BeaU-PPG: Uncertainty-aware Heart Rate Estimation from PPG signals via Belief Propagation

Abstract

We present a novel learning-based method that achieves state-of-the-art performance on several heart rate estimation benchmarks extracted from photoplethysmography signals (PPG). We consider the evolution of the heart rate in the context of a discrete-time stochastic process that we represent as a hidden Markov model. We derive a distribution over possible heart rate values for a given PPG signal window through a trained neural network. Using belief propagation, we incorporate the statistical distribution of heart rate changes to refine these estimates in a temporal context. From this, we obtain a quantized probability distribution over the range of possible heart rate values that captures a meaningful and well-calibrated estimate of the inherent predictive uncertainty. We show the robustness of our method on eight public datasets with three-different cross-validation experiments.

1 INTRODUCTION

Photoplethysmography (PPG) is an optical sensing technique that measures the blood volume pulse from the intensity of reflected light at the surface of the skin. PPG is widely used in clinical settings to monitor a patient's vital parameters, including heart rate (HR) and pulse oximetry. [Allen et al., 2021] Because PPG sensors are unobtrusive, they are a common part of wearables such as smartwatches or fitness trackers to extend monitoring of cardiac activity and health to ambulatory settings. The promise of the technology in ambulatory settings is its suitability for early diagnosis and prevention of cardiovascular diseases in the general population.

However, in practice, PPG signals are typically affected by artifacts, especially outside controlled clinical conditions and, thus, where they could prove most useful. Due to the optical sensing principle, changes in ambient light affect resolved intensities as do motion-induced artifacts from activities such as walking, running, or gesturing. These motions result in spurious peaks in the signal, causing additional candidates in the frequency spectrum of possible HR values, thus making the extraction of reliable measurements difficult (Figure 1).

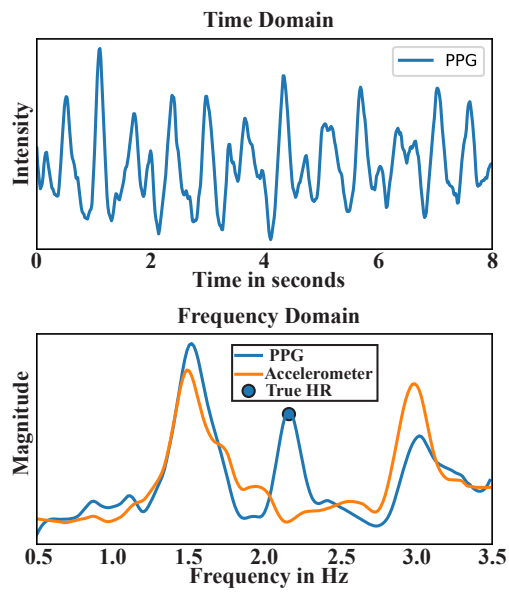

Figure 1: Example of a PPG signal captured with green LED (wavelength: 515 nm) from the IEEE test dataset [Zhang et al., 2015]. Due to motion artifacts, the PPG signal contains several dominant peaks within the range of possible heart rate frequencies.

Therefore, extracting HR from PPG signals that may be noise-afflicted is an important and challenging problem. While methods in classical signal processing have long dominated the field of estimating HR from PPG signals, recent research has shown that these algorithms tend to perform

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

poorly on unseen out-of-distribution (OOD) data [Reiss et al., 2019]. Various deep-learning based methods have since been proposed [Biswas et al., 2019, Chung et al., 2020, Chang et al., 2021]. Importantly, these methods lack the critical capability to indicate when they fail to produce a reliable result, which is essential in the healthcare sector, where erroneous readings may have severe consequences.

In this paper, we propose a novel learning-based method that simultaneously estimates HR from PPG signals alongside the uncertainty of its estimates. The key idea of our method is to consider HR prediction within the context of a time-discrete stochastic process. We model the HR sequence through a probabilistic graphical model that incorporates the statistical transition probabilities between HR classes, such that we can learn a probabilistic relation between instantaneous HR values and observed PPG signal windows through a neural network. Through message passing, our method takes the probability over all possible past HR trajectories into consideration, allowing us to predict a meaningful distribution over the range of discrete HR classes that capture the predictive uncertainty for a HR estimate at a given time step. From the mean of the distribution for each time step, we can obtain single HR estimate.

Through a series of experiments on eight public datasets with three different cross-validation schemes, we empirically demonstrate the robustness of our proposed method. For Leave-one-Subject-out (LoSo), we achieve a mean absolute error (MAE) of 3.57 beats per minute (BPM) on PPG-DaLiA [Reiss et al., 2019], the largest available PPG dataset captured in natural environments. This outperforms related approaches with state-of-the-art (SOTA) performance by 18%, even beating them when our method was only trained on samples of other datasets. We further show that our predicted output distribution is well-calibrated and demonstrate its utility in an experiment where we gradually remove samples with higher predictive uncertainty to further reduce the average prediction error.

1.1 CONTRIBUTIONS

We make the following specific contributions in this paper:

- a learning-based method that estimates heart rate as a hidden stochastic process from a sequence of observed PPG signals, combining the output of a neural network with message passing,
- a novel light-weight neural network architecture for PPG-based HR prediction, and
- a series of experiments that evaluate the calibration of the predicted output distributions and demonstrate SOTA accuracy of our HR estimates, including on out-of-distribution data.

While we demonstrate our method on the problem of PPG-based HR estimation, we believe it could show similarly encouraging results on related temporal sequence estimation problems within other domains.

2 RELATED WORK

The work in this paper is related to deep learning-based uncertainty estimation and PPG-based HR estimation.

2.1 DEEP-LEARNING BASED UNCERTAINTY ESTIMATION

The total uncertainty of predictive models can be divided into aleatoric and epistemic uncertainty [Ulmer and Cinà, 2021]. The aleatoric uncertainty captures the uncertainty inherent in the data and the epistemic uncertainty is an estimate of the uncertainty due to constraints on the model or the training process [Depeweg et al., 2017]. For a well-calibrated model, the uncertainty estimates and the model error are perfectly correlated [Band et al., 2022].

