22HLT01 QUMPHY

D3

Report on validating the uncertainties obtained for existing machine learning and deep learning models from D1 and D2. This includes the comparison of accuracy and uncertainty of at least 3 classification and 3 regression machine learning and/or deep learning models in order to identify models and methods which have high accuracy and low uncertainty for a wide range of tasks

Organisation name of the lead participant for the deliverable: PTB

Due date of the deliverable: 31st March 2026

Actual submission date of the deliverable: 26th June 2026

Confidentiality Status: PU - Public, fully open (remember to deposit public deliverables in a trusted repository)

Deliverable Cover Sheet

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EURAMET. Neither the European Union nor the granting authority can be held responsible for them.

The project has received funding from the European Partnership on Metrology, co-financed from the European Union's Horizon Europe Research and Innovation Programme and by the Participating States.

European Partnership Co-funded by the European Union

METROLOGY PARTNERSHIP EURAMET

Contents

1 Summary

In our previous reports, we described work done by the consortium on different machine learning approaches to the problems of blood pressure prediction (regression) and detection of atrial fibrillation (binary classification) using photoplethysmography (PPG) signals as input [1], which was followed by a systematic evaluation of a wide range of uncertainty quantification methods for the same two problems [2]. The aim of this work was to provide reliable uncertainties alongside machine learning predictions in order to provide guidance to clinicians in their interpretation of the trustworthiness of the model predictions. In Task 1.3, a number of distinct follow on activities are described making use of the benchmark datasets which were developed in WP2 [3]. These activities are as follows:

- A1.3.1: Analyse the full set of models which take the raw PPG signals, PPG features or PPG signals converted to images as inputs on all the benchmark datasets and summarise the results using the evaluation framework.
- A1.3.2: Analyse the results from A1.3.1 for different demographic subgroups (e.g. biological sex) and compare results with models trained on single sex data. Also, classify PPG signals by biological sex in order to determine whether there is a discernable difference in PPG signals for males and females.
- A1.3.3: Consider the robustness of the models by adding noise to the PPG signals and by considering performance on out-of-distribution datasets.
- A1.3.4: Investigate the dependence of machine learning results using PPG signals on various factors such as number of samples, added noise or filterning of the signals.
- A1.3.5: Investigate the disentangling of aleatoric and epistemic uncertainty where this is possible.
- A1.3.6: Classify PPG signals by the six-class Fitzpatrick skin tone in order to determine whether or not machine learning can distinguish between signals collected from subjects with different skin tones.

These six topics have all been investigated by the consortium and this report summarises the work done for each activity.

2 A1.3.1 – Results for all the benchmark datasets

LNE, NPL, Uni-Oldenburg and PTB will repeat the preliminary analyses carried out in A1.1.6 and A1.2.7 but this time for the full set of models from A1.1.2-A1.1.5 and for the final set of datasets provided by A2.1.4. This analysis will cover all model types from A1.1.2-A1.1.5 in combination with all model-agnostic UQ methods in addition to selected model-specific UQ methods from A1.2.1-A1.2.5. This will provide a first comprehensive assessment of performance vs. uncertainty based on in-distribution data.

2.1 Benchmark datasets

Our experiments cover both regression and classification problems across three physiological datasets. For regression, we consider blood pressure estimation and vascular age estimation on AuroraBP [4], as well as respiratory rate estimation on MIMIC-PERform [5] and OSASUD [6]. For classification, we use AuroraBP to predict blood pressure level as a three-class problem (elevated, non_elevated, hypertensive). On OSASUD, we address apnea event detection as a multi-label problem, distinguishing between obstructive apnea, hypopnea, their combination, central apnea, mixed apnea, and an aggregated "all apneas" label. Table 1 displays a summary of the benchmark datasets that were described in [3].

Table 1: Summary of the benchmark datasets used for the different regression and classification tasks.

Dataset	**Naming**	**Downstream task**	**ML task**	**Additionnal Information**
AuroraBP	BP	Blood pressure estimation (systolic)	regression	
AuroraBP	BP	Blood pressure estimation (diastolic)	regression	
AuroraBP	VA	Vascular age estimation	regression	
MIMIC-PERform	RR	Respiratory rate estimation	regression	
OSASUD	RR	Respiratory rate estimation	regression	
AuroraBP	BPL	Blood pressure level	classification (multi-class)	3 CLASSES: • ELEVATED • NON_ELEVATED • HYPERTENSIVE
OSASUD	AED	Apnea event detection	classification (multi-label)	6 LABELS: • APNEA-OBSTRUCTIVE • HYPOPNEA • APNEA-OBSTRUCTIVE & HYPOPNEA • APNEA-CENTRAL • APNEA-MIXED • ALL APNEAS

2.2 Preprocessing

Across all datasets, we first filtered out any signal–target pairs where the target value was either 0 or missing (NaN) to avoid training on invalid labels. The data processing steps for AuroraBP are detailed in Section 3.2.2. On top of the proposed pre-processing stages, we additionally applied a simple threshold-based screening step to discard suspected outliers. When needed, we standardised the data to improve numerical stability and facilitate model training. Finally, each dataset was split into training (70 %), validation (10 %), calibration (10 %), and test (10 %) subsets, where the calibration set is needed for some methods to calibrate the uncertainties. For the multi-label OSASUD dataset, folds were generated using subject-level

stratification according to AHI-based apnea severity classes, with random shuffling within each stratum.

2.3 Models

To process the different datasets, we investigated three input representations: raw-signal, feature-based, and image-based. For each representation, we used a single dedicated model, trained from scratch in all experiments. The models tested against the benchmark datasets detailed in Table 1 are described below.

Feature-based model (Wavelet + Multi-Layer Perceptron (MLP)): PPG signals were decomposed into time-frequency components using a Wavelet Packet Decomposition. Wavelet + MLP formed a pipeline where the wavelet transform served as a sophisticated feature extractor, and the MLP functioned as the predictive model utilising those features. The mother wavelet used for the decomposition was the Daubechies "db6" of order 3 [7].

Image-based model (Scalogram + ResNet18): PPG signals were transformed using the continuous wavelet transform (CWT) into time-frequency images (224×224 RGB images) derived from PPG signals that captured localised signal variations. The CWT-based images were used as inputs for a ResNet18 model.

Raw-signal-based model (Temporal Convolutional Networks (TCNs)): PPG signals were directly fed as inputs into the network that used dilated causal convolutions and residual connections to capture long-range dependencies in sequential signals.

For ease of simplicity, when referring to the models in the coming sections, the following mapping is used:

- F : Feature-based model
- I : Image-based model
- R : Raw-signal-based model

The model head and training objective were adapted to the downstream task. For classification tasks, the model output the logits and a binary cross-entropy loss with logits was used (this loss combines a sigmoid layer and the binary cross-entropy loss); for regression tasks, we used a sigmoid output with a quantile loss. Let $D = (x_i, y_i)_{i=1}^{n}$, and let the model output predicted quantiles $\hat{q}_\theta(\tau|x)$ for quantile levels τ. In our case, the model output 5 estimated quantiles (quantile loss with 5 quantiles during training), so $\tau \in \mathcal{T} := \{\tau_1, \tau_2, \tau_3, \tau_4, \tau_5\} = \{0.0228, 0.1587, 0.5, 0.8413, 0.9772\}$. The quantile loss, denoted $\mathcal{L}_{\text{QL}}$, was the averaged error over the all samples[1] and over the pinball losses (one for each predicted quantile):

$$\mathcal{L}_{\text{QL}}(\theta) = \frac{1}{\text{card}(\mathcal{T})n} \sum_{i=1}^{n} \sum_{\tau \in \mathcal{T}} \mathcal{L}_{\text{pinball}}(y_i, \hat{q}_\theta(\tau \mid x_i), \tau)\,.$$

with

$$\mathcal{L}_{\text{pinball}}(y, \hat{q}, \tau) := \begin{cases} \tau(y - \hat{q}), & (y - \hat{q}) \geq 0, \\ (\tau - 1)(y - \hat{q}), & (y - \hat{q}) < 0, \end{cases}$$

Because the benchmark required a large number of training runs, we limited the number of epochs per run. The training parameters used as a default to generate the results were: epochs (500), batch size (512), optimiser (SGD), momentum (0.9), weight decay (0.0001), learning rate (0.001), scheduler (CosineAnnealingLR ($\eta_{min} = 10^{-6}$, $T_{max} =$ epochs)).

2.4 Inference results

Tables 2 and 4 report the inference results for the benchmark datasets depicted in Table 1 for the regression and classification tasks, where the best results are shaded. The reported performance metrics are the mean

[1] Samples from the validation set during the training phase of the post-hoc calibration technique and samples from the test set to assess the performance at inference time

Table 2: Inference results for all regression tasks. Reported regression metrics are the mean absolute error (MAE) and the root mean square error (RMSE). For each task, the best model is highlighted using this shading.

					MAE / RMSE (test set)		
Task	Dataset	#train	#val	#test	F	I	R
BP	**AuroraBP (dbp)**	18,542	2,672	2,400	11.42/17.70	11.17/14.65	11.17/14.67
BP	**AuroraBP (sbp)**	18,542	2,672	2,400	16.95/28.11	16.50/22.26	17.35/23.76
VA	**AuroraBP**	9,052	1,217	1,150	8.51/10.04	8.53/10.07	8.54/10.08
RR	**MIMIC-PERform**	4,551	666	647	4.80/6.43	4.80/6.42	4.83/6.48
RR	**OSASUD**	8,987	1,641	1,724	4.91/5.75	3.62/4.30	5.26/6.17

F: Features

I: Images

R: Raw data

Table 3: Bootstrap results for regression tasks (BP: blood pressure estimation; VA: vascular age estimation; RR: respiratory rate estimation). For each configuration, the reported metric is the mean absolute error (MAE) and 50,000 bootstrap replicates have been used to report the statistics σ and 95 % confidence interval (CI).

				MAE statistics (bootstrap over test points)		
Task	Dataset	Input	Scenario	**MAE (μ)**	**MAE (σ)**	**MAE (95 % CI)**
BP	**AuroraBP**	F	dbp	11.4188	0.2716	[10.9259, 11.9928]
-	-	I	dbp	11.1651	0.1943	[10.7896, 11.5515]
-	-	R	dbp	11.1683	0.1945	[10.7906, 11.5512]
BP	**AuroraBP**	F	sbp	16.9473	0.4553	[16.1295, 17.9177]
-	-	I	sbp	16.4990	0.3036	[15.9135, 17.0885]
-	-	R	sbp	17.3491	0.3287	[16.7022, 17.9959]
VA	**AuroraBP**	F		8.5135	0.1555	[8.2067, 8.8130]
-	-	I		8.5334	0.1570	[8.2264, 8.8400]
-	-	R		8.5416	0.1595	[8.2279, 8.8542]
RR	**MIMIC-PERform**	F		4.7997	0.1670	[4.4743, 5.1310]
-	-	I		4.8003	0.1682	[4.4760, 5.1310]
-	-	R		4.8287	0.1705	[4.4966, 5.1660]
RR	**OSASUD**	F		4.9103	0.0716	[4.7677, 5.0493]
-	-	I		3.6200	0.0568	[3.5072, 3.7308]
-	-	R		5.2646	0.0788	[5.1108, 5.4185]

F: Features

I: Images

R: Raw data

Table 4: Inference results for all classification tasks. The reported classification metrics are subset accuracy (sub_ACC), micro and macro F1 scores (micro_F1, macro_F1), Hamming loss (Hamming_L) and average precision (AP). For each task, the best model is highlighted using this shading.

						CLASSIFICATION METRICS				
Task	Dataset	Input	#train	#val	#test	**sub_ACC**	**micro_F_1**	**macro_F_1**	**Hamming_L**	**AP**
AED	**OSASUD**	F	8,987	1,641	1,724	0.82	0.82	0.65	0.18	0.35
-	-	I	-	-	-	0.82	0.82	0.69	0.18	0.46
-	-	R	-	-	-	0.81	0.81	0.61	0.19	0.32
BPL	**AuroraBP**	F	18,542	2,672	2,400	0.65	0.65	0.64	0.35	0.51
-	-	I	-	-	-	0.67	0.67	0.66	0.33	0.55
-	-	R	-	-	-	0.67	0.67	0.66	0.33	0.54

F: Features
I: Images
R: Raw data

absolute error (MAE) and the root mean square error (RMSE) for regression tasks whereas for classification tasks, we rely on a set of commonly used metrics, namely: subset accuracy (sub_ACC), micro and macro F1 scores (micro_F1, macro_F1), Hamming loss (Hamming_L) and average precision (AP). Table 3 summarises the bootstrap-over-test-points results for the regression tasks and reports MAE statistics, including the bootstrap standard deviation σ and 95 % confidence intervals (CI), which quantify variability across resamples and help assess whether the differences observed between models are statistically meaningful.

Analysis of inference results for regression tasks Overall, the image-based model I provided the best test-set accuracy across most tasks, yielding the lowest or tied-lowest MAE/RMSE in four out of five settings. For blood-pressure estimation, it improved over the feature-based model F on both diastolic BP (I : 11.17/14.65; F :11.42/17.70) and systolic BP (I : 16.50/22.26; F : 16.95/28.11), and it was essentially tied with the raw-signal based model R for diastolic BP while still slightly better in RMSE. On respiration-rate estimation, the image-based approach matched the best MAE on MIMIC-PERform with a value of 4.80 and marginally reduced RMSE (I : 6.42; F : 6.43), while it delivered a substantial gain on OSASUD (I : 3.62/4.30; F : 4.91/5.75; R : 5.26/6.17), indicating a clear advantage in this more challenging setting. The only case where images did not lead was the vascular age regression task using AuroraBP dataset, where the feature-based model was slightly better (F : 8.51/10.04; I 8.53/10.07; R : 8.54/10.08), though the differences were small, suggesting broadly similar performance across representations for this target.

Analysis of inference results for classification tasks Across classification tasks, the image-based model I was the most consistent top performer, especially on the more imbalanced OSASUD AED setting. On AED, the image representation yielded the best subset accuracy and micro-F1 (I :0.819; F : 0.816; R : 0.805), while it also produced a markedly higher macro-F1 (I : 0.692; F : 0.645; R : 0.614) and average precision (I : 0.457; F : 0.347; R : 0.324). In parallel, it achieved the lowest Hamming loss (I : 0.181; F : 0.184; R : 0.195), indicating fewer label-wise errors and better minority-class sensitivity—consistent with the strong gains in macro-F1 and AP. On AuroraBP BPL, the three representations were closer: R attained the best subset accuracy and micro-F1 by a negligible margin (0.66694 vs. 0.66681 for I), whereas I provided the highest macro-F1 (0.658) and AP (0.546) with essentially tied Hamming loss ($\approx$0.333). Overall, I tended to improve class-balanced and ranking-oriented metrics (macro-F1, AP) and reduced label-wise mistakes (Hamming loss), while R remained competitive on strict exact-match accuracy in the

easier/less skewed setting.

Analysis of bootstrap results Table 3 reports the bootstrapping over test points results for each of the regression tasks that consisted in resampling indices with replacement and recomputing MAE. It was a robust approach to estimating the population mean absolute error over the input distribution, in other words, to assess the sampling variability due to a finite test set (only n test points). The bootstrap distribution of MAE gave a confidence interval (CI). This reported standard deviation, denoted MAE (σ), refers to the variability that should be observed if another test set was drawn from the same distribution by using the empirical error distribution. Across most tasks (AuroraBP dbp/sbp, VA AuroraBP, RR MIMIC-PERform), the three input modalities yielded very similar MAE estimates under bootstrap, with overlapping 95 % CI and small effect sizes. The main exception was for respiratory rate estimation using the OSASUD dataset where I provided a large and statistically clear reduction in MAE (non-overlapping CIs and a 26–31 % decrease in mean MAE compared to F and R).

