DeepPulse: An Uncertainty-aware Deep Neural Network for Heart Rate Estimations from Wrist-worn Photoplethysmography

Daniel Ray, Tim Collins, and Prasad V. S. Ponnapalli

***Abstract*— Wearable Photoplethysmography (PPG) has gained prominence as a low cost, unobtrusive and continuous method for physiological monitoring. The quality of the collected PPG signals is affected by several sources of interference, predominantly due to physical motion. Many methods for estimating heart rate (HR) from PPG signals have been proposed with Deep Neural Networks (DNNs) gaining popularity in recent years. However, the "black-box" and complex nature of DNNs has caused a lack of trust in the predicted values. This paper contributes DeepPulse, an uncertainty-aware DNN method for estimating HR from PPG and accelerometer signals, with aims of increasing the reliability, usability and interpretability of the predicted HR values. To the best of the authors' knowledge no PPG HR estimation method has considered aleatoric and epistemic uncertainty metrics. The results show DeepPulse is the most accurate method for DNNs with less than 1 million network parameters. Finally, recommendations are given to reduce epistemic uncertainty, validate uncertainty estimates, improve the accuracy of DeepPulse as well as reduce the model size for resource-constrained edge devices.**

I. INTRODUCTION

Wrist-worn reflectance mode PPG sensing is popular in many wearable devices as it provides a means of low cost, unobtrusive and continuous physiological monitoring [1]. The performance of PPG sensing is affected by several sources of interference including biological characteristics, sensor configuration and placement as well as ambient light [1]. However, the main source of interference is physical motion which distorts the collected PPG signal. The removal of motion artefacts from the signal is a challenge due to overlapping frequency bands and amplitudes much larger than the pulsatile component of the signal [1], [2].

Computational methods for estimating HR from PPG signals consist of four main steps: prepossessing, de-noising, heart rate estimation and heart rate tracking [2]. A common approach used across existing methods for de-noising is to incorporate a motion reference sensor, such as a triaxial accelerometer or gyroscope, in order to capture motion data at the measurement site and compensate for the interference the motion causes [3], [4]. Many conventional signal processing approaches to HR estimation rely on expert-tuned parameters [3] leading to difficulties in generalizing the methods [4], [5]. In order to prevent this, researchers have explored the use of deep neural networks (DNNs) for HR estimations [4]–[9]. Although the performance improvements are significant, DNNs for edge devices have their own challenges including data asymmetry, multi-modality sensing and resource constraints of edge devices [10].

Classical approaches to the fusion of heterogeneous sensing modalities rely on feature engineering to extract independent features from each sensing modality which are then fused together. This approach of extracting different features from individual sensors disregards features that use multiple sensors' data to capture information that neither has in isolation [11]. In many applications, DNNs have been adopted instead due to their ability to learn to extract features during training [11]–[13] showing improved performance in applications such as gait recognition [11], human activity recognition [11]–[13], car tracking [12], dynamic gas mixtures estimations [13] and cuffless blood pressure monitoring [13].

One major drawback to the use of DNNs is a lack of trust in the predicted values due to high complexity and uninterpretability of the generated DNNs, mainly from deep and non-linear structures [13]. In order to increase the reliability, usability and interpretability of DNNs researchers have explored ways to represent uncertainty within DNNs [13]–[16]. The two main sources of uncertainty are "aleatoric" and "epistemic". Aleatoric uncertainty describes the irreducible uncertainty in the input data due to an inherent property of the data distribution such as randomness or noise [14]. Epistemic uncertainty describes uncertainty in the model that occurs due to inadequate data which may be reduced by increasing the amount and 'diversity' of the training data [14].

Researchers have explored several methods to incorporate and quantify uncertainty in DNNs such as Monte Carlo Dropout (MCDropout), Variational Inference and Ensemble methods [14], [16]. The uncertainty framework proposed by [16] is advantageous as it requires little modification to existing DNNs [14]. The framework uses MCDropout with an aleatoric uncertainty term to simultaneously estimate aleatoric and epistemic uncertainty, showing promising results in several applications [15], [16]. MCDropout has been theorized to approximate Gaussian processes by activating dropout layers during the prediction phase to provide an ensemble of predictions [16]. The variability of the ensemble predictions distribution quantifies epistemic uncertainty [15], [16]. In order to incorporate aleatoric uncertainty, a second output unit is added to the DNN with a specially-designed loss function such as negative log likelihood (NLL). The two

This research received no external funding.

