Abstract

Photoplethysmogram (PPG) signal-based blood pressure (BP) estimation is a promising candidate for modern BP measurements, as PPG signals can be easily obtained from wearable devices in a non-invasive manner, allowing quick BP measurement. However, the performance of existing machine learning-based BP measuring methods still fall behind some BP measurement guidelines and most of them provide only point estimates of systolic blood pressure (SBP) and diastolic blood pressure (DBP). In this paper, we present a cutting-edge method which is capable of continuously monitoring BP from the PPG signal and satisfies healthcare criteria such as the Association for the Advancement of Medical Instrumentation (AAMI) and the British Hypertension Society (BHS) standards. Furthermore, the proposed method provides the reliability of the predicted BP by estimating its uncertainty to help diagnose medical condition based on the model prediction. Experiments on the MIMIC II database verify the state-of-the-art performance of the proposed method under several metrics and its ability to accurately represent uncertainty in prediction.

1. Introduction

According to the World Health Organization report (World Health Organization, 2020), cardiovascular disease (CVD) is still a leading cause of death, threatening a large number of lives. Hypertension is one of the major risk factors for CVD but most people with hypertension are unaware of the risk and does not acknowledge the necessity in controlling their blood pressure (BP). As hypertension silently damages the heart and arteries, continuous monitoring of BP is expected to play a significant role in preventing CVDs.

With traditional methods, however, it is difficult to monitor BP regularly in everyday life. For instance, cuff-based measurement is the most common method of measuring BP, which slowly deflates an inflated cuff to estimate the systolic blood pressure (SBP) and diastolic blood pressure (DBP) (Van Montfrans, 2001). However, this process is rather uncomfortable and takes a minute or two to obtain the results. In addition, the stress or anxiety of the subjects often leads to inaccurate measurements, which is known as the white coat syndrome (He et al., 2013). Another approach is a catheter-based measurement which inserts a catheter into an artery to observe the level of BP in real-time (Investigators, 2011). However, the catheter-based measurement is not adequate for regular monitoring as it is an invasive operation which requires an expert and carries the risk of infection. Thus, conventional measurement methods are typically performed as a one-off in limited situations, and more convenient BP monitoring methods are necessarily required.

Recently, photoplethysmogram (PPG) signal-based approaches have received a lot of attention for estimating BP (Hsu et al., 2020; Slapničar et al., 2019; Liang et al., 2018; Kachuee et al., 2016). Essentially, fluctuation in the PPG waveform is associated with blood circulation as the variation of blood volume affects the amount of light absorbed by the tissue (Elgendi, 2012). Also, the acquisition of PPG is relatively simple and inexpensive as follows: (1) A light-emitting diode (LED) illuminates the skin with infrared light and (2) a photodetector records the intensity of the non-absorbed light reflected from the tissue (Castaneda et al., 2018). Attributed to such convenience, the PPG has been widely used for clinical monitoring of physiologic parameters such as heart rate, oxygen saturation, and the level of hemoglobin concentration in blood (Kavsaoglu et al., 2015; Allen, 2007).

To take advantage of PPG’s simple acquisition procedure and its association with BP, we propose an elaborately designed framework to predict a continuous waveform of BP using PPG signal only. First, we adopt a variant architecture of Demucs (Defossez et al., 2019) that is suited for modeling time series data. The proposed framework employs the U-net structure to leverage PPG and BP signals’ common periodicity, and offers far better performance than the conventional methods. Secondly, we note that a simple regression loss such as mean absolute error (MAE) does not guarantee the optimal performance due to the discrep-
ancy between the training objective function and test criteria. To alleviate the mismatch, we propose an auxiliary loss function to match peak values between true and estimated BP values. Since the proposed loss function imposes more penalty on the incorrect peak prediction, the regression model prioritizes estimating SBP and DBP more accurately than other BP values. Finally, we employ deep evidential regression (DER) (Amini et al., 2019) to provide uncertainty in model prediction. Knowing the reliability of the model prediction can be very important when deploying the model in the real world as it would help diagnose a patient’s disease based on model prediction or calibrate the model estimation. Direct application of DER to our task, however, is found to cause an overfitting problem, which results in degraded performance of the BP estimation. To deal with the issue, we propose two different training techniques: i) weight initialization with deterministic regression and ii) temperature scaling. Both techniques allow the model to measure BP quite accurately and to represent the uncertainty well. In the experiments, we demonstrate that the proposed framework shows cutting-edge performance on a variety of evaluation metrics and represents uncertainty in prediction desirably.