Considering the bootstrapped mean absolute error (MAE) values, denoted MAE (μ) in the table, the relative difficulty of the regression tasks can be inferred directly from the magnitude of the reported MAE. Overall, blood pressure estimation using the AuroraBP dataset appeared to be the most challenging setting, especially the systolic case with mean MAE values in the range $[16.50, 17.35]$ across the three models (F , I , R) and mean MAE values in the range $[11.17, 11.42]$ for the diastolic case (substantially lower than sbp but still clearly higher than the remaining tasks). In contrast, vascular age regression on AuroraBP exhibited intermediate difficulty, with MAE tightly clustered in the range $[8.51, 8.54]$, suggesting a more stable prediction problem and limited sensitivity to the model choice. The lowest MAEs were observed for the respiratory rate tasks even if on OSASUD the reported MAE values suggested an effect of the selected model with a wider range (I : $\mu = 3.6200$, F : $\mu = 4.9103$, R : $\mu = 5.2646$).

The results also allow us to assess the relative sampling variability using the coefficient of variation, CV, where $\text{CV} = 100 \times \frac{\sigma}{\mu}$. Overall, the CV values were small, in the range [1.46 %, 3.53 %], indicating stable estimates across bootstrap resamples.

- For BP on AuroraBP (dbp), relative variability was moderate and decreased when moving from F (2.38 %) to I (1.74 %) and R (1.74 %).
- For BP on AuroraBP (sbp), F gsve the largest variability among the three modalities with a coefficient of variation of 2.69 %, while I and R remain below 2 %.
- For VA on AuroraBP, the three models showed nearly identical relative variability, consistent with the very similar mean MAEs observed for this task.
- The highest relative sampling variability was observed for RR on MIMIC-PERform, where all models reported a CV close to 3.5 %, despite having low absolute MAE. This reflected a larger σ relative to the small mean error level. Finally, RR on OSASUD showed the lowest relative variability overall across all models (CV < 1.6 %), suggesting particularly stable MAE estimates under bootstrap for this dataset.

2.5 Posthoc calibration results

For regression tasks, we studied five post-hoc calibration methods: GP-β, GP-Normal, g-Layers, Isotonic Regression (IR) and Variance Scaling (VS). For classification tasks, we investigated six post-hoc calibration methods: Histogram Binning (HB), Bayesian Binning (BB), Logistic Calibration (LC), β-Calibration (β-C), Isotonic Regression (IR) and Temperature Scaling (TS). Table 5 reports the investigated methods.

We evaluate regression performance using the Negative Log-Likelihood (NLL), since the calibration methods are tuned on a validation set to improve NLL. For a Gaussian predictive distribution, $p_\theta(y|x)$, with $p_\theta(y|x) =$

Table 5: Summary of the post-hoc calibration methods considered, indicating their method type (scaling, binning, Gaussian Process, hybrid or neural network) and the downstream tasks on which they were evaluated.

Methods	**Type**	**ML task**
Temperature Scaling (TS) [8]	scaling	classification
Variance Scaling (VS) [9]	scaling	regression
Logistic Calibration "Platt Scaling" (LC) [10]	scaling	classification
Beta Calibration (β-C) [11]	scaling	classification
Isotonic Regression (IR) [12]	binning	regression & classification
Histogram Binning (HB) [13]	binning	classification
Bayesian Binning (BB) [14]	binning	classification
Gaussian Process (GP-Normal) [15]	GP	regression & classification
g-Layers [16]	neural network (MLP)	regression
Beta calibration + GP (GP-β-C) [17]	hybrid (scaling + GP)	regression & classification

$\mathcal{N}(y; \mu_\theta(x), \sigma^2_\theta(x)))$, the NLL is defined as:

$$\text{NLL}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left[\log\big(2\pi\sigma^2_\theta(x_i)\big) + \frac{(y_i - \mu_\theta(x_i))^2}{\sigma^2_\theta(x_i)} \right].$$

For classification tasks, we report the Expected Calibration Error (ECE) as the primary calibration metric. Tables 6 and 7 summarise the post-hoc calibration results for the regression and classification tasks, respectively.

Analysis of post-hoc calibration results for regression tasks Table 6 summarises the post-hoc calibration results for regression tasks using the Negative Log-Likelihood (NLL), where lower values indicate better-calibrated predictive distributions.

- On AuroraBP for the blood pressure estimation task, the best NLL (diastolic case) was obtained by g-Layers for F and R ($\downarrow$ 2.74 % and $\downarrow$ 0.98 %, respectively), while IR performed best for I ($\downarrow$2.39 %).
- For the systolic case on AuroraBP, GP-β-C was consistently better across F, I, and R, yielding NLL reductions of 1.16 %, 1.70 %, and 2.61 %, respectively.
- For the vascular age estimation task (VA) on AuroraBP, g-Layers achieved the lowest NLL, improving NLL by 3.11 % (F), 1.36 % (I), and 2.15 % (R).
- For respiratory rate (RR) estimation on MIMIC-PERform, GP-β-C provided the lowest NLL, with modest gains of 1.32 % to 1.60 %. The largest improvements were observed for RR estimation on OSASUD, where g-Layers substantially outperformed the uncalibrated model, reducing NLL by 30.71 % (F), 22.79 % (I), and 29.59 % (R). In contrast, GP-Normal and Variance Scaling (VS) were generally non-competitive and NLL markedly degraded on OSASUD.

Overall, the best-performing technique varied across tasks, but it consistently improved upon the uncalibrated baseline in nearly all settings.

Table 6: Posthoc-calibration results for all datasets (regression tasks). Reported post-hoc calibration techniques are: GP-β, GP-Normal, g-Layers, Isotonic Regression (IR) and Variance Scaling (VS). Reported performance metric is the Negative Log-Likelihood (NLL), indeed the post-hoc calibration techniques aim to optimise the NLL using the calibration set. The lowest NLL value for each setting is highlighted using this shading.

					CALIBRATION TECHNIQUES					
Task	Dataset	Scenario	Inputs	Metric	**GP-β-C**	**GP-Normal**	**g-Layers**	**IR**	**VS**	**Uncalibrated**
BP	**AuroraBP**	**dbp**	F	NLL	4.1106	4.1243	4.0003	4.2671	4.1245	4.1132
-	-	-	I	NLL	4.1080	4.1172	4.1338	4.0187	4.1163	4.1172
-	-	-	R	NLL	4.0986	4.1154	4.0649	4.2518	4.1146	4.1050
BP	**AuroraBP**	**sbp**	F	NLL	4.4823	4.5379	4.5687	4.5479	4.5383	4.5348
-	-	-	I	NLL	4.4746	4.5280	4.5952	4.6064	4.5282	4.5518
-	-	-	R	NLL	4.4709	4.5935	4.4999	4.6472	4.5936	4.5905
VA	**AuroraBP**	-	F	NLL	3.7448	3.7318	3.6245	4.1610	3.7333	3.7409
-	-	-	I	NLL	3.7424	3.7338	3.6854	4.3892	3.7351	3.7364
-	-	-	R	NLL	3.7429	3.7345	3.6639	4.3363	3.7348	3.7444
RR	**MIMIC-PERform**	-	F	NLL	3.2316	3.2812	3.3221	3.5841	3.2812	3.2748
-	-	-	I	NLL	3.2355	3.2834	3.3752	3.5847	3.2832	3.2869
-	-	-	R	NLL	3.2363	3.2946	3.3700	3.3947	3.2949	3.2889
RR	**OSASUD**	-	F	NLL	4.1693	5.6111	2.3401	4.4602	5.6825	3.3774
-	-	-	I	NLL	3.9495	3.2492	2.3046	4.1034	3.2566	2.9847
-	-	-	R	NLL	3.9015	5.3040	2.9840	4.1287	5.3361	4.2378

F: Features
I: Images
R: Raw data

Table 7: Posthoc-calibration results for all datasets (classification tasks). Reported post-hoc calibration techniques are: Histogram Binning (HB), Bayesian Binning (BB), Logistic Calibration (LC), β-Calibration (β-C), Isotonic Regression (IR) and Temperature Scaling (TS). Reported performance metric is the Expected Calibration Error (ECE). The lowest ECE value for each setting is highlighted using this shading.

					CALIBRATION TECHNIQUES						
Task	Dataset	Scenario	Inputs	Metric	**Uncalibrated**	**TS**	**IR**	**HB**	**BB**	**LC**	**β-C**
BPL	**AuroraBP**	-	F	ECE	0.2949	0.0796	0.0503	0.0504	0.0497	0.0796	0.2365
-	-	-	I	ECE	0.3123	0.0479	0.0502	0.0510	0.0474	0.0479	0.2601
-	-	-	R	ECE	0.1253	0.1313	0.0587	0.0518	0.0499	0.1313	0.0796
AED	**OSASUD**	-	F	ECE	0.3560	0.4829	0.0648	0.0643	0.0416	0.4829	0.5743
-	-	-	I	ECE	0.4877	0.5064	0.0286	0.1112	0.1211	0.5064	0.6678
-	-	-	R	ECE	0.3357	0.5053	0.0210	0.0433	0.0259	0.5053	0.5966

F: Features
I: Images
R: Raw data

Analysis of post-hoc calibration results for classification tasks Table 7 shows that across both datasets and all three input settings (F , I , R), post-hoc calibration yielded substantial improvements in calibration quality, as measured by the Expected Calibration Error (ECE).

- For blood pressure level classification (BPL) on AuroraBP, the best-performing method in each setting was Bayesian Binning (BB), reducing ECE from 0.2949 to 0.0497 for F (↓ 83.15 %), from 0.3123 to 0.0474 for I (↓ 84.83 %), and from 0.1253 to 0.0499 for R (↓ 60.17 %).
- For the apnea event detection task (AED) on OSASUD, the strongest improvements were obtained with BB for F and Isotonic Regression (IR) for I and R : ECE decreased from 0.3560 to 0.0416 for F (↓ 88.31 %), from 0.4877 to 0.0286 for I (↓ 94.14 %), and from 0.3357 to 0.0210 for R (↓ 93.74 %).

Overall, these results indicated that BB and IR consistently provided large reductions in ECE (approximately 60 %–94 %), whereas Temperature Scaling and Logistic Calibration did not improve ECE in these experiments and frequently degraded calibration, particularly on OSASUD.

2.6 Evaluation of uncertainties

In this section, we present the results of running the uncertainty evaluation framework over different models, datasets, UQ approaches, and calibration methods. This is considered across eight tasks: the five regression and the two classification tasks, detailed in Table 1, and the sex classification task described in Section 3.2. These were evaluated across the post-hoc calibration methods mentioned in Table 5 as well as conformal predictions. In the case of sex classification, probabilities from uncalibrated Monte Carlo Dropout (MCD) are evaluated.

Tables 8 and 9 present results for regression calibration and conformal prediction methods, respectively. Similarly, Tables 10 and 11 show results for classification calibration and conformal predictions, respectively. Aside from UQ evaluation metrics, performance metrics are also provided in each table, so that informative comparisons can be drawn for the various calibration techniques. All of the metrics have been calculated on holdout test sets. Finally, in each column of the tables, the best values are highlighted for each task.

2.6.1 UQ evaluation in regression tasks

For regression, the following metrics were calculated:

- ENCE: Expected Normalised Calibration Error;
- CRPS: Continuous Ranked Probability Score;
- PICP: Prediction Interval Coverage Probability for a specified confidence interval;
- MPIW: Mean Predicted Interval Width for a specified confidence interval (conformal only);
- CCE: Coverage Calibration Error;
- NLL: Gaussian Negative Log Likelihood;
- MAE: Mean Absolute Error;
- RMSE: Root Mean Squared Error.

For the conformal method, where predictions are provided as intervals, appropriate conversions were applied to produce equivalent means and standard deviations to be used with metrics that do not directly operate on intervals.

Post-hoc calibration was performed based on NLL. Results in Table 8 show similar test set NLL values to those in the calibration set (Table 6), except for OSASUD respiratory rate regression, which demonstrates a slight increase on the test set. This indicates the generalisability of the performance/calibration according to NLL for in-distribution data.

AuroraBP blood pressure regression (DBP and SBP) Image-based models have the lowest MAE and CRPS for both DBP and SBP, while raw inputs provide the most optimal PICP 2σ, ENCE, and NLL. Image-based inputs provide optimal accuracy and scoring, while raw inputs provide optimal calibration and MAE. The intended use-case of the model determines which of these inputs are most suitable.

AuroraBP vascular age regression Contrary to the blood pressure regression on the same dataset, models that are trained on features presented the best CRPS, NLL, MAE and RMSE, while raw inputs had the best ENCE and PICP 1σ.

If we consider the conformal predictions, raw data appeared to produce prediction intervals that have the best metrics.

MIMIC-PERform respiratory rate regression Here, feature-based models provided the best MAE and RMSE, and CRPS and NLL. Image-based approaches had the best ENCE. Again, the optimal input depends on requirements of operational use of the model. However, conformal predictions presented few consistent trends across the input types.

OSASUD respiratory rate regression Based on both calibration and conformal results, it is very clear that with OSASUD data, image inputs outperformed others with GP-Normal as the best calibration method.

Results summary

Overall, when our results are compared to Table 2, it is evident that in both calibration and test sets, image-based models tended to have the lowest MAE and RMSE as well as better calibration, except for AuroraBP vascular age regression, which favoured features and raw data. For calibration quality, Table 6 suggests g-Layers and GP-β as the best methods as measured by NLL over the calibration set. We could not arrive at the same conclusion based on our results on the test set, even if we only considered NLL. The optimal inputs and UQ method depend on the requirements for operational use post deployment; these results provide some guidance as to which inputs and UQ methods produce given variations in predictive performance and calibration. If calibration error and nominal coverage are desired, raw inputs are strongest on BP and vascular age. If you lean on proper scoring rules (CRPS/NLL), image-based inputs are often more preferable.

2.6.2 UQ evaluation in classification tasks

For classification, the following metrics were calculated:

- ECE: Expected Calibration Error;
- ACE: Adaptive Calibration Error;
- smECE: smooth Expected Calibration Error;
- UCE: Uncertainty Calibration Error;
- VCE: Variation Calibration Error;
- NLL: Negative Log Likelihood;

- Coverage: prediction set coverage;
- AUC: Area Under the ROC Curve;
- F1 score;
- MCC sensitivity: Matthews Correlation Coefficient between targets and predictions, when predicted probability binarisation threshold is set to a value that results in a sensitivity closest to 0.8;
- MCC specificity: Matthews Correlation Coefficient between targets and predictions, when predicted probability binarisation threshold is set to a value that results in a specificity closest to 0.8;
- Sensitivity;
- Specificity.

For the conformal method, where predictions were provided as intervals, appropriate conversions were applied to produce equivalent probabilities to be used with metrics that do not directly operate on intervals.

While the AuroraBP blood pressure and OSASUD apnea multi-label classifications are multi-class tasks, here a one-vs-all approach was adopted and the problems were deemed as binary classification for each class. This is partly for compatibility with calibration methods such as Venn–Abers that are inherently binary.

We used the ECE as the main calibration quality metric. By comparing Tables 7 and 10, we noted that ECE values were different across the calibration and test sets. This discrepancy mainly arises from the fact that the UQ evaluation framework was run with the one-vs-all approach and binary classification, rather than multi-class, as opposed to distributional mismatch between the two datasets. That said, both tables still show similar relative values and trends in ECE across the different methods/inputs.

AuroraBP blood pressure classification Feature- and image-based models appeared to be better and comparable to calibrated classes on the test set, with comparable predictive performance. Conformal prediction favoured feature-based models in both categories. Conformal results (prediction intervals) were converted to probabilities before binary accuracy metrics were calculated. For both AuroraBP blood pressure classification and OSASUD apnea multi-label classification, this assigned virtually all data to class 0, resulting in unreliable accuracy metrics. For this reason, we ignored the accuracy metrics for these two tasks. This also suggests that the calibration metrics considered encode 100% inaccuracy for the positive class, which is not apparent from the global metrics used here. This motivates the adaptive-per class evaluation of calibration. Here, feature-based models produced the best calibration, while image-features gave the best AUC.

