2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) | 979-8-3503-0080-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/METROXRAINE58569.2023.10405560

Deep Learning model for video-classification of Echocardiography images

1st Michela Destito
Experimental and Clinical Medicine
University Magna Graecia
Catanzaro, Italy
michela.destito@unicz.it

2nd Paolo Zaffino
Experimental and Clinical Medicine
University Magna Graecia
Catanzaro, Italy
p.zaffino@unicz.it

3rd Jolanda Sabatino
Women's and Children's Health
University of Padua
Padua, Italy
jolesbt@hotmail.it

4th Claudia Critelli
Medical and Surgical Sciences
University Magna Graecia
Catanzaro, Italy
claudia.critelli@unicz.it

5th Arber Qoku
German Cancer Consortium (DKTK)
German Cancer Research Center (DKFZ)
Goethe University Frankfurt
Frankfurt Cancer Insititute
Frankfurt, Germany
arber.qoku@dkfz.de

6th Florian Buettner
German Cancer Consortium (DKTK)
German Cancer Research Center (DKFZ)
Goethe University Frankfurt
Frankfurt Cancer Insititute
Frankfurt, Germany
florian.buettner@dkfz.de

7th Salvatore De Rosa
Medical and Surgical Sciences
University Magna Graecia
Catanzaro, Italy
saderosa@unicz.it

8th Maria Francesca Spadea
Institute of Biomedical Engineering
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
mf.spadea@kit.edu

***Abstract*—Timely and accurate diagnosis of severe Aortic Stenosis (AS) is crucial to prevent severe clinical implications. The most commonly used parameter for diagnostic purposes is the mean transvalvular pressure gradient, measured by echocardiography ($>= 40$ mmHg). However, its use for detecting severe AS has several limitations, including technical, pathophysiological, and clinical reasons. This study aimed to develop a Deep Learning (DL) model for identifying severe AS using ColorDoppler Echocardiography video data. The new DL model used is called ViViT (Video Vision Transformers). To limit the overfitting problem, the data augmentation technique was applied during the training phase.**

The model achieved an accuracy of 87% in classifying patients with severe AS compared to healthy subjects in the testing group. Future efforts will focus on enhancing model accuracy, increasing the initial dataset, and refining the classification process by implementing multi-classification of AS with varying degrees of severity.

***Index Terms*—Doppler Echocardiography, Deep Learning, Aortic Stenosis, Video Classification, Video Vision Transformer.**

I. Introduction

Aortic Stenosis (AS) holds the highest prevalence among all valvular heart diseases in developed countries [1]. This condition, commonly associated with aging, is progressively becoming more prevalent as the average age of the population continues to rise [2]. Degenerative AS [3] is commonly attributed to chronic inflammatory injury and aggravated by endothelial damage from mechanical stress, lipid infiltration, fibrosis, and valve cusp thickening, leading to eventual calcification [4]. Although AS has a slow development and can remain symptomless for extended periods, its advancement to severe stenosis poses substantial morbidity and a notable risk of sudden cardiac death [5]. Symptomatic severe AS has an untreated annual mortality rate of 25%, with an average expected survival of 2 to 3 years.

The main approach for evaluating the severity of AS is through diagnostic imaging, notably echocardiography with Color-Doppler spectral analysis [6], [7]. This approach enables the measurement of the transvalvular pressure gradient and provides an indirect estimation of the remaining maximal valve area [8], [9]. Nevertheless, this measurement has multiple limitations of technical, pathophysiological, and clinical nature. Patients with degenerative AS often exhibit various degrees of left ventricular dysfunction at clinical presentation, which affects the transvalvular gradient and makes it challenging to assess the degree of severity accurately.

To address these challenges, Deep Learning (DL) methodologies [10] can be applied to automatically classify echocardiographic data. DL models have achieved positive outcomes in various tasks, such as view classification and disease diagnosis [11]. Furthermore, researchers in computer vision have shown growing enthusiasm for video action recognition in recent times [12], [13]. Also, within this field, some authors are focusing on developing an automatic method to classify cardiac function, Myocardial amyloidosis and, common heart diseases

979-8-3503-0080-2/23/$31.00 ©2023 IEEE

[14], [15], [16]. Additionally, certain studies concentrate on DL analysis of electrocardiography (ECG) for detecting AS, aortic regurgitation (AR), and mitral regurgitation (MR) [17], [18]. To date, there are still few studies that have addressed the challenge of classifying cardiovascular diseases on echocardiographic images [19], [20], [21], but that do not focus solely and exclusively on AS.