In summary, our main contributions are as follows:

- We present a state-of-the-art model suitable for monitoring a continuous waveform of BP using only raw PPG signals.
- We propose an auxiliary objective function called peak-to-peak matching loss to obtain better estimates of SBP and DBP.
- We propose two different training strategies to overcome the overfitting problem of DER and demonstrate that the model represents uncertainty appropriately.
- To the best of our knowledge, this is the first work to take into account the reliability of model prediction in BP measurements beyond naive regression.

2. Related Work

The basic working principle behind PPG acquisition is closely associated with the changes of blood volume. As a result, PPG signals have been widely used in calculating physiological parameters in the body. Since the blood volume variations are related to the blood flow which exerts the pressure on the vessel, PPG signals are commonly considered as a good evidence for estimating the BP (Ibtehaz & Rahman, 2020). However, the clear relation between PPG and BP is not fully understood yet.

To exploit the valuable information underlying in PPG signals for BP measurement in a data-driven manner, many recent studies have employed various machine learning algorithms (Hsu et al., 2020; Ibtehaz & Rahman, 2020; Slapničar et al., 2019; Kachuee et al., 2016). For instance, Kachuee et al. (2016) applied classical algorithms such as linear regression, decision tree, support vector machine (SVM), adaptive boosting (AdaBoost), and random forest to the handcrafted features extracted from PPG and ECG signals. The authors reported that the AdaBoost model performed best among various approaches based on the mean absolute error (MAE) criterion. Although they proposed non-invasive BP estimation methods, their models achieved low grades in British Hypertension Society (BHS) and the Association for the Advancement of Medical Instrumentation (AAMI) standards due to the limited performance of the classical machine learning algorithms.

On the other hand, Hsu et al. (2020) proposed to train a feed-forward network on the manually selected features obtained from a single cardiac cycle segmentation to output SBP and DBP values. However, their method relies on a heuristic search for feature selection and did not adopt an architecture suitable to fully utilize sequential information of PPG signal. Slapničar et al. (2019) is another line of work which directly estimates the SBP and DBP using deep learning. They employed ResNet (He et al., 2016) and GRU (Cho et al., 2014) to leverage both temporal and frequency information in PPG signal. Still, their measurements were somewhat inaccurate and cannot estimate a continuous waveform of BP. Ibtehaz & Rahman (2020) is the most similar work to our framework in that they introduced a model that predicts sequential values of BP. They employed a one-dimensional U-net model (Ronneberger et al., 2015) and used the raw PPG signal as input to perform BP regression. However, they failed to satisfy the BHS and AAMI standards in terms of SBP. In addition, none of these approaches can provide the reliability of BP measurement, which can be critical information for making a medical decision based on model prediction.

3. Proposed Method

This section introduces an elaborately designed framework appropriate for measuring a continuous waveform of BP using only the PPG signal as input. In our framework, SBP, DBP, and mean arterial pressure (MAP) values are calculated by finding the maximum, minimum, and average values of the predicted BP waveform, respectively. For better estimation of SBP and DBP, we introduce a novel peak-to-peak matching loss with a simple peak detection algorithm. Furthermore, beyond naive regression, the proposed model is optimized to faithfully represent the prediction reliability as well.

3.1. Continuous Monitoring

Most of current works still focus on directly estimating the values of SBP, DBP, and MAP given continuous PPG sig-
nals as input via traditional machine learning methods or simple feed-forward networks (Hsu et al., 2020; Mousavi et al., 2019; Kachuee et al., 2016). However, these models cannot provide invaluable information for the diagnosis and treatment of CVDs that underlies in the waveform of BP itself (Seo et al., 2015). Here, we consider a more sophisticated architecture suitable for modeling time series input and output data. Fig 1 represents the overall structure of the proposed model. The architecture is a one-dimensional adaptation of U-net (Ronneberger et al., 2015), which is similar to Demucs (Défossez et al., 2019). Though the input and the output domains are different, we adopt skip connections to facilitate the model to leverage the cardiac periodicity of PPG and BP. Each convolution layer except the last layer in the decoder is followed by gated linear unit (GLU) (Dauphin et al., 2017) or rectified linear unit (ReLU) for activation function. To stabilize the training process and achieve better test performance, batch normalization (Ioffe & Szegedy, 2015) and weight normalization (Salimans & Kingma, 2016) techniques are additionally used. Finally, a two-layer bi-directional LSTM is employed between the encoder and the decoder to capture long-term dependencies in PPG signals. For use as input to the decoder, the channels size of the bidirectional LSTM’s output is halved by a fully connected layer. Unlike PPG2ABP (Ibtehaz & Rahman, 2020) which simply outputs a sequence of point estimates ofBP, the proposed model yields a 4-dimensional temporal sequence for the parameters of the Normal Inverse-Gamma (NIG) distribution. We compute the likelihood of the ground-truth BP using the NIG parameters and train the model according to the maximum likelihood. More details about evidential regression will be explained in Section 3.3.