OSASUD apnea multi-label classification Feature- and image-based results showed better calibration, while the image-based models produced higher accuracy metrics. On the contrary, conformal predictions resulted in better calibration metrics for raw data.

Sex classification Since MCD results were not calibrated, there is no point of reference for comparison. However, results seemed to be generally in the lower ranges of calibration and performance errors, compared to other tasks studied in this section.

Overall when considered along with Table 4, our results indicated better performance and calibration for image-based models. However, we also observe similarly good results with feature-based models. Like regression, no specific conclusions can be drawn as to what calibration method performs best. As for regression, the optimal input type and UQ method depends on what is required in operational settings, as there is considerable variation in uncertainty reliability and predictive performance across these variables.

Table 8: UQ evaluation for regression calibration methods. The best metric values are highlighted per task.

Task	Input	Method	ENCE	CRPS	PICP 1σ	PICP 2σ	CCE	NLL	MAE	RMSE
AuroraBP blood pressure regression (DBP)	Feature	g-Layers	0.2427	8.2537	0.6676	0.9311	0.0008	4.1326	11.4164	17.6303
		GP-β	0.2158	8.2494	0.6867	0.9349	0.0009	4.1235	11.4154	17.6378
		GP-Normal	0.2109	8.2470	0.6705	0.9378	0.0006	4.1235	11.4174	17.6225
		IR	0.2047	8.2488	0.6871	0.9365	0.0010	4.1223	11.4156	17.6372
		VS	0.2116	8.2470	0.6705	0.9378	0.0006	4.1237	11.4174	17.6225
	Image	g-Layers	0.0910	8.1282	0.7697	0.9622	0.0173	4.1178	11.2407	14.7691
		GP-β	0.1227	8.0386	0.6739	0.9349	0.0009	4.1176	11.1619	14.6573
		GP-Normal	0.1210	8.0412	0.6768	0.9353	0.0018	4.1173	11.1623	14.6570
		IR	0.1153	8.0371	0.6776	0.9361	0.0008	4.1149	11.1621	14.6575
		VS	0.1176	8.0407	0.6813	0.9357	0.0018	4.1165	11.1623	14.6570
	Raw data	g-Layers	0.0539	8.0753	0.7402	0.9531	0.0072	4.1088	11.2107	14.7147
		GP-β	0.1170	8.0508	0.6846	0.9349	0.0019	4.1175	11.1793	14.6717
		GP-Normal	0.1057	8.0483	0.6834	0.9357	0.0006	4.1179	11.1933	14.6907
		IR	0.1036	8.0478	0.6859	0.9357	0.0009	4.1162	11.1814	14.6720
		VS	0.1023	8.0480	0.6842	0.9357	0.0007	4.1171	11.1933	14.6907
AuroraBP blood pressure regression (SBP)	Feature	g-Layers	0.3000	13.3879	0.8710	0.9772	0.0984	4.6433	17.3114	28.1863
		GP-β	0.1845	12.4375	0.7091	0.9382	0.0044	4.5343	16.9503	27.9131
		GP-Normal	0.1798	12.4396	0.7108	0.9390	0.0032	4.5360	16.9262	27.8908
		IR	0.1813	12.4389	0.7083	0.9398	0.0044	4.5335	16.9577	27.9192
		VS	0.1815	12.4391	0.7100	0.9382	0.0031	4.5364	16.9262	27.8908
	Image	g-Layers	0.2899	13.3432	0.8805	0.9793	0.1257	4.6672	17.1057	23.4553
		GP-β	0.0892	12.0735	0.7075	0.9386	0.0041	4.5254	16.5422	22.2200
		GP-Normal	0.1097	12.0748	0.7054	0.9386	0.0029	4.5286	16.5068	22.2738
		IR	0.1013	12.0729	0.7066	0.9390	0.0043	4.5256	16.5350	22.2267
		VS	0.1104	12.0745	0.7050	0.9382	0.0028	4.5288	16.5068	22.2738
	Raw data	g-Layers	0.1906	12.8343	0.8373	0.9676	0.0852	4.6012	16.9367	23.2171
		GP-β	0.0706	12.0977	0.7083	0.9382	0.0045	4.5258	16.5886	22.2644
		GP-Normal	0.0937	12.7637	0.7216	0.9311	0.1017	4.5935	17.3491	23.7629
		IR	0.0831	12.0959	0.7075	0.9390	0.0044	4.5261	16.5768	22.2728
		VS	0.0941	12.7636	0.7216	0.9311	0.1019	4.5936	17.3491	23.7629
AuroraBP vascular age regression	Feature	g-Layers	0.1685	5.8049	0.7496	0.9870	0.0261	3.7580	8.5265	10.0688
		GP-β	0.0792	5.8781	0.5496	0.9696	0.0716	3.7344	8.6569	10.0685
		GP-Normal	0.0848	5.8043	0.6261	0.9704	0.0319	3.7323	8.5292	10.0431
		IR	0.0960	5.8795	0.5409	0.9696	0.0716	3.7366	8.6384	10.0586
		VS	0.0918	5.8081	0.6148	0.9530	0.0325	3.7339	8.5292	10.0431
	Image	g-Layers	0.1581	5.8232	0.7374	0.9861	0.0260	3.7603	8.5288	10.1256
		GP-β	0.0806	5.8803	0.5487	0.9696	0.0742	3.7350	8.6556	10.0712
		GP-Normal	0.1000	5.8091	0.6809	0.9643	0.0338	3.7344	8.5319	10.0715
		IR	0.0936	5.8830	0.5391	0.9696	0.0737	3.7369	8.6447	10.0655
		VS	0.1054	5.8122	0.6765	0.9574	0.0367	3.7357	8.5319	10.0715
	Raw data	g-Layers	0.2870	6.1243	0.7974	1.0000	0.0980	3.8594	8.5825	10.4696
		GP-β	0.0772	5.8784	0.5539	0.9696	0.0718	3.7346	8.6495	10.0735
		GP-Normal	0.0756	5.8102	0.6817	0.9565	0.0300	3.7351	8.5288	10.0794
		IR	0.0901	5.8839	0.5522	0.9696	0.0716	3.7367	8.6463	10.0718
		VS	0.0779	5.8111	0.6817	0.9557	0.0303	3.7355	8.5288	10.0794
MIMIC-PERform respiratory rate regression	Feature	g-Layers	0.1818	3.6900	0.7697	0.9691	0.1003	3.3433	5.0148	6.7286
		GP-β	0.1438	3.5446	0.7635	0.9675	0.0472	3.2775	4.9723	6.4014
		GP-Normal	0.1430	3.5092	0.7991	0.9675	0.0171	3.2825	4.7954	6.4177
		IR	0.1197	3.5330	0.7558	0.9660	0.0411	3.2740	4.9544	6.3929
		VS	0.1429	3.5091	0.7991	0.9675	0.0171	3.2824	4.7954	6.4177
	Image	g-Layers	0.1610	3.7721	0.7759	0.9660	0.1278	3.3636	5.1614	6.8886
		GP-β	0.0933	3.5503	0.7465	0.9675	0.0455	3.2798	4.9857	6.4091
		GP-Normal	0.1585	3.5119	0.7991	0.9660	0.0171	3.2840	4.8042	6.4172
		IR	0.1447	3.5394	0.7419	0.9660	0.0413	3.2769	4.9670	6.3999
		VS	0.1587	3.5114	0.7991	0.9660	0.0171	3.2838	4.8042	6.4172
	Raw data	g-Layers	0.1859	3.6965	0.8377	0.9722	0.0956	3.3523	4.9757	6.6800
		GP-β	0.0992	3.5555	0.7481	0.9675	0.0490	3.2803	4.9963	6.4134
		GP-Normal	0.1289	3.5348	0.7713	0.9675	0.0242	3.2919	4.8274	6.4668
		IR	0.0997	3.5433	0.7372	0.9675	0.0439	3.2770	4.9743	6.4027
		VS	0.1297	3.5356	0.7759	0.9675	0.0245	3.2922	4.8274	6.4668
OSASUD respiratory rate regression	Feature	g-Layers	1.9206	4.0907	0.2235	0.4061	1.5352	5.9199	4.9806	5.8182
		GP-β	1.8671	3.7385	0.2551	0.4316	1.2406	5.7268	4.5473	5.4141
		GP-Normal	1.8086	4.0063	0.2420	0.4245	1.4807	5.6111	4.9103	5.7478
		IR	1.8267	4.0189	0.2366	0.4180	1.4950	5.6689	4.9186	5.7562
		VS	1.8345	4.0136	0.2366	0.4168	1.4895	5.6825	4.9103	5.7478
	Image	g-Layers	1.2902	3.6543	0.3288	0.5054	1.0939	4.4362	4.6152	5.5111
		GP-β	1.6204	3.6057	0.3002	0.4673	1.0559	5.0817	4.4454	5.3274
		GP-Normal	0.6522	2.6727	0.3995	0.6736	0.4207	3.2492	3.6200	4.3053
		IR	1.6296	3.6213	0.2967	0.4679	1.0660	5.1071	4.4620	5.3457
		VS	0.6592	2.6748	0.3989	0.6700	0.4208	3.2566	3.6200	4.3053
	Raw data	g-Layers	1.7685	3.7011	0.2800	0.4530	1.1193	5.4744	4.5177	5.4093
		GP-β	1.6124	3.5802	0.3115	0.4721	1.0396	5.0440	4.4165	5.2977
		GP-Normal	1.6433	4.2545	0.2562	0.4554	1.5006	5.3040	5.2646	6.1738
		IR	1.6079	3.6199	0.3056	0.4685	1.0626	5.0859	4.4616	5.3478
		VS	1.6572	4.2584	0.2539	0.4530	1.5033	5.3361	5.2646	6.1738

Table 9: UQ evaluation for regression conformal predictions. The best metric values are highlighted for each task (except for MPIW, which depends on PICP).

Task	Input	Confidence	ENCE	CRPS	PICP	MPIW	CCE	NLL	MAE	RMSE
AuroraBP blood pressure regression (DBP)	Feature	0.6826	0.2226	8.2676	0.6697	26.2307	0.0020	4.1279	11.4200	17.8346
		0.9544	0.1798	8.3513	0.9369	54.8907	0.0269	4.1264	11.5334	17.7999
	Image	0.6826	0.1297	8.0388	0.6672	25.8874	0.0015	4.1205	11.1625	14.6528
		0.9544	0.0899	8.0457	0.9423	54.6626	0.0041	4.1088	11.1728	14.6595
	Raw data	0.6826	0.7948	557.1321	0.6888	798.7826	0.6710	7.3437	679.9094	6801.5005
		0.9544	0.8186	587.9203	0.9656	2296.3887	0.5287	7.5111	710.3525	7089.8480
AuroraBP blood pressure regression (SBP)	Feature	0.6826	0.2240	12.5015	0.6851	40.6467	0.0109	4.5387	17.1213	28.3945
		0.9544	0.2378	13.2397	0.9324	82.7076	0.1371	4.5771	18.5272	29.5390
	Image	0.6826	0.1019	12.0957	0.6867	40.7313	0.0106	4.5257	16.6410	22.1781
		0.9544	0.1191	12.5190	0.9336	80.4524	0.0859	4.5496	17.5103	22.5666
	Raw data	0.6826	0.1562	12.5156	0.6871	40.2027	0.0636	4.5918	17.0069	23.3114
		0.9544	0.1151	12.1483	0.9344	80.0384	0.0182	4.5314	16.7554	22.2388
AuroraBP vascular age regression	Feature	0.6826	0.0561	5.8497	0.5513	19.0772	0.0580	3.7292	8.6133	10.0472
		0.9544	0.2702	6.3063	0.9348	33.0179	0.2225	3.8388	9.1672	10.4993
	Image	0.6826	0.0839	5.8737	0.5530	18.6303	0.0723	3.7338	8.6468	10.0618
		0.9544	0.1647	6.0735	0.9696	35.1870	0.1403	3.7737	8.9330	10.2637
	Raw data	0.6826	0.0812	5.8287	0.6009	19.4597	0.0473	3.7259	8.6120	10.0393
		0.9544	0.2649	6.3075	0.9348	33.1203	0.2216	3.8372	9.1737	10.5094
MIMIC-PERform respiratory rate regression	Feature	0.6826	0.1673	3.4939	0.7728	12.9883	0.0156	3.2652	4.8446	6.3605
		0.9544	0.1328	3.8951	0.9351	25.1714	0.2200	3.3530	5.5942	6.8863
	Image	0.6826	0.1144	3.4948	0.7805	12.9469	0.0173	3.2694	4.8458	6.3623
		0.9544	0.1258	3.8012	0.9490	25.4803	0.1754	3.3332	5.4543	6.7598
	Raw data	0.6826	0.1272	3.4942	0.7666	12.9364	0.0098	3.2724	4.8257	6.3764
		0.9544	0.1968	4.0047	0.9289	24.6817	0.2877	3.3899	5.7160	7.0437
OSASUD respiratory rate regression	Feature	0.6826	2.0156	3.8553	0.2325	3.6160	1.3887	6.1060	4.6668	5.5019
		0.9544	0.9768	2.8612	0.5773	8.8466	0.6993	3.6970	3.7460	4.4726
	Image	0.6826	0.3955	2.5229	0.4465	5.9895	0.3393	2.9960	3.5346	4.1858
		0.9544	0.2748	2.2320	0.9376	11.7694	0.1202	2.8140	3.2596	3.7570
	Raw data	0.6826	2.3752	4.3299	0.2188	3.5609	1.5617	7.2767	5.1432	6.0543
		0.9544	1.2877	3.0933	0.5190	8.1401	0.6762	4.2910	3.9143	4.6960

Table 10: UQ evaluation for classification calibration methods. The best metric values are highlighted for each task.

