Identification of Congenital Valvular Murmurs in Young Patients Using Deep Learning-Based Attention Transformers and Phonocardiograms

Mohanad Alkhodari, *Graduate Student Member, IEEE*,
Leontios J. Hadjileontiadis, *Senior Member, IEEE*, and Ahsan H. Khandoker, *Senior Member, IEEE*

***Abstract*—One in every four newborns suffers from congenital heart disease (CHD) that causes defects in the heart structure. The current gold-standard assessment technique, echocardiography, causes delays in the diagnosis owing to the need for experts who vary markedly in their ability to detect and interpret pathological patterns. Moreover, echo is still causing cost difficulties for low- and middle-income countries. Here, we developed a deep learning-based attention transformer model to automate the detection of heart murmurs caused by CHD at an early stage of life using cost-effective and widely available phonocardiography (PCG). PCG recordings were obtained from 942 young patients at four major auscultation locations, including the aortic valve (AV), mitral valve (MV), pulmonary valve (PV), and tricuspid valve (TV), and they were annotated by experts as absent, present, or unknown murmurs. A transformation to wavelet features was performed to reduce the dimensionality before the deep learning stage for inferring the medical condition. The performance was validated through 10-fold cross-validation and yielded an average accuracy and sensitivity of 90.23% and 72.41%, respectively. The accuracy of discriminating between murmurs' absence and presence reached 76.10% when evaluated on unseen data. The model had accuracies of 70%, 88%, and 86% in predicting murmur presence in infants, children, and adolescents, respectively. The interpretation of the model revealed proper discrimination between the learned attributes, and AV channel was found important (score > 0.75) for the murmur absence predictions while MV and TV were more important for murmur presence predictions. The findings potentiate deep learning as a powerful front-line tool for inferring CHD status in PCG recordings leveraging early detection of heart anomalies in young people. It is suggested as a tool that can be used independently from high-cost machinery or expert assessment.**

***Index Terms*—Deep learning, attention transformer, congenital heart disease, heart murmur, phonocardiography.**

I. Introduction

BIRTH heart defects, better known as congenital heart disease (CHD), are nearly affecting 1.2% of newborns worldwide with one in every four suffering from severe CHD [1]. It causes at least 220,000 deaths every year; majority during the first year after birth [2]. CHD affects blood flowing inside coronary arteries or outside the heart to the rest of the body. In severe cases, CHD can affect the function of formed parts in the heart such as valves, which require surgery in the baby's first year of life [3]. Although CHD mortality rates have heavily decreased in high-income countries owing to the advances in healthcare services, it is still a long-term increasing death factor in low- and middle-income countries [4]. These counties often suffer from the lack of capacity and advanced healthcare services to perform congenital heart surgeries; which results in a higher amount of CHD-related costs and loss of lives [5]. Therefore, timely assessment at an early stage of life with developing diagnostic techniques could potentially decrease mortality rates by almost 58% in these countries [6], which eventually can prevent many disabilities or death cases.

Currently, CHD is widely diagnosed using echocardiography, which is the current gold standard for comprehensive pre- and post-natal screening [7]. Although reliable, echo still poses difficulties in low- and middle-income countries because it requires expensive equipment. In addition, it is highly dependent on trained and experienced clinician for proper interpretation [8], which delays the identification of heart defects at an early stage especially since its interpretation still markedly vary between clinicians [9]. In many situations. patients still have to travel long distances to reach medical centers where equipment and expertise are available [10].

A potential alternative could be the use of phonocardiography (PCG) which is a non-invasive cardiac auscultation device that

Manuscript received 18 February 2023; revised 9 November 2023 and 18 December 2023; accepted 19 January 2024. Date of publication 23 January 2024; date of current version 5 April 2024. This work was supported by the Grant from Healthcare Engineering Innovation Center (HEIC), Khalifa University, Abu Dhabi, UAE, under Award 8474000132. *(Corresponding author: Mohanad Alkhodari.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by The data collection was approved by the 5192—Complexo Hospitalar HUOC/PROCAPE institution review board, under Real Hospital Portugues de Beneficiencia em Pernambuco request, and the enrollment protocol was conducted following the Declaration of Helsinki.

Mohanad Alkhodari is with the Healthcare Engineering Innovation Center (HEIC), Department of Biomedical Engineering, Khalifa University, Abu Dhabi 127788, UAE, and also with the Cardiovascular Clinical Research Facility, Radcliffe Department of Medicine, University of Oxford, OX12JD Oxford, U.K. (e-mail: mohanad.alkhodari@ku.ac.ae).

Leontios J. Hadjileontiadis is with the Healthcare Engineering Innovation Center (HEIC), Department of Biomedical Engineering, Khalifa University, Abu Dhabi 127788, UAE, and also with the Department of Electrical and Computer, Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece (e-mail: leontios.hadjileontiadis@ku.ac.ae).

Ahsan H. Khandoker is with the Healthcare Engineering Innovation Center (HEIC), Department of Biomedical Engineering, Khalifa University, Abu Dhabi 127788, UAE (e-mail: ahsan.khandoker@ku.ac.ae).

Digital Object Identifier 10.1109/JBHI.2024.3357506

carries information on the mechanical function of the heart [11], [12]. A PCG signal, recorded through an electronic stethoscope, describes the function of heart valves during blood flow inside and outside the heart including the mitral, tricuspid, aortic, and pulmonic valves [13]. The flow of blood through valves orifices produces four important heart sounds; namely S1, S2, S3, and S4. In systole, an S1 heart sound is produced while in diastole the remaining heart sounds appear [14]. CHD can cause heart murmurs which are abnormal heart sounds during systole or diastole. A systolic heart murmur appears in the interval between S1 and S2, while a diastolic heart murmur occurs after S2 and before S1 [15]. Although PCG provides reliable information on the mechanical malfunction of the heart due to CHD, it is not commonly used by cardiologists who often prefer echocardiography. However, with the current advances of computerised algorithms, i.e., artificial intelligence, and with the increasing demand for timely, continuous, and personalised healthcare services, PCG could play a pivotal role in facilitating early diagnosis protocols to reduce and prevent heart diseases [16].

The use of AI algorithms in healthcare has been rising in different aspects [17], [18], [19], and several studies have investigated its efficiency in murmur detection using PCG recordings. Machine learning (ML) algorithms such as support vector machine (SVM) and feature extraction techniques were used to identify murmurs after signal transformation using Mel-frequency cepstral coefficient (MFCC) [20]. Moreover, systole and diastole sound segmentation was utilised in several studies as input to feature extraction algorithms such as hidden Markov model (HMM) and Gammatone frequency cepstral coefficient (GFCC) [14], [21]. Most recently, deep learning algorithms including convolutional neural networks (CNN) and recurrent neural networks (RNN) were more frequently used in multiple studies [11], [16], [22] owing to their ability to extract features without any feature engineering techniques. However, there is still a research gap when it comes to the overall performance in murmur identification. The majority of these studies still require pre-processing steps and human interference such as filtration or heart sound segmentation; which in most cases are highly affected by the original quality of signals. In addition, several studies utilized basic and curated datasets that may not be representing the raw nature of PCG recordings that are often contaminated by noise sources. Moreover, deep learning models may encounter uncertainty problems due to the lack of knowledge on optimal parameters, their randomness, and meaning [23]. Therefore, there is still a need for more sophisticated approaches that could better handle complicated recordings regardless of their original quality with less interference with the signals and with better interpretability.

In this study, we investigate the use of a deep learning approach based on the latest self-attention transformer network to automatically predict the occurrence of congenital murmurs in PCG recordings. The proposed approach does not require heavy memory demands for computations due to the use of simple network architecture and a reduced dimensionality on input signals through wavelet feature transformation. Each PCG recording was transformed to wavelet-based features sequence of a short length and thus, reducing the overall complexity and dimension of the input data. Moreover, the utilisation of deep learning for the prediction of the absence or presence of heart murmurs allows for interpreting the decisions throughout network layers and thus, obtaining more insights into the impact of auscultation channels on predictions. This study has several advantages. First, our approach does not require any segmentation for heart sounds or prior interference with the input signals, and thus, raw PCG signals are enough which reduces the need for expert annotations or additional processing algorithms. Second, the use of the attention mechanism allows the network to focus on relevant parts within the inputs during training while ignoring the less relevant ones; which ensures enhanced learning especially for problems such as the one we are trying to solve where the presence of murmurs occurs at different time points within the signals and not continuously. Third, since we are including raw PCG signals, it is essential to use wavelet decomposition to capture fine details in the signals and provide simultaneous time and frequency localization. Moreover, wavelet transform allows for better analyzing the dynamic nature of the signal's frequency spectrum and thus, detecting with higher resolutions the small changes happening due to heart murmurs. Last, a two-step approach using wavelet transform and attention transformers, which was previously investigated in the 2D image and computer vision applications [24], leverages the basic design of attention networks by adding more focus on multi-scale feature maps identified through the transformation; which eventually would result in enhanced performance in solving complex problems.

