Markov-Based Neural Networks for Heart Sound Segmentation: Using Domain Knowledge in a Principled Way

Miguel L. Martins, Miguel T. Coimbra, *Senior Member, IEEE*, and Francesco Renna, *Senior Member, IEEE*

***Abstract*—This work considers the problem of segmenting heart sounds into their fundamental components. We unify statistical and data-driven solutions by introducing *Markov-based Neural Networks* (MNNs), a hybrid end-to-end framework that exploits Markov models as statistical inductive biases for an Artificial Neural Network (ANN) discriminator. We show that an MNN leveraging a simple one-dimensional Convolutional ANN significantly outperforms two recent purely data-driven solutions for this task in two publicly available datasets: PhysioNet 2016 (Sensitivity: 0.947 ± 0.02; Positive Predictive Value : 0.937 ± 0.025) and the CirCor DigiScope 2022 (Sensitivity: 0.950 ± 0.008; Positive Predictive Value: 0.943 ± 0.012). We also propose a novel gradient-based unsupervised learning algorithm that effectively makes the MNN adaptive to unseen datum sampled from unknown distributions. We perform a cross dataset analysis and show that an MNN pre-trained in the CirCor DigiScope 2022 can benefit from an average improvement of 3.90% Positive Predictive Value on unseen observations from the PhysioNet 2016 dataset using this method.**

***Index Terms*—Hybrid neural networks, Markov models, model-based deep learning, phonocardiogram, segmentation.**

I. Introduction

Cardiovascular diseases (CVDs) are the number one cause of mortality worldwide, with the total number of deaths projected to be over 23 million by 2030 given the current prevalence of CVD risk factors, such as smoking, obesity, and physical inactivity, as reported by the World Heart Federation [1]. CVDs are most prevalent in low to middle-income countries, where limited expert human resources and operational conditions constrain the effectiveness of healthcare services. To mitigate the economic and humanitarian toll of CVDs (especially in underprivileged scenarios), early discovery and treatment are only actionable if the screening becomes more reliable, inexpensive, and fast. In this context, cardiac auscultation is particularly attractive since this examination has low cost, can detect several heart conditions, and is one of the simplest cardiac screening procedures [2]. Although accurate interpretation of the sounds collected during auscultation requires access to personnel with extensive qualifications and clinical experience, collecting the necessary data in the form of a *phonocardiogram* (PCG) for retrospective examination and diagnosis is considerably simpler, thus requiring less training [3], [4], [5]. Moreover, the advent of electronic stethoscopes, coupled with significant progress in machine learning, has brought about a revolution in cardiac auscultation, since computer-assisted decision systems are now capable of extracting meaningful information from PCG recordings with significant clinical relevance [6], [7].

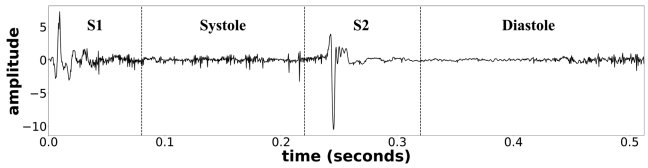

Fig. 1. Segment of 0.51 seconds of a PCG sample from the CirCor DigiScope 2022 dataset displaying an entire heart cycle.

In order to extract information from heart sounds one can first detect key events in the PCG recording, notably the two fundamental sounds present in each heart cycle: the *first sound* or S1, generated by the mitral and tricuspid valve vibrations from the systolic onset, and the *second sound* or S2, which results from the aortic and pulmonary valve closure at the diastolic onset (an example of such fundamental heart sounds is reported in Fig. 1). The task of delineating the boundaries of the S1 and S2 sounds (which consequently yields knowledge of the systolic and diastolic intervals) for every heart cycle in a PCG is called heart sound segmentation.

One should note that there are many recent solutions for CVD-related downstream tasks that do not require this explicit segmentation step (see, for example, [8], [9], [10], [11], [12]), and the community currently questions whether a dedicated segmentation component is required within CVD detection

Manuscript received 1 April 2023; revised 28 July 2023; accepted 31 August 2023. Date of publication 6 September 2023; date of current version 7 November 2023. This work was supported in part by the National Funds through the Portuguese funding agency, in part by the FCT - Fundação para a Ciência e a Tecnologia under Grant UIDB/50014/2020. The work of Miguel L. Martins was supported by the individual FCT under Reference 2021.06503.BD. *(Corresponding author: Miguel L. Martins.)*

The authors are with the Institute for Systems and Computer Engineering, Technology and Science (INESC-TEC), Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal (e-mail: miguel.l.martins@inesctec.pt; mcoimbra@fc.up.pt; francesco.renna@fc.up.pt).

Digital Object Identifier 10.1109/JBHI.2023.3312597

models [13]. We argue that precise knowledge of the four fundamental segments in each heart cycle not only enables the detection and localization of extra sound components (e.g., the third and fourth heart sounds, murmurs, ejection clicks, etc.), but it also allows analysis of the waveform morphology of the S1 and S2 sounds. Furthermore, appropriate reconstruction of the underlying heart sound sequence should constrain the response of data-driven models to physiologically plausible results, enhancing their performance and explainable value. This can be attested by the final standings in the George B. Moody PhysioNet challenge 2022 [14], since the winning classification approach for the heart murmurs task included algorithms that leveraged segmentation of the fundamental heart sounds [15].

A. Related Work

Existing literature for heart sound segmentation can be grouped into three main types of techniques: peak-picking, *ad-hoc* feature extraction coupled with statistical point-wise classifiers, and sequential models. Peak-picking strategies typically apply some transformation to the signal (either in the time or frequency domains) so that thresholding captures local maxima related to S1 and S2 sounds [16], [17], [18], [19], [20], [21].

For the second type of strategies, the signal is typically preprocessed to facilitate the discrimination of S1 and S2 components, and hence features are extracted in the time [22], frequency [23], or Wavelet [24] domains. Then, different types of classifiers have been considered to detect the S1 and S2 components, ranging from k-means clustering [25], [26], decision trees [27], Support Vector Machines (SVMs) [24], to Artificial Neural Networks (ANNs) [28], [29], [30]. However, these methods require a candidate selection step to detect S1 and S2 sounds, which is typically provided by the above-mentioned peak-picking algorithms.

Finally, sequential models exploit *a prirori* knowledge of the steady progression of the heart sounds during each heart cycle, i.e., the fact that the only permissible transitions are: S1 to systole, systole to S2, S2 to diastole, and diastole to S1. Most successful parametric sequential models for the specific task of heart sound segmentation are typically grounded on either hidden Markov models (HMM) or hidden semi-Markov models (HSMM). The Markovian/semi-Markovian process governs the latent heart cycle progression and is normally coupled with some point-wise discriminator that models the emission distribution of the signal. Gamero et al. [31] and Gill et al. [32] combined HMMs with peak-picking on an envelogram of the PCG. Expanding this line of work, [33] modelled the emission distribution explicitly through Gaussian Mixture Models.

HSMMs enable parametrization of the so-called *sojourn time* of each state, i.e., the distribution of the duration of each heart sound in the PCG sequence. Contrary to HMMs where the probability of repeating a state always decays exponentially, in HSMMs the state duration distribution is adaptive and may differ from state to state. HSMMs were first applied by Schmidt et al. for heart sound segmentation [34], [35]. Oliveira et al. [36], [37] developed methods to learn the sojourn time in the context of heart sound segmentation. In this line of work, there have also been proposals to improve the estimates for the emission distribution, such as the SVM and logistic regression proposed by Springer et al. [38], [39]. With the recent advancements in deep learning, ANNs have been successfully implemented in sequential models for heart sound segmentation. Renna et al. [40] introduced a sliding U-Net on envelopes extracted from heart sound recordings to estimate the emission distributions, which are then decoded by an HMM parameterized by the maximum likelihood estimates for the train set. In [41], Messner et al. suggested an end-to-end, fully data-driven method for simultaneous learning of the emission and latent state distributions, through a gated bidirectional Recurrent Neural Network (Bi-GRNN) over spectral and envelope features of the signals, which showed competitive results with the HSMM proposed by Springer et al. [39]. Fernando et al. [42] proposed a more capable bidirectional Long Short-Term Memory network coupled with an attention mechanism (Bi-LSTM+A) that acted on the *Mel Frequency Cepstral Components* (MFCCs) of the PCG. The attention mechanism has the ability to detect salient aspects in each sound, and its application is intended to enhance robustness against noisy or irregular recordings. Additionally, a temporal-framing adaptive (TFA) network was proposed in [43], which employs a specific transition loss function during training and can perform dynamic inference, enabling it to adapt to varying heart sound behaviours. The authors reported positive results when comparing to the HSMM method of Springer et al. [39] and the Bi-GRNN proposed by Messner et al. [41].

B. Contributions

Analysis of the literature reveals that modelling the latent state sequence assuming some statistical prior (HMM/HSMM) and the end-to-end frameworks (U-Net, Bi-GRNN, Bi-LSTM+A, TFA) yield the most competitive results. Motivated by this, we propose a hybrid approach that combines the advantages of the explicit sequential modelling of HMMs with the discriminative power of an ANN within a single end-to-end learning framework. In particular, this article presents four main contributions:

1) The formulation of a hybrid framework wherein, given *a priori* knowledge that the signal is generated by some latent HMM, the emission posteriors are estimated using the output of an ANN. In other words, a *Markov-based Neural Network* (MNN).[1]
2) A set of supervised and unsupervised loss functions tailored to exploit the *a priori* knowledge of some underlying Markovian regime.
3) Performance comparison of the proposed MNNs with recent purely data-driven approaches for the downstream task of heart sound segmentation. The experiments were performed in two publicly available PCG datasets: PhysioNet 2016 and CirCor DigiScope 2022.
4) Cross dataset performance assessment of the unsupervised learning algorithm given an MNN pre-trained in PhysioNet 2016 (CirCor DigiScope 2022) on unseen

[1]Our implementation is available at https://github.com/miguelmartins/mnn.

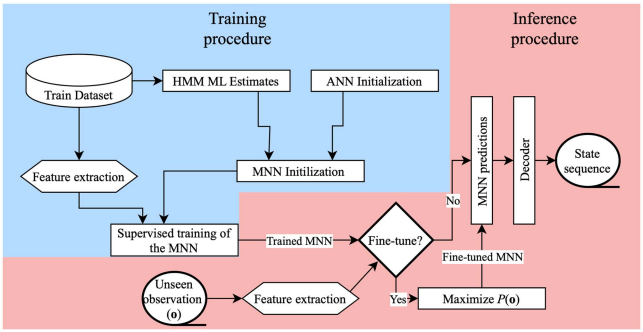

Fig. 2. Flowchart illustrating the proposed framework. The blue background signifies the training procedure, the red background marks the inference routine.

observations from CirCor DigiScope 2022 (PhysioNet 2016).