This research is the only one so far to have employed ViViT (Video Vision Transformer), pure-transformer based models, for video classification of echocardiographic images with Color Doppler, drawing, drawing on the recent successes of such models in image classification [22]. The Transformer, initially used in natural language processing, is a deep neural network primarily based on the self-attention mechanism [23], [24], [25], [26]. Transformer-based models have shown comparable or even superior performance when compared to other types of networks, including convolutional and recurrent neural networks [27].

The objective of this work was to develop a prediction DL model based Transformers model for classifying patients with severe AS versus control subjects on echocardiography images with Color Doppler.

II. Material and methods

A. Database and Image Processing

We analyzed images echocardiograms of 162 subjects from the Institutional Database at the "AOU Mater Domini" University Hospital of Catanzaro. The cohort included 73 healthy controls and 89 patients with severe AS. The demographics and clinical features of the participant group are presented in Table I.

TABLE I
DESCRIPTION OF THE PATIENT DATASET. VALUES ARE EXPRESSED AS MEDIAN AND QUARTILE, OR IN %.

	Control Subjects	Patients with severe AS
Eligible Patients (#)	73	89
Male:Female	0.65	0.43
Median Age	60 (44-68)	82 (78-85)
Median BMI	26.3 (23.97-28.28)	27.39 (24.97-30.48)
Cigarette smoking (%)	51	32
Diabetes (%)	17	44
Hypertension (%)	54	88

In Figure 1, we show example of images from a patient with severe AS and healthy subjects; distinguishing these two categories is difficult to an untrained eye. Each patient presents an echocardiographic examination with a different number of frames because it is dependent on the heart rate. For each video, all frames were extracted and subsequently preprocessed, cropping the images (to eliminate the background) and reshaping them to 224×224 pixels.

B. Data Augmentation

It is important to acknowledge that utilizing only 162 images for image classification tasks in DL may pose challenges due to the relatively small dataset size. To address this limitation, data augmentation is a widely adopted technique in computer vision tasks. This approach involves applying diverse transformations to the original data, effectively augmenting the dataset and artificially expanding the training samples. The primary objective of data augmentation is to enhance the trained model's ability to generalize and be more robust by introducing a broader range of variations in the input data [28].

In this specific case, for each individual frame of all Echocardiograms in the training dataset, three different augmentation methods were applied: "brightness adjustment", "left-right flipping", and "saturation". These methods were chosen to ensure that the changes to the images are not significant, while still increasing the dataset's diversity. The dataset was divided into Training (70%), Validation (10%), and Testing (20%) subsets. As a result, the final training dataset comprised 452 images, derived from the initial 113 images through the augmentation process. An example of an augmented image is depicted in Figure 2.

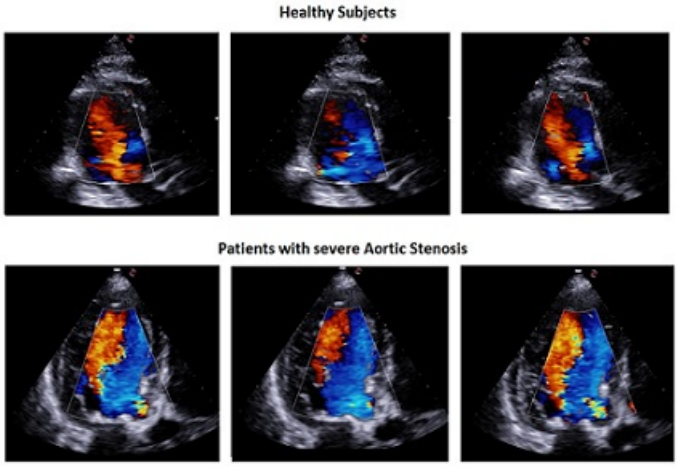

Fig. 1. Examples of echocardiograms images with ColorDoppler in Healthy Subjects and Patients with sever AS