II. Materials and Methods

A. Dataset and Patient Enrollment

The dataset used in this study was obtained from the new George B. Moody PhysioNet 2022 Challenge [25], the CirCor DigiScope dataset [26], which included pediatric patients enrolled in Northeast Brazil during two mass screening campaigns conducted in July-August 2014 and June-July 2015; the Caravana do Coração (Caravan of the Heart). The data collection protocol was approved by the 5192-Complexo Hospitalar HUOC/PROCAPE institutional review board, under the request of the Real Hospital Portugues de Beneficencia empernambuco. All young patients (aged 21 years or younger) provided a signed consent form or a legal guardian consent form if they were below 18 years old.

B. Data Preparation and Labelling

The study included a total of 1568 patients (only 942 have publicly available data – nearly 60%) who sequentially recorded one or more PCG signals from multiple auscultation locations, namely aortic valve (AV), mitral valve (MV), pulmonary valve (PV), and tricuspid valve (TV), with a sampling frequency of 4,000. Out of the 942 available data, a total of 695 had no heart murmurs, 179 had heart murmurs, and 68 were labeled as unknown. An expert annotator labeled all signals based on the presence or absence of murmurs. In addition, if the annotator was not certain about the condition of the patient, a label of "unknown" was chosen. All patients recorded at least 5 seconds on each channel and at most 65 seconds. We designed our approach first by identifying missing and very short recordings. In the case of a missing recording from a specific channel, the previous channel was duplicated to fill the gap. Moreover, we ensured that all signals are 40 seconds in length to be enough to capture the majority of sound information about heart function in each recording. Furthermore, short signals were padded, and longer signals were truncated. Signals were z-score normalized before any further analysis.

TABLE I
BASELINE CHARACTERISTICS OF PATIENTS INCLUDED IN THE STUDY WITH GROUPING BASED ON VALVULAR MURMURS CONDITION

Category	**Overall** **n = 942**	**Groups**			**p-value**
		Absent n = 695 (73.78%)	Present n = 179 (19.00%)	Unknown n = 68 (7.22%)	
Age	Neonate: 6 (0.65%) Infant: 126 (13.38%) Child: 664 (70.49%) Adolescent: 72 (7.64%) Unlabeled: 74 (7.86%)	Neonate: 4 (0.58%) Infant: 76 (10.94%) Child: 495 (71.22%) Adolescent: 53 (7.63%) Unlabeled: 67 (9.64%)	Neonate: 1 (0.56%) Infant: 25 (13.97%) Child: 132 (73.74%) Adolescent: 16 (8.94%) Unlabeled: 5 (2.79%)	Neonate: 1 (1.47%) Infant: 25 (36.76%) Child: 37 (54.41%) Adolescent: 3 (4.41%) Unlabeled: 2 (2.94%)	**<0.001**$^{\bullet +}$
Sex	Female: 486 (51.59%)	Female: 355 (51.08%)	Female: 92 (51.40%)	Female: 39 (57.35%)	0.614
Height	108 (74-130)	117 (93-134)	111 (87-130)	82 (66-123)	**<0.001**$^{\bullet +}$
Weight	18.55 (9.59-28.80)	21.70 (13.60-32.50)	18.55 (11.10-28.40)	12.15 (7.70-25.60)	**0.002**$^{\bullet}$
BMI	16.32 (14.32-18.45)	16.81 (15.20-19.10)	16.10 (14.47-17.81)	17.87 (15.88-20.00)	**0.009**$^{+}$
Pregnant	70 (7.43%)	65 (9.35%)	3 (1.68%)	2 (2.94%)	**<0.001**$^{*\bullet}$

All values are represented as median (inter-quartile range) or n (%). Bold p-values show statistically significant differences (p < 0.050) amongst the three groups using the one-way analysis of variance (ANOVA) test. *: Significant difference between absent and present groups; $^{\bullet}$: Significant difference between absent and unknown groups; $^{+}$: Significant difference between present and unknown groups. BMI = Body mass index

C. Patients Information and Statistical Analysis

We performed statistical analysis using one-way student t-test analysis of variance (ANOVA) on all available patient information (Table I). The provided information included age (neonate [birth to 27 days old], infant [28 days old to 1-year-old], child [1 to 11 years old], adolescent [12 to 18 years old], and unlabeled), sex (male or female), height (in cm), weight (in kg), body mass index (BMI, in kg/m2), and pregnancy (yes or no). A statistical significance was noted if the p-value was below or equal to 0.05.

D. Transformation to Wavelet Features

To capture internal heart function characteristics contaminated within signals, a wavelet decomposition approach (Fig. 1) was followed based on the original discrete wavelet transform (DWT) algorithm [27], [28]. Each PCG channel signal (40 seconds – 160,000 samples) was segmented into 32-sample windows that shift with a step of 32 samples to reduce overlapping (total of 5,000 windows per channel). The selection of window length, step, and wavelet type was based on an ad-hoc manner to capture enough signal information without over- or under-estimations. Symlets 8 (Sym8) wavelet family was used for decomposition into 5 main levels (J) determined based on the sampling frequency (f) as follows,

$$J = \frac{\log(f)}{2 \times \log(2)} \tag{1}$$

This decomposition level allows for capturing time and frequency characteristics of the input signals. Then, the transformed signals were converted into 105 concatenated approximation coefficient vectors. Each vector was then squared to generate the corresponding power (P) vector, and the corresponding wavelet features were then extracted from every coefficient including the energy (E), variance (σ^2), standard deviation (σ), waveform length (L), and Shannon entropy (SE). The energy, which reflects the time-scale density in the signal, was calculated as,

$$E = \sum_{i=1}^{k} P_{t_i} \tag{2}$$

where t_i corresponds to the selected coefficient and k represents the total of 15 coefficients.

The variance and standard deviation, revealing the variability in the transformed power signal, were calculated as follows,

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_{t_i} - \mu_k)^2}{k} \tag{3}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} (x_{t_i} - \mu_k)^2}{k}} \tag{4}$$

where x corresponds to the selected value.

Moreover, the waveform length and Shannon entropy, which reflects the power signal complexity, were calculated as,

$$L = \sum_{i=1}^{k} |(x_{t_i} - x_{t+1})| \tag{5}$$

$$SE = -\sum_{i=1}^{k} \mathbb{P} - ln(\mathbb{P}) \tag{6}$$

where $\mathbb{P}$ is the probability mass function given as,

$$\mathbb{P} = \frac{P_{t_k}}{E} \tag{7}$$

A total of 30 features were obtained for every 32-sample segment. Then, all features extracted per channel were concatenated to form a wavelet power transformation of the raw PCG channel. The alignment of all four channels results in having a transformed signal of 20,000 samples (4 channels × 5000 32-sample segments) that will be used for further analysis. Upon this transformation, a stronger transformation of the original PCG

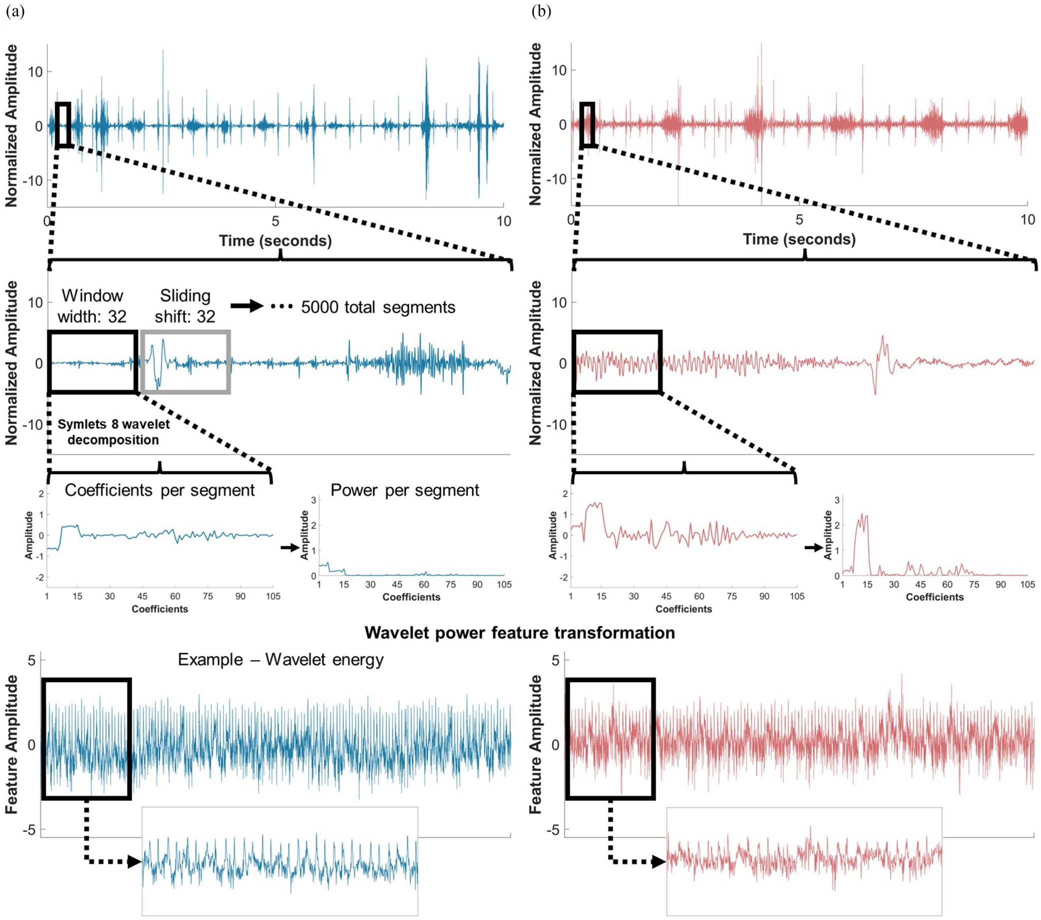


Fig. 1. Transformation of the phonocardiogram (PCG) recording (showing mitral valve (MV) channel) into stacked power features using wavelet decomposition. (a) An example of a healthy patient PCG recording (showing only 10 seconds). The signal gets decomposed sequentially for every window (a total of 5000 windows of 32 samples each) using Symlets 8 wavelet, and the corresponding decomposition coefficients are generated. The power of each coefficient segment is used to extract additional features. The final feature transformation (showing wavelet energy) is stacked for every window segment to form a new wavelet feature-based signal, which then gets stacked on a channel-by-channel basis. (b) An example of a patient with heart murmur.

