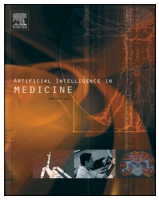
Artificial Intelligence In Medicine 153 (2024) 102867

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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



Identifying pediatric heart murmurs and distinguishing innocent from pathologic using deep learning

George Zhou [a,*], Candace Chien [b], Justin Chen [c], Lucille Luan [d], Yunchan Chen [a], Sheila Carroll [e], Jeffrey Dayton [e], Maria Thanjan [f], Ken Bayle [f], Patrick Flynn [e]

[a] Weill Cornell Medicine, New York, NY 10021, USA
[b] Children's Hospital Los Angeles, Los Angeles, CA 90027, USA.
[c] Staten Island University Hospital, Northwell Health, Staten Island, NY 10305, USA
[d] Teachers College, Columbia University, New York, NY 10027, USA
[e] Division of Pediatric Cardiology, NewYork-Presbyterian Hospital, New York, NY 10021, USA
[f] Division of Pediatric Cardiology, NewYork-Presbyterian Hospital Queens, New York, NY 11355, USA

ARTICLE INFO

Keywords:
Pediatric heart sounds
Innocent murmur
Pathologic murmur
AI
Deep learning

ABSTRACT

Objective: To develop a deep learning algorithm to perform multi-class classification of normal pediatric heart sounds, innocent murmurs, and pathologic murmurs.
Methods: We prospectively enrolled children under age 18 being evaluated by the Division of Pediatric Cardiology. Parents provided consent for a deidentified recording of their child's heart sounds with a digital stethoscope. Innocent murmurs were validated by a pediatric cardiologist and pathologic murmurs were validated by echocardiogram. To augment our collection of normal heart sounds, we utilized a public database of pediatric heart sound recordings (Oliveira, 2022). We propose two novel approaches for this audio classification task. We train a vision transformer on either Markov transition field or Gramian angular field image representations of the frequency spectrum. We benchmark our results against a ResNet-50 CNN trained on spectrogram images.
Results: Our final dataset consisted of 366 normal heart sounds, 175 innocent murmurs, and 216 pathologic murmurs. Innocent murmurs collected include Still's murmur, venous hum, and flow murmurs. Pathologic murmurs included ventricular septal defect, tetralogy of Fallot, aortic regurgitation, aortic stenosis, pulmonary stenosis, mitral regurgitation and stenosis, and tricuspid regurgitation. We find that the Vision Transformer consistently outperforms the ResNet-50 on all three image representations, and that the Gramian angular field is the superior image representation for pediatric heart sounds. We calculated a one-vs-rest multi-class ROC curve for each of the three classes. Our best model achieves an area under the curve (AUC) value of 0.92 $\pm$ 0.05, 0.83 $\pm$ 0.04, and 0.88 $\pm$ 0.04 for identifying normal heart sounds, innocent murmurs, and pathologic murmurs, respectively.
Conclusion: We present two novel methods for pediatric heart sound classification, which outperforms the current standard of using a convolutional neural network trained on spectrogram images. To our knowledge, we are the first to demonstrate multi-class classification of pediatric murmurs. Multiclass output affords a more explainable and interpretable model, which can facilitate further model improvement in the downstream model development cycle and enhance clinician trust and therefore adoption.

Abbreviations: CNN, convolutional neural network; GAF, Gramian angular field; MTF, Markov transition field; PCP, primary care provider; ROC, receiver operating curve; ViT, Vision transformer.

* Corresponding author.
E-mail address: gez4001@med.cornell.edu (G. Zhou).

https://doi.org/10.1016/j.artmed.2024.102867
Received 26 June 2023; Received in revised form 2 April 2024; Accepted 3 April 2024
Available online 4 April 2024
0933-3657/

1. Introduction

1.1. Clinical background

An estimated 66 % of all children will have heart murmurs at some point during their childhood, yet only 1.5–2 % of children are born with congenital heart disease every year [1–4]. Evaluation for a murmur is one of the most common reasons for referral to a pediatric cardiologist. Up to 60 % of the murmurs referred will be diagnosed as innocent murmurs [5]. By definition, innocent murmurs are physiologic; the presence of an innocent murmur is not indicative of an underlying structural or physiological abnormality. A significant majority of innocent murmurs will be the Still's murmur, a characteristic low-pitched, musical murmur caused by the resonation of blood in the left ventricular outflow tract [6]. Other common innocent murmurs include pulmonary and systolic flow murmurs, which are caused by normal blood flow through the heart, and venous hums, a distinct sound caused by the flow of blood returning through the veins above the heart. Pathologic murmurs, by contrast, vary widely in their identifying characteristics; they may be systolic or diastolic, harsh or quiet, have a crescendo-decrescendo quality, or be uniform in volume throughout the cardiac cycle.

The gold standard of diagnosis is echocardiogram, but auscultation is the first step that a clinician will take to evaluate a pediatric heart murmur [7]. Auscultation is a clinical skill that is highly dependent on the user. Auscultation in children is especially challenging, complicated by high heart rates which make it difficult to differentiate between systole and diastole, and by movement and crying, particularly in infants. Primary care providers (PCPs) and general practitioners, especially less experienced clinicians, often have difficulty distinguishing pediatric heart murmurs reliably and accurately. Multiple studies have shown that primary care providers have lower accuracy and wider variability in diagnosing innocent murmurs compared to pediatric cardiologists [8–10]. As a result, many PCPs will refer a child with an innocent murmur for evaluation by a pediatric cardiologist, even in the absence of symptoms. While timely diagnosis of a pediatric heart murmur is critical for the early diagnosis of congenital heart disease, prevention of anxiety and resource expenditure associated with unnecessary murmur referrals is also of high concern. Thirty to 75 % of murmur referrals will eventually be diagnosed as innocent [11]. In the United States, this amounts to up to 800,000 children referred to pediatric cardiologists for innocent heart murmurs in the US each year [12]. These referrals pose a significant burden of care, resulting in up to half a billion spent per year on unnecessary imaging [12,13].

1.2. Literature review

Automated interpretation of heart sounds has become a growing field of interest to aid accurate classification of pediatric murmurs. Previous studies on pediatric murmur classification have developed models for binary classification, typically for normal vs pathologic classification, or normal vs some targeted pathology [14–20]. For example, Liu et al. focuses on the detection of left-to-right shunts [18]. Gharehbaghi et al. focuses on the detection of ejection murmurs [19]. Wang et al. focuses on the detection of ventricular septal defects [20]. The major limitation of these previous pediatric murmur classification studies is that their datasets lack a representative distribution of pediatric heart sounds; these specific pathologies represent a minor proportion of children with murmurs. While for adult heart sounds, the distinction between no murmur versus pathologic is sufficient, pediatric heart sounds are unique in that the majority of children presenting with a murmur will have an innocent murmur, and there are several different types of innocent murmurs [6]. To address the limitations of these previous studies, we sought to create a more comprehensive dataset that captures the wide range of heart sounds that can be encountered in real world clinical practice, including normal heart sounds, innocent murmurs, and pathologic murmurs.

Only three studies so far have incorporated innocent murmurs into their dataset and attempted to differentiate the innocent murmurs as their own separate class [7,21,22]. Notably, two of the studies are from the same author group. All three studies, like the other previous studies, only propose a model for binary classification. Shekhar and Kang et al. and develop a binary classifier for differentiating Still's vs non-Stills murmurs [7,21]. The non-Still's murmur category groups together normal and pathologic sounds, which is not clinically useful. DeGroff et al. propose a binary classifier for differentiating innocent versus pathologic heart murmur but do not train the model with any normal heart sounds [22]. Additionally, all three studies lack venous hum, a common innocent murmur. We are the first to propose a model for multiclass classification of pediatric heart sounds into normal, innocent, and pathologic murmurs, which is a more clinically relevant classification problem.

The application of deep learning for the automated classification of auscultated heart sounds has been an active area of research, partially fueled by the availability of public data bases of adult heart sounds [23,24]. A variety of different deep learning models and preprocessing methods have been proposed and studied. For example, Latif et al. studied using recurrent neural networks (RNN) with Mel-frequency cepstrum coefficients (MFCC) [25]. Khan et al. studied using long-short term memory (LSTM) networks with MFCC [26]. Yang et al. studied using RNN with the 1D time series signals [27]. Raza et al. studied using LSTM networks with the 1D time series signal [28]. Chorba et al., Ryu et al., Xu et al., Humayun et al., Xiao et al., Oh et al., Baghel et al. all studied using various 1D convolutional neural networks (CNN) architectures with the 1D time series signal [29–35]. Deperlioglu et al. studied using autoencoder networks (AEN) with the 1D time series signal [36]. Sun et al. studied using a Gaussian mixture model (GMM) with features derived from a short time modified Hilbert transform [37]. Among these methods, what has defined state of the art have been the use of 2D CNN on spectrogram images, with different variations of this method studied by Demir et al., Nilanon et al., Zhou et al., Dominguez et al., Cheng et al., Maknickas et al., Alafif et al., and Rubin et al. [38–45] All these methods have been proposed and validated on the classification of adult heart sounds.

Methods specifically proposed and validated on pediatric heart sounds include the following. Kotb et al. studied hidden Markov models with MFCC [14]. Pretorius et al. studied using artificial neural networks (ANN) with the input to the model being a handcrafted feature vector consisting of features derived from the short-time Fourier transform and Shannon energy envelop [15]. Wang at el. studied ANN with a handcrafted feature vector consisting of features derived from the power spectral density and the Shannon energy envelop [16]. Xiao et al. studied using a 1D CNN with the 1D time series signal [17]. Liu et al. studied using a hybrid CNN and RNN model with the 1D time series signal [18]. Gharehbaghi proposed a novel “time growing neural network” [19]. Wang and Shekhar et al. studied using a 2D CNN on spectrogram images [7,20]. Kang and DeGroff et al. studied using ANN and support vector machine (SVM), respectively, on a handcrafted feature vector [21,22]. Sepehri et al. [62,63] employ a method that finds the frequency bands that provide the lowest error in clustering instances of disease against normal heart sounds, which the authors term the “Arash-band.” The Arash-bands are then used the feature vector input into a support vector machine for binary classification of normal versus pathologic. Determining the Arash-band is a manual procedure that must be done for each specific pathology.