Several approaches have been proposed to estimate a model's uncertainty. These include fitting a parametric probability distribution over network activations and weights [Hernández-Lobato and Adams, 2015], and sampling using dropout [Gal and Ghahramani, 2016]. Most of the previous work only evaluated their predictive performance and the quality of their uncertainty estimates on curated datasets such as CIFAR-10 and FashionMNIST [Osawa et al., 2019, Ovadia et al., 2019].

2.2 PPG-BASED HEART RATE ESTIMATION

Increasing interest into PPG-based HR prediction has led to the release of large publicly available datasets [Schmidt et al., 2018, Reiss et al., 2019, Zhang et al., 2015]. Previous research on PPG-based HR estimation has mainly focused on classical signal-processing methods such as finite state machines (FSM) [Chung et al., 2019] or adaptive filters [Guo et al., 2022, Temko, 2017] that aim to extract the underlying pulse wave component from the PPG signal before estimating HR. Zhou and Selvaraj [2020] proposed to track heart rate values in the frequency domain using rule-based decision-making algorithms. However, these methods rely on manually-tuned hyperparameters that do not generalize across datasets where participants' activities differ.

To overcome this, deep learning-based approaches have been introduced for PPG-based HR estimation [Reiss et al., 2019]. These typically either operate on the time-[Biswas et al., 2019, Rocha et al., 2020, Ray et al., 2022, Shyam et al., 2019] or time-frequency [Reiss et al., 2019, Chung et al., 2020] representation of the PPG signal. Proposed architectures include convolutional, linear and recurrent layers, and have been optimized using Network Architecture

Search [Burrello et al., 2022, 2021, Song et al., 2021]. While most of these approaches estimate the most-likely heart rate in a regression setting, Chung et al. [2020] reformulates the PPG-based HR estimation as a classification over 222 HR bins. Ray et al. [2022] propose a Bayesian neural network combining Bayesian inference with Monte Carlo dropout to obtain estimates of the aleatoric and epistemice uncertainty. Chang et al. [2021] and Sarkar and Etemad [2020] use generative machine learning to improve signal quality before applying a classical method.

Our novel deep learning method estimates the HR through a discrete set of classes over the range of possible heart rate values to better represent multi-modal distributions that capture the uncertainty across the frequency spectrum (see Figure 1). In addition, our method considers its prediction within the temporal context of the HR trajectory to decrease the effect of motion artifacts and OOD samples.

3 METHOD

In this section, we describe our method BeaU-PPG for the reliable estimation of heart rate from PPG signals as input. We represent the evolution of the heart rate in the form of a probabilistic graphical model. Using probabilistic inference, our method predicts a representative distribution that captures the inherent predictive uncertainty over the range of possible heart rate values over time.

3.1 PROBLEM FORMULATION

Our method estimates the instantaneous heart rate $y_t \in \mathbb{R}$ at time t based on a window of input signals $\boldsymbol{X}_t = \left\{\boldsymbol{x}_{t-(i-1)T_x}\right\}_{i=1}^{L_x}$ where $L_x \in \mathbb{N}$ is the length and $T_x \in \mathbb{R}^+$ is the sampling interval of the signal window. A sample of the signal window $\boldsymbol{x}_t \in \mathbb{R}^{N_x}$ generally includes the values of the channels of the PPG sensor as well as the acceleration captured by an accelerometer that is integrated into most sensing hardware for the discrimination of motion artifacts obscuring the PPG signal.

We further consider y_t in the context of a discrete-time stochastic process of PPG-based heart rate estimates sampled at an interval T_y, where each y_t corresponds to an observed signal window $\boldsymbol{X}_t$ extracted from a recorded trace of PPG and acceleration signals. We describe this sequence in the form of a Forney factor graph (see Figure 2), which reveals the structure of a hidden Markov model (HMM) assuming that the hidden heart rate state Y_t only depends on its previous state Y_{t-1} and the current observation $\boldsymbol{X}_t$. The factor graph comprises the factor vertices f_T and f_N. f_T captures the transition probability between the heart rate state Y_t and its previous heart rate state Y_{t-1}. f_N models the statistical relation between an observed signal window $\boldsymbol{X}_t$ and the heart rate state Y_t at time t, and thus, resembles the emission probability in a standard hidden Markov model.

3.2 HEART RATE ESTIMATION NETWORK

We model the factorial $f_N(Y_t, \boldsymbol{X}_t)$ capturing the statistical relation between the heart rate state Y_t and the signal window $\boldsymbol{X}_t$ at time t through a neural network trained with supervision to estimate the conditional probability $p(Y_t \mid \boldsymbol{X}_t)$. We represent $p(Y_t \mid \boldsymbol{X}_t)$ as a discrete distribution over the set of heart rate states C with c classes linearly spaced in the range $[y_{\min}, y_{\max})$ to enable the prediction of arbitrarily multimodal distributions [Van Den Oord et al., 2016]. $y_{\min}$ and $y_{\max}$ correspond to physiologically sensible limits of 30 BPM and 210 BPM respectively [Tanaka et al., 2001].

3.3 HEART RATE TRANSITION FUNCTION

We model the heart rate transition factor f_T as a conditional probability distribution using the log-normal distribution fitted to relative changes y_t/y_{t-1} across one timestep,

$$p(y_t \mid y_{t-1}) = \frac{1}{\sigma_T \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log \frac{y_t}{y_{t-1}} - \mu_T}{\sigma_T}\right)^2\right). \tag{1}$$

We discretize this conditional probability distribution to the heart rate transition function $f_T(\hat{y}_t, \hat{y}_{t-1})$ by integrating $p(y_t \mid y_{t-1})$ according to the intervals of the heart rate states C from $\hat{y}_t^- / \hat{y}_{t-1}^+$ to $\hat{y}_t^+ / \hat{y}_{t-1}^-$, where the superscript $+$ indicates the upper bound and $-$ the lower bound of the interval represented by the respective heart rate class.