C. Model architecture

ViViT was used like a Deep Learning model for the classification task. In this model, spatio-temporal tokens are extracted from the input video ((in this particular case, Echocardiography) and encoded through a series of transformer layers. The primary operation in this architecture is self-attention, computed on a sequence of spatio-temporal tokens extracted from the video. To handle a large number of spatio-temporal tokens that can be present in videos, ViViT introduces various factorization methods along spatial and temporal dimensions to enhance effectiveness and expandability. This method proposes more transformer-based architectures. Initially, we build upon an extension of ViT [29], which facilitates pairwise interactions among all spatio-temporal tokens. Subsequently, we devise more

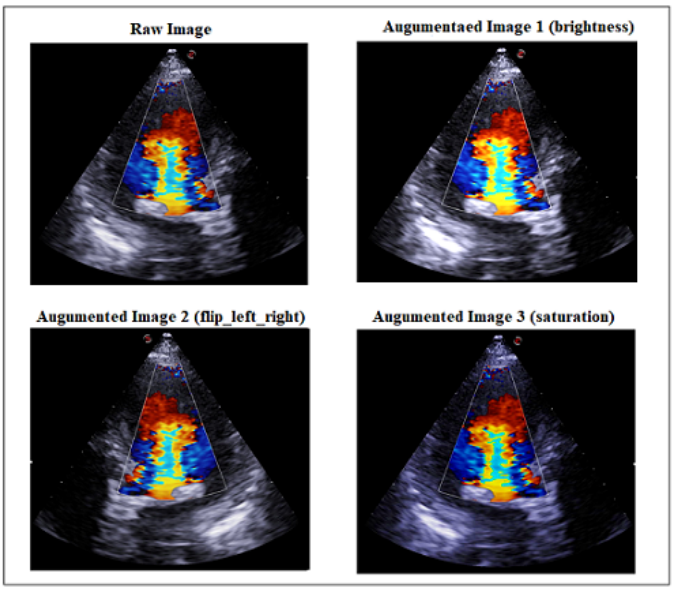

Fig. 2. Data Augumentation

efficient variations that involve factorizing the space and time dimensions of the input video at different levels of the transformer architecture. The framework of the DL algorithm, summarized in the Figure 3.

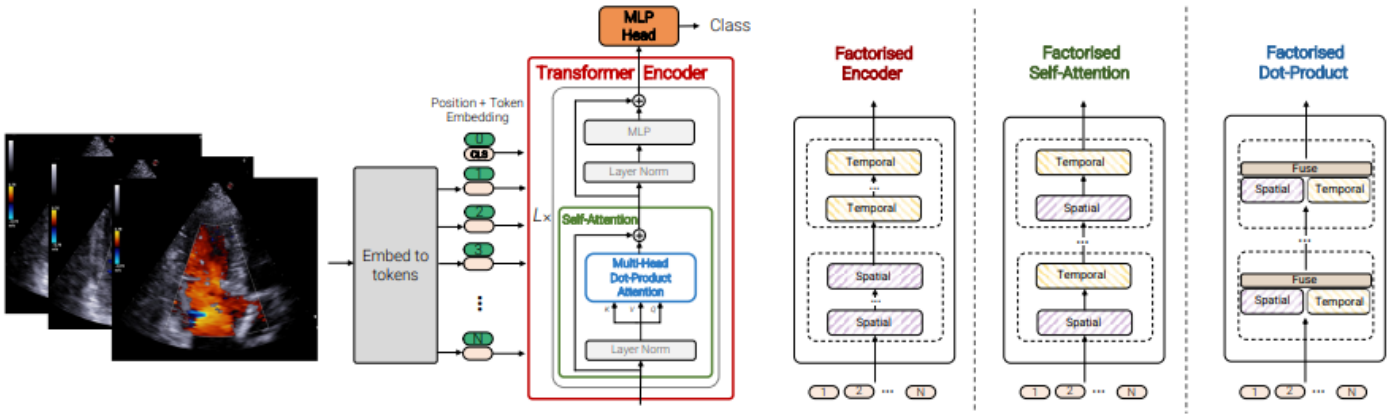

Fig. 3. ViViT (Video Vision Transformers) Architecture.