We refer the reader to Fig. 2 for a schematic representation of our framework.

We reported some preliminary results on a simpler version of this framework in [44], were we proposed to train an ANN to maximize the Maximal Mutual Information criterion under the assumption that PCG signals were governed by an underlying HMM. In this article, we encapsulate and significantly extend this idea within a broader framework, *Markov-based Neural Networks* (MNNs). We introduce a parameter re-projection step that enables gradient descent to be used for training without compromising the statistical and physiological plausibility of the underlying HMM. This not only allows the MNN to be trained completely end-to-end, but also to be fine-tuned to unseen, unlabelled datum by maximizing their likelihood given the model, as prescribed in our MNN formulation. Compared to our precursory work [44], our experiments are also substantially more thorough and are sustained by statistical hypothesis testing whenever a comparison between the outcomes of different models is made. We also compare the performance of the MNNs with the Bi-LSTM+A by Fernando et al. [42] since it employs an alternative temporal modulation of the signal without any type of statistical prior by combining a recurrent ANN with an attention mechanism.

C. Paper Structure

The remainder of this article is structured as follows: in Section II, we introduce MNNs and the respective supervised and unsupervised training routines in a principled way. Then, in Section III, we instantiate the problem of PCG fundamental heart sound segmentation, alongside a Left-to-right MNN architecture especially tailored for this problem. In Section IV, we perform a set of experiments to establish the MNN baseline using the PhysioNet 2016 dataset (Section II-A) and then compare it to the models proposed by Springer et al. [39], Renna et al. [40], and Fernando et al. [42] (Section IV-C2), both in the PhysioNet 2016 and the CirCor DigiScope 2022 datasets. Finally, in Section IV-C3 we assess cross dataset performance of the unsupervised learning algorithm given an MNN pretrained in PhysioNet 2016 (CirCor DigiScope 2022) on unseen observations from CirCor DigiScope 2022 (PhysioNet 2016). We discuss our findings in Section IV-D, and in Section V draw conclusions alongside future lines of inquiry.

II. Markov-Based Neural Networks

Suppose you have a dataset $\boldsymbol{D} = \{(\mathbf{o}^{(i)}, \mathbf{s}^{(i)})\}_{i=1}^{N}$ such that each observation sequence $\mathbf{o}^{(i)} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{T_i}]$ is a function of a state sequence $\mathbf{s}^{(i)} = [s_1, s_2, \ldots, s_{T_i}]$ that was generated by some (latent) homogeneous first-order Markov chain with discrete states $\mathbf{s}_t \in \mathcal{S}$, $\mathcal{S} = \{0, 1, \ldots, L-1\}$. The joint distribution of a pair of *emissions* $\mathbf{o}^{(i)}$ and *states* $\mathbf{s}^{(i)}$ is given by:

$$P\left(\mathbf{o}^{(i)}, \mathbf{s}^{(i)}\right) = P(s_1) \prod_{t=2}^{T_i} P(s_{t-1}|s_t) \prod_{t=1}^{T_i} P(\mathbf{o}_t|s_t). \quad (1)$$

Presume ignorance of the class of distributions to which $P(\mathbf{o}_t|s_t)$ pertains. Consider instead access to a highly discriminant artificial neural network (ANN) such that $\text{ANN}(\mathbf{o}_t) \sim P(s_t|\mathbf{o}_t)$. One can approximate (1) by using Bayes' rule to estimate the emission distribution given the posterior predicted by the ANN; we call this the *hybrid* formulation of the model, i.e., we leverage both the HMM and the ANN in order to estimate $P(\mathbf{o}^{(i)}, \mathbf{s}^{(i)})$ as:

$$P\left(\mathbf{o}^{(i)}, \mathbf{s}^{(i)}\right) = P(s_1) \prod_{t=2}^{T_i} P(s_{t-1}|s_t) \prod_{t=1}^{T_i} \frac{\text{ANN}(\mathbf{o}_t) P(\mathbf{o}_t)}{P(s_t)}. \quad (2)$$

Note that we do not assume knowledge of $P(\mathbf{o}_t)$ in (2). However, ignoring this term yields a function proportional to (2) which is sufficient for our optimization scheme.

The likelihood of $\mathbf{o}^{(i)}$ follows directly as the marginal of (2) over all possible state sequences $\mathcal{S}^{T_i}$:

$$P(\mathbf{o}^{(i)}) = \sum_{\mathbf{s} \in \mathcal{S}^{T_i}} P(s_1) \prod_{t=2}^{T_i} P(s_{t-1}|s_t) \prod_{t=1}^{T_i} \frac{\text{ANN}(\mathbf{o}_t) P(\mathbf{o}_t)}{P(s_t)}, \quad (3)$$

which can be computed efficiently without overflowing errors using a scaled forward-backward algorithm [45], [46]. Equation (2) and (3) bind the hidden Markov model (HMM) and ANN into a single, unified framework since they depend on the parameters of both models. An MNN is thus an HMM that shares the parameter space with an ANN, which models its emissions. Conversely, the MNN can also be interpreted as an ANN whose likelihood is a function of a latent Markovian state.

We denote by $\boldsymbol{\Psi} = \{\boldsymbol{\lambda}, \boldsymbol{\Theta}\}$ the set of all parameters of an MNN, where $\boldsymbol{\lambda} = \{\boldsymbol{\pi}, \boldsymbol{\Gamma}\}$ collects the parameters of the underlying Markov chain, with initial state probabilities $\boldsymbol{\pi} = (P(s_1) : s_1 \in \mathcal{S})$ such that $\boldsymbol{\pi} \in \mathbb{R}^L$, and state transition matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{L \times L}$. Finally, $\boldsymbol{\Theta}$ is the set of parameters of the ANN.

A. Training

This section will describe the training procedure assuming that the underlying Markov chain is first-order, homogeneous,

and stationary. The parameters are searched in the joint parameter space $\boldsymbol{\Psi}$ while minimizing some loss function $\mathcal{L}(\boldsymbol{D}; \boldsymbol{\Psi})$ using a customized gradient descent approach. We define several $\mathcal{L}$ using the likelihood expressions from (2) and (3).

1) Proposed Loss Functions: We start by introducing two supervised loss functions tailored toward sequence classification given knowledge of the ground truth states for a train set.

Let $P_{\boldsymbol{\Psi}}(.)$ be a density function parameterised by $\boldsymbol{\Psi}$. From (2) one can derive the *complete log-likelihood loss* ($\mathcal{L}_{\text{CL}}$):

$$\mathcal{L}_{\text{CL}}(\boldsymbol{D}; \boldsymbol{\Psi}) = -\sum_{i=1}^{N} \log P_{\boldsymbol{\Psi}} \left(\mathbf{o}^{(i)}, \mathbf{s}^{(i)} \right). \tag{4}$$

Following [47], and by noting that $\log P(\mathbf{o}^{(i)}|\mathbf{s}^{(i)}) = \log P(\mathbf{o}^{(i)}, \mathbf{s}^{(i)}) - \log P(\mathbf{o}^{(i)})$ one can use the derivations of (2) and (3) to build the *mutual information criterion loss* ($\mathcal{L}_{\text{MMI}}$):

$$\mathcal{L}_{\text{MMI}}(\boldsymbol{D}; \boldsymbol{\Psi}) = -\left(\sum_{i=1}^{N} \log P_{\boldsymbol{\Psi}} \left(\mathbf{o}^{(i)}, \mathbf{s}^{(i)} \right) - \log P_{\boldsymbol{\Psi}} \left(\mathbf{o}^{(i)} \right) \right). \tag{5}$$

One of the distinctions between $\mathcal{L}_{\text{CL}}$ and $\mathcal{L}_{\text{MMI}}$ is that the latter leverages $\log P_{\boldsymbol{\Psi}}(\mathbf{o}^{(i)})$ as to preserve the generative properties of the MNN while still accounting for discriminative performance of the model.

Finally, we also propose an unsupervised gradient-based optimization of (3) to make $\boldsymbol{\Psi}$ adaptive to co-variate shifts. Thus, we introduce the *unsupervised fine-tuning loss* ($\mathcal{L}_{\text{FT}}$):

$$\mathcal{L}_{\text{FT}}(\boldsymbol{D}; \boldsymbol{\Psi}) = -\sum_{i=1}^{N} \log P_{\boldsymbol{\Psi}} \left(\mathbf{o}^{(i)} \right). \tag{6}$$

2) Parameter Initialization: We initialize the parameters of the underlying Markov chain of an MNN by computing its maximum likelihood estimates (MLE). Given access to a labeled train set, the MLE estimate of the transition matrix, $\boldsymbol{\Gamma}_{\text{MLE}}$, can be attained directly from $\{\boldsymbol{s}^{(i)}\}_{i=1}^{N}$ by calculating the expected number of transitions between states [45]. Concerning the initial state distribution, we approximate it with the steady state probability vector $\boldsymbol{\pi}_{\text{steady}} \in \mathbb{R}^L$. Consequently, we solve the linear system:

$$\boldsymbol{\pi}_{\text{steady}} \boldsymbol{\Gamma} = \boldsymbol{\pi}_{\text{steady}}, \tag{7}$$

so that $\boldsymbol{\pi}_{\text{steady}} \geq \mathbf{0}$ and $||\boldsymbol{\pi}_{\text{steady}}||_1 = 1$, where $\geq$ is applied entry-wise and $||\cdot||_1$ is the ℓ_1-norm. Henceforth, assume $\boldsymbol{\lambda}_{\text{MLE}} = \{\boldsymbol{\Gamma}_{\text{MLE}}, \boldsymbol{\pi}_{\text{steady}}\}$.

Concerning the ANN, regardless of our choice of architecture, we initialize bias weights to zero, and the remainder of parameters using Glorot et al.'s normalized initialization [48].

3) Gradient Descent With Re-Projection: The MNN's parametrization is searched within a constrained optimization scheme. Specifically, since $\boldsymbol{\lambda}$ has probabilistic parameters, the feasible set for $\boldsymbol{\pi}$ and the rows $\boldsymbol{\Gamma}_i \in \mathbb{R}^L$ of $\boldsymbol{\Gamma}$ must be a subset of the canonical simplex $\mathcal{K}^L$ of $\mathbb{R}^L$:

$$\mathcal{K}^L = \left\{ \mathbf{x} \in \mathbb{R}^L : x_i \geq 0, i = 0, \ldots, L-1, \sum_{i=0}^{L-1} x_i = 1 \right\}. \tag{8}$$

Algorithm 1: Training in the Joint-Parameter Space of a Generic MNN.