Task	Input	Method	ECE	ACE	smECE	UCE	VCE	NLL	Cover. 95	Cover. 68	AUC	F1 score	MCC sens.	MCC spec.	Sens.	Spec.
AuroraBP blood pressure classification	Feature	BB	0.0331	0.0445	0.0331	0.1653	0.0449	0.7744	1.0000	0.8786	0.7478	0.6358	0.2945	0.4533	0.6396	0.5033
		β-C	0.1536	0.1542	0.1494	0.1079	0.3864	1.2975	0.9104	0.7789	0.6994	0.5334	0.2493	0.3071	0.4983	0.4554
		HB	0.0330	0.0344	0.0314	0.1617	0.0464	0.7741	1.0000	0.8778	0.7448	0.6337	0.2997	0.4429	0.6471	0.5125
		IR	0.0357	0.0351	0.0327	0.1626	0.0522	0.7846	0.9997	0.8796	0.7544	0.6355	0.3359	0.4422	0.6446	0.5446
		LC	0.0445	0.0584	0.0397	0.0833	0.1022	0.8844	0.9824	0.8890	0.7079	0.4479	0.2777	0.2777	0.4671	0.4875
		TS	0.0445	0.0584	0.0397	0.0833	0.1022	0.8844	0.9824	0.8890	0.7079	0.4479	0.2777	0.2777	0.4671	0.4875
	Image	BB	0.0318	0.0345	0.0316	0.1635	0.0475	0.7711	1.0000	0.8786	0.7536	0.6358	0.3200	0.4443	0.6450	0.5271
		β-C	0.1742	0.1744	0.1743	0.1203	0.4783	1.3479	0.8782	0.7775	0.7380	0.5956	0.2900	0.3944	0.5938	0.5015
		HB	0.0347	0.0329	0.0307	0.1604	0.0466	0.7892	0.9996	0.8799	0.7527	0.6326	0.3133	0.4460	0.6508	0.5279
		IR	0.0340	0.0339	0.0327	0.1603	0.0465	0.7895	0.9996	0.8783	0.7544	0.6347	0.3235	0.4439	0.6488	0.5371
		LC	0.0769	0.0847	0.0700	0.1093	0.1757	0.9028	0.9644	0.8857	0.7418	0.5906	0.3080	0.3932	0.5925	0.5219
		TS	0.0769	0.0847	0.0700	0.1093	0.1757	0.9028	0.9644	0.8857	0.7418	0.5906	0.3080	0.3932	0.5925	0.5219
	Raw data	BB	0.0333	0.0398	0.0333	0.1655	0.0450	0.7757	1.0000	0.8786	0.7408	0.6358	0.2782	0.4488	0.6529	0.4848
		β-C	0.0533	0.0649	0.0464	0.1033	0.1173	0.8552	0.9864	0.8490	0.7323	0.5707	0.2922	0.4097	0.6108	0.5040
		HB	0.0369	0.0354	0.0332	0.1631	0.0503	0.7825	0.9999	0.8774	0.7403	0.6347	0.2872	0.4425	0.6475	0.4983
		IR	0.0366	0.0363	0.0334	0.1638	0.0475	0.7761	1.0000	0.8769	0.7406	0.6355	0.3056	0.4480	0.6500	0.5192
		LC	0.0588	0.0778	0.0464	0.1261	0.0717	0.8295	0.9986	0.9169	0.7303	0.3353	0.3023	0.3876	0.5863	0.5154
		TS	0.0588	0.0778	0.0464	0.1261	0.0717	0.8295	0.9986	0.9169	0.7303	0.3353	0.3023	0.3876	0.5863	0.5154
OSASUD apnea multi-label classification	Feature	BB	0.2317	0.2222	0.2054	0.0677	0.1373	5.4068	0.8901	0.7007	0.3838	0.0926	-0.2455	-0.0594	0.1244	0.2822
		β-C	0.0879	0.1186	0.0837	0.0751	0.2806	0.9813	0.9417	0.8648	0.6258	0.2336	0.0970	0.1640	0.3868	0.3200
		HB	0.2206	0.2224	0.2099	0.0695	0.4934	5.4636	0.8901	0.7166	0.3842	0.0926	-0.2280	-0.0488	0.1445	0.3067
		IR	0.2197	0.2245	0.2331	0.0624	0.0932	5.6989	0.8901	0.7007	0.4573	0.0926	-0.0675	-0.0511	0.1445	0.5492
		LC	0.0751	0.1057	0.0681	0.0708	0.2179	0.8339	0.9570	0.8700	0.6418	0.1012	0.1351	0.1152	0.3294	0.3734
		TS	0.0751	0.1057	0.0681	0.0708	0.2179	0.8339	0.9570	0.8700	0.6418	0.1012	0.1351	0.1152	0.3294	0.3734
	Image	BB	0.2412	0.2222	0.2045	0.0671	0.1344	5.4036	0.8901	0.7007	0.3897	0.0926	-0.2455	0.0115	0.2103	0.2822
		β-C	0.1208	0.1384	0.1193	0.0766	0.3823	1.2875	0.9206	0.8380	0.6225	0.2834	0.0442	0.1988	0.4289	0.2509
		HB	0.2169	0.2231	0.1994	0.0642	0.4876	5.4023	0.8901	0.7255	0.3898	0.0951	-0.2349	-0.0110	0.1878	0.2941
		IR	0.2196	0.2227	0.2226	0.0652	0.1139	5.5508	0.8874	0.7007	0.4309	0.0926	-0.1250	-0.0734	0.1102	0.4460
		LC	0.0866	0.1075	0.0866	0.0480	0.2359	0.8449	0.9555	0.8757	0.6669	0.1131	0.1657	0.2192	0.4538	0.4175
		TS	0.0866	0.1075	0.0866	0.0480	0.2359	0.8449	0.9555	0.8757	0.6669	0.1131	0.1657	0.2192	0.4538	0.4175
	Raw data	BB	0.2296	0.2228	0.2058	0.0679	0.1386	5.4079	0.8901	0.7007	0.3837	0.0926	-0.2455	-0.0705	0.1132	0.2822
		β-C	0.1090	0.1349	0.1069	0.0756	0.3131	1.0423	0.9417	0.8336	0.6259	0.1963	0.1327	0.1244	0.3400	0.3700
		HB	0.2215	0.2227	0.2111	0.0706	0.4952	5.4325	0.8901	0.7073	0.3791	0.0926	-0.2400	-0.0622	0.1286	0.2890
		IR	0.2219	0.2252	0.2463	0.0604	0.0731	5.9049	0.8901	0.7025	0.4748	0.0926	-0.0253	-0.0641	0.1321	0.6367
		LC	0.0885	0.1158	0.0851	0.0655	0.2555	0.8674	0.9560	0.8487	0.6483	0.0878	0.1820	0.1168	0.3312	0.4411
		TS	0.0885	0.1158	0.0851	0.0655	0.2555	0.8674	0.9560	0.8487	0.6483	0.0878	0.1820	0.1168	0.3312	0.4411
Sex classification	Raw data	-	0.0321	0.0364	0.0319	0.1412	0.0407	0.8440	0.9998	0.9116	0.7618	0.7230	0.3876	0.3863	0.5804	0.5761

Table 11: UQ evaluation for classification conformal predictions. The best metric values are highlighted for each task.

Task	Input	ECE	ACE	smECE	UCE	VCE	NLL	Cover. 95	Cover. 68	AUC	F1 score	MCC sens.	MCC spec.	Sens.	Spec.
AuroraBP blood pressure classification	Feature	0.1231	0.1158	0.0504	0.1381	0.1293	0.8937	1.0000	0.9078	0.6698	-	-	-	-	-
	Image	0.1434	0.1349	0.0518	0.1555	0.1540	0.8940	1.0000	0.9185	0.6890	-	-	-	-	-
	Raw data	0.1384	0.1340	0.0525	0.1544	0.1464	0.8939	1.0000	0.9122	0.6844	-	-	-	-	-
OSASUD apnea multi-label classification	Feature	0.1551	0.1534	0.1242	0.2692	0.2958	0.6935	1.0000	0.8368	0.6385	-	-	-	-	-
	Image	0.1855	0.1831	0.1394	0.2785	0.2424	0.7018	1.0000	0.9247	0.7063	-	-	-	-	-
	Raw data	0.1346	0.1342	0.1155	0.2634	0.2605	0.6886	1.0000	0.8368	0.6056	-	-	-	-	-

3 A1.3.2 – Results for demographic subgroups and classification of biological sex

LNE will analyse the results from A1.3.1 for different demographic subgroups of the datasets provided by A2.1.4, e.g., differentiated by sex and/or age. NPL, Uni-Oldenburg and PTB will repeat the analyses of A1.3.1 for at least 2 of the datasets provided by A2.1.4 but using only subsets consisting of single-sex training data and LNE will compare performance and uncertainty with the evaluation frameworks of A1.1.6 and A1.2.6 with the results of A1.3.1.
KCL and NPL will consider classification of PPG signals (e.g. from SPAR images) by biological sex using at least two of the models developed in A1.1.2-A1.1.4 evaluating both performance and uncertainty.

Deep learning models trained to predict physiological parameters from PPG signals may fit to feature representations more commonly observed in patients with certain characteristics. E.g., hypothetically, a model trained on a dataset that has more male patients then female patients may overfit to sex-specific features. If these potential biases are unaccounted for when evaluating model performance, its effectiveness may be over or underestimated given the composition of the test data. E.g. the model mentioned previously may exhibit poor generalisability on test data with an equal representation of sexes. Knowledge of this bias may inform sensible decisions about dataset curation that improve effective use of the model in practice.

Here, we consider the role of biological sex among other demographic factors on the performance of deep learning models trained to predict physioloigical parameters from PPG signals. First, we investigate whether PPG encodes sex-specific feature representations by assessing whether a model can learn to classify sex. Then we observe whether models trained and evaluated on datasets of varying sex composition exhibit variation in predictive performance, motivating a general need to account for biological sex when optimising deep learning models.

3.1 Single sex training results

To study sex-specific generalisation, we trained the same model architecture twice under a single-sex regime: once using only male subjects (TRAIN_M) and once using only female subjects (TRAIN_F). For each training run, we tracked validation performance and retained the 10 best checkpoints (i.e., the 10 models with the best validation score observed during that single run); the first block of 10 rows in the results tables reports the corresponding test MAE values for these selected checkpoints. We then evaluated each checkpoint on three disjoint test splits, male-only (TEST_M), female-only (TEST_F), and mixed-sex (TEST_MF), to quantify within-group performance and cross-group transfer. The male-only and female-only splits were obtained by filtering the full set of test samples based on gender, whereas the mixed-sex split consisted of the entire test set for each dataset considered in this study.

The results reported in this section were obtained on the VitalDB dataset (blood pressure estimation (dbp, sbp) with the two investigated scenarios: CALIB (train and test datasets share subjects) and CALIBFREE (train and test datasets do not share subjects)), as gender information was not available across all datasets. In addition, because we considered several models (as in section 2 (A1.3.1)), this analysis required training each model twice (male-only and female-only), which substantially increased the computational cost. For this reason, we restricted the study to a single dataset. For similar reasons, image-based models were not investigated hereafter.

Tables 12 and 13 display CALIB case results for the feature-based and the raw-signal-based models, respectively. CALIBFREE case results are depicted using Tables 14 and 15. In each table, the bottom rows summarise the distribution of the 10 retained checkpoints via the mean and standard deviation, and report relative errors with respect to the overall lowest mean MAE.

Figures 1 and 2 present the distribution of MAE values obtained from the 10 checkpoints using boxplots for the feature-based and raw-signal-based models under the CALIB and CALIBFREE settings, respectively.

3.1.1 VitalDB – CALIB case

Analysis of inference results for single-sex training (CALIB)

A clear sex-conditioned generalisation pattern emerges: each model performs best when the training subset matches the test subset (TRAIN_M → TEST_M and TRAIN_F → TEST_F), while cross-sex transfer (TRAIN_M → TEST_F or TRAIN_F → TEST_M) consistently degrades MAE—most noticeably for SBP. For the feature-based model, the checkpoint-to-checkpoint variability across the 10 best validation models is extremely small, indicating very stable training, but it also shows the strongest distribution shift penalty: for DBP, TRAIN_M performs worst on TEST_F with a mean value, μ=9.129, whereas TRAIN_F performs best on TEST_F (μ=8.838); for SBP the gap is sharper, with TRAIN_F generalising poorly to males with the worst mean value, μ=14.426, while achieving the best female-only score (TEST_F μ=13.468). The raw-signal model achieves lower MAE overall (notably DBP drops to μ=8.145 on TEST_M and μ=8.349 on TEST_F, and SBP improves to μ=13.852 on TEST_M), but with much higher variability across the 10 selected checkpoints (especially for SBP, σ increases substantially), suggesting a more sensitive optimisation landscape. Mixed-sex inference typically lands between same-sex and cross-sex performance, and tends to be slightly better when trained on the subset that yields the stronger overall baseline (e.g., raw-signal SBP favors TRAIN_F on TEST_MF). Overall, these results highlight a trade-off: raw signals bring better accuracy but less stability, while features are highly stable yet more exposed to sex-specific distribution shift, particularly for SBP as clearly illustrated by the boxplots in Figure 1.

3.1.2 VitalDB – CALIBFREE case

Analysis of inference results for single-sex training (CALIBFREE)

Both representations again show a strong sex-matched generalisation effect, but with a sharper separation between stability (features) and accuracy/variance (raw signals). For the feature-based model, performance is very stable across the 10 selected checkpoints (small σ), and the best MAE is obtained when training and test sex align: for DBP, TRAIN_M achieves the minimum on TEST_M with μ=8.657, while TRAIN_F is closest to the minimum on TEST_M as well (μ=8.674) and remains competitive on TEST_F (μ=9.081). However, cross-sex transfer is clearly penalised: TRAIN_M evaluated on TEST_F increases μ to 9.475 (the worst DBP column), and for SBP the mismatch is even more pronounced, with TRAIN_M → TEST_F reaching μ=14.634, while the best SBP score is TRAIN_M → TEST_M achieving μ=13.840. The raw-signal model yields a similar qualitative pattern but with substantially larger σ and a more asymmetric behavior: it matches the best feature MAE on DBP/TEST_M (μ=8.460), (best overall in that table) and achieves the best SBP/TEST_M when trained on females (μ=13.261), yet it suffers a larger degradation under mismatch—most notably for SBP where TRAIN_M → TEST_F rises to (μ=15.607) (largest relative error). Mixed-sex inference (TEST_MF) consistently falls between the matched and mismatched cases for both models, but remains closer to the matched condition when the training sex corresponds to the lower-error group (e.g., raw-signal SBP benefits from TRAIN_F). Overall, CALIBFREE amplifies sex-related domain shift, with raw signals offering occasional improvements on the best-case columns but exhibiting stronger checkpoint sensitivity and larger cross-sex degradation, whereas features provide more reproducible performance with a smaller—but still clear—generalisation gap. Again, this analysis is further corroborated by the boxplots presented in Figure 2.

Table 12: Inference results for single sex training using feature-based model (Wavelet + MLP) – VitalDB dataset – CALIB case. Best and worst settings for each task, based on the mean MAE computed across the 10 checkpoints, are indicated by ✓and ✗, respectively.

SCENARIO – **CALIB**											
DIASTOLIC BLOOD PRESSURE (DBP)						SYSTOLIC BLOOD PRESSURE (SBP)					
TRAIN_M			TRAIN_F			TRAIN_M			TRAIN_F		
TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F
9.0732	8.9940	9.1313	8.9653	9.1243	8.8488	14.0701	14.1777	13.9912	13.9032	14.4792	13.4809
9.0779	8.9935	9.1398	8.9588	9.1201	8.8405	14.0834	14.1821	14.0111	13.8957	14.4670	13.4768
9.0528	8.9873	9.1009	8.9527	9.1116	8.8361	14.1464	14.1935	14.1118	13.8870	14.4692	13.4601
9.0648	8.9878	9.1212	8.9522	9.1102	8.8363	14.1413	14.1898	14.1058	13.8747	14.4329	13.4655
9.0606	8.9865	9.1149	8.9512	9.1061	8.8375	14.1461	14.1879	14.1155	13.8657	14.4066	13.4691
9.0721	8.9898	9.1325	8.9512	9.1092	8.8354	14.1837	14.2044	14.1685	13.8622	14.3971	13.4700
9.0836	8.9944	9.1490	8.9516	9.1085	8.8366	14.1360	14.1867	14.0988	13.8599	14.3975	13.4657
9.0685	8.9885	9.1271	8.9517	9.1108	8.8350	14.1317	14.1837	14.0935	13.8587	14.3925	13.4673
9.0765	8.9913	9.1390	8.9518	9.1111	8.8349	14.1299	14.1829	14.0910	13.8609	14.4058	13.4614
9.0747	8.9905	9.1364	8.9518	9.1115	8.8347	14.1411	14.1856	14.1084	13.8615	14.4082	13.4607
MAE – SUMMARY STATISTICS → (μ, σ)											
9.0705	8.9904	9.1292 ✗	8.9538	9.1123	8.8376 ✓	14.1310	14.1874	14.0896	13.8730	14.4256 ✗	13.4677 ✓
0.0090	0.0029	0.0139	0.0046	0.0055	0.0043	0.0324	0.0074	0.0516	0.0165	0.0338	0.0068
MAE – RELATIVE ERROR IN % → $\left[100 \times \frac{\mu - \min(\mu)}{\min(\mu)}\right]$											
2.6351	1.7287	3.2998	1.3152	3.1090	0.0000	4.9245	5.3437	4.6170	3.0088	7.1122	0.0000

Table 13: Inference results for single sex training using raw-signal-based model (TCN) – VitalDB dataset – CALIB case. Best and worst settings for each task, based on the mean MAE computed across the 10 checkpoints, are indicated by ✓and ✗, respectively.

