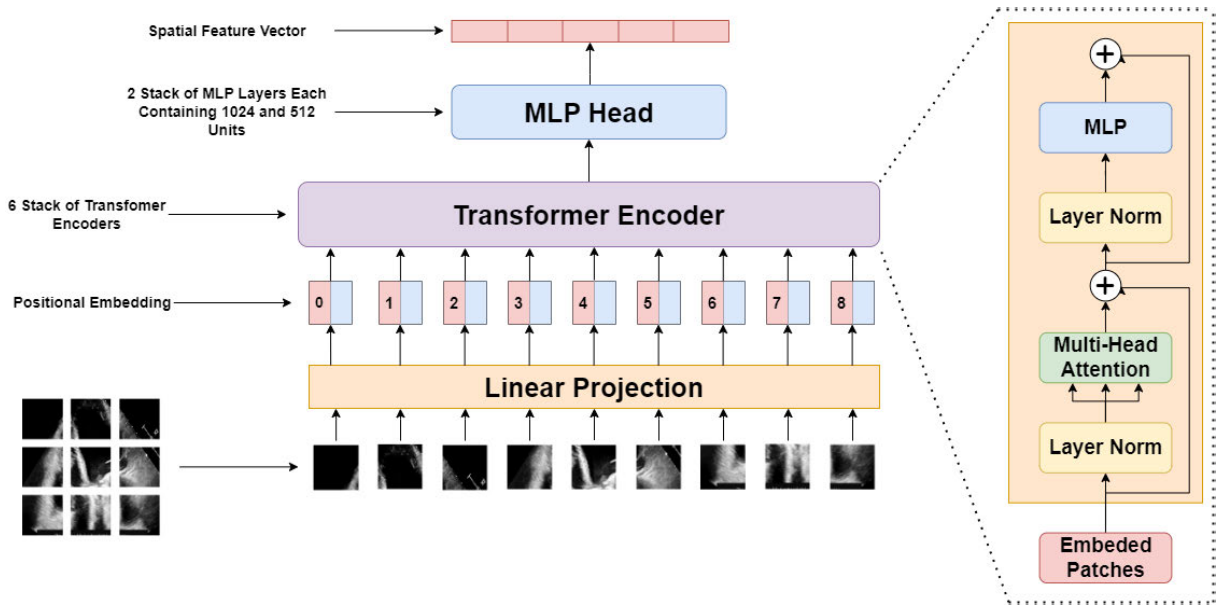


FIGURE 7. Proposed vision transformer architecture.

Note that, this vision transformer is wrapped inside the time-distributed layer which means that this same model is shared in the temporal dimension throughout all the 80 sequences of frames. Thus, the input shape for the transformer is 80×512 for sequence length and feature vector respectively.

Fig. 8 details the proposed transformer architecture. The encoded sequence of spatial feature vectors underwent

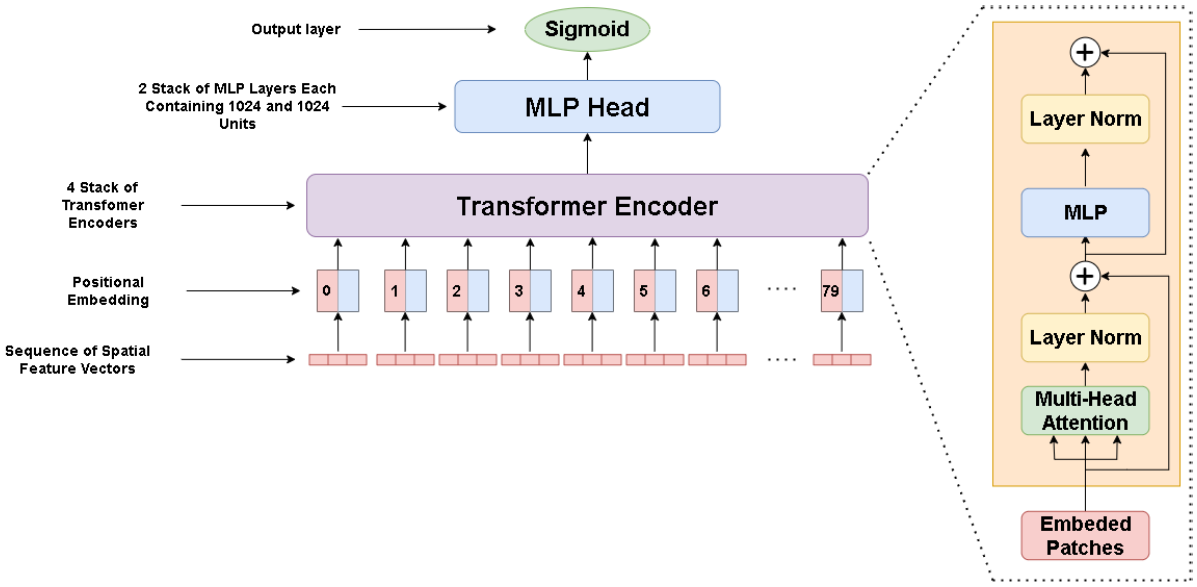


FIGURE 8. Proposed transformer architecture.

positional embedding with an embedding dimension of 512. The transformer encoder consists of four stacks each having four multi-head attentions. The feature vector is then flattened and fed to the MLP head which consists of two stacks of MLP each having 1024 units with ReLU activation function followed by a dropout layer with a dropout rate of 0.1. A single sigmoid unit is stacked on top of it for the output layer of the whole model. Lastly, the model is optimized to minimize the binary cross-entropy loss.

D. EVALUATION METRICS

We employed a confusion matrix to quantify our classification algorithm performance. Confusion matrix describes the performance of a classification algorithm by comparing its prediction result against the actual truth in the form of a matrix or table [42], [43]. This matrix consists of four elements known as True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). TP is the frequency of HF successfully predicted. FP is the frequency of normal predicted as HF. FN is the frequency of HF predicted as normal. TN is the frequency of normal predicted as normal. From these four elements, several metrics can be derived such as accuracy, F1 score, precision, recall, specificity, and Matthew correlation coefficient.

1) ACCURACY

The accuracy score is the ratio of the model's true predictions compared to all of its prediction attempts. Accuracy score can be quantified using (1).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

2) PRECISION

The precision score also known as the positive predictive value, is the ratio of the model's true positive predictions compared to all of the positive predictions attempted. Precision score can be quantified using (2).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

3) RECALL

The recall score also known as the sensitivity, is the ratio of the model's true positive predictions compared to all the positive ground truth. Recall can be quantified using (3).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

4) F1 SCORE

The F1 score sometimes called F-measure or F-beta measure is the harmonic mean of precision and recall to obtain the balance of both metrics. F1 score can be quantified using (4).

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

5) SPECIFICITY

The specificity score is the opposite of recall, it is the ratio of the model's true negative predictions compared to all the negative ground truth. Specificity can be quantified using (5).