Input: $\boldsymbol{D}$, $\mathcal{L}$, α, m
Output: $\boldsymbol{\Psi}$

$\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda}_{\text{MLE}}$;
$\boldsymbol{\Theta} \leftarrow$ `GlorotUniform()` ; `// As in [48]`
$\boldsymbol{\Psi} \leftarrow \{\boldsymbol{\lambda}, \boldsymbol{\Theta}\}$;
for *epoch* $= 1$ **to** m **do**
  $\boldsymbol{D} \leftarrow$ `shuffle`$(\boldsymbol{D})$;
  for $(\mathbf{o}^{(i)}, \mathbf{s}^{(i)}) \in \boldsymbol{D}$ **do**
    $\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} + \eta \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\mathbf{o}^{(i)}, \mathbf{s}^{(i)}; \boldsymbol{\Psi})$;
    $\boldsymbol{\Gamma} \leftarrow \boldsymbol{\Gamma} + \eta \nabla_{\boldsymbol{\Gamma}} \mathcal{L}(\mathbf{o}^{(i)}, \mathbf{s}^{(i)}; \boldsymbol{\Psi})$;
    for $\boldsymbol{\Gamma}_i \in \boldsymbol{\Gamma}$ **do**
      $\boldsymbol{\Gamma}_i \leftarrow \Phi(\boldsymbol{\Gamma}_i)$; `// Eq.` (9)
    end
    $\boldsymbol{\pi}$ s.t. $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{\Gamma}, \boldsymbol{\pi} \geq \mathbf{0}, ||\boldsymbol{\pi}||_1 = 1$; `// Eq.` (7)
    $\boldsymbol{\Psi} \leftarrow \{\{\boldsymbol{\pi}, \boldsymbol{\Gamma}\}, \boldsymbol{\Theta}\}$;
  end
end

Let $||\cdot||_2$ denote the ℓ_2-norm. Following [49], we define the projective map $\Phi : \mathbb{R}^L \longrightarrow \mathcal{K}^L$ so that:

$$\Phi(\mathbf{y}) = \arg\min_{\mathbf{x} \in \mathcal{K}^L} \left(\frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \right). \tag{9}$$

For each gradient descent update we set $\boldsymbol{\Gamma}_i = \Phi(\boldsymbol{\Gamma}_i + \eta \nabla_{\boldsymbol{\Gamma}_i} \mathcal{L})$, given some learning rate η. In this way, we guarantee that the transition matrix stays in the feasible set $\mathcal{K}^L$ throughout training. Moreover, in each epoch, we derive $\boldsymbol{\pi}$ directly by solving (7) immediately after this projection step. The pseudo-code of the framework for gradient descent in an MNN is available in Algorithm 1, given dataset $\boldsymbol{D}$, some loss function $\mathcal{L}$, learning rate $\boldsymbol{\eta}$, and maximum number of epochs m.

4) Unsupervised Fine-Tuning: The main advantage of an MNN is that it intrinsically models the (Markovian) character of the temporal regime through $\boldsymbol{\lambda}$ for some dataset $\boldsymbol{D}$. However, the training procedure outputs a fixed parametrization $\boldsymbol{\Psi}$ for all $\mathbf{o} \in \boldsymbol{D}$. The fact that the underlying model is coupled to a strong statistical prior also makes it sensitive to observations sampled from a different temporal regime from that of its training set. Herein, we propose a routine that effectively allows the model to adapt to unseen or unusual observations by means of an unsupervised loss function. Thus, we adapt Algorithm 1 so that it leverages (6) for the specific purpose of increasing the likelihood of some target observation $\mathbf{o}$ given baseline parametrization $\boldsymbol{\Psi}$. Note that we fine-tune one model per observation and re-start the procedure with baseline $\boldsymbol{\Psi}$ given a different datum. The unsupervised fine-tuning routine is described in Algorithm 2.

Hereafter, let $\boldsymbol{\Psi}_k$ be the parametrization after k rounds of fine-tuning on the same observation $\mathbf{o}$ given baseline $\boldsymbol{\Psi}$.

III. Methods

In Section II we introduced formalism for a broader family of MNNs. Herein, we will instantiate an MNN explicitly for

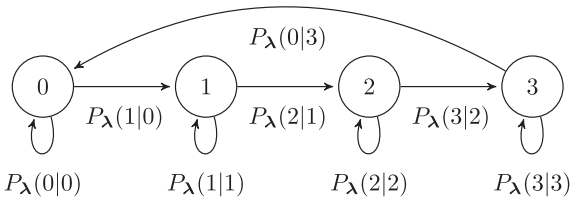

Fig. 3. Left-to-right HMM parameterized by λ for PCG segmentation containing 4 states: S1 (0), systole (1), S2 (2) and diastole (3).

Algorithm 2: Fine-Tuning Pre-Trained MNN $\boldsymbol{\Psi}$ on $\mathbf{o}$.

Input: $\boldsymbol{\Psi}$, $\mathbf{o}$, η, m
Output: $\boldsymbol{\Psi}_m$
1 $\boldsymbol{\Psi}_0 \leftarrow \boldsymbol{\Psi}$;
2 **for** $k = 1$ **to** m **do**
3 $\quad \boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} + \eta \nabla_{\boldsymbol{\Theta}} \mathcal{L}_{\text{FT}}(\mathbf{o}; \boldsymbol{\Psi}_{k-1})$;
4 $\quad \boldsymbol{\Gamma} \leftarrow \boldsymbol{\Gamma} + \eta \nabla_{\boldsymbol{\Gamma}} \mathcal{L}_{\text{FT}}(\mathbf{o}; \boldsymbol{\Psi}_{k-1})$;
5 $\quad$ Lines **9-12** of Algorithm 1;
6 $\quad \boldsymbol{\Psi}_k \leftarrow \{\{\boldsymbol{\pi}, \boldsymbol{\Gamma}\}, \boldsymbol{\Theta}\}$;
7 **end**

PCG fundamental heart sound segmentation. First, we establish an adequate feature extraction pipeline that serves the models throughout our experiments. Then we define the characteristics of the parameters in $\boldsymbol{\lambda}$ alongside their feasible set in the context of fundamental heart sound segmentation. We conclude this section with two distinct, relevant characterizations of MNNs for this task.

A. Preprocessing

Replicating [40], the signals are first filtered using a Butterworth filter of order two with pass-band [25, 400] Hz and then downsampled to 1000 Hz. After applying the spike removal algorithm proposed by [35] we extract two distinct feature maps: the *envelograms*, $\boldsymbol{X}_{\text{Env}}$, as in Renna et al. [40], and the static MFCCs, $\boldsymbol{X}_{\text{MFCC}}$, following Fernando et al. [42]. Our models adopt the same feature extraction setup described in [39]. Specifically, $\boldsymbol{X}_{\text{Env}}$ is comprised of 4 channels (one for each feature): i) the homomorphic envelogram, ii) the Hilbert envelope, iii) the wavelet envelope, and iv) the power spectral density envelope. These features are downsampled to 50 Hz and normalized so that each channel has zero mean and unit variance. Concerning $\boldsymbol{X}_{\text{MFCC}}$, following Fernando et al. [42], we extract 6 static Mel-frequency Cepstral Coefficients (MFCC), alongside their first and second order frame differences, Δ and Δ^2. Each model processes segments of 64 samples, which is equivalent to approximately 1.3 seconds. We assume that the label at the central position of each segment (i.e., $64/2 = 32$, or around 0.65 seconds) is the ground truth label for the entire frame. The value of this label is one-hot-encoded.

Henceforth, assume that $L = 4$ and that $\mathcal{S}$ maps the set of possible states: S1, systole, S2, and diastole (as in Fig. 3).

B. Left-to-Right MNN for PCG Segmentation

Heart sound segments defined by fundamental heart sounds obey three important properties: a) they occur in a cyclic fashion, b) their occurrence order is unchanged throughout time (i.e., S1, systole, S2, diastole), c) no segment lasts forever. This domain information is expressed organically by a left-to-right HMM as illustrated in Fig. 3. More formally:

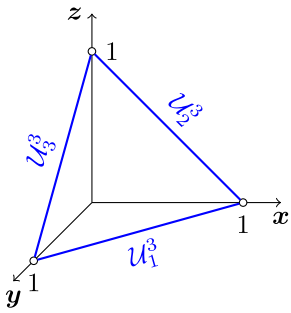

Fig. 4. Feasible set for the columns of $\boldsymbol{\Gamma} \in \mathbb{R}^{3\times3}$, where $\mathcal{U}^3 = \{\mathcal{U}_1^3, \mathcal{U}_2^3, \mathcal{U}_3^3\}$. Note how $\mathcal{U}^3 \subset \mathcal{K}^3$. More specifically, it is comprised of the edges of $\mathcal{K}^3$ while excluding its vertices.

Definition 1: Let $\mathcal{U}_i^L$ be the point set defined as:

$$\mathcal{U}_i^L = \big\{\mathbf{x} \in \mathbb{R}^L \,|\, \forall j \in \mathcal{S} : (j = i + 1 \mod L) \Rightarrow (x_i, x_j > 0, x_i + x_j = 1) \text{ else } x_j = 0\big\},$$

for some $i \in \mathcal{S}$. Then, let:

$$\mathcal{U}^L := \bigcup_{i \in \mathcal{S}} \mathcal{U}_i^L.$$

Clearly, $\mathcal{U}^L \subseteq \mathcal{K}^L$. A matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{L\times L}$ is called *non-absorbent left-to-right* if it is defined as $\boldsymbol{\Gamma} = [\boldsymbol{\Gamma}_0^{\mathrm{T}}, \boldsymbol{\Gamma}_1^{\mathrm{T}}, \ldots, \boldsymbol{\Gamma}_{L-1}^{\mathrm{T}}]^{\mathrm{T}}$, where $\boldsymbol{\Gamma}_i \in \mathcal{U}_i^L$, for all $i \in \mathcal{S}$.