3.2. Peak-to-Peak Matching Loss

Regression models are usually optimized to minimize mean squared error (MSE), mean absolute error (MAE), or negative log-likelihood (NLL) in general. For instance, PPG2ABP, which aims to monitor BP waveforms, uses the MAE and MSE loss for training the BP regression model. However, these loss functions are not optimal for the BP measurement task since some medical diagnoses at test time are conducted based on other statistics (e.g., SBP and DBP) calculated from the estimated BP waveform. The discrepancy between the training loss function and the test criteria hinders the full potential of the model from being exploited.

To mitigate the mismatch, we propose a peak-to-peak matching loss as an auxiliary objective function. Let \( y \) be a sequence of true BP values and \( \hat{y} \) be the corresponding estimate. We first divide \( y \) and \( \hat{y} \) into \( N\) segments: \( y_1, y_2, \ldots, y_N \) and \( \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N \). Then, the peak-to-peak matching loss \( \mathcal{L}_{PTP}^u \) derived from the peak points of \( y \) is computed as follows:

\[
\hat{i}_k^u = \arg\min_i y_k(i), \quad i_k^M = \arg\max_i y_k(i),
\]

\[
\mathcal{L}_{PTP}^u = \frac{1}{N} \sum_{k=1}^{N} |y_k(\hat{i}_k^u) - \hat{y}_k(\hat{i}_k^u)| + |y_k(i_k^M) - \hat{y}_k(i_k^M)|
\]

where \( y_k(i) \) and \( \hat{y}_k(i) \) are the \( i \)-th elements in the \( k \)-th frame. Similarly, the second peak-to-peak matching loss \( \mathcal{L}_{PTP}^b \) can also be obtained by replacing \( y_k(i) \) with \( \hat{y}_k(i) \) in Eq. (1) and computing Eq. (2). The total peak-to-peak matching loss is given by \( \mathcal{L}_{PTP} = \mathcal{L}_{PTP}^u + \mathcal{L}_{PTP}^b \) and we scale it with a coefficient \( \lambda_{PTP} \). Ideally, the peak-to-peak matching loss should be applied to the maximum and minimum values in every cycle. However, peak detection in an exact single cycle is somewhat tricky. Instead, we set a certain time interval and select peak values in each interval. Fig. 2 shows an example of the detected peaks of the ground-truth BP using our method. Taking into account the average heart rate, the frames are divided every 0.8 seconds. Although the method skips some peak values or incorrectly select non-peak values, most peak values are well detected. By minimizing the peak-to-peak matching loss \( \mathcal{L}_{PTP} \), the model is trained to estimate the peak values more accurately, and much reliable results can be obtain in determining hypertension based on the predicted waveforms.
3.3. Evidential Regression

In order to use the predicted BP values for medical diagnosis purposes, it is necessary not only to predict accurate values, but also to consider the reliability of the predicted value. However, most of existing works focuses solely on estimating SBP and DBP values, and overlooks the importance of the latter. In this paper, we employ deep evidential regression (DER) (Amini et al., 2019) to provide the reliability of the prediction. Furthermore, we propose two training techniques to solve an overfitting problem that arises when DER is applied to the BP measurement task.

In the DER framework, the target distribution is parameterized by a hierarchical structure with unknown mean and variance \((\mu, \sigma^2)\). The prior distribution of \((\mu, \sigma^2)\) is set to the Normal Inverse-Gamma (NIG) distribution with known parameters \((\gamma, v, \alpha, \beta)\)\(^1\). More specifically, the distribution of the \(i\)-th value of BP \(y(i)\) is assumed as follows:

\[
\sigma^2 \sim \Gamma^{-1}(\alpha, \beta), \quad \mu \sim \mathcal{N}(\gamma, \sigma^2 v^{-1}),
\]

\[
y(i) \sim \mathcal{N}(\mu, \sigma^2)
\]

where \(\gamma \in \mathbb{R}, v > 0, \alpha > 1, \beta > 0\) and \(\Gamma^{-1}(\cdot)\) is the Inverse-Gamma distribution. Eqs. (3) and (4) indicate that a single higher-order distribution yields various lower-order data distributions which in turn generate \(y(i)\). This setting allows us to define two types of uncertainty as follows:

\[
\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}, \quad \text{Var}[\mu] = \frac{\beta}{v(\alpha - 1)},
\]

where \(\mathbb{E}[\sigma^2]\) captures the innate stochasticity of data and \(\text{Var}[\mu]\) represents the uncertainty of the model arising from a lack of training data for particular data patterns.