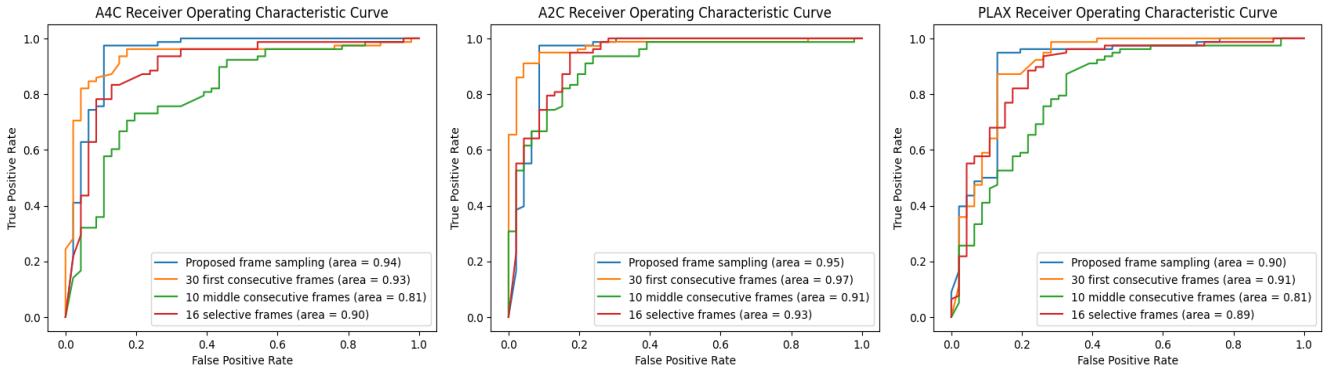
$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{5}$$

6) MATTHEW CORRELATION COEFFICIENT (MCC)

MCC is an evaluation metric that remains unaffected by imbalanced datasets. It uses all four elements of the confusion matrix so it gives a better summary of the whole classification problem. MCC score ranges from -1 to 1, with 1 as the best MCC score possible indicating a perfect alignment between prediction and ground truth while -1 as the worst MCC score possible indicating each prediction is the opposition

TABLE 1. Frame sampling comparison.

View	Frame sampling	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Norm MCC (%)	p-value
A4C	Proposed	**94.29**	94.75	**97.50**	**95.81**	88.33	**93.96**	-
	30 first consecutive [21]	89.51	**96.15**	88.23	92.02	**92.30**	86.03	0.0412
	10 middle consecutive [29]	87.09	94.87	86.04	90.24	89.47	88.70	0.0001
	16 selective [39]	86.29	93.58	85.88	89.57	87.17	85.11	0.1340
A2C	Proposed	**95.12**	**95.75**	**97.50**	**96.19**	**90.83**	**95.15**	-
	30 first consecutive [21]	92.74	93.58	94.80	94.19	89.36	92.26	0.2482
	10 middle consecutive [29]	87.09	93.58	86.90	90.12	87.50	86.00	0.0044
	16 selective [39]	90.32	94.87	90.24	92.49	90.47	89.54	0.4800
PLAX	Proposed	**91.90**	93.50	**95.00**	**93.43**	87.50	**92.22**	-
	30 first consecutive [21]	88.70	**98.71**	85.55	91.66	**97.05**	88.14	0.3428
	10 middle consecutive [29]	79.83	93.58	78.49	85.38	83.87	77.95	0.0003
	16 selective [39]	84.67	94.87	83.14	88.62	88.57	83.41	0.0080



(a) A4C (b) A2C (c) PLAX

FIGURE 9. ROC curves of various frame sampling algorithms.

of the ground truth. An MCC score of 0 indicates a random prediction. MCC score can be quantified using (6).

$$\text{MCC} = \frac{\text{TP}\times\text{FP} - \text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})\times(\text{TP}+\text{FN})\times(\text{TN}+\text{FP})\times(\text{TN}+\text{FN})}} \tag{6}$$

Based on how it is formulated, MCC will only produce a high score if a classification model can correctly predict most of the positive and most of the negative data. Even though the MCC ranges from -1 to 1, it can be scaled into a 0 to 1 range so that it can have the same value ranges as other metrics. This scaled MCC is called normalized MCC and can be quantified using (7).

$$\text{normMCC} = \frac{\text{MCC} + 1}{2} \tag{7}$$

E. IMPLEMENTATION DETAILS

The experiment in this study is done using Python programming language with the help of TensorFlow, Numpy, and Scikit-learn library. To ensure reproducibility and trackability, the random seeds for Tensorflow, Numpy, and Scikit-learn are set with a value of one. It is also done to give each trial the same random behavior so that any different result in each trial does not happen by coincidence or luck of random behavior, making our comparison design fair and square. The statistical significance is measured using McNemar's significance testing since it is suitable for categorical data [44], [45], [46].

For the sake of comprehensive evaluation, we employed the 20-fold cross-validation which resulted in a 95% training set and 5% testing set for each fold. The training set is randomly split by a percentage of 10% to obtain the validation set. The validation set is used to monitor the training process. These splitting processes are stratified, thus the percentage of samples for each class is preserved in every fold.

To help our model convergence, the learning rate is reduced by a factor of 0.5 if it is on a plateau after five epochs. Early stopping with a patience of 15 is employed to save our model from overfitting. The early stopping monitors validation loss and stops it if it doesn't improve after the patience is exhausted.

The hyperparameter is chosen by grid-searching the following hyperparameter pool:

- Optimizer: AdamW, Adam, RMSprop, and Lion;
- Batch size: 2, 4, 8, and 16;
- Learning rate: 0.001, 0.0001, and 0.00001.

In this dataset, the F1 score is not a reliable metric for model selection since it doesn't take the true negative into consideration while the negative class is the minority. The challenge in this dataset is to obtain more true negatives.

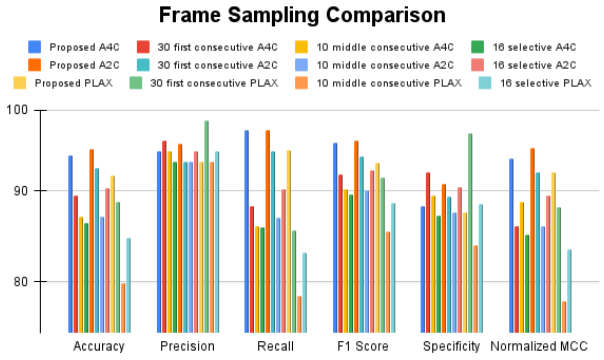


FIGURE 10. Evaluation metrics of various frame sampling algorithms.

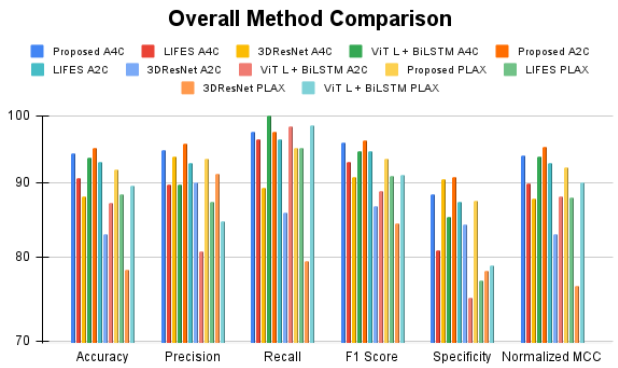


FIGURE 11. Evaluation metrics of various methods.