signals was achieved represented by the contaminated wavelet power features.

E. Development of the Self-Attention Transformer Network

Transformers have been commonly used as an effective deep learning network in natural language processing (NLP) applications [29], [30]. A transformer network is more advanced than the conventional CNN and RNN networks in its ability to capture global dependencies in the input through self-attention mechanisms [31], which forces the network to learn from regions of most significant impact on the class prediction. Most recently, transformers were used in medical applications [31], [32], [33] and provided an enhanced performance relative to conventional deep learning approaches. A transformer network (Fig. 2) consists of four major components; namely feature encoder, positional encoder, transformer unit, and decoder, which can be designed according to the problem in hand and input type.

1) Feature Encoder: A feature encoder, which is often called the CNN backbone, is the deep feature transformation block that transforms the input into an easily identifiable format. In this block, consecutive convolution operations are applied to the input followed by a series of max-pooling layers to reduce the dimensionality of the input to lower resolution, i.e., reducing the length on input signals, A convolutional operation is defined as follows,

$$C_i^{lj} = h\left(b_j + \sum_{m=1}^{M} w_m^j x_{i+m-1}^j\right) \tag{8}$$

where $x_i = [x_1, x_2, \ldots, x_n]$ is the input, n is the total number of points, l is the layer index, h is the activation function, b is the bias of the j_{th} feature map, M is the kernel size, w_m^j is the weight of the j^{th} feature map and m_{th} filter index.

In this work, and since the input (30 × 20,000) is already a wavelet-based transformation of the raw heart sound signal, we designed the feature encoder to include only two CNN blocks. The first convolution is built with a kernel of 64 points and 32 filters, whereas the second has a kernel of 32 points and 64 filters. After each convolutional step, batch normalisation and Gaussian error linear unit (GeLU) was applied. To reduce the dimensionality, i.e., make the input shorter, a max-pooling layer followed each convolutional step with a kernel size of 3 and stride of 2. Hence, the output from the feature encoder is a set

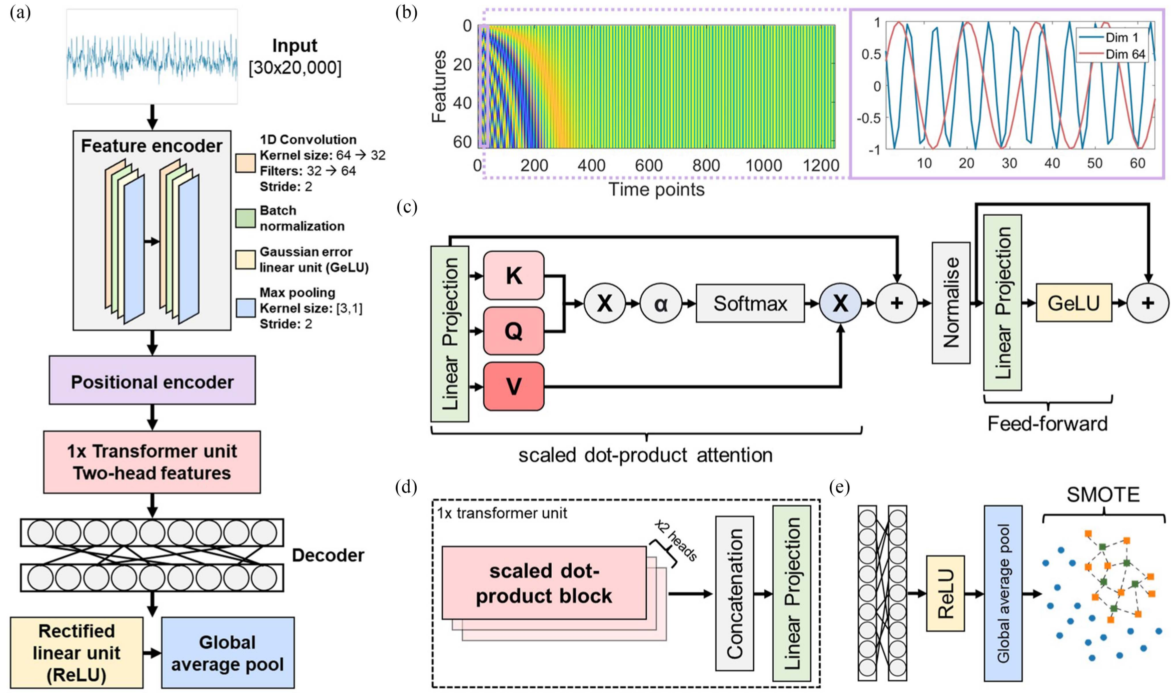


Fig. 2. Modelling of the self-attention transformer network with details on the structure and design. (a) The complete framework of the proposed model includes the four main components, namely the feature encoder, positional encoder, transformer unit, and decoder. The feature encoder starts by reducing the dimensions of the input from [30 × 20,000] to [64 × 1,250], focusing on the most important features in the input. (b) The positional encoder with a zoomed-in view of the sinusoidal signals guides the positioning of input vectors in proper time sequences. (c) The main block of the transformer includes the calculations of the scaled dot-product attention and the feed-forward network. (d) By using multiple heads within each transformer unit, the joint attention gets captured by concatenating information from multiple subspaces. (e) The last block in the network, the decoder, was designed to be simple full-connected layers. To handle data imbalance, the safe-level synthetic minority over-sampling technique (SMOTE) is applied to the attention extracted from the last layer of the network.

of DL feature vectors of 64 × 1250 size that represents the most essential features within the inputs.

2) Positional Encoder: Positional encoding is an essential step in time-sequence analysis especially before self-attention transformer layers as it describes the location of the feature vectors in a sequence (Fig. 2b) such that it is assigned with a unique representation1. In this work, the unique representation was chosen to be a set of sinusoidal signals within the positional encoder block (PE) as follows,

$$PE(i, 2j) = \sin\left(\frac{i}{10,000^{2j/d}}\right) \tag{9}$$

$$PE(i, 2j+1) = \cos\left(\frac{i}{10,000^{2j/d}}\right) \tag{10}$$

where i and j define the location in the feature and time-sequence dimensions, respectively, and d is the model dimension, i.e., number of features. In this layer, the encoding value, ω, was selected to be as low as 10,000 to ensure having a unique encoding at different positions.

The sine and cosine functions used here allows for representing complex periodic information in the input at different phases. The 10,000 allows for distinguishing elements based on their relative positions in the sequence at a wider range of positional information.

3) Transformer Unit: A transformer unit consists of the major components for the attention mechanism proposed in this study; namely the scaled dot-production attention followed by the multi-head attention.

The scaled dot-product attention (Fig. 2c) is calculated as the dot product between three inputs which are the queries (Q), keys (K), and values (V) of the input with features dimension d. First, Q's and K's are multiplied and divided by a scaling factor equal to the square root of the dimension as follows,

$$QK^T = \frac{Q \times K}{\sqrt{d/N}} \tag{11}$$

where N is the number of heads used in the model, i.e., 2 heads.

Then, a softmax function is applied to normalise the values before the second multiplication step with the V's to calculate the attention (A) as follows,

$$A(Q, K, V) = \text{softmax}(QK^T) \times V \tag{12}$$

The resulting attention from this step goes as input to the multi-head attention mechanism (Fig. 2d) which applies a series of linear projections to the input Q's, K's, and V's. Therefore, instead of having a single attention mechanism, calculating multi-attention $M(Q, K, V)$ in parallel across queries, keys, and values results in more jointly focused information from multiple subspaces with d_v dimension across the three inputs as follows,

$$M(Q, K, V) = concat(h_1, h_2, \ldots, h_N) \times W^0 \tag{13}$$

where h_i can be calculated as,

$$h_i = A\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{14}$$

where W_i is the linear projection of each input for the selected head i.