3.4 PROBABILISTIC HEART RATE INFERENCE

We infer the current heart rate state distribution using forward sum-product message passing (belief propagation),

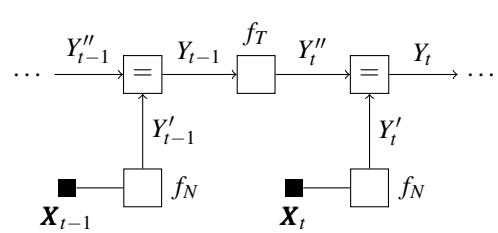

Figure 2: Forney factor graph representing the probabilistic graphical model for the evolution of the heart rate over time. The factor graph includes the factor vertices f_T and f_N. f_T captures the transition probabilities between the heart rate states Y_t across time. f_N represents the relationship between an observed PPG and acceleration signal window $\boldsymbol{X}_t$ and the heart rate state Y_t at time t.

where $\overrightarrow{\mu}$ represents the forward message in the positive time dimension,

$$\overrightarrow{\mu}_{Y_t'}(\hat{y}_t) = f_N(\hat{y}_t, \boldsymbol{X}_t), \tag{2}$$

$$\overrightarrow{\mu}_{Y_t''}(\hat{y}_t) = \sum_{\hat{y}_{t-1} \in C} f_T(\hat{y}_t, \hat{y}_{t-1}) \overrightarrow{\mu}_{Y_{t-1}}(\hat{y}_{t-1}), \tag{3}$$

$$\overrightarrow{\mu}_{Y_t}(\hat{y}_t) = \frac{1}{Z_t} \overrightarrow{\mu}_{Y_t'}(\hat{y}_t) \overrightarrow{\mu}_{Y_t''}(\hat{y}_t), \tag{4}$$

$$Z_t = \sum_{\hat{y}_i \in C} \overrightarrow{\mu}_{Y_t'}(\hat{y}_i) \overrightarrow{\mu}_{Y_t''}(\hat{y}_i), \tag{5}$$

$$p_{Y_t}(\hat{y}_t \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) = \overrightarrow{\mu}_{Y_t}(\hat{y}_t). \tag{6}$$

$p_{Y_t}(\hat{y}_t \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ is a probability mass function of the discrete heart rate state Y_t over the set of heart rate values C. Note that the forward pass for the computation of $p_{Y_t}(\hat{y}_t \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ can be computed in an online manner and efficiently implemented in standard deep-learning frameworks using matrix multiplications. The distribution over the heart rate classes $p_{Y_t}(\hat{y}_t \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ captures the inherent predictive uncertainty of our method. We can refine the granularity of the discrete heart rate class spacing through linear interpolation and normalization to approximate a more continuous distribution over the range of possible heart rate values.

It can be further reduced to a single heart rate value for time step t through the computation of the mean heart rate,

$$\tilde{y}_t = \mathbb{E}_{Y_t \sim p_{Y_t}(Y_t \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)}[Y_t] = \sum_{\hat{y} \in C} \hat{y}_t p_{Y_t}(\hat{y}_t \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t). \tag{7}$$

We compute an estimate of the predictive uncertainty using Shannon's entropy over the distribution

$$\mathscr{H}(Y_t) = -\sum_{\hat{y}_t \in C} p_{Y_t}(\hat{y}_t \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) \log p_{Y_t}(\hat{y}_t \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t). \tag{8}$$

Alternatively, we experiment with the standard deviation over $p_{Y_t}(\hat{y}_t \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ to measure how much the distribution is spread out over the heart rate classes.

In addition, we can compute an a posteriori estimate of the most likely heart rate sequence over the hidden state variables $(\hat{Y}_1, \ldots, \hat{Y}_t)$ using max-product message passing,

$$(\hat{Y}_1, \ldots, \hat{Y}_t) \mathrel{\widehat{=}} \underset{Y_1, \ldots, Y_t}{\arg\max}\, p(Y_1, \ldots, Y_t \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t), \tag{9}$$

where in the case of a unique maximizer

$$\hat{y}_k = \underset{Y_k}{\arg\max} \max_{Y_1, \ldots, Y_t \text{ except } Y_k} p(Y_1, \ldots, Y_t \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t). \tag{10}$$

Following the Viterbi algorithm [Forney, 1973], we keep track of the product maximizing previous state $\hat{y}_{t-1} \in C$ for each heart rate class $\hat{y}_t \in C$ in the forward recursion with the forward message $\overrightarrow{\mu}$ over the heart rate state variables $(Y_1, \ldots, Y_t)$,

$$\overrightarrow{\mu}_{Y_t'}(\hat{y}_t) = f_N(\hat{y}_t, \boldsymbol{X}_t), \tag{11}$$

$$\overrightarrow{\mu}_{Y_t''}(\hat{y}_t) = \max_{\hat{y}_{t-1} \in C} f_T(\hat{y}_t, \hat{y}_{t-1}) \overrightarrow{\mu}_{Y_{t-1}}(\hat{y}_{t-1}), \tag{12}$$

$$\overrightarrow{\mu}_{Y_t}(\hat{y}_t) = \overrightarrow{\mu}_{Y_t'}(\hat{y}_t) \overrightarrow{\mu}_{Y_t''}(\hat{y}_t). \tag{13}$$

We then recursively reconstruct and update the maximizing path over the state sequence in a backward pass starting from the maximum heart rate class in the last time step t,

$$\hat{y}_t = \underset{Y_t}{\arg\max}\, \overrightarrow{\mu}_{Y_t}(\hat{y}_t), \tag{14}$$

$$\hat{y}_{k-1} = \underset{\hat{y}_{k-1} \in C}{\arg\max}\, f_T(\hat{y}_k, \hat{y}_{k-1}) \overrightarrow{\mu}_{Y_{k-1}}(\hat{y}_{k-1}). \tag{15}$$

4 IMPLEMENTATION

We now describe the implementation of our method, including the representation of our input, the HR estimation network, and the HR transition function in detail.