D. Ray, T. Collins and P. V. S. Ponnapalli are with the Department of Engineering, Manchester Metropolitan University, Manchester, UK. (e-mail: Daniel.Ray@stu.mmu.ac.uk, T.Collins@mmu.ac.uk and P.Ponnapalli@mmu.ac.uk)

output units of the DNN estimate μ and σ of a distribution, where μ represents the mean value of the distribution and σ represents the standard deviation of the distribution used to quantify aleatoric uncertainty [15].

II. METHODOLOGY

A. Datasets

1) IEEE SPC 2015: consists of two datasets that employed different protocols, namely IEEE Train and IEEE Test. Both datasets were collected using a green (515 nm) reflectance mode PPG sensor as well as a single lead chest-worn ECG. IEEE Train collected 12 sessions whilst IEEE Test collected 10 sessions. Both datasets employed laboratory-based protocols with IEEE Train using a treadmill and IEEE Test focusing on arm movements, with each session duration being no longer than 15 minutes [3].

2) PPG-DaLiA: was collected using an Empatica E4 wrist-worn reflectance mode PPG sensor used green (520 nm) and red (660 nm) LEDs and a 3-lead chest-worn ECG. A total of 15 sessions were collected using a naturalistic protocol of various daily activities with each session duration being more than 1.5 hours long [4].

3) BAMI-II: was collected using a wrist-worn reflectance mode green (525 nm) PPG sensor and a medical-grade 3-lead chest-worn ECG Holter monitor. A total of 24 sessions were collected employing a laboratory-based protocol using a treadmill with each session duration being 14 minutes [17].

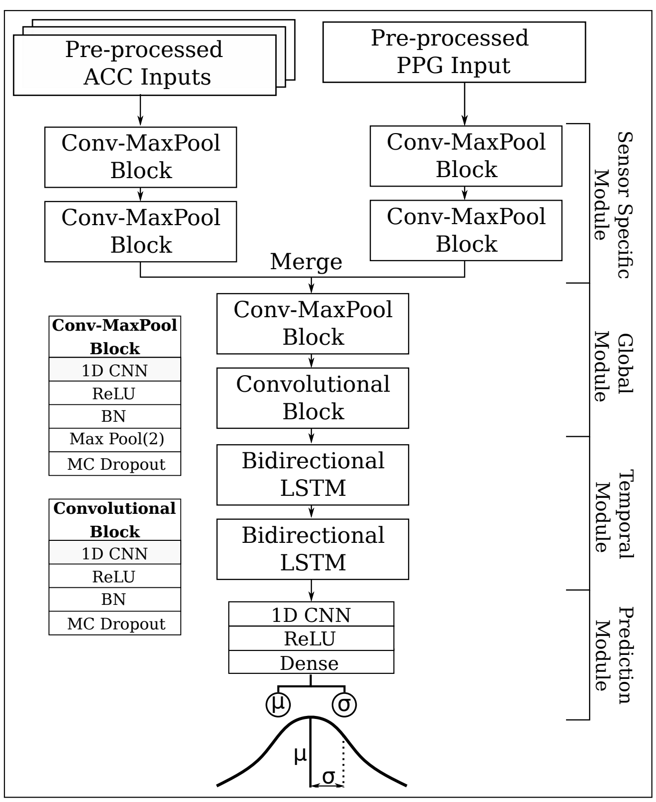

Fig. 1. The Architecture of DeepPulse.

B. Preprocessing and Learning Strategy

The PPG and accelerometer signals were first subject to a 2nd order Butterworth band-pass filter with cutoff frequencies of 0.5 Hz - 4.5 Hz to remove components of the signals outside the range of cardiac activity. The signals were then re-sampled to 64 Hz and normalized to zero mean and unit variance. Finally, a sliding window was applied to the signals with a window length of 8 seconds and a 2 second slide. To reduce the effects of data asymmetry a leave-one-session-out (LOSO) cross-validation scheme was employed [4], [6] where each session was used as test data exactly once. A more detailed explanation of the implemented LOSO scheme can be found in [4].