4) Feature Decoder: The decoder used in this study was simpler than the usual implementation of decoders by only utilising linear projections in fully connected layers (Fig. 2e). This is to ensure that the input attention is properly decoded by obtaining the features within the subspaces formed within the self-attention mechanism explained earlier.

5) Network Enhancements End Experimental Design: The transformer network was further enhanced by handling any data imbalance in the provided data during training. Thus, the learning capabilities of the network during training get improved by implementing the synthetic minority over-sampling technique (SMOTE) [34] technique. By applying the Safe-level SMOTE mechanism, the minority class gets over-sampled at the attention stage from activations extracted from the global average pooling layer by generating synthetic samples from the last layer of the network.

The training of the model was based on the adaptive moment estimation (ADAM) optimiser with an initial learning rate of 0.001 that drops by 10% at the 40th epoch (total of 60 epochs). We evaluated the trained model following a 10-fold cross-validation scheme over the whole dataset. In addition, we applied two testing scenarios; one with three classes, i.e., absent, present, and unknown murmurs, and the other with two classes, i.e., absent and present murmurs. In both scenarios, the validation was patient-independent. At each fold, 10% of the data (94 patients) were hidden and used for testing.

III. Results

A. Baseline Characteristics of Patients

The participating patients (Table I) were mostly from the child category (n = 664) followed by infants (n = 126) out of the whole dataset (n = 942). The gender of the patients was almost balanced with 51.59% for the female category (n = 486). BMI inter-quartile range was 14.32— 18.45 for the BMI with a median of 16.32, which can be considered slightly underweight. Amongst the 942 patients, 70 were pregnant (7.43%).

The statistical significance between the three groups, i.e., absent, present, and unknown, was below 0.05 in age, height, weight, BMI, and pregnancy, which demonstrates the whole characteristics except for sex (p-value: 0.614). The age group was highly significant (p-value: <0.001) between each group versus the unknown group. Moreover, BMI was significant (p-value: 0.009) between the present and unknown groups where each had a median and inter-quartile range of 16.10 (14.47–17.81) and 17.87 (15.88–20.00), respectively. Concerning pregnancy status, the majority of patients who were pregnant had no murmurs, which demonstrated a statistical significance between the absence and the other two groups (p-value: <0.001). It is worth noting that both groups were significantly different in most characteristics versus the unknown group.

B. Computation Complexity

The complexity of the computational processes in the proposed approach, i.e., multiplications or additions, includes two main parts; namely the wavelet decomposition and transformer unit. The wavelet step had a complexity of $O(N_w)$; which is dependent on the length of input (N_w). Moreover, the transformer unit had a $O(N_t^2 \times d)$ with 64 input features dimension (d). However, the input length (N_t) of the transformer gets heavily reduced from 20,000 to 1,250 because of the input decoder; thus reducing the overall complexity of the multi-head attention.

The training of the transformer model required only 5 minutes for a complete training over 60 epochs. The model was trained on an NVIDIA RTX3080 GPU with 4 GB of RAM. The training time is expected to increase slightly if simpler GPUs are used. The complexity of the approach was not a big concern due to the dimensionality reduction using wavelet features transformation which reduced the length of each signal from 160,000 to 20,000. Further, the input length gets reduced in the feature encoder stage before processing with the transformer.

C. Detection of Three-Class Heart Murmurs

The confusion matrix of the accumulated predictions after the cross-validation scheme with three classes is shown in Fig. 3a where the model had an average accuracy of 90.23%, an average sensitivity of 72.41%, and an average precision of 70.22%. Excluding the unknown group, the values become 89.49%, 87.18%, and 83.27% for the accuracy, sensitivity, and precision, respectively. The overall performance metrics breakdown is provided in Fig. 3b; highlighting that the model achieved normalized Matthews correlation coefficient (NMCC) of 82.02%, 87.45%, and 69.51% for the absent, present, and unknown groups, respectively. By excluding the unknown group, the average NMCC becomes 84.74%.

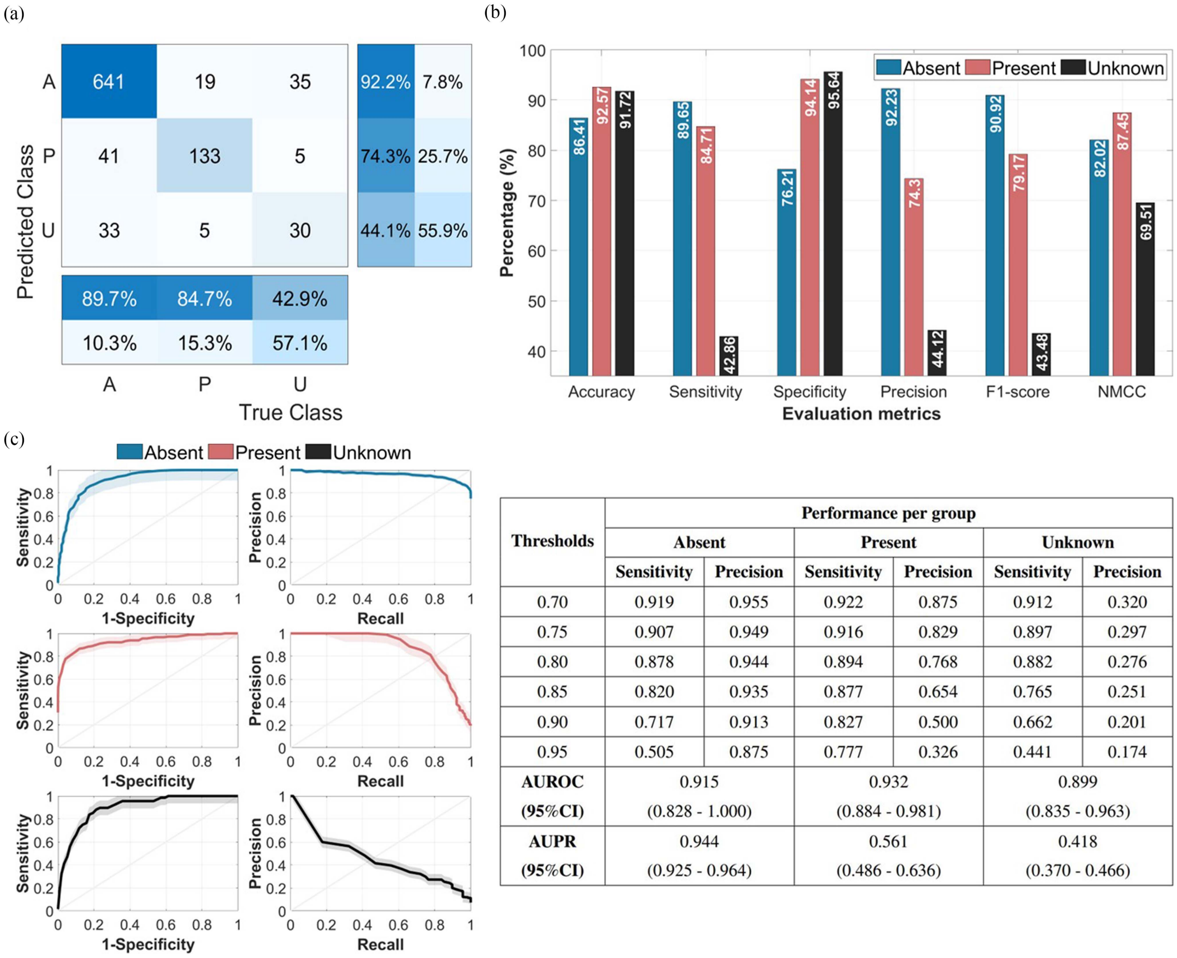
We further analyzed the performance of the model through the receiver operating characteristics (ROC) and precision-recall (PR) curves (Fig. 3c). The highest area under the ROC curve (AUROC) and the PR curve (AUPR) was achieved in predicting the absence of murmurs (AUROC: 0.915, 95% Confidence Interval: 0.828–1.000). In predicting the presence of murmurs, the model reached an AUROC of 0.932 ± 0.048 and AUPR of 0.561 ± 0.075. Moreover, the unknown class had the lowest values amongst the three groups with an overall AUROC of 0.899 (95% Confidence Interval (CI): 0.835–0.963) and AUPR of 0.418 (95% CI: 0.370–0.466).

D. Analysis of Predictions Based on Age Group

We analysed the predictions made by the developed model after the cross-validation scheme based on age groups (Fig. 7) provided in Table I. We evaluated the model for the overall dataset and when separating patients according to the murmur group, i.e., absent, present, and unknown murmurs, for five age groups including neonate, infant, child, adolescent, and unlabeled. Overall, the model had accuracies of 50%, 70%, 88%, 86%, and 95% in predictions at each age group, respectively. The absent group had roughly more than 75% correct predictions in each age group. On the other hand, the present group had a close pattern but with no correct neonate predictions (1 out of 1) and with a slight decrease in the correct child group predictions (77%). The adolescent group predictions were almost divided equally (56%) between correct and wrong predictions which were close to the unlabeled percentage in terms of miss-classification accuracy (40%). The unknown group was similar to the present group in the neonate category with no correct predictions (1 out of 1). However, the wrong predictions in the infant and child groups were higher than correct predictions by 5% and 50%, respectively. Most interestingly, the adolescent and unlabeled groups had no wrong predictions.