4.1 INPUT REPRESENTATION

Our network operates on the inputs in their time as well as frequency representation. The Fourier spectrum of a PPG signal is expected to contain a dominant peak at the heart rate frequency, facilitating a straightforward heart rate estimation for artifact-free signals. While components from motion artifacts in the PPG signals with frequencies further away from the instantaneous heart rate can be typically eliminated through a comparison with the Fourier spectra of the acceleration channels, leakage and low-resolution effects [Stoica et al., 2005] in the fast Fourier transform (FFT) can significantly obscure this filtering for motion artifacts with frequencies that are very close to the heartbeat frequency [Zhang et al., 2015, Zhang, 2015]. The latter are frequently present for wrist-worn sensors during strong movements and repetitive arm motions. In this case, the patterns in the time-domain representation of the signal might provide additional discerning information, which intuitively motivates the inherent duality of our inputs.

For the time domain representation, we z-score normalize the PPG channels of the window $\boldsymbol{X}_t$ after band-pass filtering them with corner frequencies 0.1 and 18 Hz using a Butterworth bandpass filter. Then we average across channels and resample the result to 64 Hz, obtaining $\tilde{\boldsymbol{X}}_t \in \mathbb{R}^{L_x \times 1}$.

For the Fourier representation, we first subdivide $\boldsymbol{X}_t$ into another W_s overlapping 8-second windows with 2 seconds

shift, z-score normalize and downsample them before computing the FFT with 535 points. To obtain a spectrogram for each window of $\boldsymbol{X}_t$, we extract the magnitudes of the N_s frequency components within the PPG pulse wave frequency range between 0.5 Hz to 3.5 Hz.

We repeat this process for each channel of $\boldsymbol{X}_t$ separately and then average across available PPG- and acceleration channels to obtain $\hat{\boldsymbol{X}}_t \in \mathbb{R}^{W_s \times N_s \times 2}$, where the last dimension consists of the information from the PPG sensor and the accelerometer.

4.2 HEART RATE TRANSITION FUNCTION

We fit μ_T and σ_T (see Eq. (1)) over the distribution of relative heart rate changes observed in the training set and then perform the discretization depending on the number of heart rate classes c. The supplementary material includes an example plot for the heart rate transition distribution extracted from real-world data.

4.3 HEART RATE NETWORK ARCHITECTURE

The architecture of our network consists of two input branches. One branch operates on the spectrogram $\hat{\boldsymbol{X}}_t$, while the other receives $\tilde{\boldsymbol{X}}_t$ as input.

Two consecutive 2D convolutional layers that operate across the time and frequency axes simultaneously with a kernel of size 3x3 extract an embedding for each PPG-acceleration magnitude pair in $\hat{\boldsymbol{X}}_t$. A dot-product attention layer then computes the correlation between the linear embeddings computed across the frequency dimension to obtain an attention-refined aggregate, which considers the context across all frequency components along one time step of the spectrogram. A second dot-product attention layer extracts equivalent features from the spectrogram across the time dimension for each spectrogram frequency. We add these embeddings to a linear embedding directly computed on the output of the convolutional layers. This is followed by an average pooling layer that computes the mean of all aggregate vectors along the time dimension. A 1D Attention U-Net [Oktay et al., 2018] transforms this into a logit vector with c classes. At the bottleneck of the U-Net, we fuse the latent vector $\boldsymbol{h} \in \mathbb{R}^{L_h}$ of size $L_h \in \mathbb{N}$ with the output of the time-domain branch. The time-domain branch consists of a downsized version of CorNET [Biswas et al., 2019], a two-layer 1D convolutional neural network with max-pooling layers, and a 2-layer LSTM head. The branch computes a weighting vector $\boldsymbol{v} \in \mathbb{R}^{L_h}$ and a feature vector $\boldsymbol{s} \in \mathbb{R}^{L_h}$.

We add $\boldsymbol{s}$ and $\boldsymbol{v}$ in the form of a residual through an adapted form of additive attention, obtaining the input $\hat{\boldsymbol{h}} \in \mathbb{R}^{L_h}$ for the upsampling branch of the U-Net as

$$\hat{\boldsymbol{h}} = \boldsymbol{h} + \tanh\left(\boldsymbol{W}_v\left[\boldsymbol{h};\boldsymbol{v}\right] + \boldsymbol{b}_v\right) \odot \mathrm{ReLU}\left(\boldsymbol{W}_s\left[\boldsymbol{h};\boldsymbol{s}\right] + \boldsymbol{b}_s\right). \tag{16}$$

$\boldsymbol{W}_v, \boldsymbol{W}_s \in \mathbb{R}^{L_h \times 2L_h}$ and $\boldsymbol{b}_v, \boldsymbol{b}_s \in \mathbb{R}^{L_h}$ are trainable weight matrices and bias terms respectively. ReLU is the ReLU activation function, $[\cdot;\cdot]$ is a concatenation operator and $\odot$ is the element-wise Hadamard product.

After upsampling, the softmax activation function converts the logit vector to a probability distribution estimate across the c respective classes. We included a more detailed description of the network architecture in the supplementary material.

4.4 LOSS FUNCTION

We train the network using categorical cross-entropy loss. Since we reformulate the regression of the real-numbered output y_t as a classification problem, we do not use a binary target vector but represent y_t as a discrete probability distribution over the classes c, $p\left(Y_t = \hat{y} \,|\, \boldsymbol{X}_t\right)$ with $\hat{y}_t \in C$.

We obtain $p\left(Y_t = \hat{y}_t \,|\, \boldsymbol{X}_t\right)$ by discretizing the normal distribution $f(y) = \mathcal{N}\left(y_t, \sigma_y^2\right)$ with expectation $\mathbb{E}\left[f(y)\right] = y_t$ and variance $\mathrm{Var}\left[f(y)\right] = \sigma_y^2$. Here, y_t is the ground-truth heart rate and σ_y^2 is a hyperparameter, which has a regularizing effect on the output distribution. We compute the probability mass function $p\left(Y_t = \hat{y} \,|\, \boldsymbol{X}_t\right)$ by integrating $f(y)$ from the lower $\hat{y}^-$ and upper bound $\hat{y}^+$ of the interval for the heart rate class $\hat{y}$, and normalizing for probability mass lost in the tails. We use 64 classes, resulting in a mean quantization error $\mathbb{E}\left[\left|y_t - \mathbb{E}_{Y_t \sim p(Y_t \,|\, \boldsymbol{X}_t)}\left[Y_t\right]\right|\right]$ of 0.02 and a maximum quantization error of around 0.035 for the minimum and maximum heart rate class. We implement the heart rate transition function as a matrix $T \in \mathbb{R}^{c \times c}$ to ease integration in the message-passing algorithm.