1.3. Multiclass classification using frequency spectrum image encodings

Our study aims to provide a novel, multiclass deep learning model for classifying pediatric heart sounds by introducing the Markov transition field (MTF) and Gramin Angular Field (GAF) frequency spectrum image encoding as input to the vision transformer computer vision

model. The MTF and GAF frequency spectrum image encodings represent two new preprocessing methods for producing a two-dimensional image representation of sound from the one-dimensional (i.e., univariate) audio signal. The first step for both preprocessing methods is to apply a Fourier Transform to the univariate timeseries signal to obtain the frequency spectrum. The frequency spectrum represents the audio signal in terms of its component frequencies with amplitude conveyed on the y-axis and frequency convey on the x-axis. A two-dimensional image representation is then derived from the audio frequency spectrum via a MTF or a GAF [45]. While the MTF and GAF were originally developed by Wang et al., for spatially encoding timeseries data (i.e., univariate sequence data indexed by time), our contribution is to show that the MTF and GAF can be extended to univariate sequence data indexed in the frequency domain for audio classification. The MTF treats univariate sequence data as a first-order Markov chain and depicts the transition probabilities for all pairwise sets of discretized values. The GAF visualizes a Gram matrix derived from polar encoded univariate sequence data. For computer vision tasks such as this one, CNNs have been the de facto standard. For example, Wang et al. employed tiled CNN for classifying MTF and GAF image representation of timeseries data. We propose using the Vision Transformer (ViT) [47] for classifying MTF and GAF image representations of the frequency spectrum for pediatric heart sound classification. Fig. 1 shows a schematic of our proposed method. We benchmark our results against a ResNet-50 CNN with spectrogram images as the input.

2. Material and methods

2.1. Data collection

We prospectively enrolled children under age 18 being evaluated by the Division of Pediatric Cardiology at two affiliate sites: NewYork-Presbyterian Hospital and NewYork- Presbyterian Queens Hospital. Parents provided consent to a de-identified recording of their child's heart sound. Demographic data were not collected in order to maintain the privacy of the child. Due to the rarity of some pathologies presenting to our institution, collecting demographic data such as age, and identifying the center of treatment could potentially reveal the patient's identity. Auscultated heart sounds were recorded with a 3 M Littmann Core digital stethoscope at a sampling rate of 4000 Hz. A label of "normal", "innocent", or "pathologic" was given by board-certified pediatric cardiologists. All patients referred to our pediatric cardiology center receive an echocardiogram (the gold standard) to confirm or deny the presence of pathology. We also supplemented our collected sounds with additional "normal" and "pathologic" sounds from the CirCor DigiScope dataset, a publicly available database of pediatric heart sounds collected in Brazil [48].

Our study included a total of 138 patients, 73 patients from our own data collection and 65 patients from CirCor DigiScope dataset. Distribution of heart sounds by dataset is shown in Table 1. Each recording varied between 15 and 60 s long. We split each recording into 5 s clips to maximize the number of samples. To prevent data leakage, the training, validation, and testing splits were done on the patient level, meaning samples sourced from the same patient would appear in the same split. The final dataset included 742 pediatric heart sounds in total: 366 normal heart sounds, 175 innocent murmurs, and 216 pathological

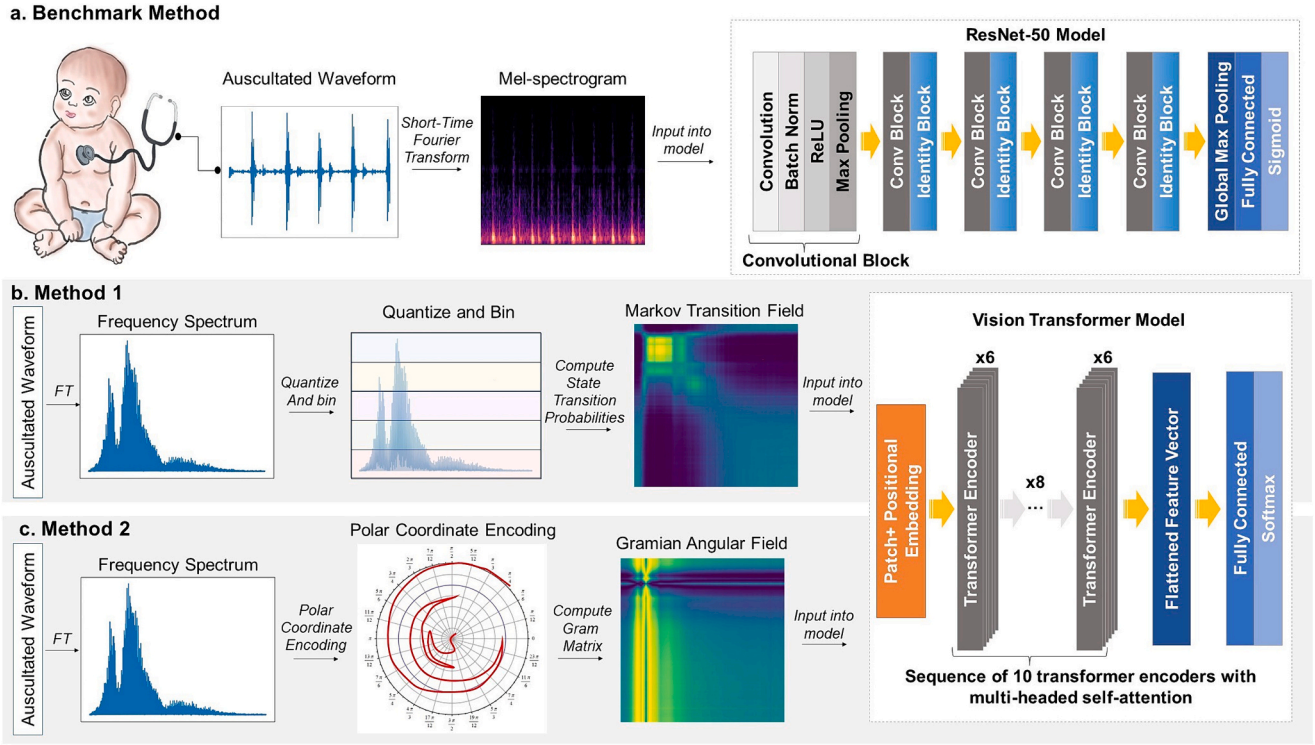


Fig. 1. Schematic of our proposed methods. We benchmark our results against **a)** a ResNet-50 CNN trained on spectrogram images. The first step of both methods is to apply the Fourier Transform (FT) to the auscultated timeseries data to produce the frequency spectrum. Then an image representation of the frequency spectrum is generated. In **b)** Method 1, the frequency spectrum is quantized and binned into discrete states. Viewing the binned frequency spectrum as a first-order Markov chain, each bin represents a distinct state. The Markov Transition Field (MTF) visualizes the Markov transition probability matrix as an image. In **c)** Method 2, the frequency spectrum is mapped onto the Polar coordinate system. The Gram matrix is calculated from the polar coordinate encoded frequency spectrum and the Gramian Angular Field (GAF) visualizes the Gram matrix as an image. The image representations are then used to train a Vision transformer (ViT) model which consists of a sequence of 10 transformer encoders followed by one fully connected layer. The final activation function used was either a sigmoid or softmax activation depending on the task being binary or multiclass classification.

Table 1
Distribution of heart sound samples by dataset.

Heart sounds	CirCor DigiScope ($N = 340$)	NewYork-Presbyterian ($N = 402$)
Normal	305 (90 %)	61 (15 %)
Innocent	0 (0 %)	175 (44 %)
Pathologic[a]	35 (10 %)	181 (45 %)

[a] Pathologic sounds from CirCor DigiScope unspecified diastolic murmurs.

murmurs. Innocent murmurs included Still's murmur, flow murmurs, and venous hums. Pathologic murmurs included ventricular septal defect (VSD), mitral regurgitation, mitral stenosis, pulmonary stenosis, pulmonary regurgitation, Tetralogy of Fallot (TOF), aortic stenosis, aortic regurgitation, and subaortic stenosis. Distribution of murmur type / underlying diagnosis is shown in Table 2. Of note, pathologic heart sounds from CirCor DigiScope were diastolic murmurs that were otherwise unspecified.

Three different classification problems are studied: binary classification of pediatric heart sounds as murmur absent vs murmur present (innocent and pathologic), binary classification of pediatric heart murmurs as innocent versus pathologic given prior information that a murmur is present (i.e., the normal sounds were excluded in this case study), and multiclass classification of pediatric heart sound as normal, innocent murmur, and pathologic murmur.

2.2. Preprocessing

We preprocess our one-dimensional audio signal timeseries data into two-dimensional image representations to spatially encode the audio features. Specifically, we study the Mel-spectrogram, Markov transition field (MTF), and the Gramian angular field (GAF) for spatially encoding the audio features [46]. A spectrogram depicts the spectrum of frequencies of a signal as it varies with time. The x-axis represents time, the y-axis represents frequency, and amplitude of a particular frequency component at a given point in time is represented by the intensity of colour. The spectrograms are generated from the pediatric heart sounds using short-time Fourier transforms as follows. First, the audio signals are windowed using a Hann window of size 512 and a hop length of 256. A 512-point fast Fourier transform is applied to each window to generate a spectrogram. The Mel-scaled, dB-scaled spectrograms are generated by logarithmic rescaling of the amplitude and frequency axis. The amplitude axis is converted to the dB scale. The frequency axis is transformed onto the Mel scale, characterized by Eq. (1),

Table 2
Distribution of prospectively collected heart sound samples by diagnosis.