E. Interpretation of Model Decisions

To interpret the model's ability to discriminate between the three groups, i.e., absent, present, and unknown murmurs, we



Thresholds	Performance per group					
	Absent		Present		Unknown	
	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
0.70	0.919	0.955	0.922	0.875	0.912	0.320
0.75	0.907	0.949	0.916	0.829	0.897	0.297
0.80	0.878	0.944	0.894	0.768	0.882	0.276
0.85	0.820	0.935	0.877	0.654	0.765	0.251
0.90	0.717	0.913	0.827	0.500	0.662	0.201
0.95	0.505	0.875	0.777	0.326	0.441	0.174
AUROC (95%CI)	0.915 (0.828 - 1.000)		0.932 (0.884 - 0.981)		0.899 (0.835 - 0.963)	
AUPR (95%CI)	0.944 (0.925 - 0.964)		0.561 (0.486 - 0.636)		0.418 (0.370 - 0.466)	

Fig. 3. Performance of the trained self-attention transformer model validated using a 10-fold cross-validation scheme with three classes. (a) Confusion matrix for the predictions of absent (A), present (P), and unknown (U) murmur classes with the corresponding one-versus-all per-class sensitivity and specificity. (b) Performance metrics calculated from the confusion matrix for the accuracy, sensitivity, specificity, precision, F1-score, and normalized Matthews correlation coefficient (NMCC). (c) Analysis of the receiver operating characteristic (ROC) and precision-recall (PR) curves with the 95% confidence interval (CI) calculations for each class.

visualise the activations at each layer of the network (Fig. 5 – top row) while including the 95% confidence interval across all patients in each group. Through visual inspection, the model discrimination ability was enhanced by going deeper with the network. In both the training and testing sets (Fig. 5 (a), and (b)), the model's best separation was in the last layers, i.e., decoder fully connected layers, and final global average pooling. Moreover, by using the t-distributed stochastic neighbor embedding (t-SNE) analysis (Fig. 5 – middle row), the clustering of the averaged activations becomes more efficient starting from the transformer layer in both sets. The analysis of each layer using the gradient-weighted class activation mapping (GradCAM) revealed that wavelet features extracted from the AV channel were highly important (score > 0.75) for the decisions of the absence of murmurs (Fig. 5 – bottom row). In addition, channels such as MV and TV had the highest importance for the prediction of the presence of murmurs. Overall, the PV and TV channels were important for the decisions of all three groups.

F. Performance With Two Classes

We tested the transformer network by training it on predicting two classes only, i.e., absent and present (Fig. 6) without the inclusion of the unknown group to further investigate the discrimination ability of the model. The accuracy has reached 91.76% with sensitivity and NMCC values of 93.81% and 86.98%, respectively. In addition, the AUROC between the two classes has reached 0.93 whereas the AUPR values was 0.97 for the absent class and 0.57 for the present class. When interpreting the model decision, similar discrimination ability was observed (Fig. 6c) between the two groups with proper attention through GradCam analysis to the AV channel for the absence group and to the TV channel for the presence group.

G. Additional Experiments

We have performed an additional comparison between the proposed transformer model with another two conventional models; namely the CNN and Bi-directional long short-term memory (Bi-LSTM). The CNN had a simple structure of three one-dimensional (1D) convolutional blocks, each followed by batch normalization and a rectified linear unit (ReLU). At every convolutional layer, the kernel sizes were 64, 32, and 16, with 32 filters each. On the other hand, the Bi-LSTM had a total of 50 hidden units operating in both directions of the input signals. The analysis of accuracy and AUC (Table II) revealed that

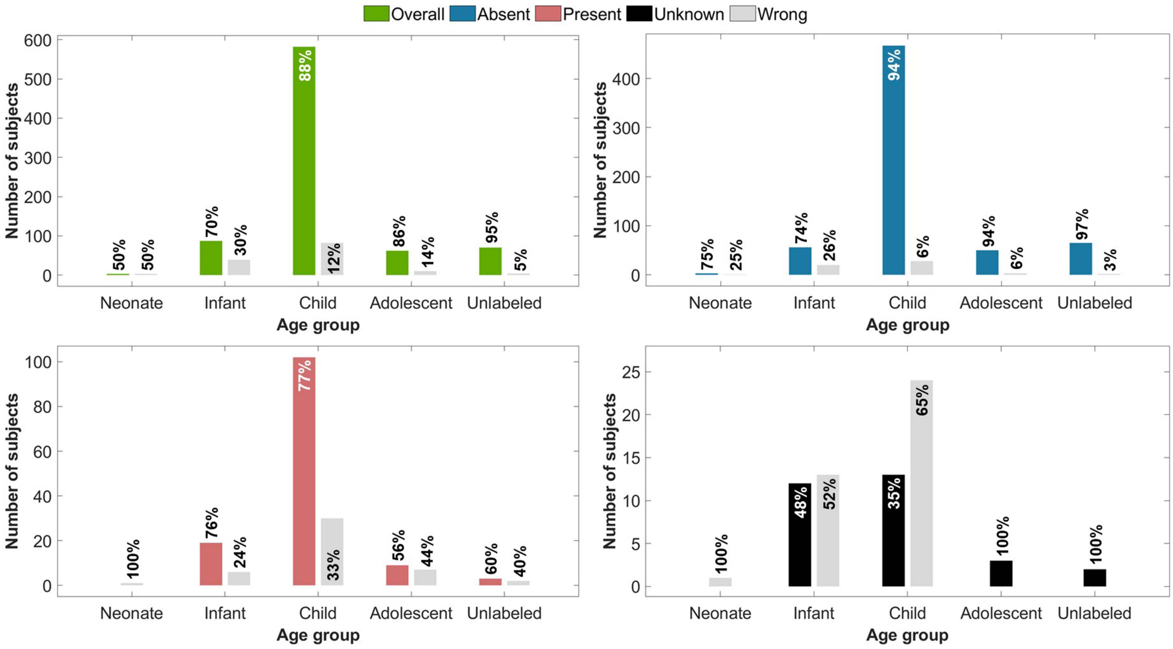


Fig. 4. Analysis based on age groups for the developed model. Showing the true and wrong (light gray) predictions for the overall dataset (green), absent (blue), present (red), and unknown (black) groups. The age groups included are neonate, infant, child, adolescent, and unlabeled. More details on the number of patients in each category is provided in Table I.

TABLE II
THREE-CLASS PERFORMANCE OF THE PROPOSED TRANSFORMER MODEL COMPARED WITH A BASIC CONVOLUTIONAL NEURAL NETWORK (CNN) AND BI-DIRECTIONAL LONG SHORT-TERM MEMORY (BI-LSTM) MODELS

Groups	*Accuracy*			*Sensitivity*			*Specificity*			*AUC*		
	Transformer	CNN	Bi-LSTM	**Transformer**	CNN	Bi-LSTM	**Transformer**	CNN	Bi-LSTM	**Transformer**	CNN	Bi-LSTM
Absent	**86.41**%	80.74%	61.89%	**89.65**%	85.29%	75.53%	**76.21**%	62.80%	30.28%	**0.92**	0.83	0.57
Present	**92.57**%	85.34%	67.62%	**84.71**%	79.73%	45.39%	**94.14**%	88.32%	63.38%	**0.93**	0.86	0.62
Unknown	**91.72**%	83.88%	90.66%	**42.86**%	36.59%	34.29%	**95.64**%	89.49%	73.01%	**0.90**	0.82	0.67
Average	**89.57**%	83.32%	73.39%	**72.41**%	67.20%	51.74%	**88.66**%	80.20%	55.56%	**0.92**	0.83	0.62

the transformer outperformed conventional models by nearly 10–20%.

To elaborate on the importance of PCG channels, we trained the model on each channel separately and observed the performance relative to the original combined PCG-based model (Fig. 7). Interestingly, we have found that the highest accuracy of 85.68% was observed for the absent group using the AV channel; which was supported by a higher AUROC of 0.876 compared to other channels. Moreover, the present group showed a variable performance across channels, but most importantly, the MV, PV, and TV channels had the highest performance. The highest accuracy of 89.85% was accounted for the TV channel with an AUROC of 0.895. Lastly, the unknown group had a balanced performance across all channels with no differences in the metrics. Overall, the performance was maximized when combining all channels which ensures enhanced learning capabilities by the model.

Moreover, we have provided a summary of the proposed approach relative to other studies in the detection of heart murmurs using the same dataset (Table III). Several studies have utilized different pre-processing steps including de-noising and heart sound segmentation. In addition, studies used feature transformation using methods such as Mel-spectrogram, Mel-frequency cepstral coefficients (MFCC), Hilbert transform, and power spectral density. However, the performance was less than our proposed approach in accuracy and AUC. Furthermore, compared with the majority of these studies, our approach did not require any pre-processing steps.