Thus, the MCC is chosen because it takes the true negative into consideration. The result of each fold is concluded by choosing the combination of hyperparameters that yield the highest normalized MCC score.

The experiment in this study is conducted with the following hardware specifications:

- RAM: 32 GB;
- GPU: Two nodes of NVIDIA GeForce GTX 1080 Ti;
- Processor: Intel (R) Core (TM) i7-6800K CPU @ 3.40GHz.

V. RESULT AND DISCUSSION

A. FRAME SAMPLING COMPARISON

In this experiment, we explore how our proposed frame sampling algorithm performs compared to existing studies. The control variable in this experiment is our proposed deep learning model which is trained on various frame sampling algorithms. Consequently, each frame sampling has the same model trained on it which makes the comparison fair and square. The result of this experiment is provided in Table 1 and Fig. 10. The receiver operating characteristic (ROC) curves along with its area under the curves (AUC) are shown in Fig. 9. These ROC curves are constructed throughout all folds and quantified by computing their area under the curves provided in its legend. Our proposed algorithm outperforms existing studies in most metrics with a few exceptions in terms of precision score, specificity score, and AUC. Our precision and specificity are second after [21] on A4C and PLAX views. However, we achieved the highest precision and specificity score on the A2C view. In addition to that, we also achieved the highest AUC on the A4C view. The p-value column in Table 1 is the result of McNemar's significance testing. Each row in the p-value column represents McNemar's test relative to our proposed frame sampling.

It is important to note that our proposed algorithm sampled 80 non-consecutive frames, [21] sampled 30 first consecutive frames, [29] sampled 10 middle consecutive frames, and [39] sampled 16 selective frames. Even though our proposed algorithm lacks precision and specificity scores, there seemed to be a linear association between the number of frames sampled and the overall classifier performance. The more frames sampled, the better the classifier performance. To quantify this, we computed the Pearson correlation between the number of frames sampled and the F1 score. The obtained Pearson correlations are 0.9786, 0.8967, and 0.8500 for A4C, A2C, and PLAX respectively. However, this obviously comes with a computational cost. Even though the time-distributed vision transformer is sequence invariant due to its sharing weight, the longer sequence still results in increased parameters for the later transformer model. Our proposed sampling algorithm results in 23.571 M parameters while [21], [29], and [39] result in 23.545 M, 23.535 M, and 23.538 M respectively.

B. OVERALL METHOD COMPARISON

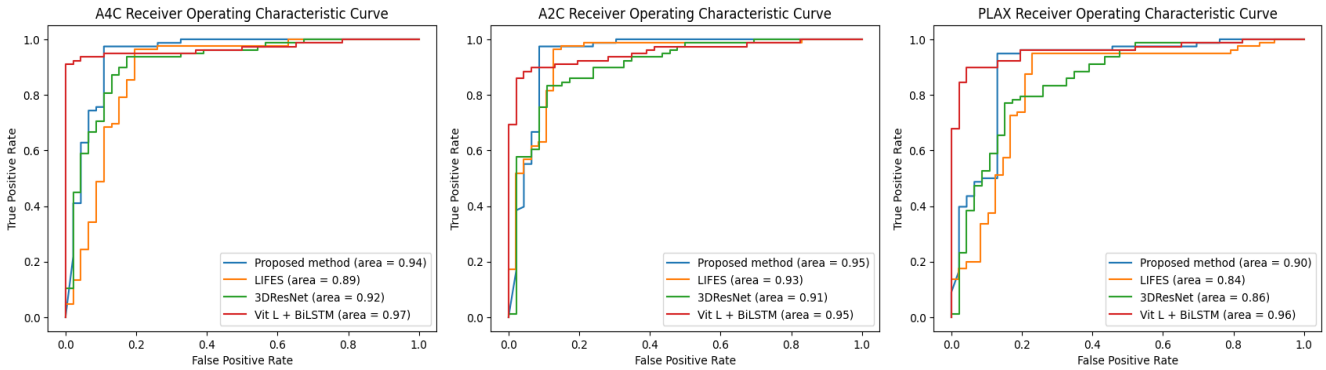
In this experiment, we explore how our overall proposed method performs compared to existing studies. The result of this experiment is provided in Table 2 and Fig. 11. It can be inferred that our proposed method outperforms [21], [29], and [39] in most metrics with recall and AUC as exceptions. In line with [21] findings, the A2C view remains to be the view with the highest F1 score. We achieved an F1 score of 95.81%, 96.19%, and 93.43% for A4C, A2C, and PLAX respectively. We also obtained better specificity scores in A2C and PLAX views than [21], [29], and [39]. However, in the A4C view, our specificity score ranks number 2 after [29]. Fig. 12 shows the ROC curves of our experiment. Our model achieved an AUC score of 94%, 95%, and 90% for A4C, A2C, and PLAX respectively. These results rank number 2 after [39] but are relatively better than [21] and [29]. In the A2C view, our proposed method tied its AUC with [39]. A possible explanation for this is that we employed a normalized MCC score for our model selection during the grid search which results in better generalization on negative class, thus our specificity score is relatively better than [39]. However, as a trade-off, our proposed method has a lower recall score compared to [39].

C. GRID SEARCH ANALYSIS

The tuned hyperparameters are provided in Table 3. These tuned hyperparameters are obtained by grid searching several

TABLE 2. Overall method comparison.

View	Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Norm MCC (%)	p-value
A4C	Proposed	**94.29**	**94.75**	97.50	**95.81**	88.33	**93.96**	-
	LIFES [21]	90.62	89.77	96.34	92.94	80.84	89.86	0.0736
	3DResNet [29]	87.97	93.75	89.25	90.79	**90.41**	87.70	0.0133
	ViT L + BiLSTM [39]	93.54	89.74	**100**	94.59	85.18	93.71	0.2280
A2C	Proposed	**95.12**	**95.75**	97.50	**96.19**	**90.83**	**95.15**	-
	LIFES [21]	92.96	92.85	96.29	94.54	87.23	92.77	0.2482
	3DResNet [29]	82.97	90.00	85.83	86.69	84.16	82.89	0.0003
	ViT L + BiLSTM [39]	87.09	80.76	**98.43**	88.73	75.00	87.98	0.0060
PLAX	Proposed	**91.90**	**93.50**	95.00	**93.43**	**87.50**	**92.22**	-
	LIFES [21]	88.28	87.35	95.00	91.01	77.03	87.83	0.0736
	3DResNet [29]	78.45	91.25	79.50	84.32	78.33	76.47	0.0001
	ViT L + BiLSTM [39]	89.51	84.61	**98.50**	91.03	78.94	89.95	0.0270



(a) A4C (b) A2C (c) PLAX

FIGURE 12. ROC curves of various methods.

hyperparameters previously explained in Section IV-E. The grid search occurs on each fold and we choose the combination of hyperparameters that yields the highest normalized MCC score on each fold.

Fig. 13 shows the frequency of batch size yielding the highest normalized MCC. The y-axis indicates the frequency of a specific batch size yielding the highest mcc score while the x-axis indicates the batch size value. It can be seen that a batch size of 8 appears more frequently than others. The frequency increases as the batch size value increases and drops as it reaches a batch size of 16. However, the trendline still indicates an uptrend with a slope of 1.4 and an intercept of 12.9 which means that the frequency tends to increase as the batch size value increases.