To compute the likelihood of \(y(i)\) given the NIG parameters \(m = (\gamma, v, \alpha, \beta)\), we should marginalize over all the possible pairs of \((\mu, \sigma^2)\):

\[
p(y(i)|m) = \int_{0}^{\infty} \int_{0}^{\infty} p(y(i)|\mu, \sigma^2) p(\mu, \sigma^2|m) d\mu d\sigma^2.
\]

Since the NIG distribution is the conjugate prior of the normal distribution, we can derive the negative log-likelihood \(L_{\text{NLL}}\) analytically in a closed form as follows:

\[
L_{\text{NLL}} = -\log p(y(i)|m) = \log \left( \frac{\pi^\frac{1}{2} \Gamma(\alpha)((y(i) - \gamma)^2v + \Omega)^{\alpha + \frac{1}{2}}}{v^\frac{1}{2} \Gamma(\alpha + \frac{1}{2}) \Omega^{\alpha}} \right),
\]

where \(\Omega = 2\beta(v + 1)\) (Murphy, 2007). To regularize high evidence on incorrect prediction, we also employ a penalty term \(L_{\text{reg}} = |y(i) - \gamma| \cdot (2\alpha + v)\) introduced by Amini et al. (2019). The total evidential loss \(L_{\text{evi}}\) is given by

\[
L_{\text{evi}} = L_{\text{NLL}} + \lambda_{\text{reg}} \cdot L_{\text{reg}},
\]

where \(\lambda_{\text{reg}} > 0\) is a regularization coefficient.

The implementation of DER using our model described in Section 3.1 is straightforward; We set the dimension of output sequence to 4 (i.e., each dimension represents \(\gamma, v, \alpha, \beta\) and \(\beta\) and train the model according to Eq. (8)). In practice, however, we found that the direct optimization of Eq. (8) results in an overfitting problem in the early training stage. Fig. 3 shows that the validation loss begins to get worse even if the MAE continues to decrease. In other words, if the model is selected based on the validation NLL, it is bound to have suboptimal prediction accuracy which leads to degraded performance on the evaluation metrics. To handle this issue, we propose two different training techniques in the following subsections.

\(^1\)We omit the index \(i\) of the NIG parameters for simplicity.
3.3.1. Weight Initialization with Deterministic Regression

The first idea is to initialize the model weights to ensure high accuracy of model prediction in the beginning. With the proposed model, a predicted value $\hat{y}(i)$ is given by $\hat{y}(i) = E[\mu] = \gamma$. Before optimizing the model according to Eq. (8), we suggest to leverage the MAE loss between $y(i)$ and $\hat{y}(i)$ in the initialization stage:

$$L_{\text{MAE}} = |y(i) - \hat{y}(i)| = |y(i) - \gamma|.$$  \hspace{1cm} (9)

Note that the gradient of $L_{\text{MAE}}$ does not backpropagate through $(\nu, \alpha, \beta)$. The initialization then is performed by selecting a model based on the MAE loss computed on the validation set. With the pretrained weights, the model is again optimized with respect to $L_{\text{evi}}$. The final model is selected based on the evidential loss calculated on the validation set. In our experiments, we verify that this simple training strategy significantly improves the prediction accuracy.

3.3.2. Temperature Scaling

As the training progresses, the model becomes too overconfident before the model prediction becomes accurate enough. To solve this problem, we retrieve the appropriate level of model confidence by scaling $\beta$ and $\nu$ with two scalar temperature parameters $T_\nu > 0$ and $T_\beta > 0$ in Eq. (5) as follows:

$$\mathbb{E}[^2] = \frac{T_\beta \beta}{\alpha - 1}, \text{Var}[^\mu] = \frac{T_\beta \beta}{T_\nu \nu (\alpha - 1)}.$$ \hspace{1cm} (10)

This strategy is motivated by a calibration method for classification networks that scales a logit vector to raise output entropy (Guo et al., 2017). When $T_\beta < 1$, both epistemic and aleatoric uncertainties are readjusted to be smaller. Likewise, as $T_\nu$ increases, the overconfidence of the model is gradually alleviated. For implementation, we first choose the best model based on the MAE loss computed on the validation set after training the model according Eq. (8). Then, $T_\beta$ and $T_\nu$ are optimized with respect to $L_{\text{evi}}$ on the validation set in a post-processing step. Note that all the model parameters are not updated during post-processing, so the model accuracy remains the same. At test time, we use the scaled parameters $\beta' = T_\beta \beta$ and $\nu' = T_\nu \nu$ for uncertainty estimation.