H. Validation on Unseen Data

The proposed approach was validated on the PhysioNet 2022 hidden validation and testing sets that were unseen during training the model [35]. The model had an accuracy level of 76.10% on the validation set that included 10% of the original dataset (156 patients). Moreover, the accuracy reached up to 75.70% when evaluated on the testing set that included 30% patients of the original dataset (470 patients). It is worth noting that the validation and testing sets were inaccessible and thus, no further analysis could be performed [25].

IV. DISCUSSION

Our study suggests that deep learning with its latest transformer network would be a powerful decision support tool to clinicians to facilitate early, timely, and continuous assessment of congenital diseases affecting the heart, i.e., valvular anomalies. To put this in perspective, mortality rates can be decreased

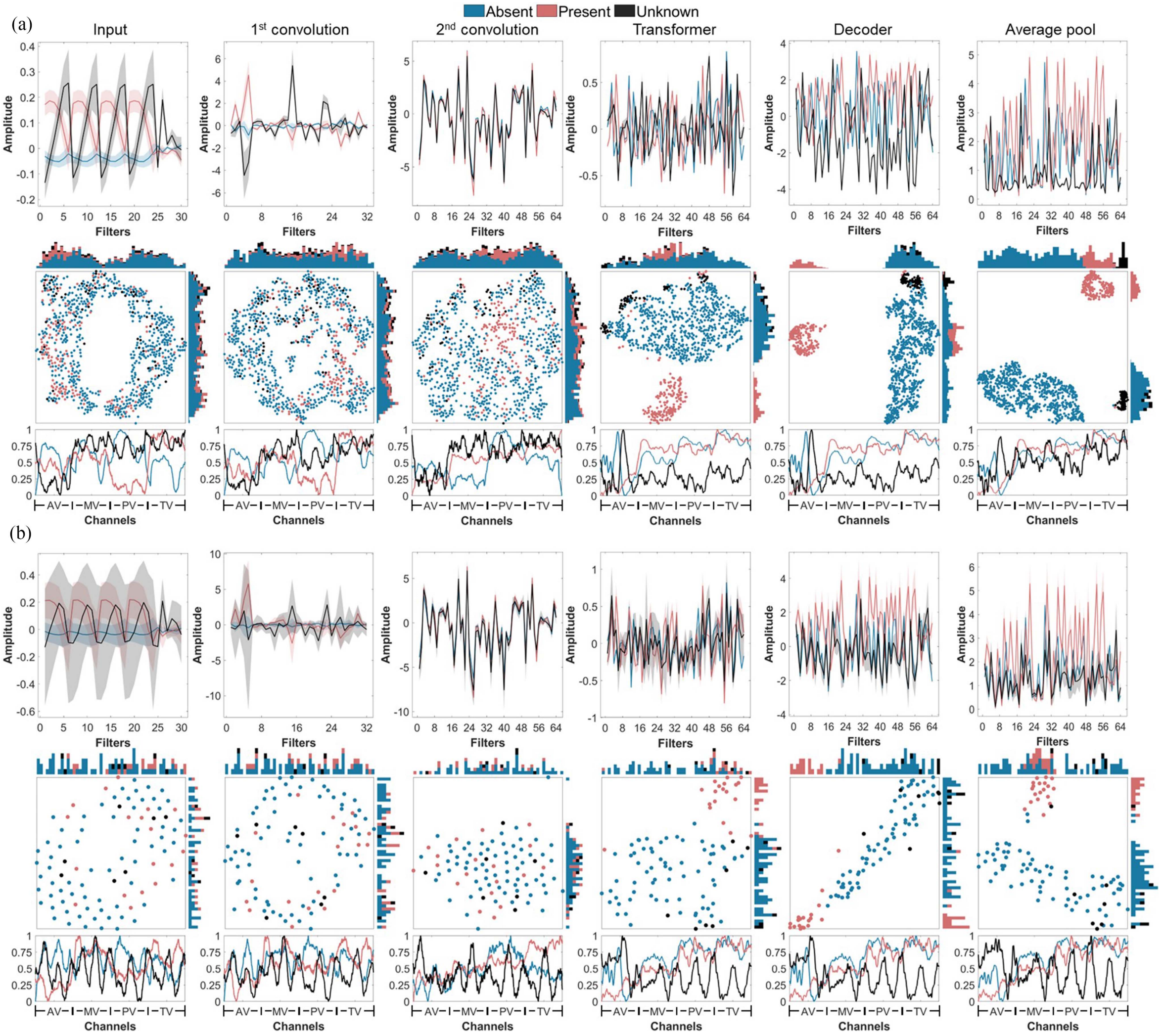


Fig. 5. Interpretation of the model's discrimination ability between the absent, present, and unknown murmurs groups. (a) Visualisation of the whole training set showing the activations of each layer in the network (top row) with the 95% confidence interval across all patients in each group. In addition, the t-distributed stochastic neighbor embedding (t-SNE) analysis (middle-row) was used to cluster the averaged activations of each layer. Lastly, the gradient-weighted class activation mapping (GradCAM) was used to visualise each auscultation channel's impact on the network's decisions at each layer. (b) Visualisation of a single fold from the testing set.

TABLE III
Comparison of Performance With Other Studies in Valvular Murmurs Detection Using the PhysioNet 2022 Dataset

Study	Year	Dataset	Pre-processing	Features	Model	Performance
Monteiro et al. [36]	2022	CirCor DigiScope [26]	Sound segmentation Band-pass filtering	Homomorphic Hilbert transform Power spectral density Wavelet envelopes	Bi-LSTM	Accuracy: 75.70%
Patwa et al. [37]			Sound segmentation Out-of-bound de-noising	Wavelet scattering	CNN	Accuracy: 83.81% AUC: up to 0.91
Summerton et al. [38]			S1 and S2 sound segmentation	MFCC	Gradient boosting Ensemble CNN	Accuracy: 75.30%
McDonald et al. [39]			S1 and S2 sound segmentation	Log-spectrogram transform	RNN Hidden semi-Markov model	Accuracy: 81.70%
Parvaneh et al. [40]			N/A - Raw signals	Mel-spectrogram	CNN (YAMNet)	Accuracy: 83.10%
Fuadah et al. [41]	2023		Sound segmentation Band-pass filtering	MFCC	k-NN	Accuracy: 76.31% AUC: up to 0.76
This study	**2023**		**N/A - Raw signals**	**Wavelet features transformation**	**Attention transformer**	**Accuracy: 89.57%** **AUC: up to 0.92**

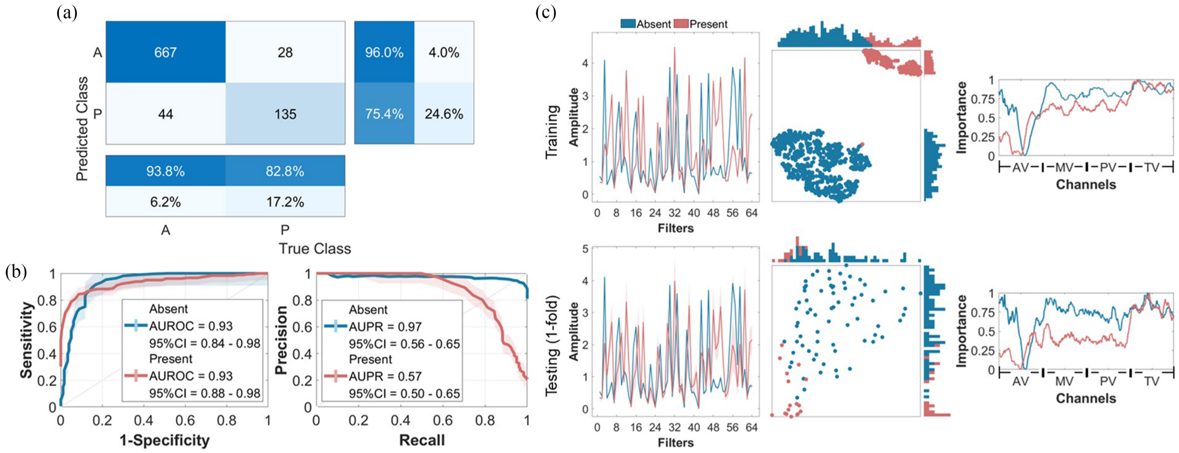


Fig. 6. Performance and interpretation of the model when trained on two classes; the absent and present heart murmurs. (a) Confusion matrix for the predictions of absent (A) and present (P). (b) Analysis of the receiver operating characteristic (ROC) and precision-recall (PR) curves with the 95% confidence interval (CI) calculations for each class. (c) Visualisation of the whole training set showing the activations of the last layer in the network (average pool — top row). Moreover, the activations of the same layer for one testing fold is provided (bottom row). The interpretation shows the t-distributed stochastic neighbor embedding (t-SNE) analysis and each auscultation channel's impact on the network's decisions when the gradient-weighted class activation mapping (GradCAM) is used.