4.5 TRAINING

We implement our model in TensorFlow. The models are trained on an *NVIDIA GeForce RTX 3090* GPU until convergence with an Early Stopping criterion obtained on a separate validation set. We use a batch size of 128 and the Adam [Kingma and Ba, 2014] optimizer with a learning rate of $2.5 \cdot 10^{-4}$. We further augment our training dataset by applying random time stretches with up to 25% on the input signals and extracted labels as well as by adding Gaussian noise n to the inputs, $n \sim \mathcal{N}\left(0, 0.25^2\right)$.

5 EXPERIMENTS

We evaluate our approach on eight publicly available datasets using three different cross-validation schemes to

test for robustness and generalization capability. Mean absolute errors (MAEs) and standard deviations are presented as the average and standard deviation across sessions.

5.1 CROSS-VALIDATION SCHEMES

Leave-one-Session-out (LoSo) cross-validation The subjects are split into $n-2$ training, one validation, and one test subject. The model is optimized on the training set until validation loss stops decreasing. The split is rotated until all sessions have been used as test set. Used on individual datasets.

Five-Fold cross-validation Each dataset is split into 20% test and 80% training sessions. One validation subject is drawn from the training set. Then, the folds from the different datasets are combined such every dataset contributes a training set.

Leave-one-Dataset-out cross-validation One dataset is taken as test data, the remainder serves as training data. One validation session is drawn from each training dataset.

Only predictions on test sets (unseen data) are reported.

5.2 DATASETS

Our evaluation considers eight datasets, including IEEE train and test [Zhang et al., 2015], PPG-DaLiA [Reiss et al., 2019], WESAD [Schmidt et al., 2018], the two datasets published by the BAMI-Labs [Chung et al., 2020, Lee et al., 2019], BIDMC PPG [Pimentel et al., 2017] as well as CLAS [Markova et al., 2019]. They jointly provide a diverse data collection comprising different subjects, activities, and recording hardware. DaLiA is not only by far the largest but also the only data collection gathered in natural environments.

5.3 HEART RATE ESTIMATION ACCURACY

On all eight datasets and under all three evaluation strategies the method achieves or outperforms state-of-the-art results among cross-validated methods. A full comparison is displayed in Table 1. Under leave-one-session-out as well as five-fold cross-validation the model achieves low mean absolute errors despite the small dataset sizes in the former and the heterogeneous data in the latter. On PPG-DaLiA the method outperforms the reported MAE of the runner-up method Q-PPG by 0.8 (18%) BPM under LoSo evaluation. The small IEEE datasets however remain challenging: Although achieving the lowest error among LoSo-evaluated deep learning methods, hand-tuned algorithms such as WFPV still perform slightly better.

During out-of-distribution evaluation (Leave-one-Dataset-out) we observe a slight decrease in accuracy on the DaLiA and BAMI datasets (+0.5, +1.7, +1.6 BPM MAE) while the performance on the other datasets remains unchanged. Notably, our method still outperforms its in-distribution-trained competition on PPG-DaLiA.

In summary, the model produces highly competitive results in terms of MAE under various data distributions and cross-validation methods.

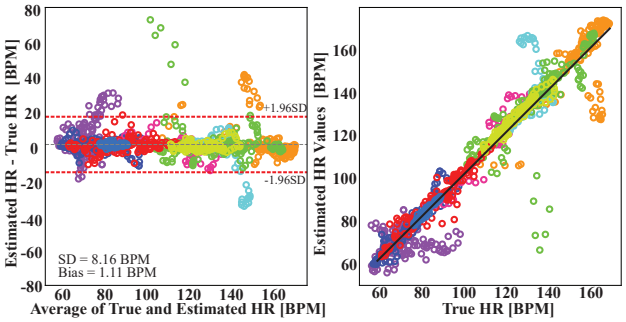

Figure 3: Bland-Altman plot of HR estimates during LoSo cross-validation on IEEE test. The colors represent subjects.

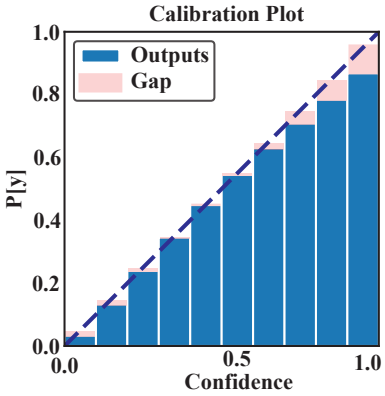

Figure 4: Model calibration visualized using the empirical probability of ground truth HR being inside the confidence region. We observe that the model is slightly overconfident for high-confidence levels, but displays good overall calibration.

Errors by activity We added a plot of errors across activities performed in PPG-DaLiA in the supplementary material. We observe the highest errors while "walking stairs" (6.6% mean absolute percentage error) and the lowest errors during sedentary activities.

Errors by subject As visible in Table 1, the error standard deviation across subjects is low in comparison with existing methods. Figure 3 further plots the error on IEEE test by subject.

Table 1: HR estimation accuracy compared to prior work. Values indicate MAE and standard deviation in BPM.