TABLE I
HYPERPARAMETERS OF DEEPPULSE ARCHITECTURE

Sensor-specific Module	
Number of Conv. Filters	64
Global Module	
Number of Conv. Filters	128
Temporal Module	
Number of LSTM Units	32
Network Parameters	
All Conv. Blocks: Convolutional Kernel Size	16
Merge Type	Concatenate Axis = 2
Dropout Rate	0.15
Optimizer	Nadam

C. DeepPulse Architecture and Implementation

DeepPulse contains four main architectural sections: sensor-specific module, global module, temporal module and prediction module (Figure 1). The convolutional blocks in the sensor-specific module extract local interactions within each sensing modality. The sensor-specific features are then merged together and passed through the convolutional blocks in the global module to extract global features. The global features are then used in the temporal module to extract temporal features using bidirectional Long Short-Term Memory (LSTM) layers. The temporal features are passed to the prediction module which contains a convolutional layer to reduce the dimensionality of the features for the fully connected layer. The selected hyper-parameters of the architectural components and network parameters can be found in Table I. A NLL loss function was used to evaluate how the DNN models the data in terms of both accuracy and aleatoric uncertainty. Each convolutional block contains a MCDropout layer used to produce an ensemble of predictions (T=10) for each input to evaluate epistemic uncertainty.

The training phase of DeepPulse was run for 200 epochs with a batch size of 32. During the training phase, early stopping was employed to reduce overfitting and the learning rate was reduced when the learning had stagnated. DeepPulse was implemented using Tensorflow (Version: 2.7.0) and

Tensorflow Probability (Version: 0.14.1). Computation was carried out using 8 Intel Broadwell CPU cores and a NVIDIA Tesla K80 GPU (CUDA Version: 11.2). The implementation of DeepPulse can be found at: `https://github.com/danielray54/DeepPulse`

D. Evaluation Metrics

Mean absolute error (MAE) was employed to assess the accuracy. Predicted values were averaged across all LOSO iteration to obtain a generalized MAE. Additionally, two uncertainty metrics were employed. $u_a(x_i)$ is the aleatoric uncertainty (Equation 1) which is the average of the squared $\sigma_{i,t}$ output unit for an ensemble of predictions, T, for each input window x_i:

$$u_a(x_i) = \frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}_{i,t}^2 \tag{1}$$

u_e is the epistemic uncertainty (Equation 2) which is the variance computed from the predicted mean values $\mu_{i,t}$ from the ensemble of predictions, T, for each input window x_i:

$$u_e(x_i) = \frac{1}{T}\sum_{t=1}^{T}\mu_{i,t}^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\mu_{i,t}\right)^2 \tag{2}$$

III. Results

A. Accuracy & Complexity

The MAE results show that DeepPulse is the second most accurate method of all DNN PPG HR estimation methods for all datasets (Table II). However, when comparing methods with less than 1 million parameters (Table III) DeepPulse is the most accurate for all datasets. This is significant as models with large complexities have not accounted for the resource constraints of edge devices [10].

TABLE II
COMPARISON OF MEAN ABSOLUTE ERRORS FOR DNN PPG HR ESTIMATION METHODS

Method	**Datasets**			
	IEEE Train	**IEEE Test**	**PPG-DaLiA**	**BAMI-II**
Deep PPG [4]	4.00 ±5.40	16.51 ±16.10	7.65 ±4.20	N/A
CorNET (LOSO) [6]	4.67 ±3.71	6.61 ±5.35	N/A	N/A
Binary CorNET [6]	6.20 ±4.95	7.31 ±6.14	N/A	N/A
PPGnet [7]	3.36 ±4.10	12.48 ±14.45	N/A	N/A
Chung *et al.* [8]	0.67 ±0.50	0.86 ±0.80	N/A	1.46 ±1.23
MH Conv-LSTM DeepPPG [9]	N/A	N/A	6.28 ±3.53	N/A
DeepPulse	2.76 ±2.95	5.05 ±5.50	2.12 ±3.09	2.38 ±2.57

All values are BPM.

B. Uncertainty

For the IEEE datasets, as the number of input windows per activity decreases the epistemic uncertainty estimates increase (Figure 2(a)). This supports the hypothesis that increasing the size and 'diversity' of the dataset will reduce the epistemic uncertainty. Assuming more intense activity or higher BPM values require more movement from the body thus more noise in the PPG signals then as either BPM values or activity intensity increase so will the aleatoric uncertainty estimates which is shown for the BAMI-II and PPG-DaLiA datasets in Figure 2(b) & 2(c). Finally, Figure 2(d) illustrates that there is little to no relationship between between aleatoric uncertainty and epistemic uncertainty estimates for the BAMI-II dataset.