SCENARIO – **CALIB**											
DIASTOLIC BLOOD PRESSURE (DBP)						SYSTOLIC BLOOD PRESSURE (SBP)					
TRAIN_M			TRAIN_F			TRAIN_M			TRAIN_F		
TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F
8.4514	8.2412	8.6055	8.6376	8.7660	8.5434	14.1875	14.2198	14.1638	14.1096	14.3978	13.8983
8.3422	8.0919	8.5258	8.4925	8.6815	8.3538	13.5326	13.5185	13.5429	13.9517	14.0621	13.8708
8.4965	8.2386	8.6856	8.3344	8.5131	8.2034	14.4647	14.1132	14.7224	13.7415	13.9642	13.5782
8.5967	8.3057	8.8101	8.6671	8.7458	8.6093	14.6819	14.3027	14.9600	13.6920	14.5095	13.0925
8.5331	8.2391	8.7486	8.6211	8.7402	8.5338	14.8986	14.4726	15.2111	13.9579	14.0548	13.8868
8.2361	7.8910	8.4891	8.8082	9.0415	8.6371	14.7367	14.3167	15.0446	13.6749	13.8915	13.5162
8.2629	7.9043	8.5258	8.3802	8.6134	8.2092	14.4563	13.9578	14.8218	13.1857	13.4770	12.9721
8.6357	8.2816	8.8953	8.2482	8.5121	8.0546	13.0273	12.7708	13.2153	13.4289	13.5047	13.3733
8.4266	8.0488	8.7037	8.1436	8.4182	7.9422	14.4912	13.8857	14.9352	13.3351	13.5879	13.1497
8.5915	8.2080	8.8727	8.6216	8.9187	8.4038	13.3409	12.9607	13.6197	13.9229	13.9121	13.9309
MAE – SUMMARY STATISTICS → (μ, σ)											
8.4573	8.1450 ✓	8.6862	8.4954	8.6950 ✗	8.3491	14.1818	13.8518	14.4237 ✗	13.7000	13.9362	13.5269 ✓
0.1332	0.1449	0.1396	0.2008	0.1816	0.2277	0.6152	0.5563	0.6894	0.2865	0.3307	0.3484
MAE – RELATIVE ERROR (w.r.t the MIN) IN % → $\left[100 \times \frac{\mu - \min(\mu)}{\min(\mu)}\right]$											
3.8336	0.0000	6.6445	4.3022	6.7530	2.5053	4.8414	2.4024	6.6298	1.2800	3.0257	0.0000

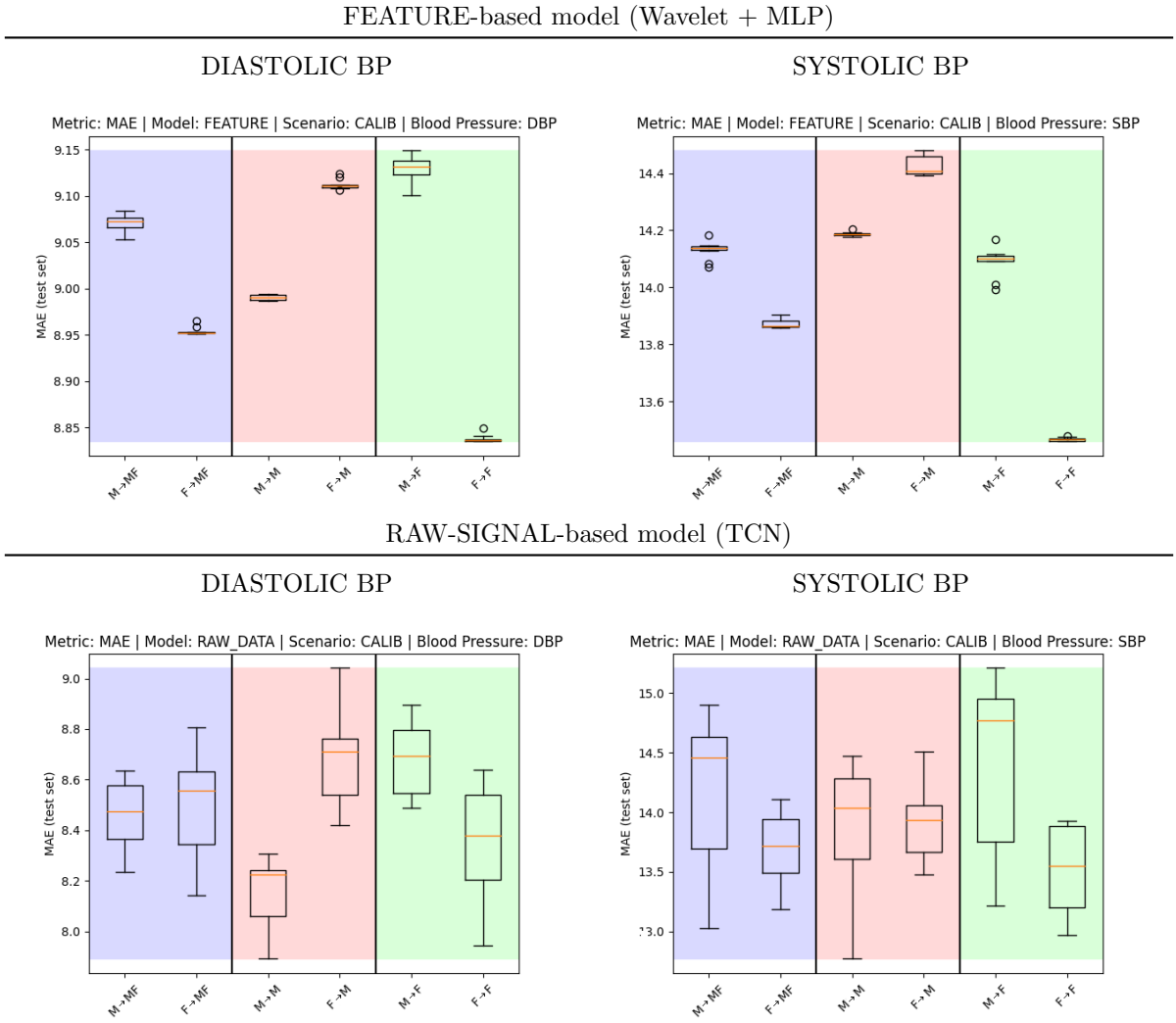


Figure 1: Inference results – MAE boxplots – VitalDB dataset – CALIB case. Top figures and bottom figures correspond to the inference results (distribution of MAE values from the 10 checkpoints per training) for the feature-based model and the raw-signal-based model, respectively. M→F notation stands for TRAIN_M → TEST_F. Blue, red and green colours are used to highlight the 3 disjoint test splits, mixed-sex, male-only and female-only, respectively.

Table 14: Inference results for single sex training using feature-based model (Wavelet + MLP) – VitalDB dataset – CALIBFREE case. Best and worst settings for each task, based on the mean MAE computed across the 10 checkpoints, are indicated by ✓and ✗, respectively.

SCENARIO – **CALIBFREE**												
DIASTOLIC BLOOD PRESSURE (DBP)						SYSTOLIC BLOOD PRESSURE (SBP)					
TRAIN_M			TRAIN_F			TRAIN_M			TRAIN_F		
TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F
9.1937	8.7135	9.5175	8.9427	8.6659	9.1294	14.4037	13.9415	14.7155	14.3454	14.1970	14.4455
9.1440	8.6619	9.4691	8.9265	8.6848	9.0896	14.3143	13.9103	14.5868	14.4623	14.4005	14.5039
9.1608	8.6433	9.5098	8.9170	8.6773	9.0787	14.3486	13.8384	14.6927	14.2885	14.0874	14.4242
9.1375	8.6663	9.4553	8.9234	8.6567	9.1033	14.2650	13.8572	14.5400	14.3068	14.1402	14.4191
9.1312	8.6599	9.4491	8.9089	8.6729	9.0682	14.2583	13.8443	14.5374	14.2751	14.0603	14.4200
9.1497	8.6436	9.4911	8.9117	8.6766	9.0702	14.2675	13.8188	14.5701	14.3030	14.1304	14.4194
9.1350	8.6446	9.4658	8.9114	8.6815	9.0665	14.3235	13.7974	14.6783	14.2805	14.0867	14.4112
9.1476	8.6441	9.4871	8.9114	8.6730	9.0721	14.3117	13.7951	14.6600	14.2813	14.0846	14.4139
9.1253	8.6491	9.4465	8.9102	8.6776	9.0671	14.2937	13.7916	14.6323	14.2740	14.0697	14.4119
9.1323	8.6454	9.4607	8.9103	8.6760	9.0684	14.3543	13.8044	14.7252	14.2677	14.0552	14.4110
MAE – SUMMARY STATISTICS → (μ, σ)											
9.1457	8.6572 ✓	9.4752 ✗	8.9174	8.6742	9.0814	14.3140	13.8399 ✓	14.6338 ✗	14.3085	14.1312	14.4280
0.0198	0.0216	0.0250	0.0107	0.0080	0.0207	0.0460	0.0510	0.0711	0.0586	0.1041	0.0285
MAE – RELATIVE ERROR (w.r.t the MIN) IN % → $\left[100 \times \frac{\mu - \min(\mu)}{\min(\mu)}\right]$											
5.6432	0.0000	9.4491	3.0056	0.1970	4.8998	3.4260	0.0000	5.7365	3.3855	2.1048	4.2493

Table 15: Inference results for single sex training using raw-signal-based model (TCN) – VitalDB dataset – CALIBFREE case. Best and worst settings for each task, based on the mean MAE computed across the 10 checkpoints, are indicated by ✓and ✗, respectively.

SCENARIO – **CALIBFREE**												
DIASTOLIC BLOOD PRESSURE (DBP)						SYSTOLIC BLOOD PRESSURE (SBP)					
TRAIN_M			TRAIN_F			TRAIN_M			TRAIN_F		
TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F	TEST_MF	TEST_M	TEST_F
9.1439	8.6482	9.4783	8.8012	8.5682	8.9584	14.7055	14.4433	14.8823	13.5928	13.1811	13.8704
9.4924	8.8379	9.9339	8.8091	8.6385	8.9242	14.2868	13.8839	14.5586	13.5655	12.9636	13.9715
9.2539	8.6791	9.6415	8.6237	8.3818	8.7869	14.3124	13.9302	14.5701	13.8077	12.9517	14.3850
9.2598	8.6789	9.6516	8.4756	8.2802	8.6074	15.5493	14.6643	16.1461	13.7587	13.0006	14.2700
9.3462	8.7194	9.7689	8.3422	8.0714	8.5249	16.2893	15.2337	17.0013	13.3835	12.7118	13.8365
9.1335	8.5343	9.5376	8.8351	8.5327	9.0390	16.1448	15.0930	16.8542	13.4039	12.6137	13.9369
8.7508	8.3002	9.0547	8.4931	8.2462	8.6597	15.0843	13.8213	15.9361	14.0346	13.3080	14.5246
8.6996	8.2432	9.0073	8.9810	8.7402	9.1434	13.9014	12.9402	14.5497	14.4383	13.7704	14.8887
8.5742	8.0557	8.9239	8.7384	8.5655	8.8550	15.0696	13.6020	16.0594	14.7146	14.0788	15.1433
8.7551	7.9016	9.3306	8.8282	8.6651	8.9382	14.7330	13.5738	15.5148	14.6558	14.0268	15.0800
MAE – SUMMARY STATISTICS → (μ, σ)											
9.0409	8.4599 ✓	9.4328 ✗	8.6928	8.4690	8.8437	15.0076	14.1186	15.6073 ✗	13.9355	13.2606 ✓	14.3907
0.3016	0.2991	0.3259	0.1907	0.2037	0.1874	0.7505	0.6876	0.8904	0.4772	0.4996	0.4777
MAE – RELATIVE ERROR IN % → $\left[100 \times \frac{\mu - \min(\mu)}{\min(\mu)}\right]$											
6.8687	0.0000	11.5011	2.7531	0.1076	4.5373	13.1743	6.4697	17.6961	5.0894	0.0000	8.5218

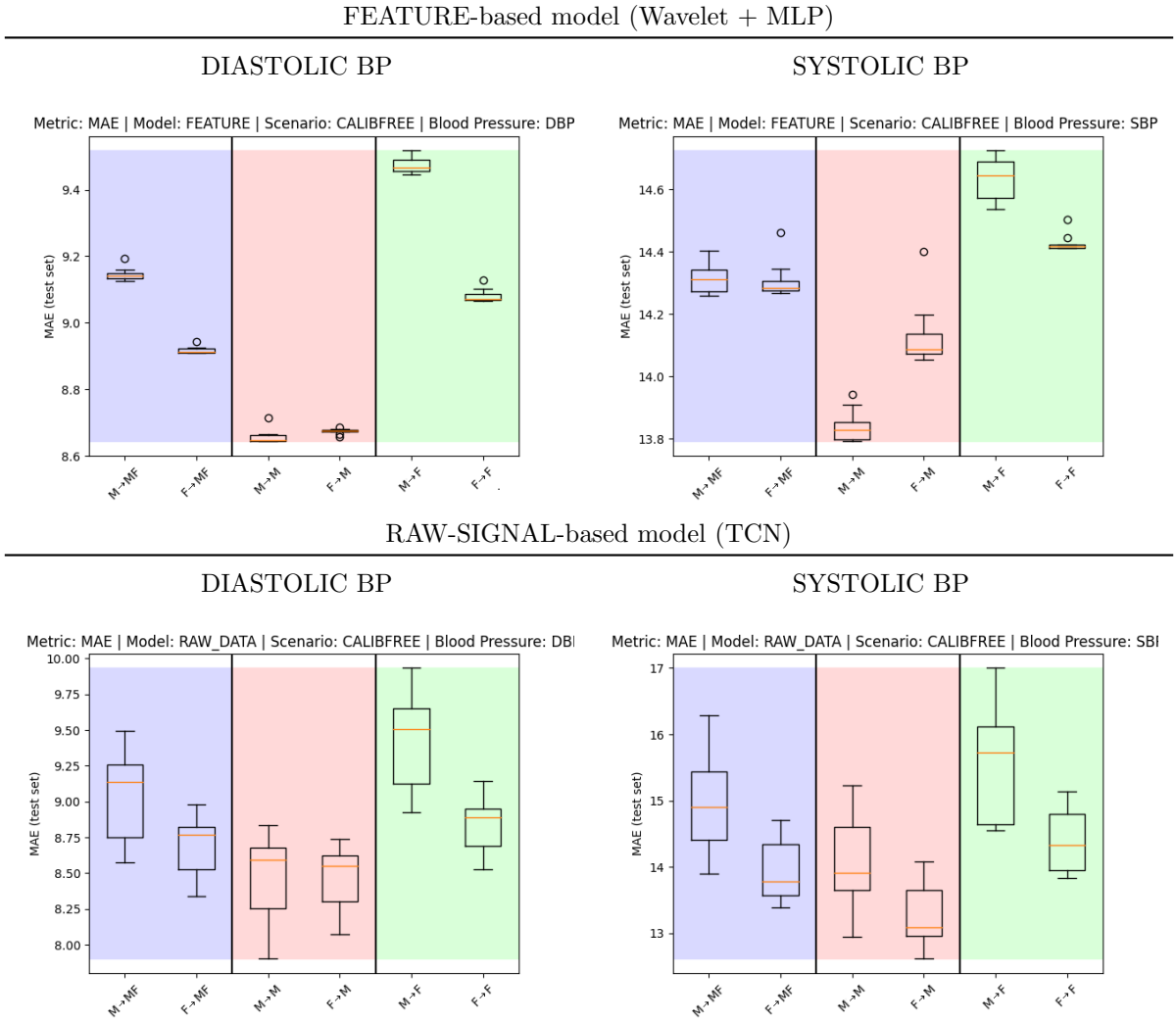


Figure 2: Inference results – MAE boxplots – VitalDB dataset – CALIBFREE case. Top figures and bottom figures correspond to the inference results (distrbution of MAE values from the 10 checkpoints per training) for the feature-based model and the raw-signal-based model, respectively. M→F notation stands for TRAIN_M → TEST_F. Blue, red and green colours are used to highlight the 3 disjoint test splits, mixed-sex, male-only and female-only, respectively.