Fig. 14 shows the frequency of learning rate yielding the highest normalized MCC. The graph shows a downtrend which means that the lower the learning rate, the higher its frequency. The trendline almost aligned perfectly with the graph and has a slope of -15 with an intercept of 35.

Fig. 15 shows the frequency of the optimizer yielding the highest normalized MCC. The graph shows no meaningful pattern since it is almost uniformly distributed across all optimizers with AdamW being the most frequent and RMSprop being the least frequent.

D. ROBUSTNESS AND SCALABILITY ANALYSIS

We analyze the robustness of our proposed method by predicting retrospective data from four other hospitals. This data acts as external validation and is collected from Pasar Minggu Hospital, Pasar Rebo Hospital, Cipto Mangunkusumo Hospital, and Universitas Indonesia Hospital. Each of these hospitals has different echocardiography machines which results in different imaging quality. Hospitals and their echocardiography machines are as follows:

- Pasar Minggu Hospital: Philips IE 33;
- Pasar Rebo Hospital: GE Vivid S70N;
- Cipto Mangunkusumo Hospital: Philips Epiq 7;
- Universitas Indonesia Hospital: Philips Epiq 7.

This data consists of 172 instances with the same inclusion criteria as explained in Section III.

The prediction is made by averaging the sigmoid output of all 20 folds of our proposed method. Then, a threshold of 0.5 is applied to obtain the final prediction. Instances with a sigmoid output that is greater than 0.5 are classified as heart failure, otherwise, they are classified as normal.

Table 4 and Fig. 16 provide the result of our external validation. It can be seen that the PLAX view outperforms A4C and A2C views in most metrics especially accuracy,

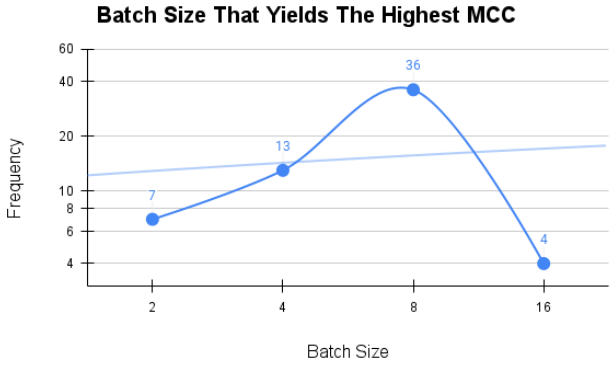


FIGURE 13. Batch size that yields the highest MCC.

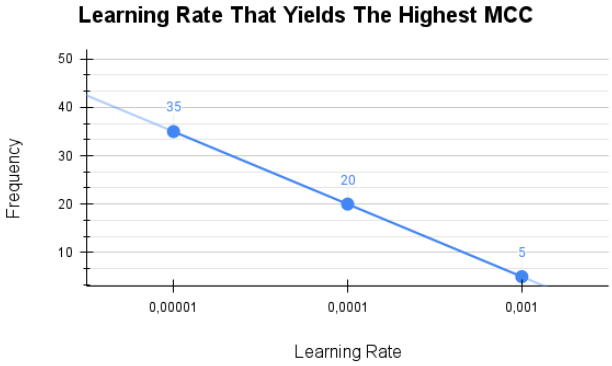


FIGURE 14. Learning rate that yields the highest MCC.

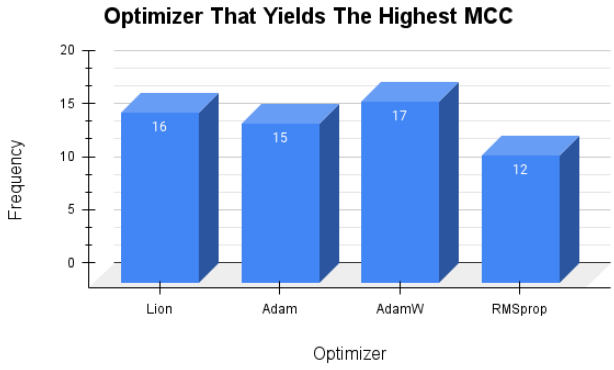


FIGURE 15. Optimizer that yields the highest MCC.

recall, F1 score, and normalized MCC. The A4C view achieved the highest specificity score, however, it lacks a recall score which indicates it misses a lot of true positives. Although the A2C view is generally better than other views in the internal validation, it doesn't generalize well on external validation data. Fig. 17 shows the ROC curves of our external validation. These ROC curves are constructed based on the averaged sigmoid output of all 20 folds of our proposed method. The PLAX view achieved the highest AUC score while A4C and A2C are tied with an AUC score of 0.58. Overall, the PLAX view has better robustness capability than the other view.

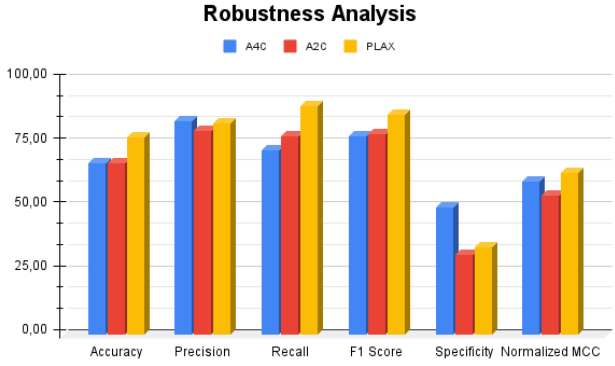


FIGURE 16. Evaluation metrics of external validation.

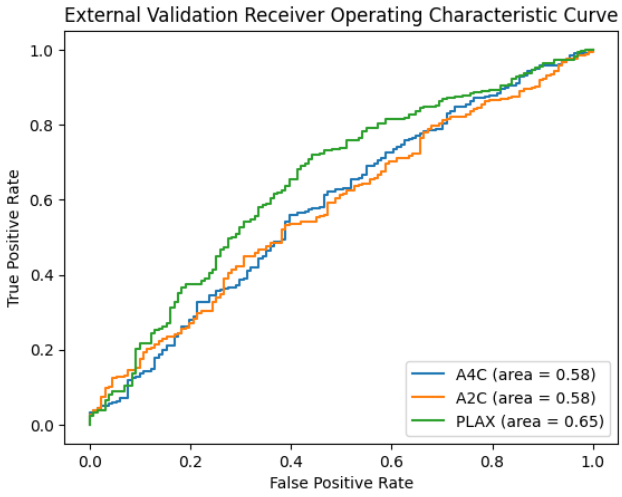


FIGURE 17. ROC curves of external validation.

We analyze the scalability of our proposed method by measuring the prediction's latency and throughput. It should be noted that the proposed architecture and input tensor are view-invariant, thus we only measure the latency and throughput on a single view. The latency is quantified by measuring the time taken by the model to finish a single prediction in seconds. The throughput is quantified by dividing the amount of data and the time taken by the model to predict the amount of data. The unit of measurement of throughput is prediction per second. However, the operating system's activity and other processes might intervene in the measurement. For this reason, we measure the latency and throughput 100 times and take the average. The average latency of our proposed method is 0.3767 seconds with an average throughput of 2.6865 predictions per second.