Approach	#Param.	IEEE train	IEEE test	PPG-DaLiA	WESAD	BAMI-1	BAMI-2	BIDMC	CLAS
CLASSICAL METHODS (not cross-validated)									
[Temko, 2017]	-	**1.02**	**1.97**	10.7 ± 3.8	8.5 ± 4.3	11.28	6.09	-	-
[Huang and Selvaraj, 2020]	-	2.5 ± 1.2	5.9 ± 3.5	4.6 ± 1.4	4.2 ± 1.4	-	-	-	-
[Zhou and Selvaraj, 2020]	-	2.2	4.5	5.0 ± 2.8	6.4 ± 1.8	-	-	-	-
DEEP LEARNING METHODS									
[Sarkar and Etemad, 2020]	-	-	-	8.03	8.06	-	-	**0.7**	-
[Chang et al., 2021]	3.3M	4.76 *(1.98[3])*	-	-	-	-	-	-	-
[Chung et al., 2020]	3.3M	*0.67[1]*	*0.86[1]*	-	-	*1.39[1]*	1.46	-	-
[Rocha et al., 2020]	257k	4.67 ± 3.7	6.61 ± 5.4	-	-	-	-	-	-
[Shyam et al., 2019]	-	3.36 ± 4.1	12.48 ± 14.5	-	-	-	-	-	-
[Reiss et al., 2019]	8.5M	4.0 ± 5.4	16.51 ± 16.1	7.65 ± 4.2	7.47 ± 3.3	-	-	-	-
[Ray et al., 2022]	730k	*2.76 ± 3.0[2]*	*5.05 ± 5.5[2]*	*2.12 ± 3.1[2]*	-	-	*2.38 ± 2.6[2]*	-	-
[Song et al., 2021]	800k	*0.82[1]*	*1.03[1]*	6.02 ± 10.6	-	-	-	-	-
[Wilkosz and Szczęsna, 2021]	680k	-	-	6.28 ± 3.5	-	-	-	-	-
[Burrello et al., 2022]	900k	3.27 ± 2.0		4.88 (*3.84[4]*)	-	-	-	-	-
[Burrello et al., 2021]	269k	-	-	4.36 (*3.61[4]*)	-	-	-	-	-
OURS									
Leave-one-Session-out Cross-Validation									
BeaU-PPG (ours)	**138k**	1.75 ± 0.8	3.78 ± 2.2	3.57 ± 1.4	4.28 ± 2.0	**2.0 ± 1.0**	1.48 ± 0.9	-	-
BeaU-PPG / viterbi	**138k**	**1.47 ± 0.6**	3.06 ± 1.9	**3.18 ± 1.3**	4.02 ± 1.9	**2.12 ± 1.1**	**1.46 ± 0.3**	-	-
Five-Fold Cross-Validation									
BeaU-PPG (ours)	**138k**	1.82 ± 0.6	4.2 ± 2.3	3.74 ± 1.8	3.81 ± 1.9	2.17 ± 1.1	**1.32 ± 0.3**	**1.11 ± 1.7**	**1.47 ± 1.0**
BeaU-PPG / viterbi	**138k**	1.49 ± 0.4	**2.61 ± 1.4**	**3.34 ± 1.7**	**3.63 ± 1.9**	2.2 ± 1.0	1.53 ± 0.2	1.4 ± 1.5	1.48 ± 0.8
Leave-one-Dataset-out Cross-Validation									
BeaU-PPG (ours)	**138k**	1.88 ± 0.7	4.39 ± 3.1	4.26 ± 1.6	3.88 ± 1.9	3.83 ± 3.8	2.9 ± 3.5	1.34 ± 1.6	**1.41 ± 0.8**
BeaU-PPG / viterbi	**138k**	1.48 ± 0.5	3.14 ± 2.1	3.83 ± 1.6	**3.7 ± 1.8**	3.67 ± 3.7	2.96 ± 3.5	1.59 ± 1.5	1.67 ± 0.8

[1] Training error
[2] Mixed training and testing error
[3] Applies post-processing technique without cross-validation
[4] Subject-specific fine-tuning

5.4 EVALUATING PREDICTIVE UNCERTAINTY

Calibration To evaluate the calibration of our predicted heart rate distributions, we assess whether the empirical probability of the ground truth heart rate falling within a heart rate region is matching the regions aggregated probability mass. In our experiment, we successively consider the intervals representing the class bins from the heart rate class with the highest assigned probability value to the heart rate class with the lowest probability value. Results are plotted in a calibration plot in Figure 4, which shows that our method is generally well-calibrated, but slightly overconfident for higher confidence levels.

Quantification of uncertainty The estimated probability distribution captures the predictive uncertainty of our model. This can be used to discard samples with high uncertainty. To that end, we quantify the predictive uncertainty using entropy and successively discard the most uncertain predictions while computing the MAE on the remaining samples. By removing the top-1% most uncertain predictions (over all datasets and experiments) we reduce MAE by 4%; by removing the top-5% we achieve a boost of 15%. Figure 5 visualizes the effect on different datasets. Standard deviation and entropy achieve almost equal (within ∼1%) results in these experiments, indicating that they both represent meaningful estimates of the predictive uncertainty.

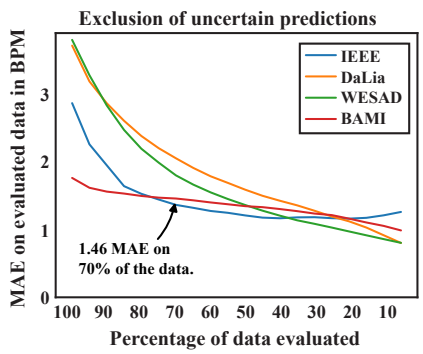

Figure 5: Evolution of MAE for five-fold cross-validation while successively excluding the most uncertain predictions, using entropy as metric.

The calibration was evaluated after linearly upsampling the probability distributions to 1000 intervals.

Table 2: Results of incremental ablation experiments using five-fold cross-validation.