The authors of ViViT introduced four distinct variants of pure transformer-based video classification models, drawing inspiration from ViT (Vision Transformer), to execute video classification tasks:

• Model-1, known as Spatio Temporal attention, implements spatio-temporal attention by tokenizing a video sample using a Tubelet embedding approach, where each Tubelet represents a token. Subsequently, each token undergoes processing through a patch embedding layer, including position encoding, and is then fed into a traditional transformer encoder.

• Model-2, named the Factorized Encoder-Decoder, diverges from utilizing a uniform encoder model for all tokens in the video sample. Instead, the video is divided into smaller clips. For the temporal transformer, each token represents a vector extracted from a specific clip, indicating diverse temporal indexes within the video. In contrast, the spatial transformer assigns each token to a Tubelet extracted from an individual clip, with all tokens originating from the same temporal index but having distinct spatial indexes.

• Model-3, named Factorized Self Attention, closely resembles the first model, with the only difference being the use of a different transformer encoder block. This modified transformer block is akin to the standard block used in original transformers, except for the inclusion of multi-headed self-attention as the sole distinction.

• Model-4, referred to as Factorized Dot Product Attention, introduces the factorization of the dot product attention heads to address spatial and temporal aspects independently.

For this research, we will be utilizing Model 1, which is the Spatio-temporal attention model. For each patient, we have not considered all frames but we have normalized based on the heart rate, taking a number of frames fixed at 10. The Adam optimization algorithm was used (Learning Rate = 0.001) [30] and the maximum training epoch was set to be 130. The whole training process was carried out using a single Nvidia Quadro RTX 5000 (16 GB).

This implementation is based on the TensorFlow library [31].

D. Experiment and Result Analysis

The evaluation of video classification performances was primarily focused on Accuracy, a widely utilized performance metric in classification tasks. Accuracy quantifies the ratio of correct classifications to the total number of samples, making it a significant indicator of the model's effectiveness. When dealing with classes that have an equal number of samples, accuracy alone suffices as a metric.

However, the dataset, which involves two classes (healthy subjects vs. patients with severe AS), is slightly unbalanced. To ensure a comprehensive evaluation of our algorithm's performance, it is considered good practice to utilize additional metrics. As such, we calculated the Confusion Matrix, which

provides a detailed overview of the model's performance, describing true positives, true negatives, false positives, and false negatives. This information aids in better understanding the classification results and allows us to gauge the algorithm's proficiency across different aspects.
Based on the confusion matrix, the following metrics can be calculated [32]:

• Precision, in the context of two-class classification, is computed as the quotient of True Positives (TP) divided by the sum of True Positives (TP) and False Positives (FP). This formula facilitates the determination of precision, a crucial metric for assessing the accuracy of positive predictions:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

• Recall, known as True Positive Rate or Sensitivity, indicates the ratio of TP to the sum of TP and FN. For Binary Classification can be calculated as:

$$Precision = \frac{TP}{TP + FN} \tag{2}$$

• The F1 Score is a metric that represents the harmonic mean of precision and recall. In the context of two-class classification, it can be expressed using the following formula:

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

III. Results

In the Testing group the classification accuracy was 87%. In Details, condidering only control subjects results of classification was: Recall (0.82), Precision (0.93) and f1-score (0.87). Instead, for patients with Severe AS, the Recall was 0.93, the Precision was 0.81, and f1-score was 0.87. In Table II are summarized all results of the classification task.

TABLE II
Classification Report of testing patients.

	Precision	**Recall**	**F1-score**
Patients with severe AS	0.93	0.82	0.87
Control Subjects	0.81	0.93	0.87
Accuracy			0.87
Macro avg[a]	0.87	0.88	0.87
Weighted avg[b]	0.88	0.87	0.87

[a] Macro-avg: The final averaged metric across all classes.
[b] Weighted-avg: Weighted contribution based on class sizes.

The Normalize Confusion Matrix is represented in Figure 4.