One can visualize $\mathcal{U}_i^L$ as an edge of the L-dimensional canonical simplex $\mathcal{K}^L$ (see Fig. 4 for an example in $\mathbb{R}^3$). We are interested in guaranteeing that the gradient updates with regards to $\boldsymbol{\lambda}$ in Algorithms 1 (lines 8 to 11) and 2 stay in $\mathcal{U}^L$. Hence, we need to compute a map Φ from any $\mathbf{x} \in \mathbb{R}^L$ to the closest point in $\mathcal{U}_i^L \subseteq \mathcal{U}^L$ for every $i \in \mathcal{S}$ (following (9)). This computation should be efficient as it will be performed during each step of gradient descent throughout the training routines. We adapted Michelot's finite projection algorithm [49] restricted to the two non-zero components for each $\boldsymbol{\Gamma}_i$. In order to guarantee a solution in $\mathcal{U}_i^L$, we add/subtract a small perturbation ϵ to the maximum/minimum component of the solution obtained with this algorithm, thus avoiding absorbing states permitted in $\mathcal{K}^L$.

C. Inference and Post-Processing

We developed the MNNs for PCG segmentation using a standard Viterbi decoder [45] to output the most likely sequence produced by the MNN. Thus, the sequence is decoded as follows:

$$\hat{\mathbf{s}}^{(j)} = \max_{s_1, s_2, \ldots, s_{T_j}} P_{\boldsymbol{\Psi}}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T_j}, s_1, s_2, \ldots, s_{T_j}), \quad (10)$$

so that $s_1, s_2, \ldots, s_{T_j} \in \mathcal{S}$.

IV. Experiments

We conducted our experiments on the 2016 PhysioNet and 2022 CirCor DigiScope datasets [50], [51]. Onwards, we will

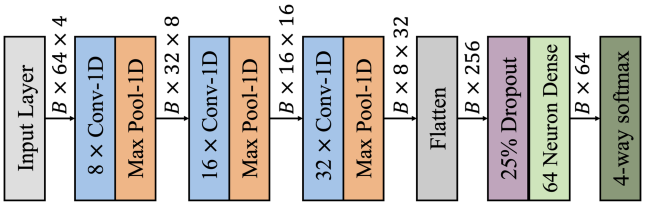

Fig. 5. Template of the one-dimensional CNN backbone for the MNNs implemented in our experiments assuming a mini-batch size B. The output layer is a 4-way softmax yielding the posterior estimates for each of the PCG states.

refer to each dataset as PhysioNet'16 and CirCor'22, respectively. Both databases contain noisy sounds collected in clinical and non-clinical environments (e.g., during screening campaigns). The raw datasets are fed independently into our preprocessing pipeline (see Section III-A), where we extract $\boldsymbol{X}_{\text{Env}}$ and $\boldsymbol{X}_{\text{MFCC}}$. We adopt a 10-fold cross-validation setup throughout our experiments, where the folds have a fixed length of roughly 10% of the length of the dataset in its entirety. We set aside 10% of the out-of-fold dataset as a validation set so that we can employ early stopping within each fold. Thus, we will have a train, validation, and test split for each of the ten dataset partitions generated for cross-validation. We also ensure that all folds are mutually exclusive with regards to their patient identifiers. The sounds that have matching patient identifiers to those of each fold are removed from the respective train and validation sets. Our experimental methodology can be separated into 3 main components: (Section IV-C1) establishment of an MNN baseline in PhysioNet'16, (Section IV-C2) performance assessment of the baseline MNN *versus* a statistical sequential model and two existing data-driven solutions in PhysioNet'16 and CirCor'22, and (Section IV-C3) cross dataset assessment of the performance of the unsupervised fine-tuning given an MNN pre-trained on either PhysioNet'16 or CirCor'22. We selected a simple one-dimensional Convolutional Neural Network (CNN) architecture (see Fig. 5) as the ANN component of our MNN. It is built from a convolutional stem comprised of three blocks of one-dimensional convolutions with kernels of size 3 and rectified linear unit (ReLU) activation functions. These are interlaced with 2×1 max-pooling layers. Each block stacks 8, 16, and 32 convolutional filters, respectively. Finally, we employ a 25% dropout bottleneck layer, followed by a fully connected layer spanning 64 neurons with the ReLU activation function. The output layer implements a 4-way softmax.

We compare our results with the well established HSMM approach by Springer et al. [39] and two recent data-driven PCG segmentation solutions: the sliding U-Net followed by Viterbi algorithm proposed by Renna et al. [40], and the Bi-LSTM+A proposed by Fernando et al. [42]. We use the preprocessing configurations associated with the best results reported by each author; hence, we use $\boldsymbol{X}_{\text{Env}}$ for the HSMM and U-Net, and $\boldsymbol{X}_{\text{MFCC}}$ for the Bi-LSTM+A. The MNNs and U-Net were trained with a mini-batch of size 1, and the Bi-LSTM+A with a mini-batch of 32 samples. We used the *Adam* gradient descent algorithm [52]. The MNNs and U-Net were trained using a learning rate of 0.001. The Bi-LSTM+A was trained with an initial learning rate of 0.002, and 0.0002 from epoch 10 onwards, as proposed by the authors [42]. Both the U-Net and the Bi-LSTM+A were trained to minimize the cross-entropy. Training was performed during 50 epochs throughout each fold and the value of the loss function in the validation set was used as the early-stopping criterion in all experiments.

A. Materials

1) The PhysioNet 2016 Dataset (PhysioNet'16): The public dataset used for the 2016 Computing in Cardiology (CinC)/PhysioNet Challenge [50] served as the initial baseline for our experiments. It is a repository of 9 different heart sound databases collected by different international research groups. It amasses a sizeable amount of PCG recordings collected from aortic, pulmonary, tricuspid, and mitral auscultation locations. It spans 2435 recordings from 1297 healthy or pathological patients, the latter spanning a variety of conditions, including heart valve and coronary artery diseases, aortic stenosis, and mitral regurgitation. It includes a fairly large amount of noisy recordings with real-world acquisition conditions. These recordings were originally re-sampled at 2000 Hz with anti-aliasing [50]. Of the set of original recordings, we use only the 792 heart sounds (181 healthy, 611 pathological from a total of 135 patients) that have an associated ECG recording.[2] The fundamental heart sound sequence was estimated through analysis of the synchronous ECG recordings following [39]. The duration of each recording in this set ranges from 1 to 35.5 seconds. We discarded 39 samples, accounting for the cases where the signal lasted less than 1 s or had noisy labels (i.e., illegal state sequences, such as those that allow transitions from state S1 directly to state S2).

2) The CirCor DigiScope Dataset (CirCor'22): The CirCor DigiScope Dataset [51] (CirCor'22) is currently the largest publicly available pediatric heart sound dataset and it was featured in the 2022 George B. Moody PhysioNet Challenge [14]. The focus of the tasks enabled by this dataset was that of cardiac murmur detection and classification. The CirCor'22 dataset was assembled using data collected from two different screening campaigns in Northeast Brazil with specific focus on the pediatric population [51]. The study included all subjects who volunteered for screening within the study period. Patients younger than 21 years of age with a parental signed consent form (when appropriate) were included, with no further exclusion criteria. Two cardiac physiologists independently analyzed the collected sounds and discarded PCG recordings based on a signal quality standard. Thus, the database spans 5282 PCG recordings collected from the aortic, pulmonary, tricuspid, and mitral locations from 1568 healthy or pathological subjects (1144 healthy, 305 pathological, 119 indiscernible). The signals were sampled at 4000 Hz with 16-bit resolution, with durations ranging from 5.3 to 80.4 seconds. The ground truth labels were provided by the same two blinded cardiac physiologists, also tasked with reviewing and correcting the automatic annotation recommendations proposed by the automatic segmentation algorithms of Springer et al. [39], Oliveira et al. [37], and Renna et al. [40].

[2]The dataset is available at https://PhysioNet.org/physiotools/hss/

Of the 5282 recordings, we used the 3279 samples publicly available in the training set for the 2022 PhysioNet Challenge.[3]

B. Metrics and Evaluation

We follow the instructions of [35] to compute the confusion matrix in our experiments. Specifically, a prediction is considered a true positive if the center of an S1 (S2) prediction is closer than 60 ms to the next S1 (S2) sound in the true sequence. All other predicted S1 or S2 segments are considered false positives. Given these assumptions, the Positive Predictive Values ($\mathbf{PPV}$) and Sensitivities ($\mathbf{S}$) are computed according to the resulting confusion matrix in a standard fashion [35].

These metrics are only tractable if no illegal transitions are present in the model prediction sequence. Thus, our implementation of the Bi-LSTM+A enforces temporal framing of the prediction sequences through Viterbi decoding using the maximum likelihood estimates of the train set (as described in Section II-A2).

C. Results

1) MNN Baseline Benchmark: We start by ablating the joint search of $\mathbf{\Psi} = \{\boldsymbol{\lambda}, \mathbf{\Theta}\}$ described in Algorithm 1 in the PhysioNet'16 dataset. Thus, we define the *static* formulation of an MNN where the parameters of the underlying HMM stay fixed to $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\text{MLE}}$ throughout training, and a *hybrid* MNN where $\boldsymbol{\lambda}$ is initialized with $\boldsymbol{\lambda}_{\text{MLE}}$ but is jointly-learned with $\mathbf{\Theta}$ throughout training. We denote the static model as $\text{MNN}^{\Theta}_{\mathcal{L}}$ and its hybrid counterpart as $\text{MNN}^{\Psi}_{\mathcal{L}}$. At the same time, we evaluate the difference between setting $\mathcal{L} = \mathcal{L}_{\text{MMI}}$ (5) and $\mathcal{L} = \mathcal{L}_{\text{CL}}$ (4). Inspecting Fig. 6, concerning $\mathcal{L} = \mathcal{L}_{\text{MMI}}$ in particular, the hybrid model $\text{MNN}^{\Psi}_{\mathcal{L}_{\text{MMI}}}$ displays less variability between folds. Moreover, this variant appears more robust in terms of sensitivity, while its static counterpart measures higher median $\mathbf{PPV}$. Notwithstanding, our pair-wise t-test ($\alpha = 0.5$) showed no statistical difference in their inter-fold performance metrics (see Fig. 7).

When one looks at the results for $\mathcal{L} = \mathcal{L}_{\text{CL}}$, the static MNN is considerably more sensitive than its hybrid counterpart while being just marginally inferior in terms of $\mathbf{PPV}$. There was a significant difference between $\text{MNN}^{\Theta}_{\mathcal{L}_{\text{CL}}}$ and $\text{MNN}^{\Psi}_{\mathcal{L}_{\text{CL}}}$ in terms of sensitivity. However, the null hypothesis could not be rejected in terms of $\mathbf{PPV}$.