Table 16: Breakdown of featurised Aurora-BP data by sex.

Raw PPG signals				Patients			
Male		Female		Male		Female	
Count	percent	Count	percent	Count	percent	Count	percent
18,490	51.9	17,116	48.1	570	50.7	554	49.3

3.2 Classification of raw PPG signals by biological sex

This task evaluates whether deep learning models are sensitive to features related to biological sex in PPG signals. Our aim is not to build sex-recognition technology, but to detect sex-linked signal differences that could bias data-driven models and to inform dataset curation (e.g., balancing representation). Where differences exist, we must assess their impact on diagnostic accuracy and care pathways and design mitigation strategies.

Sex differences in the ECG have been reported [18], but similar work for PPG is limited. However, the existing literature suggests associations between PPG features and sex [19, 20]. Here, we test whether raw PPG encodes information about reported biological sex by training deep learning classifiers to predict male versus female labels.

3.2.1 Dataset

The Aurora-BP dataset [4] was used for this task. A sample, example code, and documentation are available at the Aurora-BP GitHub page [21].

Aurora-BP is a large-scale, publicly available collection of cardiovascular sensor data with paired blood pressure measurements released by Microsoft Research [22]. Several signals were recorded simultaneously: radial tonometry, PPG, ECG, and accelerometry.

The dataset includes de-identified participant metadata (age, biological sex, weight, height, Fitzpatrick skin-tone scale value, and disease history) and a wide range of features derived from the sensor signals. Here, biological sex refers to the self-reported label (Female, Male).

The data encompass:

- Two protocols/cohorts: auscultatory and oscillometric measurements.
- Multiple phases per protocol: initial visit, return visit, ambulatory phase, and synthetic data (not used here).
- Postures and activity types: supine, seated, standing, walking, post-exercise, etc (each with variations).

An overview of the dataset and measurement counts is shown in Figure 3 [21]. We used the final quality-assessed, featurised dataset. Note that the counts refer to participants, each contributing multiple measurements. According to [4], the quality assessment steps performed when developing the Aurora-BP dataset involve computing quality scores that penalised data with artefacts, poor signal-to-noise ratio and lack of consistency between pulses within a window. A high score indicates confidence that the data and derived features are physiologic in nature.

After excluding measurements with missing data files, the featurised dataset comprised 642 auscultatory and 483 oscillometric measurement sets, each attributed to a single participant (total 1,124 participant IDs). In aggregate, this yielded 35,606 available PPG signals with a sampling frequency of 500 Hz. Table 16 summarises participant counts and raw PPG signal counts by biological sex. Given the near-balanced proportion of sex labels (roughly 51% male and 49% female), we refrained from enforcing an exactly equal representation.

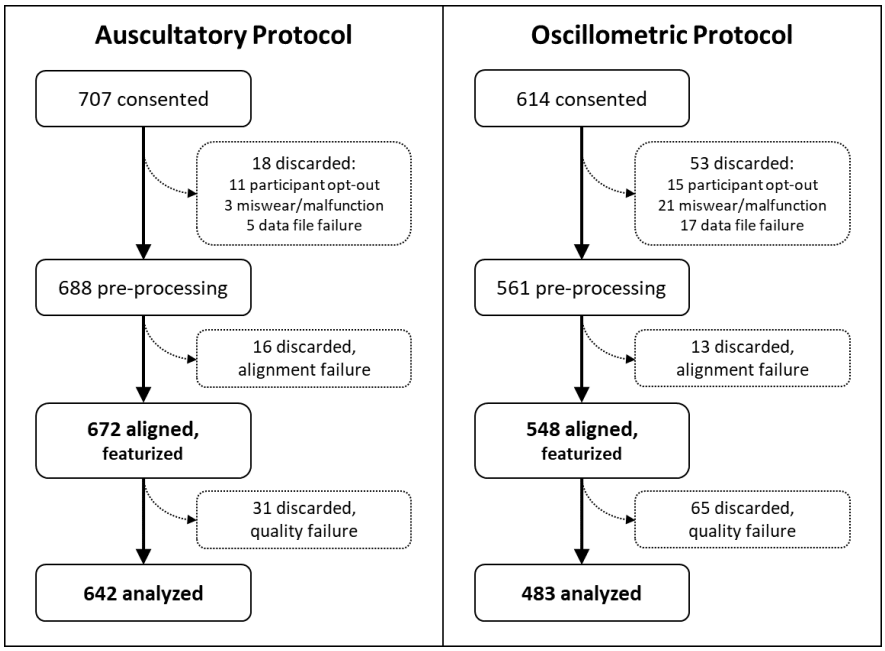


Figure 3: Breakdown of auscultatory and oscillometric measurement counts in the Aurora-BP dataset [21].

3.2.2 Data processing

Aurora-BP contains noisy and poor-quality segments, necessitating careful, comparative evaluation of data cleaning and normalisation steps. After multiple assessments, in the end, we implemented the following steps.

Signal filtering

Wearable acquisition conditions vary, making PPG signals susceptible to noise and artefacts that hinder reliable processing. Applying optimal filters is therefore important [23]. We analysed Aurora-BP PPG both in unfiltered form and after band-pass filtering. Observing the positive impact of filtering during hyperparameter sweeps (see Section 3.2.4), we adopted the detrended dataset produced by KCL.

Choice of measurement cohorts and types

Data from both auscultatory and oscillometric protocols were concatenated to create a unified dataset. Hyperparameter tuning (see Section 3.2.4) indicated that training models on specific subsets degraded performance, primarily due to the substantial reduction in dataset size.

Filtering by optical quality

Aurora-BP provides an *optical_quality* parameter for all records. We treat this as a hyperparameter: higher thresholds yield fewer, higher-quality signals, and lower thresholds lead to more, noisier signals. The Aurora-BP GitHub page [21] suggests a baseline threshold of 0.65. To assess the effect of prioritising higher quality vs. more data, we evaluated thresholds of 0.65, 0.75, 0.85, 0.90, and 0.95.

Figure 4 shows the histogram of optical quality with the chosen thresholds marked as vertical lines. Table 17 reports the number of unique participant IDs and raw PPG signals retained at each threshold. The trade-off is substantial: enforcing the highest quality markedly reduces dataset size.

Signal length and instance generation

We standardised PPG instances to a consistent timestep length. The target length was chosen as the largest value still shorter than the majority of signals, maximising usable information while minimising exclusions. Signals shorter than the target were discarded to maintain comparability across instances.

Figure 5 shows the histogram of raw signal lengths with no optical-quality thresholding. The shortest and

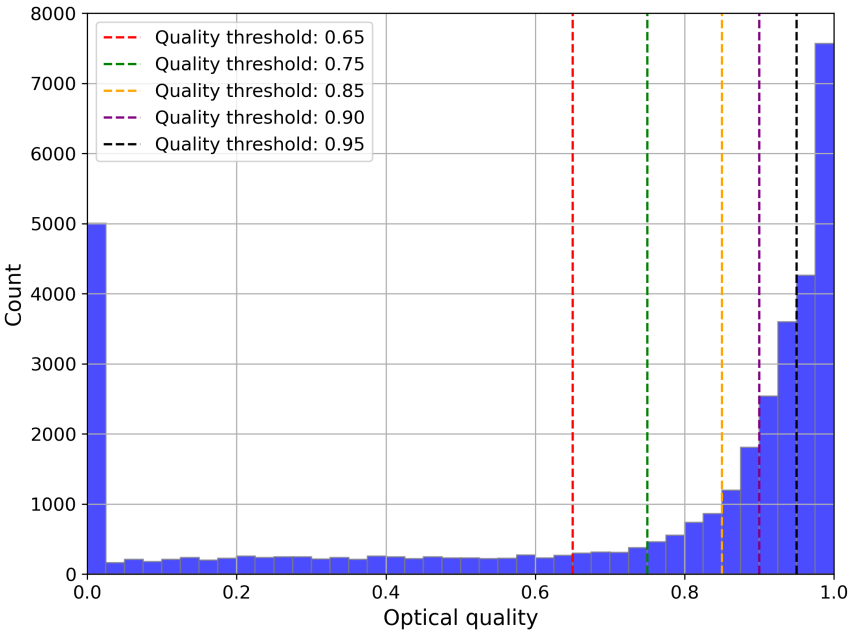


Figure 4: Histogram of optical quality values with different optical quality thresholds shown as vertical lines.

Table 17: Number of unique patient IDs and raw PPG signals after applying different quality thresholds.

Threshold	Patient IDs	Raw PPG signals	% Signals retained
0	1,124	35,606	100
0.65	1,099	24,869	69.84
0.75	1,093	23,559	66.17
0.85	1,085	20,966	58.88
0.9	1,072	17,876	50.21
0.95	1,023	11,826	33.21

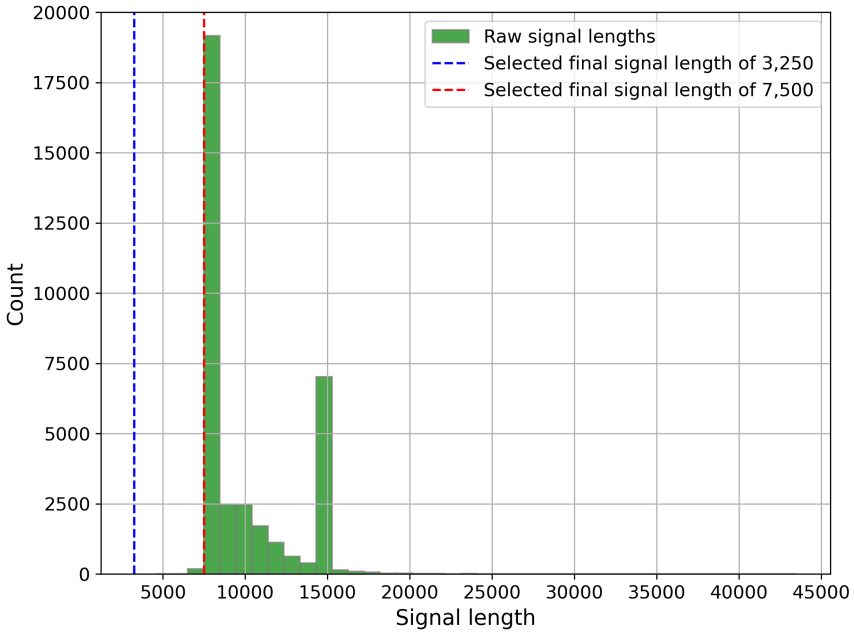


Figure 5: The histogram of raw signal lengths when no optical quality thresholding is applied. Vertical lines delineate the raw signals retained based on cutoff lengths of 3,250 and 7,500.

longest recordings are 4,561 (in the original dataset with 3,121 in the detrended dataset) and 43,543 timesteps, respectively. As the histogram indicates, most signals exceed 7,500 timesteps (15 s). A more forgiving cutoff of 3,250 (7.5 seconds) keeps virtually all recordings. Cutoff lengths of 7,500 and 3,250 timesteps retain approximately 99% and 100% of the data, respectively.

To create the training datasets, we considered three cases:

- We retained the first 7,500-step window of each signal and discarded the rest. We refer to this approach as "single."
- Starting from timestep 1, we divided each signal into multiple non-overlapping, back-to-back windows of size 7,500, discarding the tail end. For instance, if a signal was 17,000 timesteps, we took two instances from the intervals [1, 7,500] and [7,501, 15,000] and ignored the last 2,000 timesteps. We call this scheme "multiple."
- We set the target length to 3,250 to generate more training data, taking as many non-overlapping, back-to-back instances as possible from each signal similar to the above case (multiple approach).

During final instance extraction, some data were discarded or unavailable: (1) the raw signal was shorter than the target length; (2) referenced files were non-existent; (3) files were empty; or (4) files contained unreadable values (e.g., NaNs). The latter two issues primarily arose in the detrended dataset.

Table 18 reports the number of final signals recovered under each scheme. To maximise the number of training instances, one can densely sample 3,250-step windows (see the third bullet point above) while using a lower optical-quality threshold (0.65). This yields approximately six times as many instances as using single sampling with 7,500-step windows (see the first bullet point above) and a high quality threshold (0.95).

Data split for training

The dataset is well-balanced by biological sex. We, therefore, did not enforce equal male-female representation before splitting the dataset. We randomly shuffled and partitioned the data into training (75%), validation (12.5%), and test (12.5%) sets. Splits were stratified by biological sex to preserve the overall male-female

Table 18: The number of signals recovered when different final signal lengths (L) are considered. *Single* and *Multiple* refer to whether a single instance is extracted from the beginning of each raw PPG signal or multiple ones are taken back-to-back, respectively.

Threshold	Raw PPG signals	$L = 7,500$ (Single)	$L = 7,500$ (Multiple)	$L = 3,250$ (Multiple)
0.65	24,869	24,340	28,848	64,401
0.75	23,559	23,086	27,263	60,994
0.85	20,966	20,560	24,134	54,098
0.9	17,876	17,550	20,394	45,645
0.95	11,826	11,595	13,259	29,464

proportions in subsets and grouped by participant ID to prevent subject-level leakage across sets. Class labels were encoded as $Female = 1$ and $Male = 0$ (binary). It should be noted that enforcing equal sex representation and stratification with respect to sex are two different steps. The latter establishes the same number of males and females in the entire dataset (typically by removing data belonging to the sex that is overrepresented), while the former ensures that the same male-female proportions as in the whole dataset are preserved across all data splits (not necessarily equal male-female numbers).

3.2.3 Deep learning models

Two deep learning models used in QUMPHY and a new transformer-based model were trained and tested as classifiers of PPG signals by biological sex:

1. QUMPHY model: 1D variant of AlexNet [24],
2. QUMPHY model: 1D variant of XResNet (xresnet1d50) [25],
3. New model: a 1D Vision Transformer (ViT) classifier [26].

The ViT uses the encoder part of a standard transformer [27] and applies it to embeddings created from an image (here a 1D signal). These embeddings are derived by flattening and linearly projecting windows within each signal. We used a window size of 25 timesteps, with each window projected onto a 50-dimensional space. Our ViT comprises five transformer layers with five heads in each attention layer. The Multi-Layer Perceptron (MLP) part of transformer layers includes 100 nodes. After the transformer block, the model uses a CNN-based classifer (similar to a small classification AlexNet) to develop the outputs. Following some empirical tuning, we used dropout rates of 0.2, 0.3, and 0.5 in attention, MLP, and CNN layers, respectively.

All models produced logits, which could then be converted to class probabilities using a *Softmax* function. Alternatively, since this is a binary classification problem, a single *sigmoid* activation function could be employed to directly produce the probability of the positive class.

3.2.4 Hyperparameter sensitivity and tuning approach

Classification performance was sensitive to hyperparameter choices across data processing, model architecture, and training. Table 19 reports the validation performance for selected configurations. To manage computational cost and avoid a full grid search, we adopted a staged (block-wise) tuning procedure: each block compared related hyperparameters (e.g., normalisation methods), the best option was fixed, and subsequent blocks tuned the next set. Where interactions were likely, we explored targeted sub-grids, e.g., jointly varying dropout rates while comparing different normalisation settings.