Diagnosis	Total (N = 402)
Normal	61
Innocent	175
Still's murmur	141 (81 %)
Flow murmur	24 (14 %)
Venous hum	10 (5.7 %)
Pathologic	181
Ventricular septal defect	57 (31 %)
Mitral regurgitation	35 (19 %)
Aortic stenosis	27 (15 %)
Mitral valve prolapse	12 (6.6 %)
Tetralogy of Fallot (pulmonary stenosis, pulmonary regurgitation)	9 (5.0 %)
Sub-Aortic membrane with aortic stenosis	8 (4.4 %)
Mitral stenosis	7 (3.9 %)
Tricuspid regurgitation	6 (3.3 %)
Pulmonary stenosis, pulmonary regurgitation, unspecified	5 (2.8 %)
Pulmonary stenosis	5 (2.8 %)
Sub-Aortic Stenosis	5 (2.8 %)
Hypoplastic left heart syndrome with aortic insufficiency	5 (2.8 %)

$$Mel = 2595^{*}log\left(1+\frac{f}{500}\right) \tag{1}$$

where f is frequency in Hz. The resulting Mel-scaled, dB-scaled spectrograms resized to be 100 × 100 (time resolution x frequency resolution) in size using bicubic interpolation. Here, brighter colors correspond to greater intensity or amount of a given frequency component, and darker colors correspond to lower intensity or amount.

The MTF treats one-dimensional sequence data as a first-order Markov chain and depicts the transition probabilities for all pairwise sets of discretized values. For our pediatric heart sounds, we generate MTF image representations of the audio signal in the frequency domain. First, we apply the Fourier transform to the pediatric heart sound timeseries data to obtain the frequency spectrum. The frequency spectrum is discretized into $Q = 5$, *10,* and *15* distinct bins along the different possible values that can be assumed, with the first and last bin corresponding to the highest and lowest possible frequency value ranges, respectively. We use a quantile binning strategy so that each bin contains the same number of points. Viewing the discretized frequency spectrum as a first-order Markov chain, each bin represents a distinct state. A Q x Q Markov transition matrix is computed by quantifying the number of state transitions between all pairwise sets of states. (i.e. the diagonal of the Markov transition matrix represents self-transition probabilities). Mathematically, this can be stated as follows. Let $F = \{f_0, f_1, f_2, \ldots f_i \ldots f_N\}$ represent the discretized points over which the frequency spectrum spans N timestamps such that the frequency at time t_i with is given by the value f_i. Each value f_i is mapped to a bin or state q_j, where $j \in [1,Q]$. The QxQ Markov transition matrix M is defined by Eq. (2):

$$M = \begin{bmatrix} m_{1,1} = P(f_i \in q_1 | f_{i-1} \in q_1) & \ldots & m_{1,Q} = P(f_i \in q_1 | f_{i-1} \in q_Q) \\ m_{2,1} = P(f_i \in q_2 | f_{i-1} \in q_1) & \ldots & m_{2,Q} = P(f_i \in q_2 | f_{i-1} \in q_Q) \\ \vdots & \ddots & \vdots \\ m_{Q,1} = P(f_i \in q_Q | f_{i-1} \in q_1) & \ldots & m_{Q,Q} = P(f_i \in q_Q | f_{i-1} \in q_Q) \end{bmatrix} \tag{2}$$

where $m_{i,j}$ represents the frequency count with which a frequency value in bin q_j is followed by a frequency value in the bin q_i. Transition probabilities are derived by normalizing the Markov transition matrix: $\sum_{i=1}^{Q}\sum_{j=1}^{Q} m_{i,j} = 1$. Finally, the MTF is a visual depiction of the Markov transition probabilities where brighter colors correspond to higher transition probabilities and darker colors correspond to lower transition probabilities. The resulting MTF images are resized to be 100 × 100 using bicubic interpolation.

The GAF visualizes a quasi-Gram matrix derived from one-dimensional sequence data. For our pediatric heart sounds, we generate GAF image representations of the audio signal in the frequency domain. A Gram matrix is a matrix of all possible pairwise sets of inner products. The term "quasi-Gram matrix" is used here because the resulting matrix that is visualized is a version of the Gram matrix that uses a modified definition of the inner product as explained below. First, we apply the Fourier transform to the pediatric heart sound timeseries data to obtain the frequency spectrum. The Gram matrix calculates inner products of vectors in a 2D space; therefore, the frequency spectrum is first mapped onto the Polar coordinate system. Again, let $F = \{f_0, f_1, f_2, \ldots f_i \ldots f_N\}$ represent the discretized points over which the frequency spectrum spans N timestamps such that the frequency at time t_i with is given by the value f_i. The frequency spectrum is mapped onto Polar coordinate system according to Eq. (3):

$$\begin{cases} \theta_i = cos^{-1}(f_i) \\ r = \dfrac{t_i}{N}, i \in N \end{cases} \tag{3}$$

Now in 2D space, the Gram matrix can be derived. One of the limitations of the inner product in 2D polar space is that the norm of each

vector is adjusted for the frequency dependency, meaning the inner product will be biased towards the higher frequency component. To address this issue, the original authors proposed using either a trigonometric sum or difference between each vector pair. We study both variations where the final matrix that is derived uses either the trigonometric difference of two vector pairs $sin(\theta_i - \theta_j)$ or the trigonometric summation of two vector pairs $cos(\theta_i + \theta_j)$, *where* $i,j \in N$ (hence the term "quasi-Gram matrix"). The NxN quasi-Gram matrix G is defined by Eq. (4):

$$G = \begin{bmatrix} < f_1, f_1 > & \cdots & < f_1, f_N > \\ < f_2, f_1 > & \ldots & < f_1, f_1 > \\ \vdots & \ddots & \vdots \\ < f_n, f_1 > & \ldots & < f_n, f_n > \end{bmatrix} \quad (4)$$

where the inner product $< u, v >$ is redefined to be $< u, v >= \sqrt{1-u^2} \bullet v - u \bullet \sqrt{1-v^2}$ for the trigonometric difference of two vectors or $< u, v >= u \bullet v - \sqrt{1-u^2} \bullet \sqrt{1-v^2}$ for the trigonometric summation of two vectors. The full mathematical derivation for how to arrive at this modified inner product definition for the trigonometric difference of two vector pairs is shown in Appendix A. The modified inner product definition for the trigonometric summation of two vector pairs is derived analogously. Finally, the GAF visualizes this quasi-Gram matrix with brighter colors corresponding to larger inner products and darker colors corresponding to smaller inner products. The resulting GAF images are resized to be 100 × 100 using bicubic interpolation. All image representations (spectrograms, MTF, GAF) are normalized prior to input into the model into the range [−1,1].

2.3. Models

We study the ResNet-50 convolutional neural network (CNN) and a vision transformer (ViT) for classifying the pediatric heart sounds based on the preprocessed image representations. Briefly, the ResNet-50 consists of 5 blocks, with each block consisting of a convolutional layer, batch normalization layer, ReLu activation layer, a max pooling layer, and residual connections that allow activations from earlier layers to be propagated down to deeper layers [46]. The final output from the last layer is reshaped into a flattened feature vector using global max pooling, which is fed into a fully connected layer for classification. In the case of binary classification, the final fully connected layer consists of a single node with sigmoid activation function. In the case of multi-class classification, the fully connected layer consists of three nodes with softmax activation function. The number of parameters for the ResNet-50 CNN is 23.6 M. The model is trained using an adaptive moment estimation (Adam) optimizer at a learning rate of 1×10^{-3} over the binary cross entropy loss function in the case of binary classification and over the categorical cross entropy loss function in the case of multi-class classification. A batch size of 64 is used. We investigate the ResNet-50 model performance with both randomly initialized weights and with transfer learning using ImageNet pretrained weights.

For this ViT model, first the input image is tokenized into 10 by 10 patches [47]. The patches are flattened and linearly projected (i.e., multiplied by a learnable weight matrix) into a feature vector. A positional encoding is added to each linear projected patch; the positional encoding is a learnable embedding. The linearly projected patches with their corresponding positional encodings are fed into a sequence of 10 transformer encoder layers [49]. Each transformer encoder layer is comprised of 2 subcomponents. The first subcomponent consists of a layer normalization followed by the multi-headed self-attention layers. For the ViT in this study, we use 6 attention heads. The second subcomponent of each transformer encoder consists of another layer normalization followed by a 2-layer fully connected network using ReLU activation function. Skip or residual connections are used to propagate feature vector representations between each subcomponent of each transformer encoder layer. The final output from the last transformer encoder layer is reshaped into a flattened feature vector, which is then fed into a fully connected layer for classification. In the case of binary classification, the final fully connected layer consists of a single node with sigmoid activation function. In the case of multi-class classification, the fully connected layer consists of three nodes with softmax activation function. The number of parameters for the ViT-B16 is 87.5 M. The model is trained using Adam optimizer at a learning rate of 1×10^{-3} over the binary cross entropy loss function in the case of binary classification and over the categorical cross entropy loss function in the case of multi-class classification. A batch size of 64 is used. We investigate the ViT-B16 model performance with both randomly initialized weights and with transfer learning using ImageNet pretrained weights.

All code was written in Google Colab notebooks using Python version 3.10.12 and Pytorch version 2.1.0 + cu118. Model training is completed on A100 GPUs. Generating the various frequency spectrum image encodings is compute intensive and requires at least 25GM CPU RAM.

2.4. Data splitting and training procedure

For training and evaluating the models, we use 5-fold cross-validation. The dataset is split into 5 equal folds. For each iteration, one fold serves as the test set (20 %) and the remaining four parts serves as the training set (80 %). Within the training set, 15 % of the data is randomly held out to serve as the validation set and remaining 65 % is utilized to directly train the model. Early stopping is us employed such that training is terminated if the validation loss does not improve after 20 epochs. The model weights corresponding with the lowest validation loss is used for evaluation on the test set. This procedure is repeated until each fold serves as the testing set exactly once.

3. Results

3.1. Dataset

Tables 1 and 2 display the quantity of each diagnosis type included in our dataset used to train and evaluate our models.

3.2. Frequency spectrum image encodings illustrative examples

Fig. 2 displays illustrative examples of the Mel-spectrogram, MTF, and GAF image representations for various pediatric heart sounds. Additional examples can be found in the supplemental information (Appendix B).