Finally, a general view of the pair-wise tests in Fig. 7 (in conjunction with the box-plots of Fig. 6) also reveals a statistical difference between $\text{MNN}^{\Theta}_{\mathcal{L}_{\text{CL}}}$ and $\text{MNN}^{\Theta}_{\mathcal{L}_{\text{MMI}}}$ in terms of sensitivity. Concerning $\mathbf{PPV}$, both models optimizing $\mathcal{L}_{\text{CL}}$ were statistically different to $\text{MNN}^{\Theta}_{\mathcal{L}_{\text{MMI}}}$. Thus, $\text{MNN}^{\Theta}_{\mathcal{L}_{\text{CL}}}$ appears to be the superior configuration regarding sensitivity, while $\text{MNN}^{\Theta}_{\mathcal{L}_{\text{MMI}}}$ is better suited for optimal $\mathbf{PPV}$. Henceforth, we choose $\text{MNN}^{\Theta}_{\text{CL}}$ as the baseline for the remaining experiments, since it offers the most efficient paradigm, as $\mathcal{L}_{\text{CL}}$ can be calculated in $\mathcal{O}(T)$, where T is the length of the sound, without the forward-backward algorithm. Moreover, a static MNN does not need the costly gradient re-projection step.

[3]The dataset is available at https://physionet.org/content/circor-heart-sound/1.0.3/

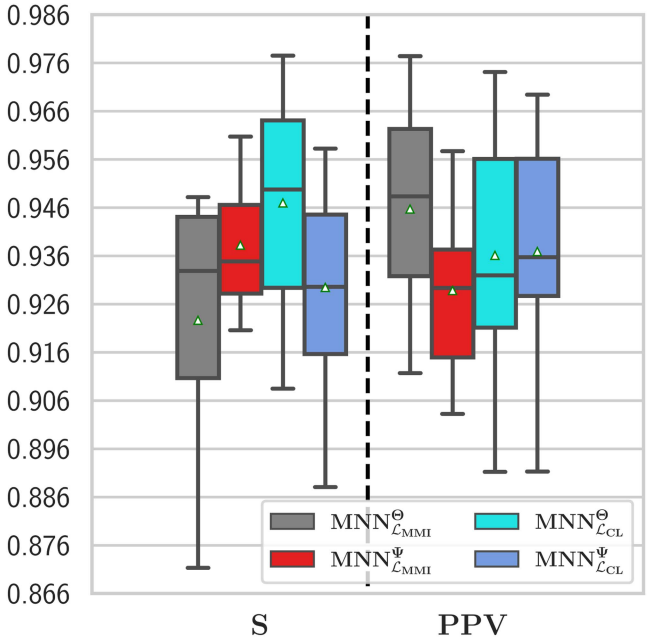

Fig. 6. Measured sensitivity ($\mathbf{S}$) and positive predictive value ($\mathbf{PPV}$) in the 10-fold cross-validation experiment of the PhysioNet'16 dataset for different static/hybrid MNNs, under different loss functions $\mathcal{L} = \mathcal{L}_{\text{MMI}}/\mathcal{L} = \mathcal{L}_{\text{MMI}}$.

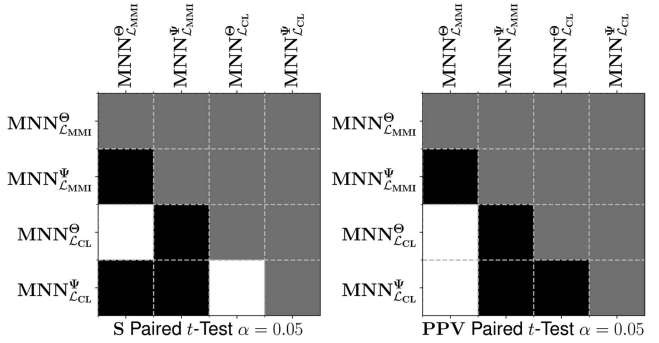

Fig. 7. Pairwise t-test ($\alpha = 0.05$) between $\mathbf{S}$ and $\mathbf{PPV}$ of $\text{MNN}^{\Theta}_{\mathcal{L}_{\text{MMI}}}$, $\text{MNN}^{\Psi}_{\mathcal{L}_{\text{MMI}}}$, $\text{MNN}^{\Theta}_{\mathcal{L}_{\text{CL}}}$ and $\text{MNN}^{\Psi}_{\mathcal{L}_{\text{CL}}}$ in the PhysioNet'16. White cells signify that we can reject the null hypothesis for a pair of models, black cells mark the converse. The grey colour denotes uninformative cells.

2) Model Performance Comparison: We now compare $\text{MNN}^{\Theta}_{\mathcal{L}_{\text{CL}}}$ to the HSMM with logistic regression by Springer et al. [39], the U-Net proposed by Renna et al. [40] and the Bi-LSTM+A introduced by Fernando et al. [42]. We will be performing measurements on 10-fold cross-validations of the PhysioNet'16 and CirCor'22 datasets. We recorded the average performance on each fold in the box-plots of the 10-fold cross-validated experiments in Figs. 8 and 10, alongside pairwise t-tests with a significance level $\alpha = 0.05$ (Figs. 9 and 11).

Concerning PhysioNet'16, an inspection of the box-plots (Fig. 8) reveals that the models that leverage ANNs are substantially more performant than the HSMM. Our proposed MNN displays superior robustness with regards to both $\mathbf{S}$ and $\mathbf{PPV}$. In concrete terms, the proposed $\text{MNN}^{\Theta}_{\mathcal{L}_{\text{CL}}}$ scored an average $\mathbf{S}$ of 0.950 ± 0.022 and $\mathbf{PPV}$ of 0.937 ± 0.025. It scored higher

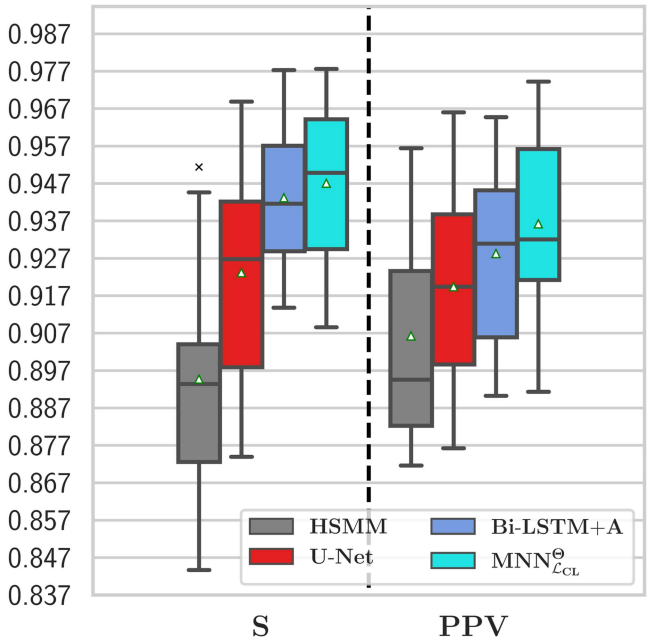

Fig. 8. Measured sensitivity (**S**) and positive predictive value (**PPV**) statistics in 10-fold cross-validation on the CirCor'22 dataset between the HSMM by Springer et al. [39], the U-Net by Renna et al. [40], the Bi-LSTM+A by Fernando et al. [42], and MNN$^{\Theta}_{\mathcal{L}_{CL}}$. The green edged triangles mark the mean of the distribution, while the black crosses signify outliers. Note how the MNN with fixed λ during gradient descent using the complete likelihood loss (MNN$^{\Theta}_{\mathcal{L}_{CL}}$) yields the best trade-off between the two metrics.

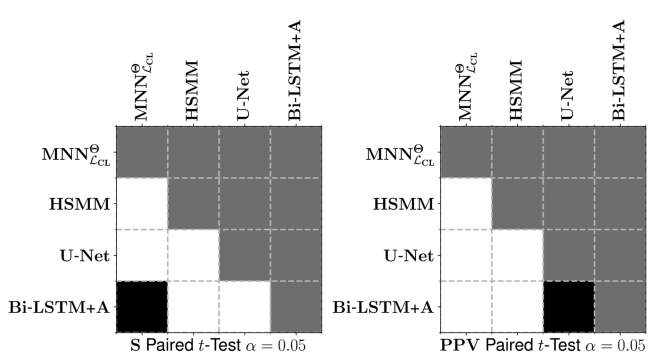

Fig. 9. Pairwise t-test between MNN$^{\Theta}_{\mathcal{L}_{CL}}$, the HSMM by Springer et al. [39], the U-Net by Renna et al. [40], and the Bi-LSTM+A by Fernando et al. [42] with significance $\alpha = 0.05$ in the PhysioNet'16 dataset. White cells signify that we can reject the null hypothesis for a pair of models, black cells mark the converse. The grey colour denotes uninformative cells.

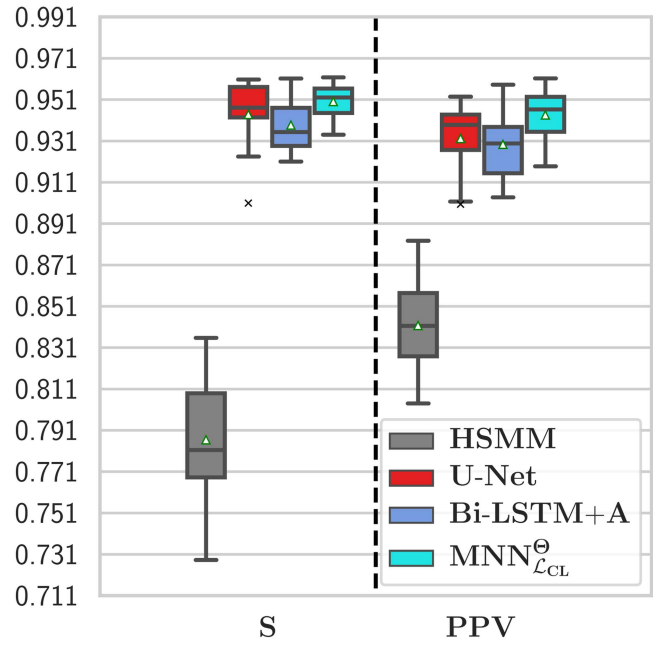

Fig. 10. Measured sensitivity (**S**) and positive predictive value (**PPV**) statistics in 10-fold cross-validation on the CirCor'22 dataset between the HSMM by Springer et al. [39], the U-Net by Renna et al. [40], the Bi-LSTM+A by Fernando et al. [42], and MNN$^{\Theta}_{\mathcal{L}_{CL}}$. The green edged triangles mark the mean of the distribution, while the black crosses signify outliers. Note how the MNN with fixed λ during gradient descent using the complete likelihood loss (MNN$^{\Theta}_{\mathcal{L}_{CL}}$) yields the best trade-off between the two metrics.