IV. Conclusion

Although patients with severe AS may not display symptoms, they face an unfavorable prognosis with a high event rate and a risk of swift functional deterioration [33]. The mean

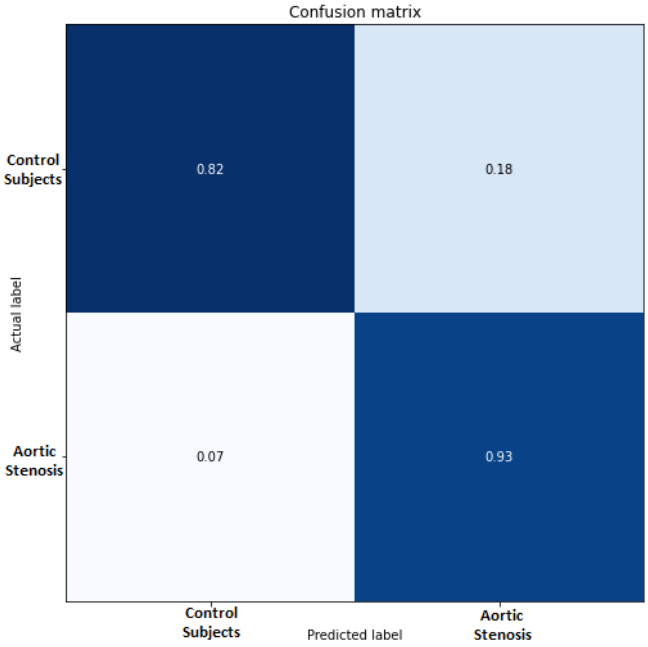

Fig. 4. Confusion Matrix of testing subjects.

transvalvular pressure gradient, assessed through echocardiography, is commonly utilized for diagnosing severe AS. However, in certain cases, this parameter may not be conclusive due to various clinical factors. Hence, the development of an automated model becomes crucial, providing clinicians with an additional assessment during the diagnostic process. Compared with other studies in the literature [19], [20], [21] that have used echocardiographic videos none have focused exclusively on AS but rather on a range of cardiovascular diseases. In addition, we used a new method (ViViT) to analyze the Echocardiography that has multiple advantages over the better-known DL techniques [34], such as considering the time variable of the videos and not considering each frame separately.
This study has certain limitations that warrant discussion. Firstly, the relatively small number of patients might significantly influence the learning process, potentially leading to suboptimal prediction performances and overfitting, despite employing various techniques like Data Augmentation [35]. By augmenting the data, the trained model becomes exposed to a larger and more varied dataset during training. This exposure can potentially lead to improved performance and better generalization, despite the initial limitation of having a small number of images. However, it is essential to recognize that data augmentation alone might not be sufficient for handling more complex tasks, and other strategies such as transfer learning or acquiring additional data could be considered to further enhance the model's capabilities.
Our future objective is to further expand the dataset in order to validate and reinforce these promising results.
Secondly, the ViViT architecture requires a fixed number of

input frames and this can be a problem when each patient has a different number of frames. We have tried to overcome this problem by normalizing all frames with heart rate. However, in a future development we would like to be able to analyze patient echocardiographic images with a variable number of frames.

The DL algorithm employed has the potential to automate of the clinical workflow for screening Echocardiographic images for the presence Sever Aortic Stenosis and in the feature for quantifying metrics of disease severity.

REFERENCES

[1] Nkomo, Vuyisile T., et al. "Burden of valvular heart diseases: a population-based study." The lancet 368.9540 (2006): 1005-1011.

[2] Carabello, el al. "Aortic stenosis." The lancet 373.9667 (2009): 956-966.

[3] Ramaraj, Radhakrishnan, et al. "Degenerative aortic stenosis." Bmj 336.7643 (2008): 550-555.

[4] Joseph, Jessica, et al. "Aortic stenosis: pathophysiology, diagnosis, and therapy." The American journal of medicine 130.3 (2017): 253-263.

[5] Campo, John, et al. "Prognosis of severe asymptomatic aortic stenosis with and without surgery." The Annals of thoracic surgery 108.1 (2019): 74-79.

[6] Perry, Gilbert J., et al. "Evaluation of aortic insufficiency by Doppler color flow mapping." Journal of the American College of Cardiology 9.4 (1987): 952-959.

[7] Sachpekidis, Vasileios, et al. "A novel handheld echocardiography device with continuous-wave Doppler capability: Implications for the evaluation of aortic stenosis severity." Journal of the American Society of Echocardiography 35.12 (2022): 1273-1280.