TABLE III
COMPARISON OF NETWORK COMPLEXITIES FOR DNN PPG HR ESTIMATION METHODS

Method	**Number of Parameters**
Deep PPG [4]	8.5M
CorNET [5]	250K
PPGnet [7]	765K
Chung *et al.* [8]	3.3M
MH Conv-LSTM DeepPPG [9]	680K
DeepPulse	730K

IV. Future Work

The performance of PPG sensing is affected by several sources of interference and inaccuracies. Some of these sources such as skin tone, skin temperature, age, sex and BMI have not been fully considered in the datasets used. Increasing the size and 'diversity' of the data will be beneficial in improving the accuracy, robustness and generalizability [1] as well as epistemic uncertainty of DNN PPG HR estimation algorithms. Moreover, ensuring that the collected "truth values" are an accurate depiction of the cardiac activity is essential which can achieved by using medically validated chest-worn ECG devices [1]. In order to improve the performance and reduce the model size of DeepPulse, hyperparameter optimization and network architecture search should be carried out [18]. Additionally, weight clustering and model quantization may prove to be effective methods to further reduce the model size. Finally, further improvement to the accuracy of DeepPulse may be made by introducing a post-processing step that averages predicted values of several input windows when the aleatoric uncertainty is high. To better evaluate aleatoric uncertainty, an accurate signal-to-noise ratio method should be developed to eliminate assumption made based on activity type. Additionally, to validate the epistemic uncertainty estimates, training DeepPulse on subsets of the datasets would provide more insight. Similarly, adding noise to the input windows would enable the validation of the aleatoric uncertainty estimates.

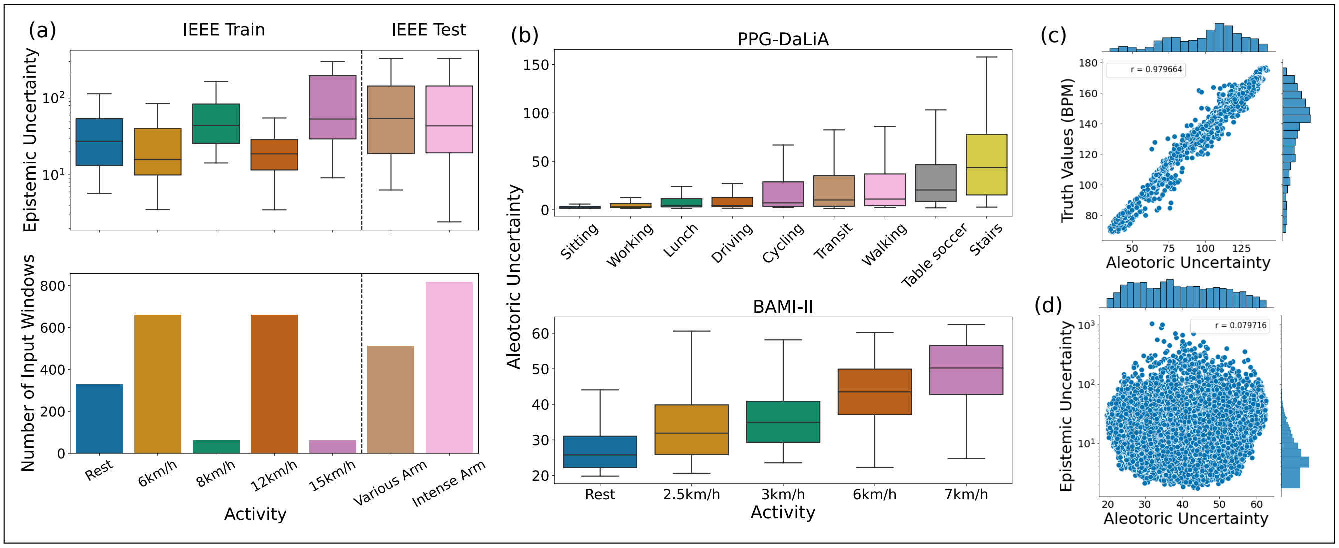

Fig. 2. (a) shows the relationship between epistemic uncertainty and the number of input windows for each activity in both of the IEEE datasets, (b) shows the relationship between aleatoric uncertainty and activity in the BAMI-II and PPG-DaLiA datasets, (c) shows the relationship between aleatoric uncertainty and truth values in BAMI-II dataset and (d) shows the relationship between aleatoric and epistemic uncertainty in the BAMI-II dataset.

V. CONCLUSION

Wearable Photoplethysmography (PPG) has gained prominence as a method for physiological monitoring but is subject to several sources of interference making the estimation of HR challenging. DNNs have gained popularity in recent years with promising results. However, the "black-box" and complex nature of DNNs has caused a lack of trust in the predicted values. This paper contributes DeepPulse, a multimodal uncertainty-aware DNN method for estimating HR from PPG and accelerometer signals. The results show DeepPulse is the most accurate method for DNNs with less than 1 million network parameters. Finally, recommendations have been given to improve the accuracy and reduce the complexity of DeepPulse for resource-constrained edge devices as well as reduce and validate uncertainty estimates.