4. Experiments

We conducted a set of experiments using the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC) II (Saeed et al., 2011) database to evaluate the proposed framework. We utilized the refined version of the MIMIC II provided by Kachuee et al. (2015) and followed the same pre-processing steps as in Ibtehaz & Rahman (2020) to sample well-distributed BP signals between 50mmHg and 200mmHg. We used 10 second long signals with sample rate of 125Hz for training and validation sets. Since the evaluation metrics used in this work require BP values computed from a single cardiac cycle, we constructed a test set consisting of 2 second long segments to ensure that each test data contains at least one cardiac cycle. The total duration of the training, validation, and test sets was 250 hours, 27.8 hours, and 75.7 hours, respectively.

We constructed the encoder of the model with 4 blocks of convolution networks ($N = 4$). Each block consisted of a convolution layer with kernel size of 6 and stride of 2 ($k_i = 6$ and $s_i = 2$), a 1x1 convolution layer, and a batch normalization layer. We also applied a weight normalization to each convolution layer. The output channel size of the first block was set to 64 and the subsequent blocks doubled it ($c_1 = 64$ and $c_{i+1} = 2c_i$). The same configuration was symmetrically used for the decoder. Between the encoder and the decoder, a two-layer bidirectional LSTM with hidden size of 512 and a fully-connected layer were employed. The fully connected layer converted the dimension of the output of the bidirectional LSTM back to 512.

We trained the models for 500K iterations with a batch size of 512 using a single 2080-Ti GPU. We used the Adam optimizer (Kingma & Ba, 2014) and set the learning rate to $5 \times 10^{-4}$. When the weight initialization was conducted using $L_{\text{MAE}}$, the model was further fine-tuned 50K steps more according to Eq. (8) where $\lambda_{\text{reg}}$ was set to 1.5. When we performed a post-process to calibrate the model confidence, we adjusted the learning rate to 0.02 and optimized the temperature parameters for 1K steps using the validation set. In this case, we set $\lambda_{\text{reg}}$ to 0.1 for both training and post-processing.

4.1. BP Measurement

4.1.1. Validation of Proposed Framework

We trained the proposed model with various configurations and report the results of BP measurement in Table. 1. SBP, MAP, and DBP were calculated by finding the maximum, average, and minimum values in the test segments, respectively. Since the proposed model is capable of measuring an entire BP waveform, we also report the MAE of arterial blood pressure (ABP) at all time steps. First of all, it can be observed that the additional optimization of $L_{\text{PTP}}$ improved the accuracy of SBP measurement in all cases (Model 2, Model 4, and Model 6). Moreover, the use of $L_{\text{PTP}}$ did not hurt the performance of predicting ABP and MAP values. These results suggest that the optimization of $L_{\text{PTP}}$ efficiently leads the model to prioritize estimating peak values more accurately than other BP values while maintaining the accuracy of other BP predictions. Secondly, we can ver-
10mmHg, and (iii) 15mmHg) and evaluates the accuracy of the prediction errors of our model in Table 3. It is noteworthy that the proposed model meets the requirements of the AAMI standard in all cases. Even in the case of SBP, it is very meaningful to get a grade A score since most literature could not achieve it on the MIMIC II dataset (Ibtehaz & Rahman, 2020; Mousavi et al., 2019). The results demonstrate that the proposed model provides accurate BP measurement.

### 4.1.3. AAMI STANDARD

The AAMI standard is another evaluation metric that has been widely used in the literature for benchmark. The AAMI standard requires BP measuring methods to simultaneously meet the following criteria: (i) mean error (ME) is less than 5mmHg, (ii) the standard deviation (STD) of errors is less than 8mmHg, and (iii) evaluation is performed on at least 85 subjects. We report the average and standard deviation of the prediction errors of our model in Table 3. It is noteworthy that the proposed model meets the requirements of the AAMI standard in all cases. Even in the case of SBP, which is the most difficult BP value to predict, the proposed model satisfies the AAMI criteria by recording 0.270mmHg and 7.058mmHg for ME and STD, respectively. These results demonstrate that the proposed model provides quite helpful for solving an overfitting problem and result in much reliable measurements. when the model parameters were initialized according to \( \mathcal{L}_{\text{MAE}} \) and then updated to optimize \( \mathcal{L}_{\text{evi}} \), the model converged to a way better point for BP measurements (Model 3 and Model 4). Likewise, when the model was selected based on \( \mathcal{L}_{\text{MAE}} \) of the validation set and the model confidence was scaled with temperature parameters, the BP estimation faithfully followed the true BP values (Model 5 and Model 6). Thus, the overall results demonstrate the proposed methods work well in practice. In the following experiments, we evaluate the performance of the proposed framework using Model 6.