E. TRUST ANALYSIS

We map the resulting sigmoid probability using a 1D scatter plot as shown in Fig. 18 where blue instances are normal while orange instances are HF. The blue vertical line lies on the 0.5 of the x-axis, it is the threshold used to transform the sigmoid probability into the final prediction. The instances that lie on the right side of the blue vertical line are the

TABLE 3. Tuned hyperparameters.

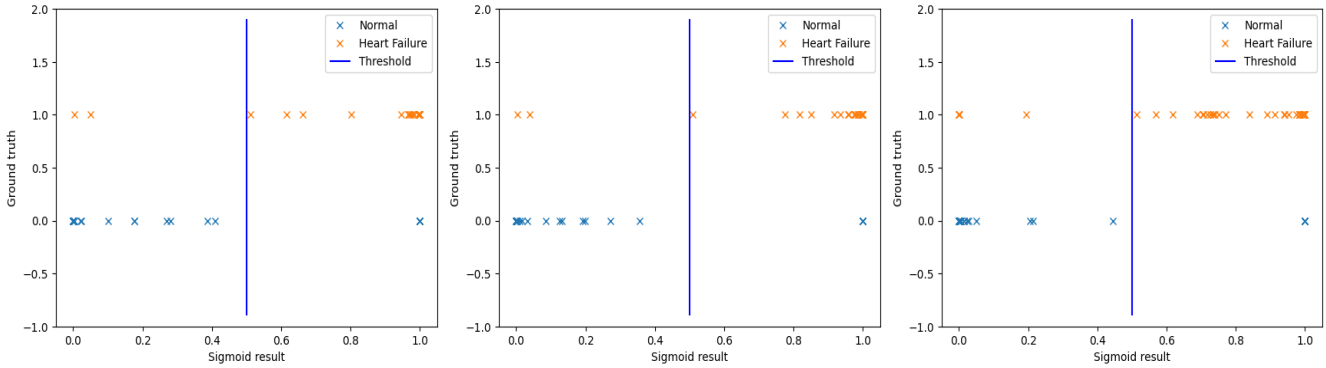
View	Fold	Hyperparameters			Metrics					
		Batch	Lr	Optimizer	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Norm MCC (%)
A4C	1	8	0.00001	Lion	100	100	100	100	100	100
	2	8	0.0001	Adam	100	100	100	100	100	100
	3	8	0.00001	AdamW	100	100	100	100	100	100
	4	8	0.00001	AdamW	85.71	80.00	100	88.89	66.67	86.51
	5	2	0.001	Adam	100	100	100	100	100	100
	6	8	0.00001	Lion	100	100	100	100	100	100
	7	4	0.0001	AdamW	100	100	100	100	100	100
	8	8	0.0001	AdamW	100	100	100	100	100	100
	9	4	0.00001	RMSprop	100	100	100	100	100	100
	10	4	0.0001	Lion	100	100	100	100	100	100
	11	2	0.00001	RMSprop	83.33	80	100	88.89	50.00	81.62
	12	4	0.00001	Adam	100	100	100	100	100	100
	13	2	0.0001	RMSprop	100	100	100	100	100	100
	14	8	0.001	AdamW	83.33	100	75.00	85.71	100	85.36
	15	2	0.0001	Adam	100	100	100	100	100	100
	16	2	0.00001	Lion	83.33	80.00	100	88.89	50.00	81.62
	17	8	0.00001	Lion	100	100	100	100	100	100
	18	8	0.00001	Adam	66.67	75.00	75.00	75.00	50.00	62.50
	19	8	0.00001	Lion	100	100	100	100	100	100
	20	8	0.00001	RMSprop	83.33	80.00	100	88.89	50.00	81.62
A2C	1	16	0.0001	Adam	100	100	100	100	100	100
	2	4	0.00001	AdamW	85.71	100	75.00	85.71	100	87.50
	3	8	0.00001	AdamW	100	100	100	100	100	100
	4	4	0.00001	Lion	100	100	100	100	100	100
	5	8	0.00001	AdamW	83.33	75.00	100	85.71	66.67	85.36
	6	8	0.0001	Adam	100	100	100	100	100	100
	7	8	0.00001	AdamW	100	100	100	100	100	100
	8	4	0.0001	AdamW	83.33	100	75.00	85.71	100	85.36
	9	8	0.00001	AdamW	100	100	100	100	100	100
	10	8	0.0001	AdamW	100	100	100	100	100	100
	11	4	0.0001	AdamW	100	100	100	100	100	100
	12	8	0.0001	AdamW	100	100	100	100	100	100
	13	8	0.00001	RMSprop	100	100	100	100	100	100
	14	8	0.00001	Adam	83.33	80.00	100	88.89	50.00	81.62
	15	2	0.0001	Lion	83.33	80.00	100	88.89	50.00	81.62
	16	4	0.0001	Lion	83.33	80.00	100	88.89	50.00	81.62
	17	8	0.00001	Lion	100	100	100	100	100	100
	18	16	0.0001	Lion	100	100	100	100	100	100
	19	8	0.00001	Lion	100	100	100	100	100	100
	20	16	0.0001	Lion	100	100	100	100	100	100
PLAX	1	4	0.00001	Lion	100	100	100	100	100	100
	2	8	0.00001	RMSprop	85.71	80.00	100	88.89	66.67	86.51
	3	8	0.00001	RMSprop	100	100	100	100	100	100
	4	8	0.00001	RMSprop	85.71	80.00	100	88.89	66.67	86.51
	5	8	0.0001	RMSprop	83.33	75.00	100	85.71	66.67	85.36
	6	8	0.001	RMSprop	100	100	100	100	100	100
	7	8	0.00001	Adam	100	100	100	100	100	100
	8	2	0.00001	Lion	100	100	100	100	100	100
	9	8	0.00001	Adam	100	100	100	100	100	100
	10	8	0.001	Adam	100	100	100	100	100	100
	11	4	0.0001	Adam	100	100	100	100	100	100
	12	8	0.00001	Adam	100	100	100	100	100	100
	13	8	0.00001	Adam	83.33	80.00	100	88.89	50.00	81.62
	14	8	0.00001	Lion	100	100	100	100	100	100
	15	8	0.0001	AdamW	66.67	75.00	75.00	75.00	50.00	62.50
	16	8	0.00001	RMSprop	66.67	100	50.00	66.67	100	75.00
	17	4	0.00001	AdamW	100	100	100	100	100	100
	18	16	0.0001	RMSprop	100	100	100	100	100	100
	19	4	0.001	AdamW	83.33	100	75.00	85.71	100	85.36
	20	8	0.00001	Adam	83.33	80.00	100	88.89	50.00	81.62

instances that are predicted as HF while the instances that lie on the left side of the blue vertical line are the instances that are predicted as normal. The more the orange instances to the right corner the better the model performance. On the other hand, the more the blue instances to the left corner the better the model performance.

It can be seen that the false positives have a close to 1 probability and at the same time, there are many true

TABLE 4. The result of external validation.