3.3. Murmur present versus murmur absent binary classification

Fig. 3 displays the five-fold cross-validation ROC and PR curves for murmur absent versus murmur present (innocent or pathologic) binary classification for the ResNet-50 models and the ViT-B16 models using transfer learning with ImageNet pretrained weights trained on each image representation: Mel-spectrogram, GADF, and MTF with a bin count of 10, respectively. Table 3 lists the mean area under the receiver-operating characteristic curve (AuROC) and mean area under the precision-recall curve (AuPRC) for each model and preprocessing combination, with and without transfer learning using ImageNet pretrained weights. The AuROC and AuPRC values for each individual fold and additional experiments investigating the Gramian angular summation field, MTF with bin counts of 5 and 15, along with models initialized using random weights, can be found in Appendix C.

3.4. Innocent versus pathologic murmur binary classification

Fig. 4 displays the five-fold cross-validation ROC and PR curves for innocent versus pathologic murmur binary classification given prior information that a murmur exists (i.e. normal heart sounds were excluded from the dataset) for the ResNet-50 models and the ViT-B16

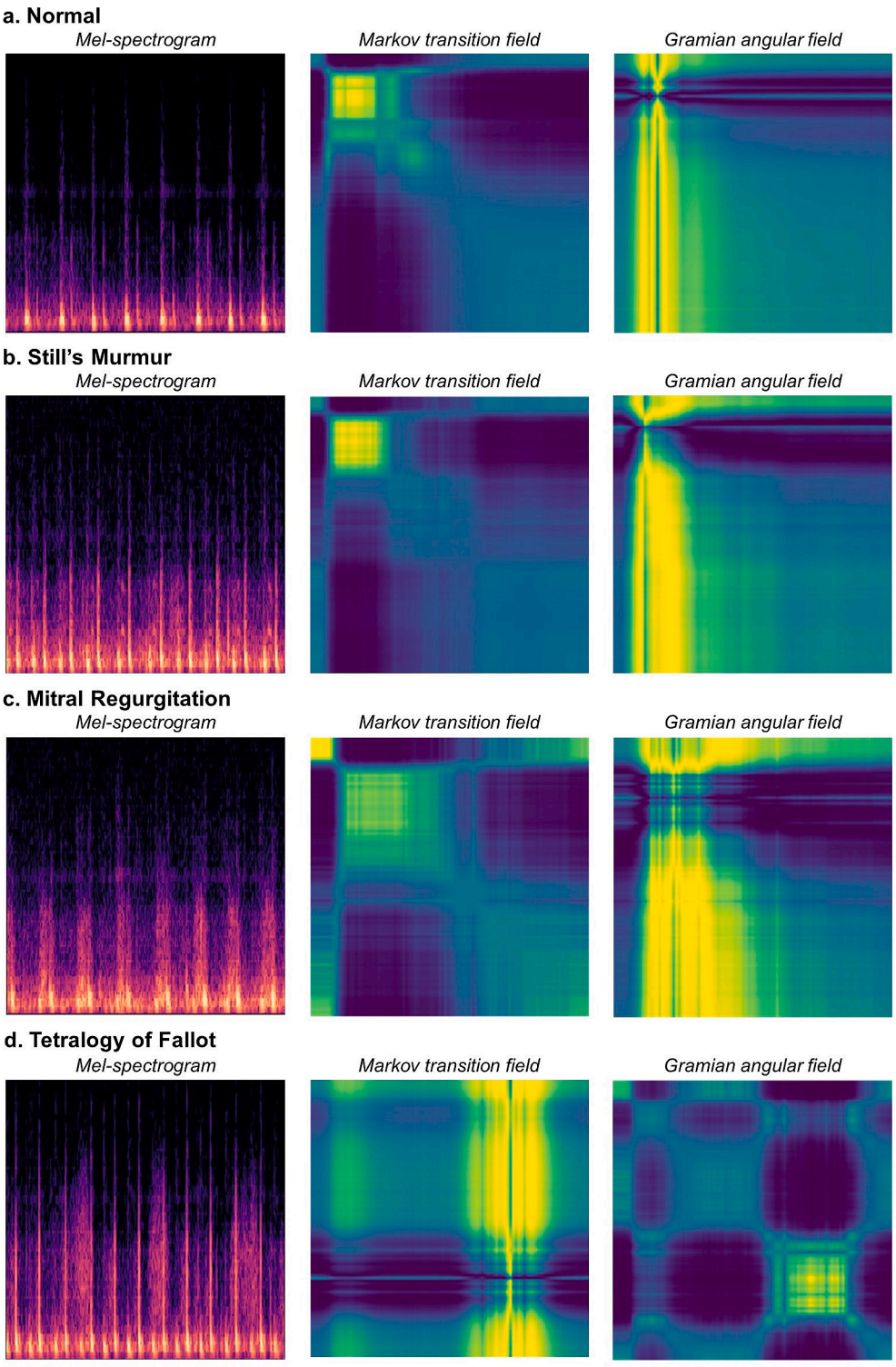


Fig. 2. The Mel-spectrogram (left), Markov transition field (middle), and Gramian angular field (right) image representations for a) normal pediatric heart sound, b) Still's murmur (innocent), c) mitral regurgitation (pathologic), and d) pulmonary stenosis and regurgitation due to Tetralogy of Fallot (pathologic).

models using transfer learning with ImageNet pretrained weights trained on each image representation: Mel-spectrogram, GADF, and MTF with a bin count of 10, respectively. Table 4 lists the mean AuROC and mean AuPRC for each model and preprocessing combination, with and without transfer learning using ImageNet pretrained weights. The AuROC and AuPRC values for each individual fold and additional experiments investigating the Gramian angular summation field, MTF with bin counts of 5 and 15, along with models initialized using random weights, can be found in Appendix C.

3.5. *Innocent versus pathologic versus no murmur multiclass classification*

Fig. 5 displays the five-fold cross-validation ROC and PR curves extended to one-vs-rest multiclass classification (innocent vs pathologic vs none) for the ResNet-50 models and the ViT-B16 models using transfer learning with ImageNet pretrained weights trained on each

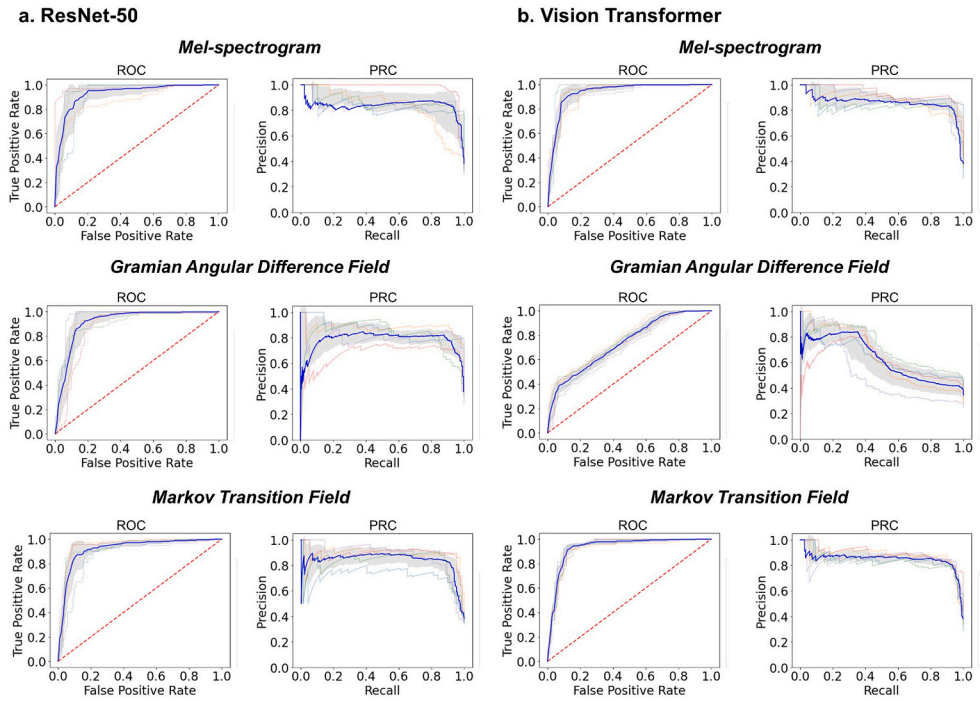


Fig. 3. Binary classification of pediatric heart sounds as murmur absent or murmur present (innocent or pathologic). Five-fold cross-validation ROC and PR curves are shown for the **a)** ResNet-50 model pretrained on ImageNet (left) and the **b)** ViT-B16 model pretrained on ImageNet (right). Each model is trained on each image representation: the Mel-spectrogram (top), the Gramian angular difference field (middle), and the Markov transition field (MTF) with a bin count of 10 (bottom). For the ROC curves, the line of no-discrimination is shown as a dotted red line, and for both curves ±1 standard deviations are shown as the gray shaded region. The "murmur present" class is treated as the positive class. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Summary statistics for murmur present versus murmur absent binary classification.

Preprocessing	Transfer learning with ImageNet pretrained weights	Model architecture			
		ResNet-50		ViT-B16	
		Mean AuROC	Mean AuPRC	Mean AuROC	Mean AuPRC
Mel-Spectrogram	*No*	0.90 ± 0.06	0.84 ± 0.11	0.92 ± 0.04	0.81 ± 0.09
	Yes	0.93 ± 0.04	0.84 ± 0.06	0.94 ± 0.02	0.86 ± 0.03
Gramian Angular Difference Field	*No*	0.88 ± 0.06	0.81 ± 0.15	0.90 ± 0.03	0.82 ± 0.06
	Yes	0.92 ± 0.03	0.78 ± 0.07	0.74 ± 0.03	0.63 ± 0.07
Markov Transition Field *(Bins = 10)*	*No*	0.94 ± 0.05	0.77 ± 0.10	0.92 ± 0.04	0.83 ± 0.05
	Yes	0.92 ± 0.03	0.84 ± 0.07	0.93 ± 0.01	0.85 ± 0.02

image representation: Mel-spectrogram, GADF, and MTF with a bin count of 10, respectively. Tables 5, 6, 7 list the mean AuROC and mean AuPRC for each model and preprocessing combination, with and without transfer learning using ImageNet pretrained weights, for when pathological, innocent, or normal is treated as the positive class, respectively. The AuROC and AuPRC values for each individual fold and additional experiments investigating the Gramian angular summation field, MTF with bin counts of 5 and 15, along with models initialized using random weights, can be found in Appendix C.