REFERENCES

[1] D. Ray, T. Collins, S. Woolley and P. Ponnapalli, "A Review of Wearable Multi-wavelength Photoplethysmography," in IEEE Reviews in Biomedical Engineering, doi: 10.1109/RBME.2021.3121476.

[2] Pankaj, A. Kumar, R. Komaragiri, and M. Kumar, "A Review on Computation Methods Used in Photoplethysmography Signal Analysis for Heart Rate Estimation," Archives of Computational Methods in Engineering, vol. 1. p. 3, 2021, doi: 10.1007/s11831-021-09597-4.

[3] 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 Trans. Biomed. Eng., vol. 62, no. 2, pp. 522–531, 2015, doi: 10.1109/TBME.2014.2359372.

[4] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks," Sensor, vol. 19, no. 14, pp. 1–27, 2019, doi: 10.3390/s19143079.

[5] D. Biswas et al., "CorNET: Deep Learning Framework for PPG-Based Heart Rate Estimation and Biometric Identification in Ambulant Environment," IEEE Trans. Biomed. Circuits Syst., vol. 13, no. 2, pp. 282–291, 2019, doi: 10.1109/TBCAS.2019.2892297.

[6] L. G. Rocha et al., "Binary CorNET: Accelerator for HR Estimation from Wrist-PPG," IEEE Trans. Biomed. Circuits Syst., vol. 14, no. 4, pp. 715–726, Aug. 2020, doi: 10.1109/TBCAS.2020.3001675.

[7] A. Shyam, V. Ravichandran, S. P. Preejith, J. Joseph, and M. Sivaprakasam, "PPGnet: Deep Network for Device Independent Heart Rate Estimation from Photoplethysmogram," Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, pp. 1899–1902, 2019, doi: 10.1109/EMBC.2019.8856989.

[8] 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, vol. 8, pp. 63390–63402, 2020, doi: 10.1109/ACCESS.2020.2981956.

[9] M. Wilkosz and A. Szczęsna, "Multi-Headed Conv-LSTM Network for Heart Rate Estimation during Daily Living Activities," Sensors, vol. 21, no. 15, p. 5212, Jul. 2021, doi: 10.3390/s21155212.

[10] X. Zeng, "Mobile sensing through deep learning," in MobiSys 2017 PhD Forum - Proceedings of the 2017 Workshop on MobiSys 2017 Ph.D. Forum, co-located with MobiSys 2017, Jun. 2017, vol. 9908 LNCS, pp. 5–6, doi: 10.1145/3086467.3086476.

[11] V. Radu et al., "Multimodal Deep Learning for Activity and Context Recognition," Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol., vol. 1, no. 4, pp. 1–27, Jan. 2018, doi: 10.1145/3161174.

[12] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing," 26th Int. World Wide Web Conf. WWW 2017, pp. 351–360, Nov. 2016, doi: 10.1145/3038912.3052577.

[13] S. Yao et al., "RDeepSense: Reliable Deep Mobile Computing Models with Uncertainty Estimations," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 1, no. 4, Jan. 2018, doi: 10.1145/3161181.

[14] O. Dürr, B. Sick, and E. Murina, Probabilistic Deep Learning: with Python, Keras and Tensorflow Probability. Manning Publications, 2020.

[15] K. Fang, D. Kifer, K. Lawson, and C. Shen, "Evaluating the Potential and Challenges of an Uncertainty Quantification Method for Long Short-Term Memory Models for Soil Moisture Predictions," Water Resour. Res., vol. 56, no. 12, p. e2020WR028095, Dec. 2020, doi: 10.1029/2020WR028095.

[16] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in 33rd International Conference on Machine Learning, ICML 2016, Jun. 2016, vol. 3, pp. 1651–1660, Accessed: Jan. 10, 2022. [Online]. Available: https://arxiv.org/abs/1506.02142v6.

[17] H. Lee, H. Chung, and J. Lee, "Motion Artifact Cancellation in Wearable Photoplethysmography Using Gyroscope," IEEE Sens. J., vol. 19, no. 3, pp. 1166–1175, Feb. 2019, doi: 10.1109/JSEN.2018.2879970.

[18] D. Ray, T. Collins, and P. Ponnapalli, "Deep Neural Network Architecture Search for Wearable Heart Rate Estimations," Stud. Health Technol. Inform., vol. 281, pp. 1106–1107, May 2021, doi: 10.3233/SHTI210366.