View	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Norm MCC (%)
A4C	67.44	**83.62**	72.38	77.59	**50.00**	59.91
A2C	67.44	80.00	77.61	78.78	31.57	54.43
PLAX	**77.32**	82.75	**89.55**	**86.02**	34.21	**63.55**



(a) A4C (b) A2C (c) PLAX

FIGURE 18. Sigmoid prediction probability.

TABLE 5. NetTrustScore of our proposed method.

View	NetTrustScore
A4C	0.9712
A2C	**0.9767**
PLAX	0.9527

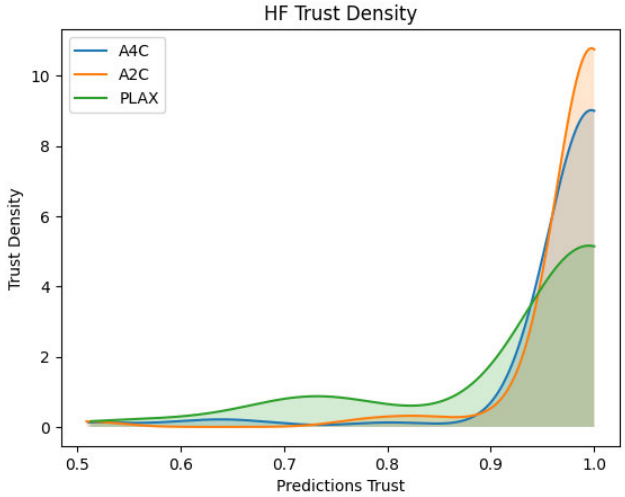


FIGURE 19. HF trust density.

negatives that are not close to 0. A possible explanation for this is that the majority class in this dataset is positive which makes the model biased toward the positive class more often. We quantify Fig. 18 by computing the binary cross-entropy loss and obtain a loss of 0.5270, 0.4812, and 0.7396 for A4C, A2C, and PLAX respectively.

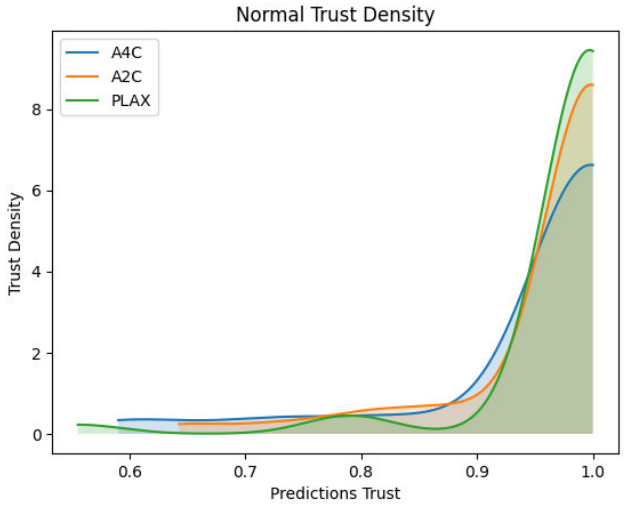


FIGURE 20. Normal trust density.

To elaborate on that, we computed the trust metrics proposed by [47]. However, [47] tested the trust metric on the softmax activation function with 1 said to be a confidence prediction. Thus, in this study, the normal trust density is computed by inverting the sigmoid probability such that 1 is said to be a confidence prediction for the normal class.

Fig. 19 shows the trust density for HF while Fig. 20 shows the trust density for normal. The more the trust density concentrated to the right, the better the trust of the model. It can be inferred that our model with the A2C view predicts HF more trustworthy than other views, while our model with the PLAX view predicts normal more trustworthy than other views. The trust spectrum which is essentially the quantification of Fig. 19 and Fig. 20 is visualized using a bar

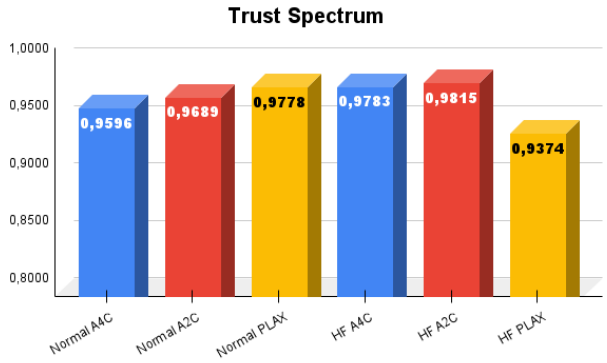


FIGURE 21. Trust spectrum.

chart as can be seen in Fig. 21. Lastly, the NetTrustScore of our proposed model is provided in Table 5.

VI. CONCLUSION

In this study, we proposed a novel deep learning model to detect heart failure based on an echocardiography video. The problem we discussed in this study is the low specificity score of other existing studies, the CNN-based spatial feature extractor that doesn't capture the global context of an image, and the frame sampling that doesn't guarantee a full cardiac cycle. We fixed this problem by employing the normalized MCC score for our model selection, we proposed a time-distributed vision transformer stacked with a transformer, and we proposed a frame sampling algorithm.

The experimental result shows that our proposed method has relatively better metrics than its comparison, especially the specificity score. We achieved an F1 score of 95.81%, 96.19%, and 93.43% for A4C, A2C, and PLAX respectively. We also quantified the overall trustworthiness of our proposed method and achieved a NetTrustScore of 0.9712, 0.9767, and 0.9527 for A4C, A2C, and PLAX respectively. These findings answer our research question.

Even though our proposed method has a good classification performance, the dataset used in this study is obtained from a single hospital. Future studies may extend this work by training this model on a dataset that is acquired from various hospitals. The implementation of data augmentation or class weighting algorithm is also a possible choice to solve the imbalance class problem. A more computationally cheaper model such as Mamba [48] which runs 5× faster than the transformer is also a possible extension of our study. The implementation of this model on an online system so that it can be used by the medical workforce in remote areas is also in our interest. Lastly, we also consider improving our sampling algorithm by quantifying information in each frame and only retaining the frame with a high information score.

As of now, we hope this study paves the way for future research and development of AI in echocardiography. It is important to note that the existence of AI in echocardiography is by no means to replace the cardiologist, instead, it is to help reach the remote area especially the area with a cardiologist shortage to perform early diagnosis or screening of HF. The diagnosis results of certified and experienced cardiologists should always overshadow AI diagnosis results.

ACKNOWLEDGMENT

The authors thank all the colleges in the Intelligent Robots and Systems (IRoS) Laboratory and the Tokopedia-UI AI Center for providing them with infrastructure.