	IEEE train	IEEE test	PPG-DaLiA	WESAD	BAMI-I	BAMI-II
U-Net (8-sec)	2.65 ± 1.9	11.16 ± 8.1	7.93 ± 4.2	6.89 ± 2.7	3.21 ± 1.5	1.84 ± 0.7
CorNET Light (8-sec)	9.85 ± 7.7	19.32 ± 9.9	5.31 ± 2.7	4.24 ± 2.0	9.86 ± 6.7	3.16 ± 1.8
Fused Model (8-sec)	1.69 ± 1.2	13.27 ± 10.0	4.86 ± 2.3	4.6 ± 2.2	3.32 ± 1.6	1.97 ± 0.9
BeaU-PPG	1.62 ± 1.1	8.56 ± 8.9	4.02 ± 1.9	4.17 ± 2.0	2.43 ± 1.4	1.5 ± 0.8
+ train-time aug.	$\mathbf{1.62 \pm 0.8}$	7.61 ± 5.2	4.02 ± 1.9	3.98 ± 2.0	2.42 ± 1.2	1.4 ± 0.3
+ belief propagation	1.67 ± 0.6	$\mathbf{4.47 \pm 2.9}$	$\mathbf{3.59 \pm 1.7}$	3.94 ± 2.1	2.08 ± 1.0	1.34 ± 0.3
+ 4x higher resolution	1.66 ± 0.8	4.53 ± 4.0	3.65 ± 1.4	$\mathbf{3.89 \pm 1.9}$	$\mathbf{1.96 \pm 1.0}$	$\mathbf{1.24 \pm 0.4}$

5.5 ABLATION STUDY

We perform an ablation study on the different parts of our method.

Network Architecture Performance degrades when successively disassembling the proposed architecture. Table 2 shows that the omission of train-time augmentations leads to worse results. The additional removal of the attention layer that combines the seven frequency-timesteps increases errors by more than 20% on PPG-DaLiA. Notably, this reduces the overall considered signal window to 8s. If we further decompose the resulting network into the downsized CorNET and the frequency-domain U-Net, MAEs increase by another 9% and 63% respectively.

Effect of message passing Table 2 also shows how temporal modeling in the form of message passing improves performance. The effect is particularly strong on datasets with larger errors such as IEEE test and PPG-DaLiA.

Output HR resolution We validate the chosen tradeoff between complexity and accuracy in representing the output space with 64 intervals (2.81 BPM width) by running experiments against a higher resolution of 256 intervals (0.70 BPM width). The input spectra and the U-Net are scaled by the same factor, resulting in a non-trivial increase in the computational cost. Results indicate little improvement in accuracy.

Regularization with σ_y To reduce the quantization error in the label we use a Gaussian distribution with a standard deviation of $\sigma_y = 1.5$. Ablations (provided in the supplementary material) indicate that the choice of σ_y has very little impact on performance.

5.6 OFFLINE DECODING

If online estimates are not required, Viterbi decoding can further decrease error. On PPG-DaLiA, this results in an additional reduction of 11% MAE compared to online decoding. Corresponding results are displayed in Table 1.

6 LIMITATIONS

The HR discretization introduces quantization errors, albeit small, and imposes additional hyperparameters on the model such as the number of bins, label variance σ_y^2, and the fixed range. Experiments, however, demonstrate low sensitivity towards these parameters. Despite not being specific to HR estimation, the framework is limited to low-variate regression problems due to the exponential scale-up of the number of output classes.

7 CONCLUSION

We introduced a novel deep-learning method for PPG-based heart rate estimation. Our method joins the time and the time-frequency domains with a dual, U-Net-based neural network, producing discrete probability distributions over HR intervals. These distributions are contextualized with a transition prior modeling belief propagation over time, allowing online inference of expected heart rate.

Through extensive experiments on eight publicly available datasets, we have demonstrated state-the-art performance and robustness of our method, surpassing previous method on PPG-DaLiA by 18% in accuracy. Even in our out-of-distribution evaluation, our model outperforms existing methods.

The predictive uncertainty of our method is well-calibrated, in particular in lower certainty levels. Measuring uncertainty by entropy enables incremental MAE reduction by rejecting uncertain predictions, with 5% rejections resulting in a 15% boost on average. Alternatively, performance can be improved by decoding the probabilistic model offline.

Our work provides a new baseline for PPG-based HR estimation during daily life activities. We believe that our probabilistic, context-aware HR inference contributes a relevant approach to the domain of cardiac monitoring. We also believe that our dual representation approach will be suitable for creating models to process other physiological signals more reliably.

References

J. Allen, D. Zheng, P. Kyriacou, and M. Elgendi. Photoplethysmography (ppg):: State-of-the-art methods and applications. *Physiological Measurement*, 42(10), Nov. 2021. ISSN 0967-3334. doi: 10.1088/1361-6579/ac2d82.

ANSI/AAMI. Cardiac monitors, heart rate meters, and alarms. *Arlington*, 2002.

C. T. Association. Physical activity monitoring for heart rate, ansi/cta 2065, 2018. Accessed: 2023-02-17.

N. Band, T. G. J. Rudner, Q. Feng, A. Filos, Z. Nado, M. W. Dusenberry, G. Jerfel, D. Tran, and Y. Gal. Benchmarking bayesian deep learning on diabetic retinopathy detection tasks. *ArXiv*, abs/2211.12717, 2022.

D. Biswas, L. Everson, M. Liu, M. Panwar, B.-E. Verhoef, S. Patki, C. H. Kim, A. Acharyya, C. Van Hoof, M. Konijnenburg, and N. Van Helleputte. Cornet: Deep learning framework for ppg-based heart rate estimation and biometric identification in ambulant environment. *IEEE Transactions on Biomedical Circuits and Systems*, 13(2): 282–291, 2019. doi: 10.1109/TBCAS.2019.2892297.

A. Burrello, D. J. Pagliari, M. Risso, S. Benatti, E. Macii, L. Benini, and M. Poncino. Q-ppg: Energy-efficient ppg-based heart rate monitoring on wearable devices. *IEEE Transactions on Biomedical Circuits and Systems*, 15(6): 1196–1209, 2021. doi: 10.1109/TBCAS.2021.3122017.

A. Burrello, D. J. Pagliari, P. M. Rapa, M. Semilia, M. Risso, T. Polonelli, M. Poncino, L. Benini, and S. Benatti. Embedding temporal convolutional networks for energy-efficient ppg-based heart rate monitoring. *ACM Trans. Comput. Healthcare*, 3(2), mar 2022. ISSN 2691-1957. doi: 10.1145/3487910.

X. Chang, G. Li, G. Xing, K. Zhu, and L. Tu. Deepheart: A deep learning approach for accurate heart rate estimation from ppg signals. *ACM Trans. Sen. Netw.*, 17(2), jan 2021. ISSN 1550-4859. doi: 10.1145/3441626.