4. Discussion

4.1. Frequency spectrum image encodings

Using computer vision models to classify spectrogram image representations of sound has been the state-of-the-art method in audio classification. [38,50–52] In our study, we present two novel methods for classifying audio data: generating Markov transition field (MTF) and Gramian angular field (GAF) 2D image representations from an audio signal's 1D frequency spectrum for input into a neural network. The MTF views the 1D frequency spectrum as a first-order Markov chain and visualizes the transition probabilities between all pairwise set of frequency states of the (discretized) audio signal. The GAF visualizes a quasi-Gram matrix of all pairwise set of vectors after mapping the 1D frequency spectrum into polar coordinates. Convolutional neural networks (CNNs) have long been the de factor standard for computer vision tasks including spectrogram classification. In our study, we also explore how the newer vision transformer model, which implements the self-attention mechanism, performs in classifying image representations of sound, benchmarking the results against the ResNet-50 CNN.

We find that the MTF and GAF image representations either perform comparably or outperform the spectrogram image representation when

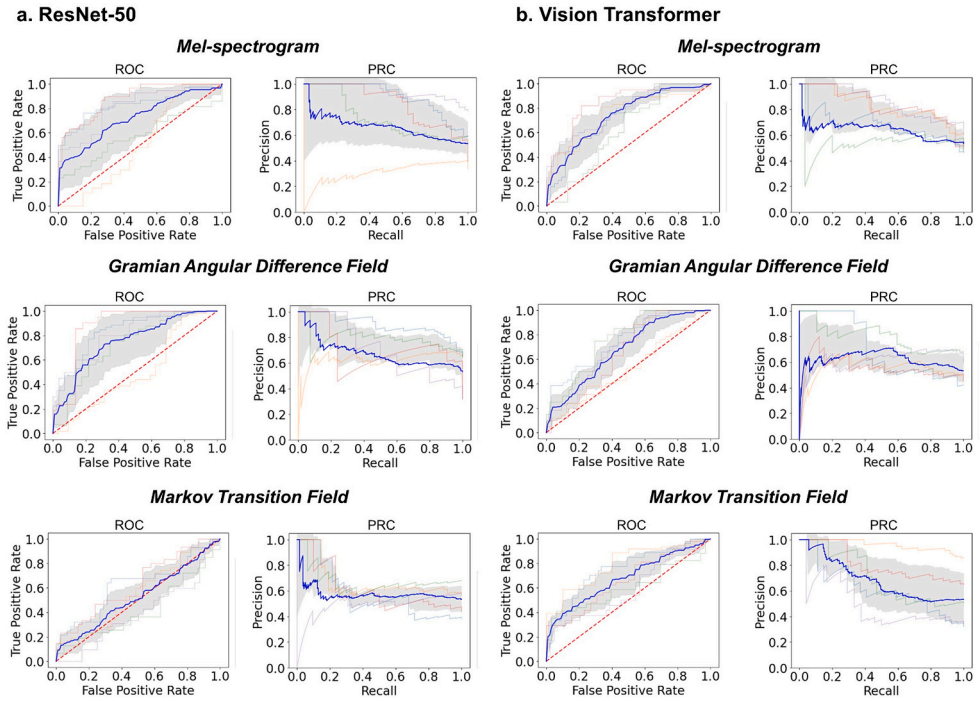


Fig. 4. Binary classification of pediatric heart sounds as innocent versus pathologic murmur. Five-fold cross-validation ROC and PR curves are shown for the **a)** ResNet-50 model pretrained on ImageNet (left) and the **b)** ViT-B16 model pretrained on ImageNet (right). Each model is trained on each image representation: the Mel-spectrogram (top), the Gramian angular difference field (middle), and the Markov transition field (MTF) with a bin count of 10 (bottom). For the ROC curves, the line of no-discrimination is shown as a dotted red line, and for both curves ±1 standard deviations are shown as the gray shaded region. The "pathologic murmur" class is treated as the positive class. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Summary statistics for innocent murmur versus pathologic murmur binary classification.

Preprocessing	Transfer learning with ImageNet pretrained weights	Model architecture			
		ResNet-50		ViT-B16	
		Mean AuROC	Mean AuPRC	Mean AuROC	Mean AuPRC
Mel-Spectrogram	*No*	0.66 ± 0.15	0.63 ± 0.12	0.72 ± 0.06	0.58 ± 0.10
	Yes	0.72 ± 0.15	0.66 ± 0.23	0.74 ± 0.07	0.64 ± 0.12
Gramian Angular Difference Field	*No*	0.72 ± 0.06	0.58 ± 0.10	0.71 ± 0.10	0.71 ± 0.18
	Yes	0.75 ± 0.14	0.68 ± 0.09	0.68 ± 0.08	0.62 ± 0.10
Markov Transition Field *(Bins = 10)*	*No*	0.48 ± 0.14	0.48 ± 0.08	0.52 ± 0.12	0.59 ± 0.10
	Yes	0.62 ± 0.10	0.63 ± 0.13	0.68 ± 0.08	0.68 ± 0.16

used to train the vision transformer (ViT) model. The spectrogram representation, which depicts frequency on the y-axis as it varies with time on the x-axis, inherently has a time-frequency resolution tradeoff. Higher frequency resolution results in less time resolution and vice versa. In contrast, the MTF and GAF image representation are generated from the audio signal in the frequency domain. Thus, the MTF and GAF image representation offer full frequency resolution at the expense of zero-time resolution. For the case of pediatric heart sound classification, this is beneficial: the frequency content is what strictly determines which class a heart sound belongs to (normal vs innocent murmur vs pathologic murmur), rather than *when* certain frequencies occur.

For cases such as deriving semantic information from speech, the order of the frequency components absolutely matters. For heart sound classification, however, temporal information is not important in determining the class to which the heart sound belongs, given the rhythmic nature of heart sounds, which has repeating frequency components (i.e., S1 and S2). Thus, the spectrogram representation has a lot of redundancy as a result of preserving temporal information due to the cyclic nature of heart sounds. The vast majority of pediatric heart sounds, regardless of class, will have an S1 and S2 component, which is not useful for differentiating between these heart sounds. In the spectrogram representation, repeating S1 and S2 frequency components visually occupies multiple regions of the image representation. Due natural variances such as recording start times and variations in heart rate, the regions occupied by S1 and S2 frequency components are different from sample to sample, which likely hinders the performance of the computer vision models. The MTF and GAF give full frequency resolution with no temporal information. Higher frequency resolution in and of itself likely improves model performance. Additionally, the S1 and S2 frequency components will more consistently occupy similar regions in the image representation; thus, the computer vision model will have an easier time learning to ignore certain regions while focusing on other regions of higher importance (i.e., the ones that provide discriminatory information). We find that transfer learning with

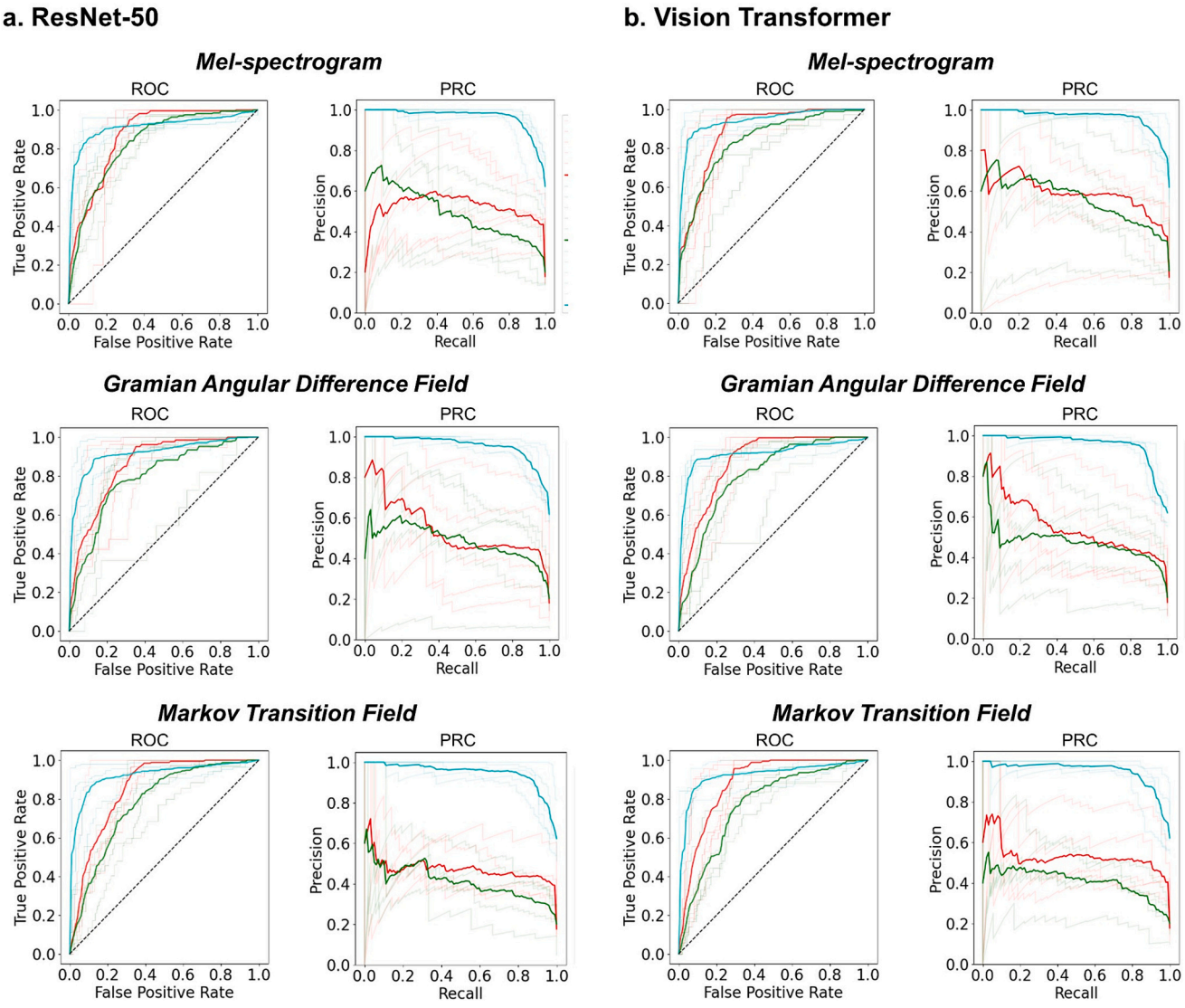


Fig. 5. One-vs-rest multiclass classification of pediatric heart sounds as innocent murmur vs. rest, pathologic murmur vs. rest, and no murmur vs. rest. Five-fold cross-validation ROC and PR curves are shown for the **a)** ResNet-50 model pretrained on ImageNet (left) and the **b)** ViT-B16 model pretrained on ImageNet (right). Each model is trained on each image representation: the Mel-spectrogram (top), the Gramian angular difference field (middle), and the Markov transition field (MTF) with a bin count of 10 (bottom). For the ROC curves, the line of no-discrimination is shown as a dotted black line, and ±1 standard deviations are shown as the gray shaded region. The ROC and PR curves are red, green, blue for when pathological, innocent, or normal is treated as the positive class, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Summary statistics for one-vs-rest multiclass classification with pathologic murmur as the positive class.