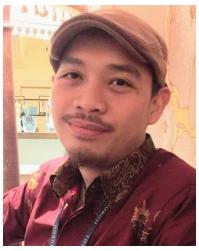
REFERENCES

[1] B. Shahim, C. J. Kapelios, G. Savarese, and L. H. Lund, "Global public health burden of heart failure: An updated review," *Cardiac Failure Rev.*, vol. 9, p. e11, Jul. 2023.
[2] R. S. Velagaleti and R. S. Vasan, "Heart failure in the twenty-first century: Is it a coronary artery disease or hypertension problem?" *Cardiol. Clinics*, vol. 25, no. 4, pp. 487–495, Nov. 2007.
[3] M. S. Khan, I. Shahid, A. Bennis, A. Rakisheva, M. Metra, and J. Butler, "Global epidemiology of heart failure," *Nature Rev. Cardiol.*, vol. 21, no. 10, pp. 717–734, Jun. 2024, doi: 10.1038/s41569-024-01046-6.
[4] J. Hung, K. Shahzad, R. Beeri, and R. A. Levine, "Ventricular remodeling and secondary valvular dysfunction in heart failure progression," in *Heart Failure*. Boca Raton, FL, USA: CRC Press, 2004, pp. 115–134.
[5] B. Bozkurt et al., "Heart failure epidemiology and outcomes statistics: A report of the heart failure Society of America," *J. Cardiac Failure*, vol. 29, no. 10, pp. 1412–1451, Sep. 2023.
[6] M. Vaduganathan, G. A. Mensah, J. V. Turco, V. Fuster, and G. A. Roth, "The global burden of cardiovascular diseases and risk," *J. Amer. College Cardiol.*, vol. 80, no. 25, pp. 2361–2371, Dec. 2022, doi: 10.1016/j.jacc.2022.11.005.
[7] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[8] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[9] C. J. Lee, H. Lee, M. Yoon, K.-H. Chun, M. G. Kong, M.-H. Jung, I.-C. Kim, J. Y. Cho, J. Kang, J. J. Park, H. C. Kim, D.-J. Choi, J. Lee, and S.-M. Kang, "Heart failure statistics 2024 update: A report from the Korean society of heart failure," *Int. J. Heart Failure*, vol. 6, no. 2, pp. 56–69, 2024.
[10] M. M. Oo, C. Gao, C. Cole, Y. Hummel, M. Guignard-Duff, E. Jefferson, J. Hare, A. A. Voors, R. A. de Boer, C. S. P. Lam, I. R. Mordi, J. Tromp, and C. C. Lang, "Artificial intelligence-assisted automated heart failure detection and classification from electronic health records," *ESC Heart Failure*, vol. 11, no. 5, pp. 2769–2777, Oct. 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ehf2.14828
[11] P. A. Heidenreich et al., "2022 AHA/ACC/HFSA guideline for the management of heart failure: A report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines," *Circulation*, vol. 145, no. 18, pp. 895–1032, 2022, doi: 10.1161/CIR.0000000000001063.
[12] C. S. P. Lam, A. A. Voors, R. A. de Boer, S. D. Solomon, and D. J. van Veldhuisen, "Heart failure with preserved ejection fraction: From mechanisms to therapies," *Eur. Heart J.*, vol. 39, no. 30, pp. 2780–2792, Aug. 2018.
[13] K. J. Clerkin, D. M. Mancini, and L. H. Lund, *Diagnosis of Heart Failure*. Cham, Switzerland: Springer, 2019, pp. 83–101, doi: 10.1007/978-3-319-98184-0_6.
[14] G. Hanton, V. Eder, G. Rochefort, P. Bonnet, and J.-M. Hyvelin, "Echocardiography, a non-invasive method for the assessment of cardiac function and morphology in preclinical drug toxicology and safety pharmacology," *Exp. Opinion Drug Metabolism Toxicol.*, vol. 4, no. 6, pp. 681–696, Jun. 2008.
[15] V. Bello, I. Nicastro, V. Barletta, L. Conte, I. Fabiani, A. Morgantini, and G. Lastrucci, "Professional education, training and role of the cardiac sonographer in different countries," *J. Cardiovascular Echography*, vol. 23, no. 1, pp. 18–23, 2013.
[16] F. C. Noya, S. E. Carr, and S. C. Thompson, "Attracting, recruiting, and retaining medical workforce: A case study in a remote province of Indonesia," *Int. J. Environ. Res. Public Health*, vol. 20, no. 2, p. 1435, Jan. 2023.