In Table 19, each row corresponds to a model and rows with bold entries belong to the top-performing models. Cells within hyperparameter columns are colour-coded to indicate whether that hyperparameter

was evaluated for the model and the outcome of that evaluation. As noted above, some models assess multiple hyperparameters simultaneously. Within each hyperparameter column:

- Green denotes preferred (accepted) values.
- Red denotes discarded (rejected) values.
- Grey denotes values varied case by case, since no range consistently outperformed alternatives.

Given the descriptions already provided, what each column represents should be self-explanatory. It is only worth adding that overall and individual normalisation columns refer to the cases where PPG signals were normalised altogether (as a whole) or on an individual basis (separately, one-by-one), respectively.

Models were trained by an *AdamW* optimiser [28], minimising *Binary Cross-Entropy (BCE)*. The *learning rate* was decayed from an initial value of 10^{-4} to a minimum value of 10^{-8} by a cosine annealing schedule [29]. The weight decay factor was also set to 0.01. These settings were found to be consistently effective for all models.

Table 19: Validation performance vs. hyperparameter choices. Green, red and grey fill colours indicate accepted, rejected, and decided per case, respectively. Green text means the hyperparameter is fixed as optimal, based on the experiments in the previous rows.

Goal	Dataset hyperparameters								Model hyperparameters			Training hyperparameters				Performance: validation (12.5%)				
	Dataset	Wave filtering	Quality threshold	Time steps	Multiple instances	Cohort	Measur. Type	Num waves	Type	Size	Dropout rate	Overall normal.	Individ. normal.	Batch size	Epochs	BCE	Accuracy	Precision	Recall	F1 score
Filtering	Original	No	0.65	7,500	No	Both	All	24,604	AlexNet	Smaller	0.4	None	None	64	10	0.6736	0.5575	0.6073	0.3295	0.4235
	Original	Yes	0.65	7,500	No	Both	All	24,604	AlexNet	Smaller	0.4	None	None	64	10	0.6450	0.6312	0.6236	0.6684	0.6427
Dropout rate	Original	Yes	0.65	7,500	No	Both	All	24,604	AlexNet	Smaller	0.3	None	None	64	10	0.6337	0.6516	0.6784	0.5793	0.6219
	Original	Yes	0.65	7,500	No	Both	All	24,604	AlexNet	Smaller	0.5	None	None	64	10	0.6385	0.6472	0.6662	0.6013	0.6287
	Original	Yes	0.65	7,500	No	Both	All	24,604	AlexNet	Smaller	0.6	None	None	64	10	0.6499	0.6303	0.6735	0.5108	0.5761
Baseline	Detrended	Yes	0.65	7,500	No	Both	All	24,340	AlexNet	Smaller	0.5	None	None	64	10	0.6244	0.6679	0.6520	0.6596	0.6528
Normalis-ation	Detrended	Yes	0.65	7,500	No	Both	All	24,340	AlexNet	Smaller	0.5	Z-score	None	64	10	0.6075	0.6757	0.6741	0.6336	0.6504
	Detrended	Yes	0.65	7,500	No	Both	All	24,340	AlexNet	Smaller	0.5	Min-Max	None	64	10	0.6927	0.5187	0.0000	0.0000	0.0000
	Detrended	Yes	0.65	7,500	No	Both	All	24,340	AlexNet	Smaller	0.5	None	Z-score	64	10	0.5932	0.6891	0.7130	0.5952	0.6460
	Detrended	Yes	0.65	7,500	No	Both	All	24,340	AlexNet	Smaller	0.5	None	Min-Max	64	10	0.6250	0.6582	0.6612	0.5925	0.6216
Optical quality threshold	Detrended	Yes	0.75	7,500	No	Both	All	23,086	AlexNet	Smaller	0.5	None	Z-score	64	10	0.5747	0.6975	0.7116	0.5607	0.6224
	Detrended	Yes	0.75	7,500	No	Both	All	23,086	AlexNet	Smaller	0.5	Z-score	None	64	10	0.6053	0.6808	0.6486	0.6587	0.6483
	Detrended	Yes	0.85	7,500	No	Both	All	20,560	AlexNet	Smaller	0.6	None	Z-score	64	10	0.6327	0.6603	0.6978	0.5353	0.5953
	Detrended	Yes	0.85	7,500	No	Both	All	20,560	AlexNet	Smaller	0.5	None	Z-score	64	10	0.6329	0.6641	0.6871	0.5696	0.6127
	Detrended	Yes	0.85	7,500	No	Both	All	20,560	AlexNet	Smaller	0.5	Z-score	None	64	10	0.6352	0.6484	0.6442	0.6084	0.6197
	Detrended	Yes	0.9	7,500	No	Both	All	17,550	AlexNet	Smaller	0.5	None	Z-score	64	10	0.6229	0.6730	0.6459	0.6520	0.6440
	Detrended	Yes	0.9	7,500	No	Both	All	17,550	AlexNet	Smaller	0.5	Z-score	None	64	10	0.6332	0.6748	0.6593	0.6185	0.6342
	Detrended	Yes	0.95	7,500	No	Both	All	11,595	AlexNet	Smaller	0.5	None	Z-score	64	10	0.5924	0.6969	0.6857	0.6294	0.6497
	Detrended	Yes	0.95	7,500	No	Both	All	11,595	AlexNet	Smaller	0.5	Z-score	None	64	10	0.6382	0.6444	0.6045	0.6148	0.6039
Multiple L=7,500	Detrended	Yes	0.65	7,500	Yes	Both	All	28,848	AlexNet	Smaller	0.8	None	Z-score	512	24	0.5936	0.6924	0.7023	0.6721	0.6867
	Detrended	Yes	0.75	7,500	Yes	Both	All	27,263	AlexNet	Smaller	0.75	None	Z-score	64	16	0.5825	0.6972	0.6568	0.6671	0.6584
	Detrended	Yes	0.85	7,500	Yes	Both	All	24,134	AlexNet	Smaller	0.75	None	Z-score	64	16	0.6219	0.6607	0.7011	0.6282	0.6600
	Detrended	Yes	0.9	7,500	Yes	Both	All	20,394	AlexNet	Smaller	0.65	None	Z-score	128	16	0.5839	0.7121	0.6903	0.6760	0.6816
	Detrended	Yes	0.95	7,500	Yes	Both	All	13,259	AlexNet	Smaller	0.75	None	Z-score	256	24	0.5784	0.6966	0.7107	0.6983	0.7038
Multiple L=3,250	**Detrended**	**Yes**	**0.65**	**3,250**	**Yes**	**Both**	**All**	**64,401**	**AlexNet**	**Smaller**	**0.6**	**None**	**Z-score**	**128**	**16**	**0.5858**	**0.7019**	**0.7157**	**0.7135**	**0.7130**
	Detrended	Yes	0.65	3,250	Yes	Both	All	64,401	AlexNet	Smaller	0.5	Z-score	None	64	16	0.5891	0.7027	0.7338	0.6767	0.7012
	Detrended	**Yes**	**0.75**	**3,250**	**Yes**	**Both**	**All**	**60,994**	**AlexNet**	**Smaller**	**0.5**	**None**	**Z-score**	**64**	**16**	**0.5594**	**0.7081**	**0.6865**	**0.7443**	**0.7108**
	Detrended	Yes	0.75	3,250	Yes	Both	All	60,994	AlexNet	Smaller	0.5	Z-score	None	64	16	0.5676	0.7131	0.7140	0.6903	0.6994
	Detrended	Yes	0.85	3,250	Yes	Both	All	54,098	AlexNet	Smaller	0.75	None	Z-score	512	24	0.6071	0.6776	0.6799	0.6907	0.6842
	Detrended	Yes	0.85	3,250	Yes	Both	All	54,098	AlexNet	Smaller	0.65	Z-score	None	128	24	0.6241	0.6647	0.6784	0.6367	0.6549
	Detrended	**Yes**	**0.9**	**3,250**	**Yes**	**Both**	**All**	**45,645**	**AlexNet**	**Smaller**	**0.75**	**None**	**Z-score**	**1024**	**32**	**0.5639**	**0.7184**	**0.7137**	**0.6905**	**0.7017**
	Detrended	Yes	0.9	3,250	Yes	Both	All	45,645	AlexNet	Smaller	0.6	Z-score	None	512	32	0.5847	0.7056	0.6935	0.6938	0.6934
	Detrended	Yes	0.95	3,250	Yes	Both	All	29,464	AlexNet	Smaller	0.75	None	Z-score	256	24	0.6012	0.6822	0.6813	0.7099	0.6948
	Detrended	Yes	0.95	3,250	Yes	Both	All	29,464	AlexNet	Smaller	0.75	Z-score	None	32	24	0.6114	0.6770	0.6766	0.7077	0.6863
Cohort	Detrended	Yes	0.65	3,250	Yes	Oscillom.	All	37,545	AlexNet	Smaller	0.6	None	Z-score	128	16	0.5758	0.6965	0.6907	0.6812	0.6846
	Detrended	Yes	0.65	3,250	Yes	Auscult.	All	26,856	AlexNet	Smaller	0.6	None	Z-score	128	16	0.5603	0.7215	0.7182	0.6811	0.6980
Type	Detrended	Yes	0.75	3,250	Yes	Both	Seated	14,665	AlexNet	Smaller	0.75	None	Z-score	128	24	0.5886	0.7010	0.6653	0.6980	0.6796
Model Size	**Detrended**	**Yes**	**0.65**	**3250**	**Yes**	**Both**	**All**	**64,401**	**AlexNet**	**Larger**	**0.75**	**None**	**Z-score**	**64**	**16**	**0.5757**	**0.7069**	**0.7204**	**0.7203**	**0.7173**
	Detrended	**Yes**	**0.75**	**3250**	**Yes**	**Both**	**All**	**60,994**	**AlexNet**	**Larger**	**0.5**	**None**	**Z-score**	**64**	**16**	**0.5556**	**0.7128**	**0.6987**	**0.7281**	**0.7096**
	Detrended	**Yes**	**0.9**	**3250**	**Yes**	**Both**	**All**	**45,645**	**AlexNet**	**Larger**	**0.8**	**None**	**Z-score**	**1024**	**32**	**0.5623**	**0.7210**	**0.7262**	**0.6730**	**0.6983**
Other models	**Detrended**	**Yes**	**0.65**	**3,250**	**Yes**	**Both**	**All**	**64,401**	**XResNet**	-	**0.25**	**None**	**Z-score**	**128**	**32**	**0.5787**	**0.7040**	**0.7368**	**0.6742**	**0.7024**
	Detrended	Yes	0.65	3,250	Yes	Both	All	64,401	ViT	-	0.2-0.5	None	Z-score	128	16	0.6167	0.6705	0.7014	0.6442	0.6707

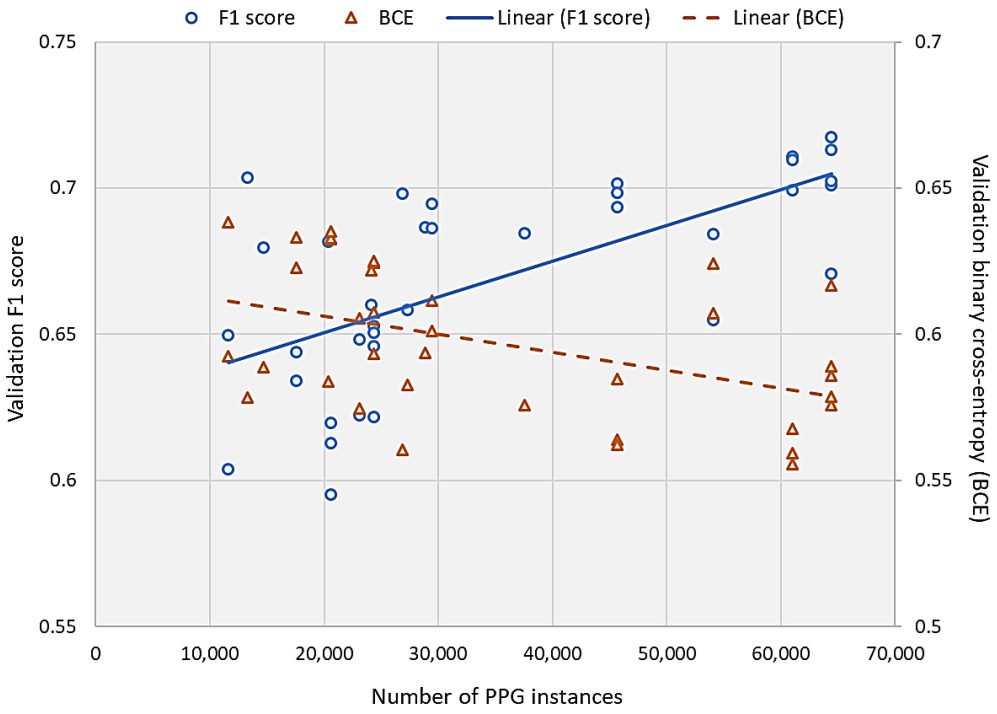


Figure 6: The average effect of dataset size on the validation loss (*BCE*) and *F1 score* .

Given the breadth of hyperparameters, the final configuration reflects empirical judgement, which was also informed by additional experiments beyond those shown here. For instance, Table 19 presents only a subset of the XResNet and ViT runs actually performed. In nearly all trials, however, AlexNet consistently outperformed these models. Consequently, the experiments underpinning the final hyperparameter conclusions are not fully captured in the table.

Hyperparameter effects are not absolute. Here, we focused on average behaviours. For example, in specific configurations, subject to other hyperparameters, a higher optical quality threshold (e.g., 0.90) may outperform 0.65. Nevertheless, we adopted 0.65 because, on average, it yielded better validation performance.

To further demonstrate this, Figure 6 illustrates the relationship between dataset size and validation loss (*BCE*) and *F1* score. Although the points are scattered due to variation in other hyperparameters, performance, on average, improves clearly as more data are introduced.

For some factors such as *dropout rate* and *batch size*, no single value proved uniformly optimal. Accordingly, we varied such parameters per experiment and selected values from a range that performed best on a case-by-case basis.

3.2.5 Uncertainty evaluation

Once the optimal modelling approach and hyperparameters were identified, we introduced uncertainty quantification into the model using the Monte Carlo Dropout (*MCD*) approach [30]. The procedures for training the uncertainty-aware classifier and quantifying the uncertainties are explained in Section 6 (A1.3.5), where uncertainties are disentangled. In summary, we:

- Included two output logits per class, representing a mean (μ) and a variance (σ^2), i.e., four outputs in total. In our binary problem, a *sigmoid* activation function is more straightforward as its mean and variance directly correspond to probabilities of the "true" class. However, we adopt a *softmax* head and two logits per class for generalisability to cases with more than two classes.

- Replaced *BCE* loss with Gaussian Negative Log-likliehood (*GNLL*), considering both mean and variance of probability.
- During training, sampled from the Gaussian logit distribution for each instance and converted those to class probabilities using the *softmax* function, allowing for calculation of the loss.
- During inference, activated Monte Carlo layers, and performed multiple stochastic forward passes per instance to approximate the posterior predictive distribution. Logit samples were then converted to class probabilities. Finally, *entropies* were calculated across MCD passes, leading to total vs. aleatoric uncertainties.

For a classification problem with n classes and predictive probabilities $p_1, \ldots, p_n$, the predictive entropy is

$$\mathcal{H}(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Entropy provides a principled measure of predictive uncertainty by quantifying the dispersion of the class probabilities. High predictive entropy indicates that probability is spread, signalling low model confidence, whereas low entropy corresponds to more certain and narrow predictions. As such, entropy provides a useful summary value for comparing uncertainty across samples, detecting out-of-distribution inputs, and supporting risk-aware decision-making in deep-learning classifiers.