### 4.1.2. BHS STANDARD

The BHS standard is a protocol of requirements for the evaluation of BP measuring devices and methods (O’Brien et al., 1993). The BHS standard counts the cumulative number of predictions belonging to three intervals (i.e., whether the absolute error of a prediction is lower than (i) 5mmHg, (ii) 10mmHg, and (iii) 15mmHg) and evaluates the accuracy of measurement according to the tabulated grading criteria. There are four types of grade in the BHS standard: grade A, grade B, grade C, and grade D. To get a specific grade, the cumulative error percentage should satisfy three thresholds simultaneously. If a measurement method cannot fulfill even the grade C thresholds, it acquires a grade D score. Table 2 presents the grading criteria of the BHS standard and the cumulative error percentage of the proposed model. Surprisingly, the proposed model acquires grade A scores in all assessments. In other words, most of our model’s predictions fit well with the corresponding true BP values within the 15mmHg error range. In particular, for the result of SBP, it is very meaningful to get a grade A score since most literature could not achieve it on the MIMIC II dataset (Ibtehaz & Rahman, 2020; Mousavi et al., 2019).

### Table 1. Comparison of BP measurements performance using different model configurations. MAE values are provided with their 95% confidence interval. The results demonstrate that the auxiliary loss \( \mathcal{L}_{\text{FP}} \) and the two training techniques for DER play an important role in achieving high accuracy of BP measurements.

| \( \lambda_{\text{FP}} \) | TRAINING TECHNIQUE | MEAN ABSOLUTE ERROR (mmHg) |
|---|---|---|
|  | SBP | MAP | DBP | ABP |
| MODEL 1 | 0.0 | NOT APPLIED | 3.870 ± 0.034 | 1.956 ± 0.021 | 1.998 ± 0.021 | 3.161 ± 0.002 |
| MODEL 2 | 1.0 | | 3.469 ± 0.032 | 1.949 ± 0.020 | 1.973 ± 0.021 | 3.170 ± 0.002 |
| MODEL 3 | 0.0 | WEIGHT INITIALIZATION | 3.443 ± 0.034 | 1.871 ± 0.021 | 1.904 ± 0.021 | 2.831 ± 0.002 |
| MODEL 4 | 1.0 | | 3.040 ± 0.033 | 1.811 ± 0.021 | 1.776 ± 0.022 | 2.817 ± 0.002 |
| MODEL 5 | 0.0 | TEMPERATURE SCALING | 3.098 ± 0.034 | 1.761 ± 0.021 | 1.756 ± 0.022 | 2.688 ± 0.002 |
| MODEL 6 | 1.0 | | 3.044 ± 0.035 | 1.811 ± 0.022 | 1.815 ± 0.023 | 2.768 ± 0.002 |

### Table 2. Cumulative error percentage of BP predictions obtained from our model and grading criteria of the BHS standard.

| | \( \leq 5\text{mmHg} \) | \( \leq 10\text{mmHg} \) | \( \leq 15\text{mmHg} \) |
|---|---|---|---|
| SBP | 82.83 | 91.73 | 95.34 |
| MAP | 90.40 | 96.27 | 98.24 |
| DBP | 91.36 | 96.62 | 98.33 |
| ABP | 85.03 | 93.28 | 96.44 |

| GRADING CRITERIA | A | B | C | D |
|---|---|---|---|---|
| SBP | 60 | 85 | 95 |
| MAP | 50 | 75 | 90 |
| DBP | 40 | 65 | 85 |

| | ME (mmHg) | STD (mmHg) | # OF SUBJECTS |
|---|---|---|---|
| SBP | 0.337 | 7.058 | |
| MAP | 0.270 | 4.364 | 942 |
| DBP | 0.200 | 4.508 | |
| ABP | 0.270 | 6.311 | |

### Table 3. Comparison with the AAMI standard.

Criterion \( \leq 5 \) \( \leq 8 \) \( \geq 85 \)
4.2. Reliability of Model Prediction

4.2.1. Uncertainty Estimation

To efficiently show the relation between the reliability of model prediction and its accuracy, we aim to visualize the level of model uncertainty. For this goal, we first calculated the model uncertainty $\text{Var}[\mu]$ of each measurements. Then, we colored the area around the BP measurements in proportion to the corresponding model uncertainty. In Fig. 4, we present several visualized examples of model reliability obtained through this process. By comparing the first and second rows, we can observe the significant correlation between the accuracy of model prediction and the estimated uncertainty. When the estimated uncertainty is high, as shown in the first row, the predicted BP values have some obvious errors compared to the actual BP values. In contrast, when the model uncertainty is comparatively low as in the second row, the model predictions are highly trustworthy and almost perfectly fit the true BP values. The experimental results demonstrate that we have appropriately applied the DER framework to the BP measurement task along with the proposed training methods and managed to properly estimate the reliability of the predictions.