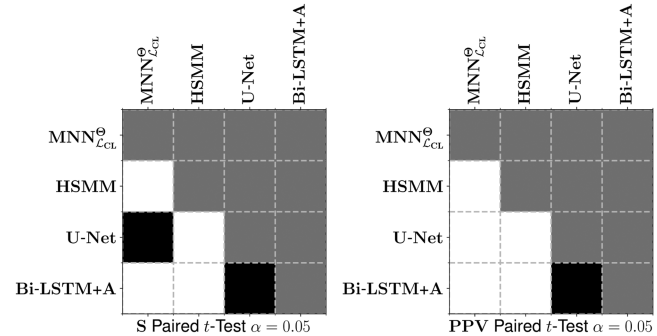

Fig. 11. Pairwise t-Test between MNN$^{\Theta}_{\mathcal{L}_{CL}}$, the HSMM by Springer et al. [39], the U-Net by Renna et al. [40], and the Bi-LSTM+A by Fernando et al. [42] with significance $\alpha = 0.05$ in the CirCor'22 dataset. White cells signify that we can reject the null hypothesis for a pair of models, black cells mark the converse. The grey colour denotes uninformative cells.

on average then all other models. However, in terms of **S**, we could not find a significant difference between the Bi-LSTM+A, which measured an average of 0.944 ± 0.020 **S**. On the other hand, with respect to **PPV**, it was significantly superior to all other approaches by a considerable margin, with an average 1.5% improvement over the 0.929 ± 0.026 **PPV** observed in the Bi-LSTM+A, the second best scoring model. With respect to the CirCor'22 dataset, the gap between the HSMM and the data-driven solutions becomes even wider. The measurements in Fig. 10 reveal that the superiority of our model is even more pronounced in this (arguably) more challenging dataset (see Fig. 10). Our model scored 0.950 ± 0.008 **S** and 0.943 ± 0.012 **PPV**. Concerning **S**, we could not prove statistical difference with the U-Net (see pairwise t-tests in Fig. 11), which scored 0.944 ± 0.018. However, one should note that the mean and variance of MNN$^{\Theta}_{\mathcal{L}_{CL}}$ are still higher. Moreover, the U-Net has a sizeable outlier in one of the folders, where it scored just 0.901 **S**. Concerning **PPV**, the MNN$^{\Theta}_{\mathcal{L}_{CL}}$ is significantly superior to all other models, scoring an additional 1.1% **PPV** on average over the U-Net, its runner-up, which registered an average 0.932 ± 0.017 **PPV**.

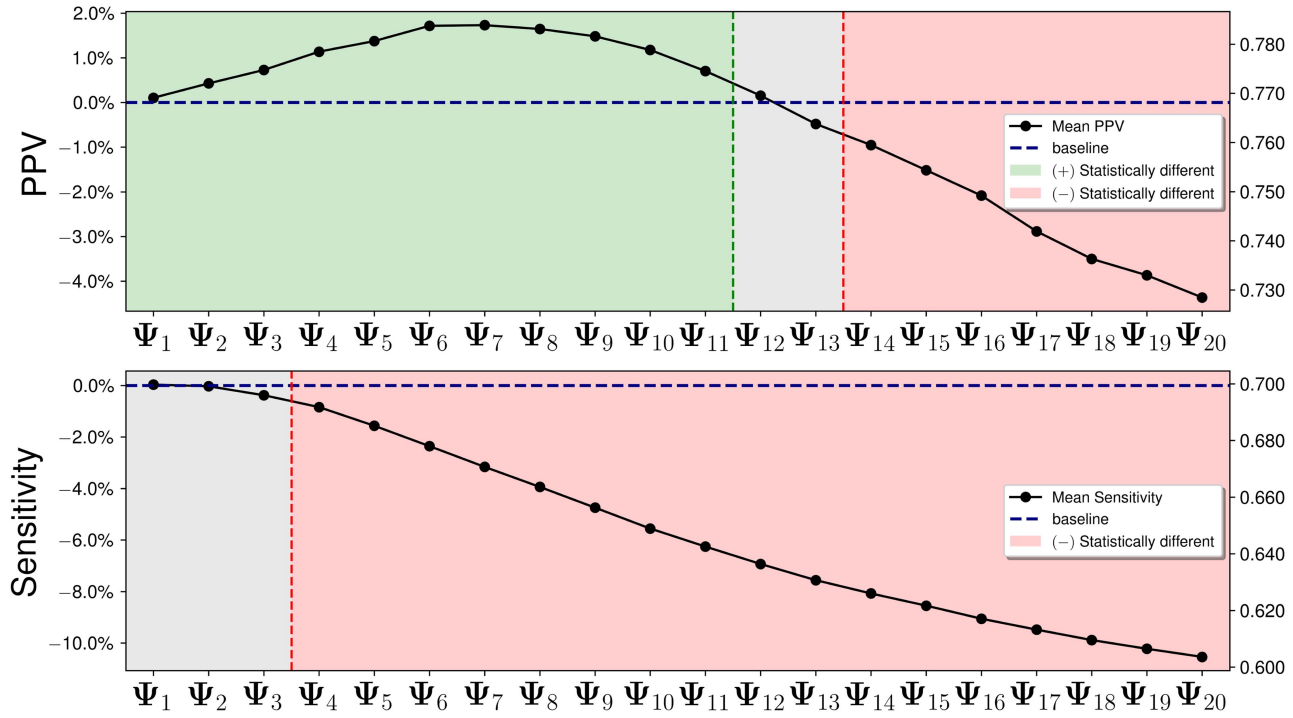

Fig. 12. Fine-tuning average performance of $\text{MNN}^{\Psi_0}_{\mathcal{L}_{\text{CL}}}$ compared with $\text{MNN}^{\Psi_i}_{\mathcal{L}_{\text{FT}}}$, for $1 \leq i \leq 20$. The baseline is pre-trained on PhysioNet'16 and fine-tuned throughout different number of epochs to the CirCor'22 dataset. (Top) $\mathbf{PPV}$. (Bottom) Sensitivity. A pairwise t-test between the baseline metric distribution and the outcome after i epochs was made to assess statistically significant differences. The left vertical axis displays average relative changes in percentage and the right horizontal the average measured values.

3) Fine-Tuning MNN on Unseen Data: In this section we measure the effects of a pre-trained static MNN in the PhysioNet'16/CirCor'22 as the *source* dataset when fine-tuned to CirCor'22/PhysioNet'16 as the *target* dataset. For each experiment, we set $\mathcal{L} = \mathcal{L}_{\text{CL}}$ and use random holdout with 80/10/10 split with early-stopping at the best loss value. The parameters of the model attained at this pre-training stage are denoted as $\mathbf{\Psi}_0$, so that $\mathbf{\Psi}_0 = \{\boldsymbol{\lambda}_{\text{MLE}}, \boldsymbol{\Theta}\}$ depends on the train set split of the source dataset. Then, using Algorithm 2, we fine-tuned the model in a hybrid fashion on each observation and recorded the impact of each additional round of fine-tuning until $k = 20$ epochs (Figs. 12 and 13). We measure the mean performance of each metric in the same folders of the previous experiments for each $\text{MNN}^{\Psi_i}_{\mathcal{L}_{\text{FT}}}$. Pairwise t-tests between $\text{MNN}^{\Psi_0}_{\mathcal{L}_{\text{CL}}}$ and $\text{MNN}^{\Psi_i}_{\mathcal{L}_{\text{FT}}}$, $1 \leq i \leq k$ with significance $\boldsymbol{\alpha} = 0.05$ (Fig. 12) were performed to grasp the statistical significance in $\mathbf{PPV}$ and $\mathbf{S}$.

When PhysioNet'16 is used as the source and CirCor'22 as the target datasets, the fine-tuning procedure tends to result in an increase in $\mathbf{PPV}$ at the cost of $\mathbf{S}$. The results in Fig. 12 show that $\mathbf{PPV}$ increases or stays statistically the same for $i \leq 13$. Afterwards, the overall performance degrades, and each subsequent round of fine-tuning is completely detrimental to the performance of the model. More specifically, the baseline $\text{MNN}^{\Psi_0}_{\mathcal{L}_{\text{CL}}}$ mean scores are 0.769 $\mathbf{PPV}$ and 0.704 $\mathbf{S}$. At iteration $i = 2$ it is possible to have a small increase in average $\mathbf{PPV}$, 0.777 (0.80% statistically significant increase), compromising $\mathbf{S}$ to 0.703 (0.01% statistically non-significant decrease). For larger numbers of i, the trade-off in $\mathbf{S}$ is always significant.

The results are more promising when CirCor'22 is the source and PhysioNet'16 the target dataset (see Fig. 13). A steady increase in $\mathbf{PPV}$ is observed throughout all i with a positive or negligible impact in $\mathbf{S}$. In fact, baseline $\text{MNN}^{\Psi_0}_{\mathcal{L}_{\text{CL}}}$ scored a mean $\mathbf{PPV}$ of 0.847 and mean $\mathbf{S}$ of 0.891, while the model $\text{MNN}^{\Psi_{20}}_{\mathcal{L}_{\text{FT}}}$ scores the best average $\mathbf{PPV}$ at 0.886 (3.90% statistically significant increase) with mean sensitivity of 0.889 (0.20% statistically non-significant decrease), which is a substantial improvement in performance.

D. Discussion

The fact that the proposed MNN algorithm supported by a very simple CNN (18785 parameters) consistently outperforms models with 9.52 and 3.40 times the number of parameters, such as the U-Net (178828 parameters) and Bi-LSTM+A (63908 parameters), highlights the impact the Markovian inductive bias has for the task of PCG heart sound segmentation. Note that this is true for both PhysioNet'16 and CirCor'22, regardless of the sizeable co-variate shift in age groups, the heart sound duration statistics (see Fig. 15), and the overall average length of the recordings between the two datasets. Furthermore, even for CirCor'22, where one would expect the larger dataset (roughly 4.12 times the number of samples of PhysioNet'16) to benefit more complex networks, our experiments show that the MNNs are still superior.