P. H. Charlton, K. Kotzen, E. Mejía-Mejía, P. J. Aston, K. Budidha, J. Mant, C. Pettit, J. A. Behar, and P. A. Kyriacou. Detecting beats in the photoplethysmogram: benchmarking open-source algorithms. *Physiological Measurement*, 43(8):085007, aug 2022. doi: 10.1088/1361-6579/ac826d.

H. Chung, H. Lee, and J. Lee. Finite state machine framework for instantaneous heart rate validation using wearable photoplethysmography during intensive exercise. *IEEE Journal of Biomedical and Health Informatics*, 23 (4):1595–1606, 2019. doi: 10.1109/JBHI.2018.2871177.

H. Chung, H. Ko, H. Lee, and J. Lee. Deep learning for heart rate estimation from reflectance photoplethysmography with acceleration power spectrum and acceleration intensity. *IEEE Access*, 8:63390–63402, 2020. doi: 10.1109/ACCESS.2020.2981956.

S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, 2017.

G. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973. doi: 10.1109/PROC.1973.9030.

Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1050–1059. JMLR.org, 2016.

J. Guo, X. Chen, J. Zhao, X. Zhang, and X. Chen. An effective photoplethysmography heart rate estimation framework integrating two-level denoising method and heart rate tracking algorithm guided by finite state machine. *IEEE Journal of Biomedical and Health Informatics*, 26 (8):3731–3742, 2022. doi: 10.1109/JBHI.2022.3165071.

J. M. Hernández-Lobato and R. P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1861–1869. JMLR.org, 2015.

N. Huang and N. Selvaraj. Robust ppg-based ambulatory heart rate tracking algorithm. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5929–5934, 2020.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. URL `https://arxiv.org/abs/1412.6980`.

H. Lee, H. Chung, and J. Lee. Motion artifact cancellation in wearable photoplethysmography using gyroscope. *IEEE Sensors Journal*, 19(3):1166–1175, 2019. doi: 10.1109/JSEN.2018.2879970.

V. Markova, T. Ganchev, and K. Kalinkov. Clas: A database for cognitive load, affect and stress recognition. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*, pages 1–4, 2019. doi: 10.1109/BIA48344.2019.8967457.

O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention u-net: Learning where to look for the pancreas, 2018. URL `https://arxiv.org/abs/1804.03999`.

K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota, and M. E. Khan. *Practical Deep Learning with Bayesian Principles*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. *Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift*. Curran Associates Inc., Red Hook, NY, USA, 2019.

M. A. F. Pimentel, A. E. W. Johnson, P. H. Charlton, D. Birrenkott, P. J. Watkinson, L. Tarassenko, and D. A. Clifton. Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Biomedical Engineering*, 64(8):1914–1923, 2017. doi: 10.1109/TBME.2016.2613124.

D. Ray, T. Collins, and P. Ponnapalli. DeepPulse: An Uncertainty-aware Deep Neural Network for Heart Rate Estimations from Wrist-worn Photoplethysmography. 3 2022. doi: 10.36227/techrxiv.19368140.v2.

A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14), 2019. ISSN 1424-8220. URL `https://www.mdpi.com/1424-8220/19/14/3079`.

L. G. Rocha, D. Biswas, B.-E. Verhoef, S. Bampi, C. Van Hoof, M. Konijnenburg, M. Verhelst, and N. Van Helleputte. Binary cornet: Accelerator for hr estimation from wrist-ppg. *IEEE Transactions on Biomedical Circuits and Systems*, 14(4):715–726, 2020. doi: 10.1109/TBCAS.2020.3001675.

P. Sarkar and A. Etemad. Cardiogan: Attentive generative adversarial network with dual discriminators for synthesis of ecg from ppg, 2020. URL `https://arxiv.org/abs/2010.00104`.

P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 400–408, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356923. URL `https://doi.org/10.1145/3242969.3242985`.

A. Shyam, V. Ravichandran, S. Preejith, J. Joseph, and M. Sivaprakasam. Ppgnet: Deep network for device independent heart rate estimation from photoplethysmogram. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1899–1902, 2019. doi: 10.1109/EMBC.2019.8856989.

S. B. Song, J. W. Nam, and J. H. Kim. Nas-ppg: Ppg-based heart rate estimation using neural architecture search. *IEEE Sensors Journal*, 21(13):14941–14949, 2021. doi: 10.1109/JSEN.2021.3073047.

P. Stoica, R. L. Moses, et al. *Spectral analysis of signals*, volume 452. Pearson Prentice Hall Upper Saddle River, NJ, 2005.

H. Tanaka, K. D. Monahan, and D. R. Seals. Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology*, 37(1):153–156, 2001. doi: 10.1016/S0735-1097(00)01054-8.

A. Temko. Accurate heart rate monitoring during physical exercises using ppg. *IEEE Transactions on Biomedical Engineering*, 64(9):2016–2024, 2017. doi: 10.1109/TBME.2017.2676243.

D. Ulmer and G. Cinà. Know your limits: Uncertainty estimation with relu classifiers fails at reliable ood detection. In C. de Campos and M. H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1766–1776. PMLR, 27–30 Jul 2021. URL `https://proceedings.mlr.press/v161/ulmer21a.html`.

A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016.

M. Wilkosz and A. Szczęsna. Multi-headed conv-lstm network for heart rate estimation during daily living activities. *Sensors*, 21(15), 2021. ISSN 1424-8220. doi: 10.3390/s21155212.

Z. Zhang. Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction. *IEEE Transactions on Biomedical Engineering*, 62:1902 – 1910, 08 2015. doi: 10.1109/TBME.2015.2406332.

Z. Zhang, Z. Pi, and B. Liu. TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Transactions on Biomedical Engineering*, 62(2):522–531, feb 2015. doi: 10.1109/tbme.2014.2359372.

M. Zhou and N. Selvaraj. Heart rate monitoring using sparse spectral curve tracing. volume 2020, pages 5347–5352, 07 2020. doi: 10.1109/EMBC44109.2020.9175349.