Preprocessing	Transfer learning with ImageNet pretrained weights	Model architecture			
		ResNet-50		ViT-B16	
		Mean AuROC	Mean AuPRC	Mean AuROC	Mean AuPRC
Mel-Spectrogram	*No*	0.85 ± 0.06	0.56 ± 0.25	0.87 ± 0.03	0.45 ± 0.09
	Yes	0.87 ± 0.04	0.52 ± 0.19	0.89 ± 0.06	0.59 ± 0.27
Gramian Angular Difference Field	*No*	0.71 ± 0.12	0.31 ± 0.11	0.62 ± 0.11	0.25 ± 0.06
	Yes	0.86 ± 0.07	0.54 ± 0.22	0.88 ± 0.04	0.55 ± 0.18
Markov Transition Field *(Bins = 10)*	*No*	0.85 ± 0.05	0.46 ± 0.11	0.85 ± 0.05	0.42 ± 0.14
	Yes	0.85 ± 0.05	0.47 ± 0.10	0.88 ± 0.05	0.53 ± 0.13

Table 6
Summary statistics for one-vs-rest multiclass classification with innocent murmur as the positive class.

Preprocessing	Transfer learning with ImageNet pretrained weights	Model architecture			
		ResNet-50		ViT-B16	
		Mean AuROC	Mean AuPRC	Mean AuROC	Mean AuPRC
Mel-Spectrogram	*No*	0.89 ± 0.16	0.54 ± 0.19	0.89 ± 0.16	0.54 ± 0.19
	Yes	0.83 ± 0.04	0.49 ± 0.18	0.83 ± 0.04	0.49 ± 0.18
Gramian Angular Difference Field	*No*	0.59 ± 0.15	0.28 ± 0.13	0.59 ± 0.15	0.28 ± 0.13
	Yes	0.79 ± 0.12	0.48 ± 0.23	0.79 ± 0.12	0.48 ± 0.23
Markov Transition Field *(Bins = 10)*	*No*	0.73 ± 0.08	0.36 ± 0.13	0.73 ± 0.08	0.36 ± 0.13
	Yes	0.78 ± 0.06	0.41 ± 0.09	0.78 ± 0.06	0.41 ± 0.09

Table 7
Summary statistics for one-vs-rest multiclass classification with no murmur as the positive class.

Preprocessing	Transfer learning with ImageNet pretrained weights	Model architecture			
		ResNet-50		ViT-B16	
		Mean AuROC	Mean AuPRC	Mean AuROC	Mean AuPRC
Mel-Spectrogram	*No*	0.89 ± 0.08	0.95 ± 0.24	0.89 ± 0.08	0.95 ± 0.24
	Yes	0.91 ± 0.03	0.96 ± 0.26	0.91 ± 0.03	0.96 ± 0.26
Gramian Angular Difference Field	*No*	0.76 ± 0.09	0.85 ± 0.29	0.76 ± 0.09	0.85 ± 0.29
	Yes	0.92 ± 0.05	0.95 ± 0.28	0.92 ± 0.05	0.95 ± 0.28
Markov Transition Field *(Bins = 10)*	*No*	0.92 ± 0.02	0.95 ± 0.28	0.92 ± 0.02	0.95 ± 0.28
	Yes	0.92 ± 0.05	0.95 ± 0.25	0.92 ± 0.05	0.95 ± 0.25

ImageNet pretrained weights do not afford any significant boost in performance compared to random weight initialization, consistent with the idea that the inductive biases learned from natural images do not transfer to audio data image representations.

We find that the GAF is a better image representation than the MTF for pediatric heart sound classification. We can attribute this to that fact that generating the MTF is a subjective process while generating the GAF is a bijective process. In other words, multiple different sounds can result in the same MTF image, but a GAF image representation will correspond to one and only one sound. The larger inverse image space of MTF likely hinders its performance relative to GAF. In the main manuscript we present the results from the Gramian angular difference field and the MTF with a bin count of 10. We find that performance does not significantly change when using variations on these image representations such as the Gramian angular summation field and the MTF with bin counts of 5 and 15 (Appendix C).

The strengths of using the GAF and MTF image representations of pediatric heart sounds is that it captures differences in the frequency domain dynamics which holds key diagnostic information, it is translation invariant (i.e. robust to shifts in time and phase of the heart sound signals which may vary in timing and intensity), and it affords dimensionality reduction compared to the raw signal which can reduce the computational complexity and memory requirements of models. The weaknesses of such encodings are that they are sensitive to hyperparameter choice (i.e. number of bins) which would need to be fine-tuned for a given signal and these image representations may struggle to generalize to unseen variations in murmurs or background noise patterns if the training data is limited in diversity.

4.2. Model architectures

Finally, we find that the ViT consistently outperforms the ResNet-50 across all three image representations. The convolution operator aggregates information via spatial sliding windows or kernels which use the same learned weights as it slides across an image. This architecture structurally introduces two important inductive biases inherent to CNN: translational equivariance and locality. Pooling layers, used in conjunction with convolutional layers in our models, help the model achieve translational invariance. Translational equivalence and invariance mean that an object can be detected irrespective of its location in the image. The locality bias is the notion that closely spaced pixels are more correlated than pixels that are far away.

While these image representations of sound (spectrograms, MTF, GAF) and natural images are both images from a data structure point of view (i.e., a grid of pixel values), the two images represent fundamentally different natural phenomena. The inductive biases of translational invariance and locality structurally built into the CNN architecture are not as suitable for processing and interpreting image representations of sound. While translation invariance is a good assumption for natural images whose axes convey a measure of physical distance (i.e., a cat in the upper left corner is the same as a cat in the lower right corner), the same is not true for these images that depict frequency or frequency derived information along their axes. For example, a spectrogram conveys time on the x-axis and frequency on the y-axis. It may be a fair assumption that translational invariance applies to the time axis (i.e., a sound event happening at 5 s is the same as one happening at 10 s), but it does not make much sense to uphold translational invariance to the frequency axis because semantic meaning is encoded in the frequency domain. Furthermore, the spectral properties of sound are non-local. The pitch of a sound is determined by the fundamental frequency,

while the quality or timbre of a sound is determined by its harmonics (the n^{th} harmonic has a frequency $F_n = nF_1$, where F_1 is the fundamental frequency). The fundamental frequency and its harmonics are not locally grouped despite originating from the same sound source. For example, if the fundamental frequency is 100 Hz, then its harmonics are 200 Hz, 300 Hz, etc. The locality bias, again while useful for natural images, is not a good inductive bias for image representations of sound because the frequencies associated with a given sound event are non-locally distributed.

The ViT, by using the self-attention mechanism, structurally lack these two inductive biases of translational invariance and locality, which are usually quite useful biases for natural images. Typically, the ViT are known to be "data hungry" because ViT must learn these inductive biases from the data itself; however, for image representation of sound, it makes good sense to disregard these biases as they do not pertain to these images. Since the ViT is not structurally constrained to the inductive biases of translational invariance and locality like the CNN, the model can explore the parameter space more freely to find a better set of generalizable rules for classifying image representations of sound. Furthermore, the vision transformer has a global receptive field; it can more easily model non-locally distributed spectral properties. This explains the superior performance of the ViT over the convolution-based neural networks in classifying image representations of pediatric heart sounds.

4.3. Strengths

The first strength of our study is our novel technical approach to classifying pediatric heart sounds, which performs comparably or outperforms current state-of-the-art methods. Our presented algorithm may generalize to various other bioacoustics signals and such use cases warrants investigation. The second strength of our study is our original, curated dataset that includes a large number and variety of innocent murmurs validated by echocardiogram. Many of the existing studies [14–20] do not include innocent murmurs in their dataset used to train their models. While one group of authors [7] does examine the Still's murmur, the most common innocent murmur, their study is limited by the binary classification of Still's versus all other types. The classification of normal heart sounds, pathological murmurs, and other innocent murmurs into a broad non-Still's class is not clinically relevant. By contrast, our study's database was able to include the three most common types of innocent murmurs: Still's murmur, flow murmurs, and venous hum.

Additionally, we are the first to demonstrate multiclass classification of pediatric heart sounds into three classes: innocent murmurs, pathologic murmurs, and no murmur. We note that since the meaningful clinical end point is decreasing the number of innocent murmur referrals, having a multiclass output of innocent versus pathologic versus normal would not change management relative to a binary classifier of pathologic versus non-pathologic (i.e. grouping the normal heart sounds and innocent murmurs together in one class). However, even though classifying an innocent murmur as normal heart sound (and vice versa) does not change clinical management, it still technically represents an incorrect classification. Our multiclass output creates a more interpretable and explainable model which can facilitate further model improvement and refinement in the downstream model development cycle as well as enhance clinician trust and therefore adoption. Additionally, a multiclass model has implications for medical education as if can help trainees learn to differentiate innocent murmurs from pathologic.