[8] Clavel, Marie-Annick, et al. "Low-gradient aortic stenosis." European heart journal 37.34 (2016): 2645-2657.

[9] Clavel, Marie-Annick,et al. "Cardiac imaging for assessing low-gradient severe aortic stenosis." JACC: Cardiovascular Imaging 10.2 (2017): 185-202.

[10] LeCun, Yann, et al. "Deep learning." nature 521.7553 (2015): 436-444.

[11] Lv, Feiya, et al. "Fault diagnosis based on deep learning." 2016 American control conference (ACC). IEEE, 2016.

[12] Feichtenhofer, Christoph, et al. "Spatiotemporal multiplier networks for video action recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[13] Pareek, Preksha, et al. "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications." Artificial Intelligence Review 54 (2021): 2259-2322.

[14] Liu, Bohan, et al. "A deep learning framework assisted echocardiography with diagnosis, lesion localization, phenogrouping heterogeneous disease, and anomaly detection." Scientific Reports 13.1 (2023): 3.

[15] Zhang, Xiaofeng, et al. "Deep learn-based computer-assisted transthoracic echocardiography: approach to the diagnosis of cardiac amyloidosis." The International Journal of Cardiovascular Imaging (2023): 1-11.

[16] Ueda, Daiju, et al. "Artificial intelligence-based model to classify cardiac functions from chest radiographs: a multi-institutional, retrospective model development and validation study." The Lancet Digital Health (2023).

[17] Elias, Pierre, et al. "Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease." Journal of the American College of Cardiology 80.6 (2022): 613-626.

[18] Kwon, Joon-Myoung, et al. "Deep learning–based algorithm for detecting aortic stenosis using electrocardiography." Journal of the American Heart Association 9.7 (2020): e014717.

[19] Wu, Haotang, et al. "The predictive value of deep learning-based cardiac ultrasound flow imaging for hypertrophic cardiomyopathy complicating arrhythmias." European Journal of Medical Research 28.1 (2023): 36.

[20] Yang, Feifei, et al. "Automated analysis of doppler echocardiographic videos as a screening tool for valvular heart diseases." Cardiovascular Imaging 15.4 (2022): 551-563.

[21] Zhang, Jeffrey, et al. "Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy." Circulation 138.16 (2018): 1623-1635.

[22] Arnab, Anurag, et al. "Vivit: A video vision transformer." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[23] Han, Kai, et al. "A survey on vision transformer." IEEE transactions on pattern analysis and machine intelligence 45.1 (2022): 87-110.

[24] Lin, Tianyang, et al. "A survey of transformers." AI Open (2022).

[25] Khan, Salman, et al. "Transformers in vision: A survey." ACM computing surveys (CSUR) 54.10s (2022): 1-41.

[26] Tay, Yi, et al. "Efficient transformers: A survey." ACM Computing Surveys 55.6 (2022): 1-28.

[27] Cao, Mingdeng, et al. "Vdtr: Video deblurring with transformer." IEEE Transactions on Circuits and Systems for Video Technology 33.1 (2022): 160-171.

[28] Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." Journal of big data 6.1 (2019): 1-48.

[29] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[30] Jais, Imran Khan Mohd, Amelia Ritahani Ismail, and Syed Qamrun Nisa. "Adam optimization algorithm for wide and deep neural network." Knowledge Engineering and Data Science 2.1 (2019): 41-46.

[31] Singh, Pramod, et al. "Introduction to tensorflow 2.0." Learn TensorFlow 2.0: Implement Machine Learning and Deep Learning Models with Python (2020): 1-24.

[32] Luque, Amalia, et al. "The impact of class imbalance in classification performance metrics based on the binary confusion matrix." Pattern Recognition 91 (2019): 216-231.

[33] Rosenhek, Raphael, et al. "Natural history of very severe aortic stenosis." Circulation 121.1 (2010): 151-156.

[34] Mathew, Amitha, P. Amudha, and S. Sivakumari. "Deep learning techniques: an overview." Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020 (2021): 599-608.

[35] Taqi, Arwa Mohammed, et al. "The impact of multi-optimizers and data augmentation on TensorFlow convolutional neural network performance." 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2018.