We surmise that MNN performance could be further unlocked by increasingly more complex discriminators or feature extraction components. We believe this to be especially likely for the

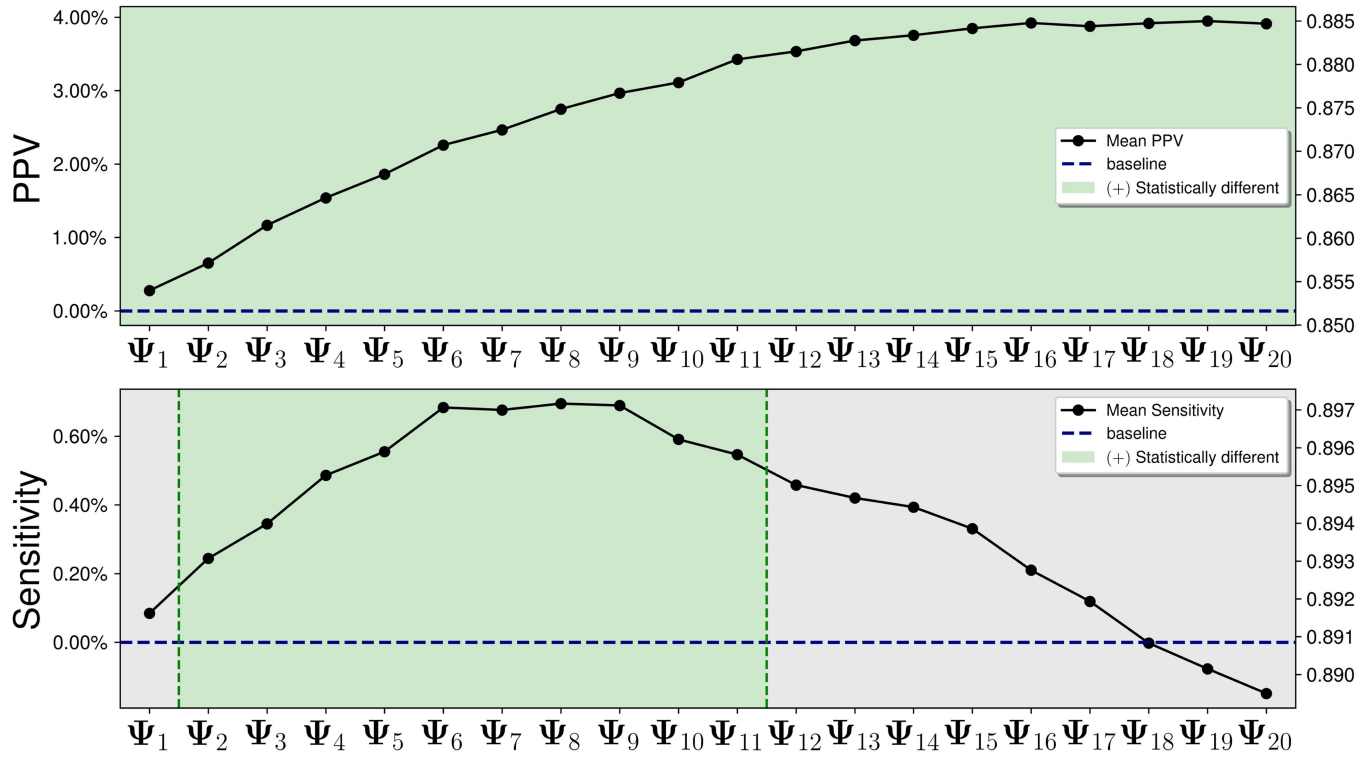

Fig. 13. Fine-tuning average performance of $\text{MNN}^{\Psi_0}_{\mathcal{L}_{\text{CL}}}$ compared with $\text{MNN}^{\Psi_i}_{\mathcal{L}_{\text{FT}}}$, for $1 \leq i \leq 20$. The baseline is pre-trained on CirCor'22 and fine-tuned throughout different number of epochs to the PhysioNet'16 dataset. (Top) $\mathbf{PPV}$. (Bottom) Sensitivity. A pairwise t-test between the baseline metric distribution and the outcome after i epochs was made to assess statistically significant differences. The left vertical axis displays average relative changes in percentage and the right horizontal the average measured values.

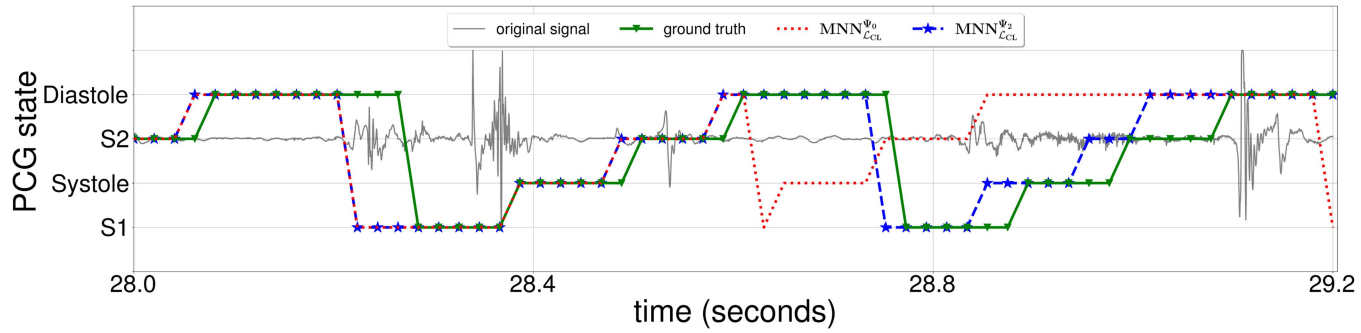

Fig. 14. Predicted state sequences of an unseen CirCor'22 sound using a pre-trained $\text{MNN}^{\Psi_0}_{\mathcal{L}_{\text{CL}}}$ on PhysioNet'16 as baseline (dotted red line) and an MNN fine-tuned before the sensitivity significance threshold $\text{MNN}^{\Psi_2}_{\mathcal{L}_{\text{FT}}}$ (dashed blue line with star markers). Note how $\text{MNN}^{\Psi_0}_{\mathcal{L}_{\text{CL}}}$ deviates from the underlying regime, specifically from the diastole in the interval [28.4, 28.8] seconds onward. This behaviour is mitigated by $\text{MNN}^{\Psi_2}_{\mathcal{L}_{\text{FT}}}$, which effectively reduces the differences in rhythmic structure between the real and predicted state sequences.

unsupervised fine-tuning case since we found that the gradient of $\mathcal{L}_{\text{CL}}$ depended mostly on the CNN, with the underlying HMM parameters only undergoing very subtle changes during this type of training. Algorithm 2 effectively allows MNNs to adapt to the rhythmic structure of unseen sounds sampled from different latent distributions, as it can be observed in Fig. 14. On the other hand, our experiments reveal that the success of this procedure heavily depends on the source and target datasets.

For the case when CirCor'22 is used as the source domain, we did not observe any significant decrease in performance as the number of epochs increased. We suppose that pre-training in a larger source dataset with more variation in its observations may yield a more robust discriminator, which in turn results in a more stable fine-tuning procedure. When PhysioNet'16 was used as the source domain, the model quickly became over-tuned as the number of epochs increased. It is still unclear how this over-tuning phenomenon could be avoided in real-world scenarios where access to labeled data might be impossible.

We note that the parameter set $\boldsymbol{\lambda}$ can be seen as a global descriptor of the statistics of the sequences produced by the MNN, but lacks local information. We envision that statistics on

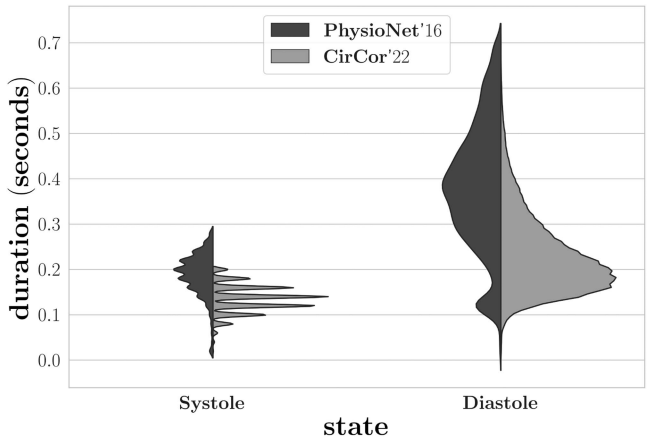

Fig. 15. Distributions of the systole and diastole heart duration in the PhysioNet'16 (dark gray) and CirCor'22 (light gray). In order to improve readability, we only accounted for values within two standard deviations of the mean when plotting the distributions.

local invariances of the signal could be used as additional prior information to the MNN in order to mitigate over-tuning. Other limitations of this study are: the fact that, in section III-B, we assume that fundamental heart sounds occur in an unchanged cyclic fashion, which is true from a strictly physiologically perspective. However, these assumptions are too strong once we appreciate that mistakes may occur during sound collection (e.g., accidentally lifting the stethoscope). Note that the proposed MNN framework is already sufficiently generic to account for this phenomenon. One needs only to include an extra state for such an occasion, and change the underlying automaton from left-to-right to a more densely connected first order HMM. Furthermore, we only studied HMMs as the inductive bias for the time modulation of PCG heart signals. Notwithstanding, our framework is easily extended to support more sophisticated priors, such as an adaptation of the HSMM model proposed by Springer et al. [39].

Finally, we note that our framework has some computational performance barriers, especially during training. The proposed loss functions (4), (6) and (5) require sequential computation of the exact likelihood estimates, which requires mini-batches of size 1. Consequently, although the model is less complex and its training is theoretically more time efficient, it is nonetheless less parallelizable, hence its GPU training is lengthy compared to the more complex Bi-LSTM+A, since the latter can implement mini-batch sizes greater than 1. On the other hand, our MNN instantiation is faster during inference than the U-Net and the Bi-LSTM+A given its very simple discriminator and feature extraction pipeline. We note that in order to guarantee physiological valid predictions, all models share the same limiting factor: the decoder. In fact, we are contemplating non-sequential extensions of our framework for real-time screening, since the Viterbi algorithm requires complete observation of a sound in order to decode its output sequence. A possible next step in this direction is coupling our MNN with an online short-time Viterbi decoder [53], thus allowing valid sequences to be emitted by the MNN in real time.

V. Conclusion

In this article, we introduced MNNs by formalizing a set of principles that allow the embedding of some underlying HMM with a highly discriminant data-driven neural network through a unifying loss function. We proposed algorithms encapsulating gradient-descent strategies for supervised and unsupervised learning that guarantee the Markovian assumption holds using a projective strategy of the gradient updates. We instantiated left-to-right (non-absorbent) MNNs for the downstream task of phonocardiogram fundamental heart sound segmentation, where we showed the superiority of the novel framework compared to two recent fully data-driven architectures. The experiments were based on the results measured in two publicly available datasets: PhysioNet'16 and CirCor'22. For the supervised case, we showed that the simplest MNN leveraging one-dimensional feature maps is superior to both the U-Net and Bi-LSTM+A. Was also showed that an MNN is adaptive to new datum sampled from dissimilar distribution by means of the proposed unsupervised fine-tuning procedure.