4.4. Limitations

The main limitation of our study is the volume of data collected. Training on a limited dataset can lead to overfitting and poor generalization. We attempted to overcome this limitation by collecting data from two demographically distinct locations to create a more heterogenous dataset. We prospectively collected predominantly innocent and pathological murmurs from two affiliated sites in New York City and supplemented our dataset with additional sounds from the CirCor Digiscope dataset, a publicly available dataset of pediatric heart sounds from Brazil [48]. By the same token, because our data is only sourced from New York City and Brazil, our model may not generalize to other regions with different demographic and environmental factors. Outside of the CirCor Digiscope database, existing publicly available heart sound databases have an adult focus [53,54]. Adult heart sounds are not applicable for pediatric heart sound classification. Children have much higher heart rates, and therefore shorter diastolic period relative to adult heart sounds, which impacts the interpretation. Furthermore, the physiology underlying murmurs in children differs greatly from that in adults. We attempted to create a comprehensive dataset that reflects the range of innocent and pathological pediatric murmurs potentially encountered in clinical practice. While our dataset captures the vast majority of common innocent and pathologic pediatric murmurs, our dataset is most notably missing atrial septal defects and peripheral pulmonic stenosis. However, it is worth noting that these two murmurs have similar qualities to pulmonic stenosis, which is included in our dataset, but a larger dataset that includes these congenital heart defects would potentially increase our model's generalizability. Additionally, our data was collected under quiet, controlled conditions (i.e. outpatient pediatric offices). Our models may not generalize to other care settings that may have variations in ambient noise levels or recording techniques such as emergency rooms, inpatient floors, or specialized pediatric clinics.

Atrial septal defects are the second most common congenital heart defect in children; they frequently go undiagnosed until adulthood, as they are often asymptomatic [55]. The characteristic murmur is a soft systolic murmur, similar to common innocent murmurs, albeit with a distinct splitting of the second heart sound. While small defects may spontaneously resolve, large ones can cause complications such as dysrhythmias, pulmonary hypertension, or in severe cases right-sided heart failure. Therefore, it is important for a deep learning algorithm to distinguish between this common defect versus innocent murmurs. Peripheral pulmonic stenosis is a subtype of pulmonary stenosis, the fifth most common congenital heart defect [56]. Peripheral pulmonic stenosis is a common murmur in infants and is caused by a narrowing in a distal branch of the pulmonic artery. While other types of pulmonary stenosis (i.e., valvular and sub-valvular pulmonary stenosis) are pathologic and often require intervention, peripheral pulmonic stenosis is considered an innocent murmur with a benign clinical course [57]. It is also important to note that our dataset is intended to reflect what can be encountered in the general pediatric office, so murmurs that would be encountered in the perinatal period or in the neonatal intensive care unit (i.e., patent ductus arteriosus, coarctation of the aorta) are not included. Our dataset is also missing certain types of critical congenital heart defects, such as truncus arteriosus, transposition of the great arteries, total anomalous pulmonary vein return, and Ebstein's anomaly, but each of these pathologies make up 1 to 3 % of congenital heart disease with incidences as low as <1 in 100,000 [58–61]. A larger, more comprehensive data set will likely result in better and more generalizable models. However, for practical purposes, each of these rarer lesions will require specialized pediatric cardiology evaluation for precise diagnosis. The greater value to the population of deep learning-based evaluation of heart murmurs will be the identification of innocent murmurs, which do not require subspecialty referral, from pathologic murmurs, which do require investment in subspecialty care.

4.5. Conclusion and future directions

In summary, we presented two novel methods for pediatric heart sound classification. Our methodology involves creating either an MTF or GAF image representation of the heart sound's frequency spectrum

and using the image-based representation of sound to train a ViT. We find that this methodology outperforms the current state-of-the-art of using a CNN trained on spectrogram images as well as a ViT trained on spectrogram images. Our deep learning model has application in both resource-rich and resource-limited settings. Resource-rich areas may benefit more from preventing over-referrals of innocent murmurs and over-utilization of echocardiography for common benign murmurs. Resource-limited areas that may lack easy access to subspecialists or to echocardiography may benefit more from the use of this model to detect pathological murmurs.

The best way to combat poor generalization is through curating a larger, more diverse dataset. Future work should involve optimizing our model on an expanded dataset to include examples of innocent and pathologic sounds missing from our data set. With an expanded dataset, it may also be possible to achieve more granular multiclass classification of pediatric heart sounds (i.e., distinguishing aortic stenosis from mitral regurgitation). From a clinical perspective, it would be interesting to have a breakdown of model performance by individual pathology and should be investigated in future work. Most importantly, the meaningful clinical endpoint is decreasing the number of innocent murmur referrals. Thus, to ascertain the true clinical benefit, a prospective, multicenter study should be conducted to study how a pediatricians decision making is affected by our models' predictions and if the number of innocent murmur referrals is decreased as a result.

CRediT authorship contribution statement

George Zhou: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Candace Chien:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Justin Chen:** Software, Investigation, Formal analysis. **Lucille Luan:** Visualization, Validation, Investigation, Formal analysis. **Yunchan Chen:** Software. **Sheila Carroll:** Supervision, Data curation. **Jeffrey Dayton:** Supervision, Data curation. **Maria Thanjan:** Supervision, Data curation. **Ken Bayle:** Supervision, Data curation. **Patrick Flynn:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

Research support

This research received no external financial or non-financial support.

Relationships

There are no additional relationships to disclose.

Patents and intellectual property

Authors George Zhou, Candace Chien, and Yunchan Chen have a patent application pending to Cornell University.

Other activities

This research received no external funding. There are no additional relationships or activities to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2024.102867.

References

[1] The Children's Hospital of Philadelphia. Heart murmur in children. Retrieved January 30, 2023, from, https://www.chop.edu/conditions-diseases/heart-murmur; 2014, March 26.

[2] Liu Y, et al. Global birth prevalence of congenital heart defects 1970–2017: updated systematic review and meta-analysis of 260 studies. Int J Epidemiol 2019;48:455–63.

[3] Chen M, Riehle-Colarusso T, Yeung LF, Smith C, Farr SL. Children with heart conditions and their special health care needs — United States, 2016. MMWR Morb Mortal Wkly Rep 2018;67:1045–9. https://doi.org/10.15585/mmwr.mm6738a1external icon.

[4] GBD 2017 Congenital Heart Disease Collaborators. 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 2020 Mar;4(3):185–200. https://doi.org/10.1016/S2352-4642(19)30402-X. Epub 2020 Jan 21. Erratum in: Lancet Child Adolesc Health. 2020 Feb 7;: PMID: 31978374; PMCID: PMC7645774.

[5] McCrindle BW, Shaffer KM, Kan JS, Zahka KG, Rowe SA, Kidd L. Factors prompting referral for cardiology evaluation of heart murmurs in children. Arch Pediatr Adolesc Med 1995 Nov;149(11):1277–9. https://doi.org/10.1001/archpedi.1995.02170240095018 [PMID: 7581765].

[6] Mejia E, Dhuper S. Innocent Murmur. [Updated 2022 Sep 5]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan.. Available from: https://www.ncbi.nlm.nih.gov/books/NBK507849.

[7] Shekhar R, Vanama G, John T, Issac J, Arjoune Y, Doroshow RW. Automated identification of innocent Still's murmur using a convolutional neural network. Front Pediatr 2022 Sep 21;(10):923956. https://doi.org/10.3389/fped.2022.923956. PMID: 36210944; PMCID: PMC9533723.

[8] Haney I, Ipp M, Feldman W, McCrindle BW. Accuracy of clinical assessment of heart murmurs by office based (general practice) paediatricians. Arch Dis Child 1999;81(5):409–12.

[9] Kumar K, Thompson WR. Evaluation of cardiac auscultation skills in pediatric residents. Clin Pediatr 2013;52(1):66–73.

[10] Kostopoulou E, Dimitriou G, Karatza A. Cardiac murmurs in children: a challenge for the primary care physician. Curr Pediatr Rev 2019;15(3):131–8. https://doi.org/10.2174/1573396315666190321105536 [PMID: 30907325].

[11] Wen J, Snyder C. Prevalence of innocent murmurs in pediatric patients. 2019.

[12] John T, Doroshow RW, Shekhar R. A smartphone stethoscope and application for automated identification of innocent still's murmur. In: Frontiers in biomedical devices. 40789. American Society of Mechanical Engineers; 2018, April. p. V001T01A011.

[13] Bensky AS, Covitz W, DuRant RH. Primary care physicians' use of screening echocardiography. Pediatrics 1999;103(4):e40.

[14] Kotb Magd Ahmed, Elmahdy Hesham Nabih, El Mostafa Fatma El Zahraa, Falaki Mona, Shaker Christine William, Refaey Mohamed Ahmed, et al. Improving the recognition of heart murmur. Int J Adv Comput Sci Appl 2016;7(7). https://doi.org/10.14569/IJACSA.2016.070740.

[15] Pretorius E, Cronje ML, Strydom O. Development of a pediatric cardiac computer aided auscultation decision support system. Annu Int Conf IEEE Eng Med Biol Soc 2010;2010:6078–82. https://doi.org/10.1109/IEMBS.2010.5627633 [PMID: 21097128].

[16] Wang J, You T, Yi K, Gong Y, Xie Q, Qu F, et al. Intelligent diagnosis of heart murmurs in children with congenital heart disease. J Healthc Eng 2020 May 9;(2020):9640821. https://doi.org/10.1155/2020/9640821. PMID: 32454963; PMCID: PMC7238385.

[17] Xiao B, Xu Y, Bi X, Li W, Ma Z, Zhang J, et al. Follow the sound of Children's heart: a deep-learning-based computer-aided pediatric CHDs diagnosis system. IEEE Internet Things J 2020;7:1994–2004.

[18] Liu J, Wang H, Yang Z, Quan J, Liu L, Tian J. Deep learning-based computer-aided heart sound analysis in children with left-to-right shunt congenital heart disease. Int J Cardiol 2022 Feb 1;(348):58–64. https://doi.org/10.1016/j.ijcard.2021.12.012. Epub 2021 Dec 10. PMID: 34902505.

[19] Gharehbaghi A, Sepehri AA, Lindén M, Babic A. A hybrid machine learning method for detecting cardiac ejection murmurs. In: Eskola H, Väisänen O, Viik J, Hyttinen J, editors. EMBEC & NBC 2017. EMBEC NBC 2017 2017. IFMBE proceedings. vol. 65. Singapore: Springer; 2018. https://doi.org/10.1007/978-981-10-5122-7_197.