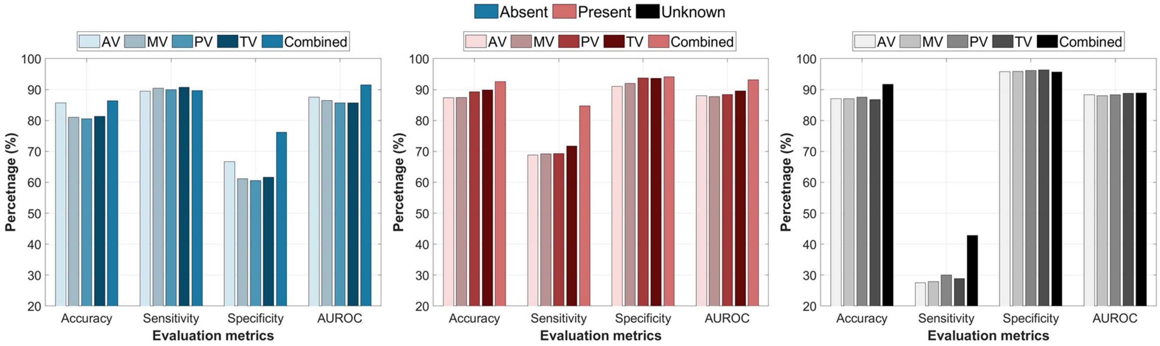


Fig. 7. Evaluation of the model's performance when trained on each channel separately relative to the original combined PCG-based model. PCG channels are the aortic valve (AV), mitral valve (MV), pulmonary valve (PV), and tricuspid valve (TV). The combined model performance is the last bar on each plot.

by more than 50% in low- and middle-income countries if cost-effective and accurate technological tools have been utilised for diagnosis [6]. Because we include patients who are from a young cohort, a murmur prediction accuracy of 90% at an early stage of life and by an automated approach could leave its impact on guiding medication, treatment, and intervention plans with less dependency on experts who have variable interpretation for murmurs [9]; all of which would improve the quality of life to these patients and reduce missed diagnosis.

The findings of this study have several implications. First, the accurate discrimination between the absence or presence of congenital murmurs in PCG recordings is of significance in promoting automated cardiac auscultation through deep learning as a tool to complement echocardiography that requires expert interpretation and expensive machinery. The use of physiological signals to assess cardiovascular diseases has been widely used in literature [42], however, due to its lengthy nature, it could be hard to interpret visually by clinicians. Therefore, signals and their extracted features serve as a solid material to build deep learning tools that are capable of handling them, learning from thousands of patient clinical details and hidden characteristics, and providing automated predictions of medical conditions. We keep in mind that there is a necessity to develop such tools to serve as a first-stage screening before requesting additional tests that are more expensive and not universally available. Moreover, we tried to utilize PCG signals while being as raw as possible without any human interference or prior sound segmentation. Therefore, the proposed approach ensures more robustness in clinical application; reducing the need for expert annotations or automated techniques that may lack the required accuracy.

Secondly, although the expert annotator was uncertain of the medical condition in some PCG recordings, i.e., labeled as unknown, training a model to infer the condition after learning deep characteristics from big multi-class patient data is advantageous as it gives more insights on this ambiguous group. For example, out of the 68 unknown cases, only 30 were found to have common patterns that isolate them from the other two groups, i.e., absence and presence of murmurs, whereas 33 were found to carry patterns that characterise them similarly to the absence of murmurs signals. This could raise a question as to whether to consider the unknown group as a separate entity on its own; more specifically as a transition from the absence to the presence of murmurs, or just an uncertain diagnosis. Several studies have shown that there could be some innocent murmur types that are still ambiguous to experts [43], especially when unusual changes were found in some cases of mitral valve prolapse [44]. Fig. 7 shows that some of the unknown cases were separated clearly from the other two groups through deep learning, yet they carry both importance patterns of absent and present when evaluating the importance of auscultation channels. This was supported by the observations when models were trained using each channel separately; which revealed specific channels to have a higher impact on detecting the presence of murmurs or being unknown. We note that a 100% correct prediction was made for the neonate, adolescent, and unlabelled age group; which could be either driven by the low number of subjects in each group (max of 3) or by a specific pattern for these participants that deviates from the patterns of absence or presence of murmurs. To the best of our knowledge, our study is the first to touch on this point through PCG recordings and with discoveries of a trained deep learning tool.

Thirdly and most importantly, inferring the condition of CHD according to heart murmurs is of high importance, especially at a young stage in life, i.e., infant and child. This is supported by the findings in this study (Fig. 3) where the majority of correct predictions for the absence and presence of murmurs were obtained in these two age groups. Several studies [45], [46] have suggested timely clinical assessment for CHD during the first few weeks of life, i.e., neonate to infant, as they can heavily reduce morbidity and mortality rates in newborns. However, it could be a hard task for clinicians to detect murmurs at an early stage [47] and reduce false positive rates in the assessment. Therefore, a trained tool based on deep learning could aid in decision-making and provide some guidance to clinicians when evaluating murmurs in a timely fashion. Interestingly and as mentioned in the previous paragraph, the unknown group had more wrong classifications for the infant and child group which could be because it is not a separate entity on its own when compared to the absence and presence of murmurs groups. However, it had a 100% correct prediction rate in the adolescent group which could provide insight on the nature of this group in older patients.

Fourthly, the performance and efficacy of the proposed tool could pave the way for applying new techniques in clinical settings to aid clinicians in decision-making. While the current gold-standard technique, echocardiography, is considered the first line of defence in evaluating heart murmurs, the proposed tool could assist in performing cost-effective and continuous screening for patients with reliable results. Heart sounds in nature, being continuous acoustic waves, could benefit from frequent monitoring applications with a simple, yet effective, tool that is driven by AI technology; especially at an early stage or before advanced assessment. However, AI tools in general, especially those relying on deep learning, require wider exploration of their performance when the source of training data is varying. The diversity in such tools and the ability to generalize over a wider patient cohort should be carefully taken into consideration before implementing them in clinical settings.

Despite the presented study's potential, some limitations could be considered before applying our study in the clinic. First, we included patient data which was publicly available by the challenge to train and evaluate our model. We also evaluated the performance of unseen data taken from the same database. However, it will be interesting to explore the reliability of the model with additional data and extra validation through external datasets. Moreover, a multi-center evaluation across different hospitals with varying patient characteristics is needed. Second, we provided insights on the separate group, i.e., the unknown murmur, with a high accuracy level, therefore, more studies, yet, are needed to verify this group clinically to whether consider it as a special entity between absence and presence of congenital murmurs, as it was originally annotated by experts in the used dataset, or just an uncertainty in diagnosis. This would add more insights into the reliability of the proposed approach and AI tools in general in being considered as an assistant to doctors in clinical decision-making. Third, our approach transforms raw PCG recordings to their corresponding wavelet-domain features which reduces the noise impact on the signals. However, the model may also benefit from additional features describing the segments within the PCG recordings, i.e., S1 and S2. Last, the proposed structure of the deep learning model with attention mechanism was basic yet efficient, however, with the recent advancement of transformers, an enhanced performance could be achieved when utilizing advanced structures; taking into consideration the trade-off between performance and model complexity which was relatively low in this study.

IV. Conclusion

Overall, we have developed a deep-learning tool to provide an automated prediction for CHD-induced murmurs. The findings suggest deep learning as a powerful front-line tool that could potentially guide clinical decision-making for early detection of heart anomalies in young people. As a complementary tool to echocardiography, PCG and deep learning can leverage accurate discrimination between the absence and presence of congenital murmurs with less dependency on expert assessment and high-cost machinery, especially in low- and middle-income countries. Therefore, future research should be directed toward evaluating the clinical cost-effectiveness of the approach while maintaining high performance.

Conflict of Interest

All authors have reported that they have no relationships relevant to the contents of this paper to disclose.

Acknowledgment

The authors would like to acknowledge Syafiq Azman from the AIQ, ADNOC H.Q., Abu Dhabi, United Arab Emirates, for his advice and review while developing the transformer network. This work was part of the PhysioNet/CinC 2022 challenge and achieved the 4th place during the event. All authors had final responsibility for the decision to submit for publication. Data used in this study were fully accessed by the corresponding authors.

Data availability

The dataset used in this study was part of the George B. Moody PhysioNet 2022 challenge which is publicly available at: https://moody-challenge.physionet.org/2022/. Wavelet decomposition method can be obtained from here: https://github.com/RamiKhushaba/getmswtfeat. The transformer network was developed in MATLAB for the first time and can be obtained from here: https://github.com/malkhodari/Transformer_MATLAB.git.

References

[1] W. Wu, J. He, and X. Shao, “Incidence and mortality trend of congenital heart disease at the global, regional, and national level, 1990–2017,” *Medicine*, vol. 99, no. 23, 2020, Art. no. e20593.

[2] M. S. Zimmerman et al., “Global, regional, and national burden of congenital heart disease, 1990–2017: A systematic analysis for the global burden of disease study 2017,” *Lancet Child Adolesc. Health*, vol. 4, no. 3, pp. 185–200, 2020.