References

[1] World Heart Federation, "Cardiovascular diseases - global facts and figures," 2021.[Online]. Available: https://world-heart-federation.org/resource/cardiovascular-diseases-cvds-global-facts-figures/

[2] S. Mendis, P. Puska, and B. Norrving, *Global Atlas on Cardiovascular Disease Prevention and Control*. Geneva, Switzerland: World Health Organization, 2011.

[3] S. Mangione, "Cardiac auscultatory skills of physicians-in-training: A comparison of three english-speaking countries," *Amer. J. Med.*, vol. 110, no. 3, pp. 210–216, 2001.

[4] A. A. Ishmail, S. Wing, J. Ferguson, T. A. Hutchinson, S. Magder, and K. M. Flegel, "Interobserver agreement by auscultation in the presence of a third heart sound in patients with congestive heart failure," *Chest*, vol. 91, no. 6, pp. 870–873, 1987.

[5] H. Vermarien, "Phonocardiography," in *Encyclopedia of Medical Devices and Instrumentation*, 2nd ed., J. G. Webster, Ed., vol. 5. Hoboken, NJ, USA: Wiley, 2006, pp. 278–290.

[6] I. Bank, H. W. Vliegen, and A. V. Bruschke, "The 200th anniversary of the stethoscope: Can this low-tech device survive in the high-tech 21st century?," *Eur. Heart J.*, vol. 37, no. 47, pp. 3536–3543, 2016.

[7] S. Leng, R. S. Tan, K. T. C. Chai, C. Wang, D. Ghista, and L. Zhong, "The electronic stethoscope," *Biomed. Eng. Online*, vol. 14, no. 1, pp. 1–37, 2015.

[8] W. Zhang, J. Han, and S. Deng, "Abnormal heart sound detection using temporal quasi-periodic features and long short-term memory without segmentation," *Biomed. Signal Process. Control*, vol. 53, 2019, Art. no. 101560.

[9] P. T. Krishnan, P. Balasubramanian, and S. Umapathy, "Automated heart sound classification system from unsegmented phonocardiogram (PCG) using deep neural network," *Phys. Eng. Sci. Med.*, vol. 43, pp. 505–515, 2020.

[10] B. Xiao et al., "Follow the sound of children's heart: A deep-learning-based computer-aided pediatric CHDs diagnosis system," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1994–2004, Mar. 2020.

[11] V. Maknickas and A. Maknickas, "Recognition of normal–abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients," *Physiol. Meas.*, vol. 38, no. 8, 2017, Art. no. 1671.

[12] T. Nilanon, J. Yao, J. Hao, S. Purushotham, and Y. Liu, "Normal/abnormal heart sound recordings classification using convolutional neural network," in *Proc. IEEE Comput. Cardiol. Conf.*, 2016, pp. 585–588.

[13] J. Oliveira, D. Nogueira, F. Renna, C. Ferreira, A. M. Jorge, and M. Coimbra, "Do we really need a segmentation step in heart sound classification algorithms?," in *Proc. IEEE 43rd Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2021, pp. 286–289.

[14] M. A. Reyna et al., "Heart murmur detection from phonocardiogram recordings: The George B. Moody PhysioNet challenge 2022," in *Proc. IEEE Comput. Cardiol.*, 2022, pp. 1–4.

[15] 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.

[16] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," in *Proc. IEEE Comput. Cardiol.*, 1997, pp. 105–108.

[17] S. Ari, P. Kumar, and G. Saha, "A robust heart sound segmentation algorithm for commonly occurring heart valve diseases," *J. Med. Eng. Technol.*, vol. 32, no. 6, pp. 456–465, 2008.

[18] A. Moukadem, A. Dieterlen, N. Hueber, and C. Brandt, "A robust heart sounds segmentation module based on S-transform," *Biomed. Signal Process. Control*, vol. 8, no. 3, pp. 273–281, 2013.

[19] S. Sun, Z. Jiang, H. Wang, and Y. Fang, "Automatic moment segmentation and peak detection analysis of heart sound pattern via short-time modified Hilbert transform," *Comput. Methods Programs Biomed.*, vol. 114, no. 3, pp. 219–230, 2014.

[20] L. Huiying, L. Sakari, and H. Iiro, "A heart sound segmentation algorithm using wavelet decomposition and reconstruction," in *Proc. IEEE 19th Annu. Int. Conf. Eng. Med. Biol. Soc. 'Magnificent Milestones Emerg. Opportunities Med. Eng.'*, 1997, pp. 1630–1633.

[21] A. Castro, T. T. V. Vinhoza, S. S. Mattos, and M. T. Coimbra, "Heart sound segmentation of pediatric auscultations using wavelet analysis," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc.*, 2013, pp. 3909–3912.

[22] H. Naseri and M. Homaeinezhad, "Detection and boundary identification of phonocardiogram sounds using an expert frequency-energy based metric," *Ann. Biomed. Eng.*, vol. 41, no. 2, pp. 279–292, 2013.

[23] D. Kumar et al., "Detection of S1 and S2 heart sounds by high frequency signatures," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc.*, 2006, pp. 1410–1416.

[24] J. Vepa, "Classification of heart murmurs using cepstral features and support vector machines," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2009, pp. 2539–2542.

[25] C. N. Gupta, R. Palaniappan, S. Swaminathan, and S. M. Krishnan, "Neural network classification of homomorphic segmented heart sounds," *Appl. Soft Comput.*, vol. 7, no. 1, pp. 286–297, 2007.

[26] T. Chen, K. Kuan, L. A. Celi, and G. D. Clifford, "Intelligent heartsound diagnostics on a cellphone using a hands-free kit," in *Proc. AAAI Spring Symp. Ser.*, 2010, pp. 26–31.

[27] A. C. Stasis, E. Loukis, S. Pavlopoulos, and D. Koutsouris, "Using decision tree algorithms as a basis for a heart sound diagnosis decision support system," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc.*, 2003, pp. 354–357.

[28] J. E. Hebden and J. N. Torry, "Neural network and conventional classifiers to distinguish between first and second heart sounds," in *Proc. IEE Colloq. Artif. Intell. Methods Biomed. Data Process.*, 1996, pp. 1–6.

[29] A. A. Sepehri, A. Gharehbaghi, T. Dutoit, A. Kocharian, and A. Kiani, "A novel method for pediatric heart sound segmentation without using the ECG," *Comput. Methods Programs Biomed.*, vol. 99, no. 1, pp. 43–48, 2010.

[30] T.-E. Chen et al., "S1 and S2 heart sound recognition using deep neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 372–380, Feb. 2017.

[31] L. G. Gamero and R. Watrous, "Detection of the first and second heart sound using probabilistic models," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc.*, 2003, pp. 2877–2880.

[32] D. Gill, N. Gavrieli, and N. Intrator, "Detection and identification of heart sounds using homomorphic envelogram and self-organizing probabilistic model," in *Proc. IEEE Comput. Cardiol.*, 2005, pp. 957–960.

[33] Y.-J. Chung, "Pattern recognition and image analysis, Iberian conference," in *ch. Classification of Continuous Heart Sound Signals Using the Ergodic Hidden Markov Model*. Berlin, Heidelberg: Springer, 2007, pp. 563–570.

[34] S. E. Schmidt, E. Toft, C. Holst-Hansen, C. Graff, and J. J. Struijk, "Segmentation of heart sound recordings from an electronic stethoscope by a duration dependent Hidden-Markov Model," in *Proc. IEEE Comput. Cardiol.*, 2008, pp. 345–348.

[35] S. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, "Segmentation of heart sound recordings by a duration-dependent Hidden Markov Model," *Physiol. Meas.*, vol. 31, no. 4, pp. 513–529, 2010.

[36] J. Oliveira, T. Mantadelis, F. Renna, P. Gomes, and M. Coimbra, "On modifying the temporal modeling of HSMMs for pediatric heart sound segmentation," in *Proc. IEEE Int. Workshop Signal Process. Syst.*, 2017, pp. 1–6.

[37] J. Oliveira, F. Renna, and M. Coimbra, "A subject-driven unsupervised hidden semi-Markov model and Gaussian mixture model for heart sound segmentation," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 323–331, May 2019.

[38] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Support vector machine hidden semi-Markov model-based heart sound segmentation," in *Proc. IEEE Comput. Cardiol. Conf.*, 2014, pp. 625–628.

[39] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-HSMM-based heart sound segmentation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 822–832, Apr. 2016.

[40] F. Renna, J. H. Oliveira, and M. T. Coimbra, "Deep convolutional neural networks for heart sound segmentation," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 6, pp. 2435–2445, Nov. 2019.

[41] E. Messner, M. Zöhrer, and F. Pernkopf, "Heart sound segmentation—An event detection approach using deep recurrent neural networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1964–1974, Sep. 2018.

[42] T. Fernando, H. Ghaemmaghami, S. Denman, S. Sridharan, N. Hussain, and C. Fookes, "Heart sound segmentation using bidirectional LSTMs with attention," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 6, pp. 1601–1609, Jun. 2020.

[43] X. Wang, C. Liu, Y. Li, X. Cheng, J. Li, and G. D. Clifford, "Temporal-framing adaptive network for heart sound segmentation without prior knowledge of state duration," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 2, pp. 650–663, Feb. 2021.

[44] F. Renna, M. L. Martins, and M. Coimbra, "Joint training of hidden Markov model and neural network for heart sound segmentation," in *Proc. IEEE Comput. Cardiol.*, 2021, pp. 1–4.

[45] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[46] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer, 2006.

[47] L. Fritz and D. Burshtein, "Simplified end-to-end MMI training and voting for ASR," 2017, *arXiv:1703.10356*.

[48] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 30th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[49] C. Michelot, "A finite algorithm for finding the projection of a point onto the canonical simplex of α^n," *J. Optim. Theory Appl.*, vol. 50, no. 1, pp. 195–200, 1986.

[50] C. Liu et al., "An open access database for the evaluation of heart sound algorithms," *Physiol. Meas.*, vol. 37, no. 12, pp. 2181–2213, 2016.

[51] 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, Jun. 2022.

[52] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, San Diega, CA, USA, 2015.

[53] J. Bloit and X. Rodet, "Short-time Viterbi for online HMM decoding: Evaluation on a real-time phone recognition task," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 2121–2124.