[20] Wang JK, Chang YF, Tsai KH, Wang WC, Tsai CY, Cheng CH, et al. Automatic recognition of murmurs of ventricular septal defect using convolutional recurrent neural networks with temporal attentive pooling. Sci Rep 2020 Dec 11;10(1):21797. https://doi.org/10.1038/s41598-020-77994-z. PMID: 33311565; PMCID: PMC7732853.

[21] Kang S, Doroshow R, McConnaughey J, Shekhar R. Automated identification of innocent still's murmur in children. IEEE Trans Biomed Eng 2017 Jun;64(6):1326–34. https://doi.org/10.1109/TBME.2016.2603787. Epub 2016 Aug 26. PMID: 27576242.

[22] DeGroff CG, Bhatikar S, Hertzberg J, Shandas R, Valdes-Cruz L, Mahajan RL. Artificial neural network-based method of screening heart murmurs in children. Circulation 2001 Jun 5;103(22):2711–6. https://doi.org/10.1161/01.cir.103.22.2711 [PMID: 11390342].

[23] Chen W, Sun Q, Chen X, Xie G, Wu H, Xu C. Deep learning methods for heart sounds classification: a systematic review. Entropy (Basel) 2021 May 26;23(6):667. https://doi.org/10.3390/e23060667. PMID: 34073201; PMCID: PMC8229456.

[24] Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, et al. An open access database for the evaluation of heart sound algorithms. Physiol Meas 2016 Dec;37(12):2181–213. https://doi.org/10.1088/0967-3334/37/12/2181 [Epub 2016 Nov 21. PMID: 27869105; PMCID: PMC7199391].

[25] Latif S, Usman M, Rana R, Qadir J. Phonocardiographic sensing using deep learning for abnormal heartbeat detection. IEEE Sens J 2018;18:9393–400.

[26] Khan FA, Abid A, Khan MS. Automatic heart sound classification from segmented/unsegmented phonocardiogram signals using time and frequency features. Physiol Meas 2020;41:055006.
[27] Yang T-C, Hsieh H. Classification of acoustic physiological signals based on deep learning neural networks with augmented features. In: Proceedings of the 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 11–14 September; 2016. p. 569–72.
[28] Raza A, Mehmood A, Ullah S, Ahmad M, Choi GS, On BW. Heartbeat sound signal classification using deep learning. Sensors 2019;19:4819.
[29] Chorba JS, Shapiro AM, Le L, Maidens J, Prince J, Pham S, et al. Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform. J Am Heart Assoc 2021 May 4;10(9):e019905. https://doi.org/10.1161/JAHA.120.019905 [Epub 2021 Apr 26. PMID: 33899504; PMCID: PMC8200722].
[30] Ryu H, Park J, Shin H. Classification of heart sound recordings using convolution neural network. In: Proceedings of the 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 11–14 September; 2016. p. 1153–6.
[31] Xu Y, Xiao B, Bi X, Li W, Zhang J, Ma X. Pay more attention with fewer parameters: a novel 1-D convolutional neural network for heart sounds classification. In: Proceedings of the Computing in Cardiology Conference (CinC), Maastricht, The Netherlands, 23–26 September. Volume 45; 2018. p. 1–4.
[32] Humayun AI, Ghaffarzadegan S, Ansari I, Feng Z, Hasan T. Towards domain invariant heart sound abnormality detection using learnable filterbanks. IEEE J Biomed Health Inform 2020;24:2189–98.
[33] Xiao B, Xu Y, Bi X, Zhang J, Ma X. Heart sounds classification using a novel 1-D convolutional neural network with extremely low parameter consumption. Neurocomputing 2020;392:153–9.
[34] Oh SL, Jahmunah V, Ooi CP, Tan R-S, Ciaccio EJ, Yamakawa T, et al. Classification of heart sound signals using a novel deep WaveNet model. Comput Methods Programs Biomed 2020;196:105604.
[35] Baghel N, Dutta MK, Burget R. Automatic diagnosis of multiple cardiac diseases from PCG signals using convolutional neural network. Comput Methods Programs Biomed 2020;197:105750.
[36] Deperlioglu O, Kose U, Gupta D, Khanna A, Sangaiah AK. Diagnosis of heart diseases by a secure internet of health things system based on autoencoder deep neural network. Comput Commun 2020;162:31–50.
[37] Sun S, Huang T, Zhang B, He P, Yan L, Fan D, et al. A novel intelligent system based on adjustable classifier models for diagnosing heart sounds. Sci Rep 2022 Jan 25; 12(1):1283. https://doi.org/10.1038/s41598-021-04136-4. PMID: 35079025; PMCID: PMC8789933.
[38] Demir F, Şengür A, Bajaj V, Polat K. Towards the classification of heart sounds based on convolutional deep neural network. Health Inf Sci Syst 2019;7:1–9.
[39] Nilanon T, Yao J, Hao J, Purushotham S. Normal/abnormal heart sound recordings classification using convolutional neural network. In: Proceedings of the Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 11–14 September; 2016. p. 585–8.
[40] Zhou G, Chen Y, Chien C. On the analysis of data augmentation methods for spectral imaged based heart sound classification using convolutional neural networks. BMC Med Inform Decis Mak 2022 Aug 29;22(1):226. https://doi.org/10.1186/s12911-022-01942-2. PMID: 36038901; PMCID: PMC9421122.
[41] Dominguez-Morales JP, Jimenez-Fernandez AF, Dominguez-Morales MJ, Jimenez-Moreno G. Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors. IEEE Trans Biomed Circuits Syst 2018;12:24–34.
[42] Cheng X, Huang J, Li Y, Gui G. Design and application of a laconic heart sound neural network. IEEE Access 2019;7:124417–25.
[43] Maknickas V, Maknickas A. Recognition of normal abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefcients. Physiol Meas 2017;38:1671–84.
[44] Alafif T, Boulares M, Barnawi A, Alafif T, Althobaiti H, Alferaidi A. Normal and abnormal heart rates recognition using transfer learning. In: Proceedings of the 2020 12th International Conference on Knowledge and Systems Engineering (KSE), Can Tho, Vietnam, 12–14 November; 2020. p. 275–80.
[45] Rubin J, Abreu R, Ganguli A, Nelaturi S, Matei I, Sricharan K. Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients. In: Proceedings of the 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 11–14 September; 2016. p. 813–6.
[46] Wang Z, Oates T. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In: Proceedings of the 2015 international conference on artificial intelligence and statistics (AISTATS); 2015. p. 626–34.
[47] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Computer Vision (ICCV); 2021. p. 3149–58.
[48] Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, et al. The CirCor DigiScope dataset: from murmur detection to murmur classification. IEEE J Biomed Health Inform 2022 Jun;26(6):2524–35. https://doi.org/10.1109/JBHI.2021.3137048 [Epub 2022 Jun 3. PMID: 34932490; PMCID: PMC9253493].
[49] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 770–8.
[50] Tariq Z, Shah SK, Lee Y. Feature-based fusion using CNN for lung and heart sound classification. Sensors (Basel) 2022 Feb 16;22(4):1521. https://doi.org/10.3390/s22041521. PMID: 35214424; PMCID: PMC8875944.
[51] Kudriavtsev V, Polyshchuk V, Roy DL. Heart energy signature spectrogram for cardiovascular diagnosis. Biomed Eng Online 2007 May 4;(6):16. https://doi.org/10.1186/1475-925X-6-16. PMID: 17480232; PMCID: PMC1899182.
[52] Huai X, Kitada S, Choi D, Siriaraya P, Kuwahara N, Ashihara T. Heart sound recognition technology based on convolutional neural network. Inform Health Soc Care 2021 Sep 2;46(3):320–32. https://doi.org/10.1080/17538157.2021.1893736. Epub 2021 Apr 4. PMID: 33818274.
[53] Tuchinda C, Thompson WR. Cardiac auscultatory recording database: delivering heart sounds through the internet. Proceedings/AMIA annual symposium 2001: 716–20.
[54] Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, et al. An open access database for the evaluation of heart sound algorithms. Physiol Meas 2016;37(9).
[55] Menillo AM, Lee LS, Pearson-Shaver AL. Atrial septal defect. [Updated 2022 Aug 8]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan. Available from: https://www.ncbi.nlm.nih.gov/books/NBK535440/.
[56] Keane John, Fyler Donald. Pulmonary stenosis. 2006. https://doi.org/10.1016/B978-1-4160-2390-6.50036-2.
[57] Heaton J, Kyriakopoulos C. Pulmonic stenosis. [Updated 2023 Jan 4]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan. Available from: https://www.ncbi.nlm.nih.gov/books/NBK560750/.
[58] Bhansali S, Phoon C. Truncus arteriosus. [Updated 2022 Aug 8]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan. Available from: https://www.ncbi.nlm.nih.gov/books/NBK534774/.
[59] Szymanski MW, Moore SM, Kritzmire SM, et al. Transposition of the great arteries. [Updated 2023 Jan 15]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan. Available from: https://www.ncbi.nlm.nih.gov/books/NBK538434/.
[60] Konduri A, Aggarwal S. Partial and total anomalous pulmonary venous connection. [Updated 2022 Aug 16]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan. Available from: https://www.ncbi.nlm.nih.gov/books/NBK560707/.
[61] Attenhofer Jost CH, Connolly HM, Dearani JA, Edwards WD, Danielson GK. Ebstein's anomaly. Circulation 2007;115(2):277–85.
[62] Sepehri AA, Kocharian A, Janani A, Gharehbaghi A. An intelligent phonocardiography for automated screening of pediatric heart diseases. J Med Syst 2016 Jan;40(1):16. https://doi.org/10.1007/s10916-015-0359-3. Epub 2015 Oct 30. PMID: 26573653.
[63] Bordbar A, Kashaki M, Vafapour M, Sepehri AA. Determining the incidence of heart malformations in neonates: a novel and clinically approved solution. Front Pediatr 2023 Mar 15;(11):1058947. https://doi.org/10.3389/fped.2023.1058947. PMID: 37009269; PMCID: PMC10050760.