[3] M. E. Oster et al., “Temporal trends in survival among infants with critical congenital heart defects,” *Pediatrics*, vol. 131, no. 5, pp. e1502–e1508, 2013.

[4] M. Zimmerman and C. Sable, “Congenital heart disease in low-and-middle-income countries: Focus on sub-saharan africa,” *Amer. J. Med. Genet. Part C: Seminars Med. Genet.*, vol. 184, pp. 36–46, 2020.

[5] S. Rahman et al., “Linking world bank development indicators and outcomes of congenital heart surgery in low-income and middle-income countries: Retrospective analysis of quality improvement data,” *BMJ Open*, vol. 9, no. 6, 2019, Art. no. e028307.

[6] H. Higashi et al., “The burden of selected congenital anomalies amenable to surgery in low and middle-income regions: Cleft lip and palate, congenital heart anomalies and neural tube defects,” *Arch. Dis. Childhood*, vol. 100, no. 3, pp. 233–238, 2015.

[7] R. Lytzen et al., “Live-born major congenital heart disease in Denmark: Incidence, detection rate, and termination of pregnancy rate from 1996 to 2013,” *JAMA Cardiol.*, vol. 3, no. 9, pp. 829–837, 2018.

[8] G. Mcleod et al., “Echocardiography in congenital heart disease,” *Prog. Cardiovasc. Dis.*, vol. 61, no. 5/6, pp. 468–475, 2018.

[9] J. S. Chorba et al., “Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform,” *J. Amer. Heart Assoc.*, vol. 10, no. 9, 2021, Art. no. e019905.

[10] Z. Hoodbhoy et al., “Diagnostic accuracy of machine learning models to identify congenital heart disease: A meta-analysis,” *Front. Artif. Intell.*, vol. 4, 2021, Art. no. 708365.

[11] M. Alkhodari and L. Fraiwan, “Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings,” *Comput. Methods Programs Biomed.*, vol. 200, 2021, Art. no. 105940.

[12] J. Burns, M. Ganigara, and A. Dhar, “Application of intelligent phonocardiography in the detection of congenital heart disease in pediatric patients: A narrative review,” *Prog. Pediatr. Cardiol.*, vol. 64, 2022, Art. no. 101455.

[13] S. Aziz et al., “Phonocardiogram signal processing for automatic diagnosis of congenital heart disorders through fusion of temporal and cepstral features,” *Sensors*, vol. 20, no. 13, 2020, Art. no. 3790.

[14] A. A. Sepehri, A. Kocharian, A. Janani, and A. Gharehbaghi, “An intelligent phonocardiography for automated screening of pediatric heart diseases,” *J. Med. Syst.*, vol. 40, pp. 1–10, 2016.

[15] J.-K. Wang et al., “Automatic recognition of murmurs of ventricular septal defect using convolutional recurrent neural networks with temporal attentive pooling,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020.

[16] B. Bozkurt, I. Germanakis, and Y. Stylianou, “A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection,” *Comput. Biol. Med.*, vol. 100, pp. 132–143, 2018.

[17] P. Hamet and J. Tremblay, “Artificial intelligence in medicine,” *Metabolism*, vol. 69, pp. S36–S40, 2017.

[18] O. Tutsoy, “Pharmacological, non-pharmacological policies and mutation: An artificial intelligence based multi-dimensional policy making algorithm for controlling the casualties of the pandemic diseases,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9477–9488, Dec. 2022.

[19] M. Alkhodari et al., “Deep learning predicts heart failure with preserved, mid-range, and reduced left ventricular ejection fraction from patient clinical profiles,” *Front. Cardiovasc. Med.*, vol. 8, 2021, Art. no. 755968.

[20] J. J. G. Ortiz, C. P. Phoo, and J. Wiens, “Heart sound classification based on temporal alignment techniques,” in *Proc. Comput. Cardiol. Conf.*, 2016, pp. 589–592.

[21] M. Daibo, “Toroidal vector-potential transformer,” in *Proc. 11th Int. Conf. Sens. Technol.*, 2017, pp. 1–4.

[22] M. Gjoreski et al., “Machine learning and end-to-end deep learning for the detection of chronic heart failure from heart sounds,” *IEEE Access*, vol. 8, pp. 20313–20324, 2020.

[23] O. Tutsoy and M. Y. Tanrikulu, “Priority and age specific vaccination algorithm for the pandemic diseases: A comprehensive parametric prediction model,” *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, 2022, Art. no. 4.

[24] T. Yao et al., “Wave-Vit: Unifying wavelet and transformers for visual representation learning,” in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 328–345.

[25] M. A. Reyna et al., “Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022,” *PLoS Digit. Health*, vol. 2, no. 9, 2023, Art. no. e0000324.

[26] J. Oliveira et al., “The circor digiscope dataset: From murmur detection to murmur classification,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 6, pp. 2524–2535, 2021.

[27] P. Abry and P. Flandrin, “On the initialization of the discrete wavelet transform algorithm,” *IEEE Signal Process. Lett.*, vol. 1, no. 2, pp. 32–34, Feb. 1994.

[28] A. Silik, M. Noori, W. A. Altabey, and R. Ghiasi, “Selecting optimum levels of wavelet multi-resolution analysis for time-varying signals in structural health monitoring,” *Struct. Control Health Monit.*, vol. 28, no. 8, 2021, Art. no. e2762.

[29] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[30] T. Wolf et al., “Transformers: State-of-the-art natural language processing,” in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, 2020, pp. 38–45.

[31] R. Hu, J. Chen, and L. Zhou, “A transformer-based deep neural network for arrhythmia detection using continuous ecg signals,” *Comput. Biol. Med.*, vol. 144, 2022, Art. no. 105325.

[32] C. Che et al., “Constrained transformer network for ECG signal processing and arrhythmia classification,” *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–13, 2021.

[33] P. Lu et al., “Improving classification of tetanus severity for patients in low-middle income countries wearing ecg sensors by using a CNN-transformer network,” *IEEE Trans. Biomed. Eng.*, vol. 70, no. 4, pp. 1340–1350, Apr. 2023.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[35] M. Alkhodari, S. K. Azman, L. J. Hadjileontiadis, and A. H. Khandoker, “Ensemble transformer-based neural networks detect heart murmur in phonocardiogram recordings,” in *Proc. Comput. Cardiol.*, 2022, pp. 1–4.

[36] S. Monteiro, A. Fred, and H. P. da Silva, “Detection of heart sound murmurs and clinical outcome with bidirectional long short-term memory networks,” in *Proc. Comput. Cardiol.*, 2022, pp. 1–4.

[37] A. Patwa, M. M. U. Rahman, and T. Y. Al-Naffouri, “Heart murmur and abnormal PCG detection via wavelet scattering transform & a 1D-CNN,” 2023, *arXiv:2303.11423*.

[38] S. Summerton et al., “Two-stage classification for detecting murmurs from phonocardiograms using deep and expert features,” in *Proc. Comput. Cardiol.*, 2022, pp. 1–4.

[39] A. McDonald, M. J. Gales, and A. Agarwal, “Detection of heart murmurs in phonocardiograms with parallel hidden semi-Markov models,” in *Proc. Comput. Cardiol.*, 2022, pp. 1–4.

[40] S. Parvaneh et al., “Heart murmur detection using ensemble of deep learning classifiers for phonocardiograms recorded from multiple auscultation locations,” in *Proc. Comput. Cardiol.*, 2022, pp. 1–4.

[41] Y. N. Fuadah, M. A. Pramudito, and K. M. Lim, “An optimal approach for heart sound classification using grid search in hyperparameter optimization of machine learning,” *Bioengineering*, vol. 10, no. 1, 2022, Art. no. 45.

[42] K. Bayoumy et al., “Smart wearable devices in cardiovascular care: Where we are and how to move forward,” *Nature Rev. Cardiol.*, vol. 18, no. 8, pp. 581–599, 2021.

[43] D. A. Danford, A. Martin, S. Fletcher, and C. H. Gumbiner, “Echocardiographic yield in children when innocent murmur seems likely but doubts linger,” *Pediatr. Cardiol.*, vol. 23, pp. 410–414, 2002.

[44] S. Honda, M. Yamano, and T. Kawasaki, “Unusual change in murmurs in a case of mitral valve prolapse,” *Cureus*, vol. 14, no. 8, 2022, Art. no. e28411.

[45] D. Laohaprasitiporn, T. Jiarakamolchuen, P. Chanthong, K. Durongpisitkul, J. Soongswang, and A. Nana, “Heart murmur in the first week of life: Siriraj hospital,” *J. Med. Assoc. Thailand*, vol. 88, no. Suppl 8, pp. S163–S168, 2005.

[46] M. Mirzarahimi et al., “Heart murmur in neonates: How often is it caused by congenital heart disease?,” *Iranian J. Pediatrics*, vol. 21, no. 1, 2011, Art. no. 103.

[47] Q.-m. Zhao et al., “Accuracy of cardiac auscultation in detection of neonatal congenital heart disease by general paediatricians,” *Cardiol. Young*, vol. 29, no. 5, pp. 679–683, 2019.