4.2.2. BP Measurement on Selected Samples

The uncertainty of model prediction can be exploited to produce more reliable BP measurements. For example, if the model uncertainty for a particular PPG input is high, we can remeasure it or optimize the model for more steps using PPG signals similar to that input pattern. Here, we choose to skip BP measurements on low-reliability PPG signals and re-evaluate BP measurement performance. We computed the model uncertainty of all test samples and excluded 20% of them with the highest model uncertainty. Using this subset of the test set, we assessed the BP measurement performance of our model again. As shown in Table 4, we can observe that the MAE values of SBP, MAP, DBP, and ABP decreased by 0.666 mmHg, 0.433 mmHg, 0.468 mmHg, and 0.539 mmHg, respectively. The experimental results well illustrate that the estimated reliability is highly consistent with the model accuracy and it can be further utilized in a wide range of applications.

4.2.3. Hypertension Classification

As a real-world application, we can diagnose hypertension based on the model prediction. The BP classification criteria for SBP and DBP are well established (Holm et al., 2006), and for this experiment, we classified the levels of BP into three categories based on SBP: hypertension, prehypertension, and normotension. In order for more accurate diagnosis, we also conducted BP classification on the subset of the test set filtered by reliability in the same way as in Section 4.2.2. We present the resultant confusion matrices of BP classification in Fig. 5. Due to the sophisticated architecture and the proposed methods, the model achieved high-performance by detecting hypertension with a proba-
Continuous Monitoring of Blood Pressure with Evidential Regression

Table 5. Comparison of overall performance with other approaches. The proposed model meets the BHS and AAMI standards, and achieves state-of-the-art performance on the MAE criteria in SBP and DBP measurements. In addition, the proposed model is the only one capable of monitoring both the continuous BP waveform and the prediction reliability.

| INPUT | MAE (MMHG) | BHS STANDARD | AAMI STANDARD | WAVEFORM MONITORING | PREDICTION RELIABILITY |
|-------|------------|--------------|---------------|----------------------|------------------------|
|       | SBP | DBP | SBP | DBP | SBP | DBP | SBP | DBP | SBP | DBP |
| KACHUEE ET AL. | PPG, ECG | 8.21 | 4.31 | D | A | × | √ | × | × |
| IBTEHAZ ET AL. | PPG | 5.73 | 3.45 | B | A | × | √ | √ | × |
| LI ET AL. | PPG, ECG | 6.73 | 2.52 | B | A | × | √ | × | × |
| HSU ET AL. | PPG | 3.21 | 2.23 | A | A | √ | √ | × | × |
| OURS | PPG | **3.10** | **1.76** | A | A | √ | √ | √ | √ |

4.3. Comparison with Other Works

Although there have been lots of studies using PPG signals to measure BP, it is difficult to compare them directly with our work since each study used different configurations in the experiments, and some of them even assessed the performance using private data. In order to evaluate the model as fairly as possible, we selected the papers that conducted training and evaluation using the MIMIC II dataset, and present the overall performance comparison in Table 5. First of all, it can be observed that the estimation of SBP is so difficult that Kachuee et al. (2016) and Li et al. (2020) obtained grade D and grade B from the BHS standard, respectively, even with additional electrocardiogram (ECG) input. Nevertheless, our model achieved grade A for both SBP and DBP, recording the lowest MAEs. Though the BP measurement performance of Hsu et al. (2020) is close to ours, their model cannot estimate other BP values except SBP and DBP. Ibtehaz & Rahman (2020) proposed to predict a whole BP waveform using a raw PPG signal, but their measurement accuracy was somewhat insufficient for the BHS standard. Most importantly, none of the existing works has adopted a toolbox to provide the reliability of prediction, which may play a critical role in real-world deployments. These facts strongly support that our model achieved cutting-edge performance with the attractive features and paved a new way for continuous monitoring of BP waveforms.

5. Conclusion and Future Work

In this paper, we have introduced an elaborately designed framework for monitoring a continuous waveform of BP using the raw PPG signal as input. We experimentally demonstrated that the proposed model is capable of measuring the BP values with high accuracy and satisfies the BHS and AAMI standards even in SBP measurement. To go further, we proposed two training techniques to adequately apply the DER framework to the BP measurement task. These techniques ensure strong correlation between the model accuracy and the estimated reliability, and this is experimentally demonstrated through the uncertainty visualization, BP measurement and BP classification on the high-reliability subset of the test set. We believe that the estimated reliability can be utilized to help determine whether the measured BP values should be trusted or provide more informative training data to optimize robust models. We also expect that the proposed approach can be employed for other safety-critical applications as well.