[17] R. Cribb and M. Ford, *Indonesia As an Archipelago: Managing Islands, Managing the Seas*. Singapore: ISEAS-Yusof Ishak Institute, Jan. 2009. [Online]. Available: http://hdl.handle.net/2123/16146
[18] M. D. Sutrisno, "Shortages of medical doctors in Indonesia, is it true?" *Asian J. Health Res.*, vol. 2, no. 2, pp. 1–2, Aug. 2023.
[19] K. Seetharam, S. Raina, and P. P. Sengupta, "The role of artificial intelligence in echocardiography," *Current Cardiol. Rep.*, vol. 22, no. 9, p. 50, Sep. 2020. [Online]. Available: https://www.mdpi.com/2313-433X/9/2/50
[20] J. Zhou, M. Du, S. Chang, and Z. Chen, "Artificial intelligence in echocardiography: Detection, functional evaluation, and disease diagnosis," *Cardiovascular Ultrasound*, vol. 19, no. 1, p. 29, Aug. 2021.
[21] L. D. Liastuti, B. B. Siswanto, R. Sukmawan, W. Jatmiko, I. Alwi, B. Wiweko, A. Kekalih, Y. Nursakina, R. Y. I. Putri, G. Jati, M. M. L. Ramadhan, E. Govardi, and A. A. Nur, "Learning intelligent for effective sonography (LIFES) model for rapid diagnosis of heart failure in echocardiography," *Acta Medica Indonesiana*, vol. 54, no. 3, pp. 428–437, 2022.
[22] B. Healy, A. Khan, H. Metezai, I. Blyth, and H. Asad, "The impact of false positive COVID-19 results in an area of low prevalence," *Clin. Med.*, vol. 21, no. 1, pp. 54–56, Jan. 2021.
[23] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Mining*, vol. 16, no. 1, p. 4, Feb. 2023.
[24] D. Chicco, M. J. Warrens, and G. Jurman, "The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and brier score in binary classification assessment," *IEEE Access*, vol. 9, pp. 78368–78381, 2021.
[25] O. Uparkar, J. Bharti, R. K. Pateriya, R. K. Gupta, and A. Sharma, "Vision transformer outperforms deep convolutional neural network-based model in classifying X-ray images," *Proc. Comput. Sci.*, vol. 218, pp. 2338–2349, Jan. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923002090
[26] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Appl. Sci.*, vol. 13, no. 9, p. 5521, Apr. 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/9/5521
[27] A. P. Akerman, M. Porumb, C. G. Scott, A. Beqiri, A. Chartsias, A. J. Ryu, W. Hawkes, G. D. Huntley, A. Z. Arystan, G. C. Kane, S. V. Pislaru, F. Lopez-Jimenez, A. Gomez, R. Sarwar, J. O'Driscoll, P. Leeson, R. Upton, G. Woodward, and P. A. Pellikka, "Automated echocardiographic detection of heart failure with preserved ejection fraction using artificial intelligence," *JACC, Adv.*, vol. 2, no. 6, Aug. 2023, Art. no. 100452. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772963X23003125
[28] K. Brown, P. Roshanitabrizi, J. Rwebembera, E. Okello, A. Beaton, M. G. Linguraru, and C. A. Sable, "Using artificial intelligence for rheumatic heart disease detection by echocardiography: Focus on mitral regurgitation," *J. Amer. Heart Assoc.*, vol. 13, no. 2, Jan. 2024, Art. no. e031257. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/JAHA.123.031257
[29] J. A. Naser, E. Lee, S. V. Pislaru, G. Tsaban, J. G. Malins, J. I. Jackson, D. M. Anisuzzaman, B. Rostami, F. Lopez-Jimenez, P. A. Friedman, G. C. Kane, P. A. Pellikka, and Z. I. Attia, "Artificial intelligence-based classification of echocardiographic views," *Eur. Heart J.-Digit. Health*, vol. 5, no. 3, pp. 260–269, May 2024, doi: 10.1093/ehjdh/ztae015.
[30] H. Liao, S. K. Zhou, and J. Luo, "Chapter 2—Deep learning basics," in *Deep Network Design for Medical Image Computing* (The MICCAI Society Book Series), H. Liao, S. K. Zhou, and J. Luo, Eds. USA: Academic Press, 2023, pp. 11–23. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128243831000095
[31] D. Kim, H. Cho, H. Shin, S.-C. Lim, and W. Hwang, "An efficient three-dimensional convolutional neural network for inferring physical interaction force from video," *Sensors*, vol. 19, no. 16, p. 3579, Aug. 2019.
[32] X. Huang and Z. Cai, "A review of video action recognition based on 3D convolution," *Comput. Electr. Eng.*, vol. 108, May 2023, Art. no. 108713. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790623001374
[33] L. Hedegaard, N. Heidari, and A. Iosifidis, "Chapter 14—Human activity recognition," in *Deep Learning for Robot Perception and Cognition*, A. Iosifidis and A. Tefas, Eds. USA: Academic Press, 2022, pp. 341–370. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323857871000191
[34] K. A. Athira and J. D. Udayan, "Temporal fusion of time-distributed VGG-16 and LSTM for precise action recognition in video sequences," *Proc. Comput. Sci.*, vol. 233, pp. 892–901, Jan. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924006380
[35] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Mach. Vis. Appl.*, vol. 24, no. 5, pp. 971–981, Jul. 2013, doi: 10.1007/s00138-012-0450-4.
[36] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.
[37] S. Xingjian, C. Zhourong, W. Hao, Y. Dit-Yan, W. Wai-Kin, and W. Wang-Chun, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 1. Cambridge, MA, USA: MIT Press, 2015, pp. 802–810.
[38] 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," *J. Med. Artif. Intell.*, vol. 3, pp. 1–4, Mar. 2020.
[39] Y. Liu, X. Han, T. Liang, L. Chen, B. Dong, J. Yuan, H. Wang, Z. Zhang, L. Zhao, and Y. Zhang, "Intelligent detection of left ventricular hypertrophy from pediatric echocardiography videos," *Int. J. Imag. Syst. Technol.*, vol. 34, no. 3, May 2024, Art. no. e23086.
[40] C. Vasile and X. Iriart, "Embracing AI: The imperative tool for echo labs to stay ahead of the curve," *Diagnostics*, vol. 13, no. 19, p. 3137, Oct. 2023.
[41] L. G. Klaeboe and T. Edvardsen, "Echocardiographic assessment of left ventricular systolic function," *J. Echocardiography*, vol. 17, no. 1, pp. 10–16, Nov. 2018.
[42] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Chapter 5—Diagnosing of disease using machine learning," in *Machine Learning and the Internet of Medical Things in Healthcare*, K. K. Singh, M. Elhoseny, A. Singh, and A. A. Elngar, Eds. USA: Academic Press, 2021, ch. 5, pp. 89–111. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128212295000033
[43] M. S. Sandeep, K. Tiprak, S. Kaewunruen, P. Pheinsusom, and W. Pansuk, "Shear strength prediction of reinforced concrete beams using machine learning," *Structures*, vol. 47, pp. 1196–1211, Jan. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352012422011791
[44] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/bf02295996.
[45] M. M. L. Ramadhan, G. Jati, and W. Jatmiko, "Building damage assessment using feature concatenated Siamese neural network," *IEEE Access*, vol. 12, pp. 19100–19116, 2024.
[46] R. Riffenburgh, "Chapter 9—Tests on categorical data," in *Statistics in Medicine*, 3rd ed., R. Riffenburgh, Ed., San Diego, CA, USA: Academic Press, 2012, ch. 9, pp. 175–202. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123848642000093
[47] A. Wong, X. Yu Wang, and A. Hryniowski, "How much can we really trust you? Towards simple, interpretable trust quantification metrics for deep neural networks," 2020, *arXiv:2009.05835*.
[48] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.

MGS M. LUTHFI RAMADHAN received the bachelor's degree from the Faculty of Computer Science, University of Sriwijaya, and the master's degree from the Faculty of Computer Science, University of Indonesia. He is a Research Assistant with the Faculty of Computer Science, University of Indonesia. His research interests include deep learning, pattern recognition, and computer vision.

ADYATMA W. A. NUGRAHA YUDHA is currently pursuing the combined bachelor's and master's degrees in computer science with the University of Indonesia. He is a Laboratory Assistant with the Faculty of Computer Science, University of Indonesia. His research interests include deep learning, computer vision, and speech processing.



MUHAMMAD FEBRIAN RACHMADI is a Faculty Member with the Faculty of Computer Science, University of Indonesia. He is also affiliated as a Postdoctoral Research Scientist with Brain Image Analysis (BIA) Unit, RIKEN Center of Brain Science (RIKEN CBS), Wako, Japan, working together with Dr. Henrik Skibbe, the Team Leader of BIA Unit to develop machine/deep learning techniques for processing and analyzing marmoset brain image data. His main area of interests include medical image analysis and computation using data-driven methods, such as deep learning algorithms.

KEVIN MOSES HANKY JR TANDAYU received the Cardiologist and Medical Doctor degrees from the Faculty of Medicine, University of Indonesia. He is a National Board Certified General Cardiologist. His research interests include echocardiography, structural heart disease, and artificial intelligence clinical applications in cardiovascular medicine.

LIES DINA LIASTUTI received the Medical degree and cardiology specialization from the Faculty of Medicine, University of Indonesia, and the Doctor of Medical Sciences degree with a focus on artificial intelligence applications in cardiovascular diagnostics, in 2022. Further, she advancing her expertise with a subspecialty in echocardiography from Austin and Repatriation Medical Centre, Melbourne. Her leadership roles span over two decades, including the President and the Director of the Cipto Mangunkusumo Hospital, where she led major transformations. Her research interests include artificial intelligence in cardiovascular diagnostics and improving healthcare systems at the national level.

WISNU JATMIKO (Senior Member, IEEE) received the B.S. degree in electrical engineering and the M.Sc. degree in computer science from the University of Indonesia, Depok, Indonesia, in 1997 and 2000, respectively, and the Dr.Eng. degree from Nagoya University, Japan, in 2007. Currently, he is a Full Professor with the Faculty of Computer Science, University of Indonesia. His research interests include autonomous robots, optimization, real-time traffic monitoring systems, machine learning, and artificial intelligence. He is the Chairperson of the IEEE Indonesia Section, from 2019 to 2020.