Although we presented the state-of-the-art model for BP estimation, there are still some room for improvement. First, the proposed approach may accompany biased uncertainty estimation due to the two-stage training procedure. To enhance our approach, one may combine the weight initialization and temperature scaling together or develop another end-to-end model for exact likelihood estimation. Another issue is that the threshold for model reliability is rather heuristically determined in this work. We believe that the boundaries can be theoretically derived to produce more convincing reliability estimates and improved results. In order to develop safe medical applications, we plan to conduct extensive research to address these issues.
Continuous Monitoring of Blood Pressure with Evidential Regression

References

Allen, J. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1, 2007.

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. *arXiv preprint arXiv:1910.02600*, 2019.

Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C., and Nazeran, H. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International journal of biosensors & bioelectronics*, 4(4):195, 2018.

Cho, K., Van Merriënoor, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017.

Défossez, A., Usunier, N., Bottou, L., and Bach, F. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*, 2019.

Elgendi, M. On the analysis of fingertip photoplethysmogram signals. *Current cardiology reviews*, 8(1):14–25, 2012.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, X., Goubran, R. A., and Liu, X. P. Evaluation of the correlation between blood pressure and pulse transit time. In *2013 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 17–20. IEEE, 2013.

Holm, S. W., Cunningham, L. L., Bensadoun, E., and Madsen, M. J. Hypertension: classification, pathophysiology, and management during outpatient sedation and local anesthesia. *Journal of oral and maxillofacial surgery*, 64(1):111–121, 2006.

Hsu, Y.-C., Li, Y.-H., Chang, C.-C., and Harfiya, L. N. Generalized deep neural network model for cuffless blood pressure estimation with photoplethysmogram signal only. *Sensors*, 20(19):5668, 2020.

Ibtehaz, N. and Rahman, M. S. Ppg2abp: Translating photoplethysmogram (ppg) signals to arterial blood pressure (abp) waveforms using fully convolutional neural networks. *arXiv preprint arXiv:2005.01669*, 2020.

Investigators, S. H.-. Catheter-based renal sympathetic denervation for resistant hypertension: durability of blood pressure reduction out to 24 months. *Hypertension*, 57(5):911–917, 2011.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Kachuee, M., Kiani, M. M., Mohammadzade, H., and Shabany, M. Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. In *2015 IEEE international symposium on circuits and systems (ISCAS)*, pp. 1006–1009. IEEE, 2015.

Kachuee, M., Kiani, M. M., Mohammadzade, H., and Shabany, M. Cuffless blood pressure estimation algorithms for continuous health-care monitoring. *IEEE Transactions on Biomedical Engineering*, 64(4):859–869, 2016.

Kavsaoğlu, A. R., Polat, K., and Hariharan, M. Non-invasive prediction of hemoglobin level using machine learning techniques with the ppg signal’s characteristics features. *Applied Soft Computing*, 37:983–991, 2015.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Li, Y.-H., Harfiya, L. N., Purwandari, K., and Lin, Y.-D. Real-time cuffless continuous blood pressure estimation using deep learning model. *Sensors*, 20(19):5606, 2020.

Liang, Y., Chen, Z., Ward, R., and Elgendi, M. Photoplethysmography and deep learning: enhancing hypertension risk stratification. *Biosensors*, 8(4):101, 2018.

Mousavi, S. S., Firouzman, M., Charmi, M., Hemmati, M., Moghadam, M., and Ghorbani, Y. Blood pressure estimation from appropriate and inappropriate ppg signals using a whole-based method. *Biomedical Signal Processing and Control*, 47:196–206, 2019.

Murphy, K. P. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2):16, 2007.

O’Brien, E., Petrie, J., Littler, W., de Swiet, M., Padfield, P. L., Altman, D., Bland, M., Coats, A., Atkins, N., et al. The british hypertension society protocol for the evaluation of blood pressure measuring devices. *J hypertens*, 11(Suppl 2):S43–S62, 1993.
Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39 (5):952, 2011.

Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *arXiv preprint arXiv:1602.07868*, 2016.

Seo, J., Pietrangelo, S. J., Lee, H.-S., and Sodini, C. G. Noninvasive arterial blood pressure waveform monitoring using two-element ultrasound system. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 62 (4):776–784, 2015.

Slapničar, G., Mlakar, N., and Luštrek, M. Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. *Sensors*, 19(15):3420, 2019.

Van Montfrans, G. A. Oscillometric blood pressure measurement: progress and problems. *Blood pressure monitoring*, 6(6):287–290, 2001.

World Health Organization. World health statistics 2020: monitoring health for the SDGs, sustainable development goals, Geneva, 2020.