As mentioned, uncertainty quantification results are presented in Section 6 (A1.3.5), where total (entropy of average probabilities across MCD runs) and aleatoric (average of entropies across MCD runs) uncertainties are evaluated.

In the following subsection, the details of the final model along with its testing performance are presented.

3.2.6 Final uncertainty-aware model's performance

Based on hyperparameter tuning results (Table 19), the final deep-learning classifier used the following:

- Dataset:
  - Detrended dataset (filtered signals).
  - Data with optical quality ≥ 0.65.
  - Data from all measurement types in both auscultatory and oscillometric cohorts.
  - PPG signal length of 3,250, allowing for multiple instances to be extracted from each signal.
  - Dataset size: $64,401$ (75% training, 12.5% validation, and 12.5% testing). It should be noted that dataset size is not treated as a hyperparameter in this work. It is only reported in Table 19 (*Num waves* column) because it may change depending on the choice of other hyperparameters, such as the optical quality threshold.
- Model:
  - AlexNet (the larger version). Note that XResNet also performed well and may have the ability to outperform AlexNet with a more comprehensive hyperparameter tuning.
  - A dropout rate of 0.75.
- Training:
  - Normalising signals individually by z-score.
  - Batch size of 64.
  - Epochs of 32, saving the best-performing model.

Table 20: Classification performance metrics of the optimal AlexNet for training, validation, and testing sets.

Dataset	Loss	Accuracy	Precision	Recall (sensitivity)	F1 score	Specificity	AUC
Training	0.6007	0.7123	0.7094	0.6836	0.6916	0.7380	0.7838
Validation	0.6167	0.7044	0.7214	0.7054	0.7108	0.7019	0.7635
Testing	0.5787	0.7101	0.7290	0.7349	0.7287	0.6817	0.7714

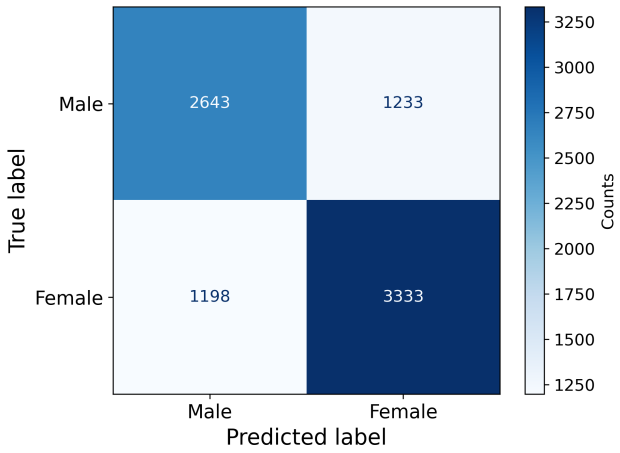


Figure 7: Confusion matrix of predictions for the testing set.

– Initial learning rate of 10^{-4} decayed to a minimum value of 10^{-8} by cosine annealing. The optimal learning rate schedule was selected based on a few less formal hyperparameter tuning attempts not shown in Table 19 for brevity.

Table 20 reports performance metrics for the trained classifier. For each instance, class probabilities were estimated by taking 50 samples from the predicted Gaussian logit distributions, applying *softmax*, and averaging over 50 Monte Carlo Dropout (MCD) passes. Final labels refer to classes that had the highest probability per instance. Taken together, the metrics indicate that the model attains approximately 72% accuracy on the test set, with little to no evidence of overfitting. The proportion of the true and false male and female predictions are presented as a confusion matrix in Figure 7. The model is accurate in recognising females and males 73.6.2% and 68.2% of the times, respectively.

3.2.7 Discussion

Our results show that deep learning models can classify PPG signals by biological sex with accuracy exceeding 70%, despite the noisy nature of the Aurora-BP dataset. This suggests that PPG signals embed features associated with biological sex and, consequently, sex must be handled carefully when training models for downstream tasks.

In related studies, it is advisable to examine: (1) how sex imbalance in datasets affects performance; (2) whether models trained on pooled data generalise equally across sexes; and (3) whether disparities are more pronounced in subpopulations with specific conditions or characteristics.

Any ability to classify by sex may reflect a combination of physiological differences and measurement artefacts. For example, colder peripheral circulation, which is more common in women, can degrade PPG quality. Regardless of origin, such differences must be accounted for to prevent bias in diagnostic accuracy and care pathways.

Complementary studies may consider:

- Assessing sex-classification accuracy in data subsets, e.g., across age groups or measurement types.
- Applying more aggressive filtering (e.g., excluding participants with pre-existing medical conditions) and optimising preprocessing to improve training data quality.
- Investigating the effect of biological sex on relevant downstream tasks.
- Conducting more comprehensive hyperparameter tuning to leverage the higher capacity of *XResNet*.
- Employing larger datasets to better exploit transformer-based models, which can potentially outperform CNNs.

3.3 Classification of SPAR images derived from PPG signals by biological sex

This task evaluates whether a simple deep learning model is sensitive to features related to biological sex in Symmetric Projection Attractor Reconstruction (SPAR) images generated from PPG signals. The work in this section follows what has been presented in Section 3.2.

3.3.1 Dataset and data processing

The dataset of choice for this task was the Aurora-BP dataset. An in-depth explanation has already been given in Section 3.2.1, together with the introduction of the quality threshold criteria applied to data selection. Data pre-processing was carried out in the same way as described in Section 3.2.2, with the exception that the full length of each signal was considered and analysed as a whole, without considering multiple segments from a single recording. This decision was due to the nature of the Symmetric Projection Attractor Reconstruction (SPAR) method, which will be explained more in depth in Section 3.3.2.

3.3.2 The SPAR method

The Symmetric Projection Attractor Reconstruction (SPAR) method [31] is a non-fiducial points-based method that combines mathematics and cardiovascular physiology to quantify pulse waveform morphology and variability. Given a raw pulse wave signal, the average cycle length is first estimated for the selected time window. Next, a three-dimensional (3D) attractor is generated using three delay coordinates, each separated by one-third of the average cycle length. The 3D attractor is then projected onto a two-dimensional (2D) plane normal to the unit vector and converted into a density plot (Figure 8a). This plot is the final attractor image used to train and validate CNNs in this work.

3.3.3 Deep learning model

The CNN used in this study was based on a simplified TinyVGG architecture originally presented by Wang et al. [32]. It comprises two convolutional blocks followed by a final classification layer, as shown in Figure 8b.

3.3.4 Model training

AuroraBP data from both protocols was considered for analysis, and the data was stratified according to the *"optical_quality"* value. Thresholds of 0.65, 0.75, 0.85 and 0.95 were considered separately for four different training scenarios, which were then combined with a varying number of training epochs (namely 25, 50, 75 and 100 epochs of training). The aim of these different settings was to assess model performance in relation to data quality and training time. Results for the different scenarios are presented in Figures 9 and 10

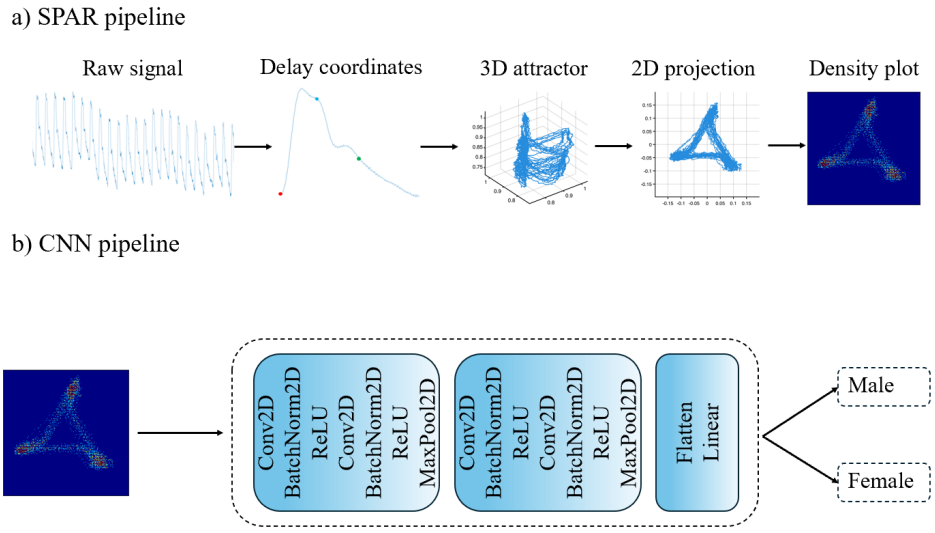


Figure 8: Symmetric-Projection-Attractor-Reconstruction (SPAR) and Convolutional Neural Network (CNN) pipelines. (a) The SPAR pipeline processes a raw signal into a SPAR attractor image (density plot). (b) Schematic of the CNN architecture used for classifying SPAR attractor images as hypertensive or non elevated.

3.3.5 Discussion

This study employed predefined training and test splits of the AuroraBP dataset, which were stratified by sex, to evaluate sex classification from photoplethysmography (PPG) signals using the SPAR method. Although our results indicate reasonable classification performance of around 70%, these findings should be interpreted with caution. The dataset splits were not controlled or disaggregated with respect to several key confounding variables, including age, disease status (hypertensive, controlled hypertensive, or non-hypertensive), active medication use, time of day and body posture during acquisition, and sampling environment (laboratory versus community settings).

As a result, it is not possible to determine whether one or more of these factors were unevenly distributed across the training and test sets, potentially contributing to the observed classification performance.

Furthermore, visual inspection of the resulting SPAR attractor images did not reveal clear or systematic differences between sexes, leaving the specific signal characteristics driving classification performance uncertain. Consequently, the present analysis cannot establish whether the model is detecting sex-specific physiological differences or exploiting correlated confounding factors. Future work should explicitly control for, or stratify by, these variables to isolate the contribution of sex to PPG-derived features. Only through such rigorously controlled analyses can robust conclusions regarding sex-specific differences in PPG signals be drawn.

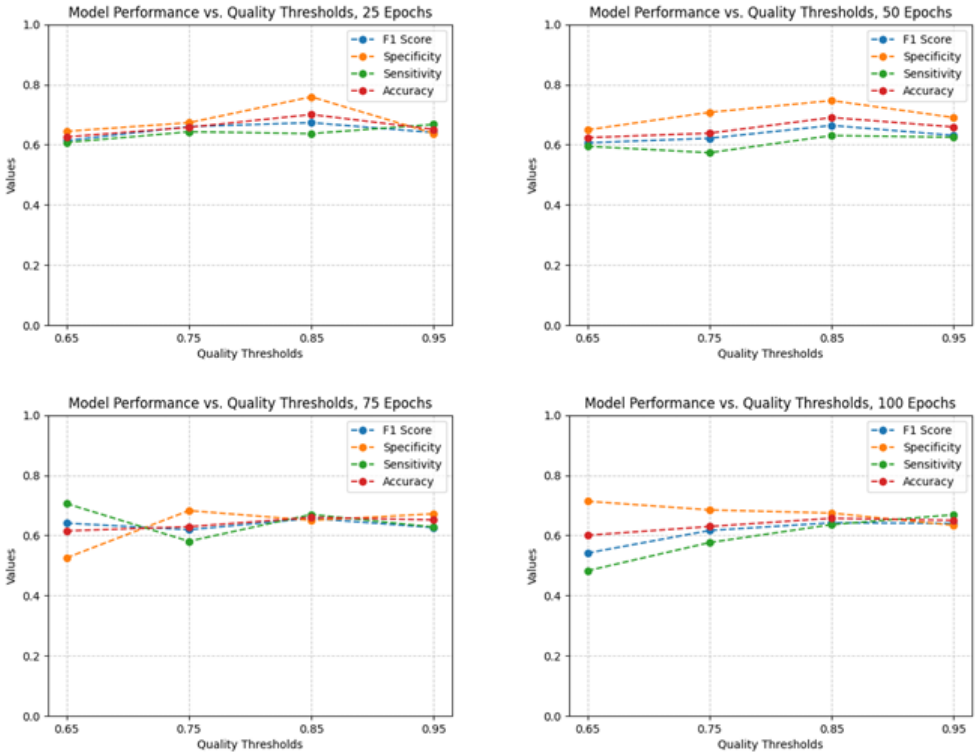


Figure 9: Model performance for different quality thresholds, increasing training time. Top to bottom, left to right: 25 epochs, 50 epochs, 75 epochs and 100 epochs.

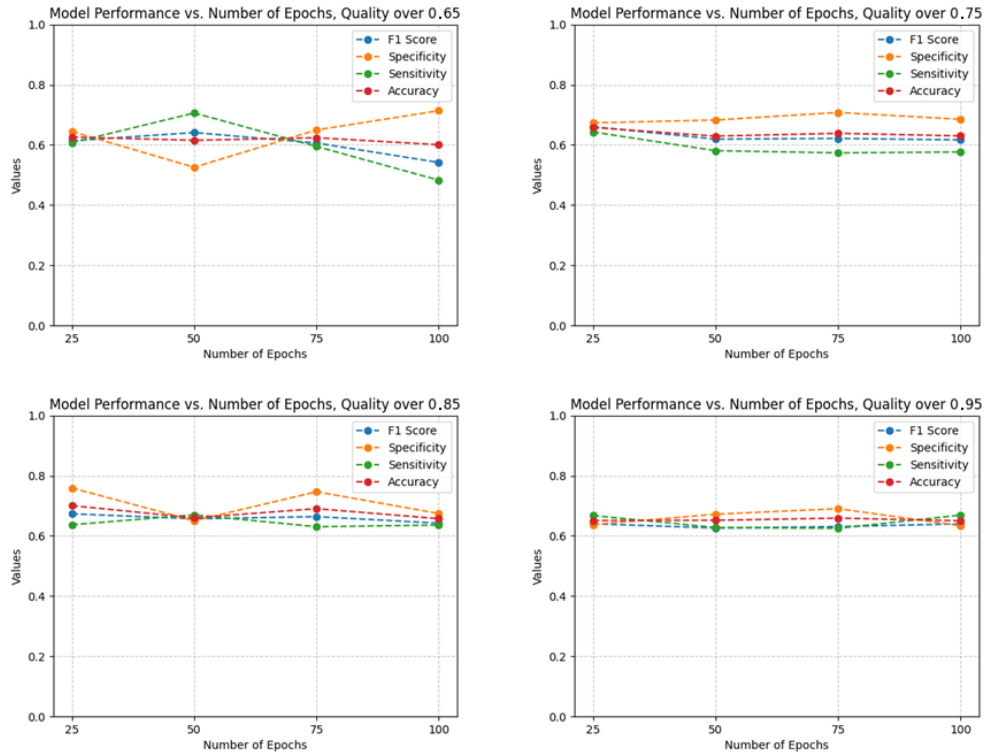


Figure 10: Model performance for different training times, increasing quality thresholds. Top to bottom, left to right: optical quality over 0.65, optical quality over 0.75, optical quality over 0.85 and optical quality over 0.95.

4 A1.3.3 – Out-of-distribution evaluation

KTU, with support from NPL, UCAM and KCL will identify datasets from A2.1.4 that are suitable for out-of-distribution evaluation. KTU will develop parametric noise models for PPG data that can be used to artificially increase the noise level of the original datasets. For both kinds of out-of-distribution data, NPL and Uni-Oldenburg will carry out the following analyses:
First, NPL will repeat the analyses from A1.3.1 for evaluation on out-of-distribution data to study the robustness of both accuracy as well as uncertainty with the metrics of A1.1.6 and A1.2.6, respectively, for at least two selected models from A1.1.2-A1.1.5. Second, Uni-Oldenburg will implement a common evaluation framework to assess the quality of uncertainty estimates based on the ability to distinguish in-distribution from out-of-distribution data.

While there are several scalable uncertainty quantification approaches that can be implemented on BP regression models [2, 33, 34], the practical utility of uncertainty estimates depend on their reliability (i.e. how well their magnitude correlates with the underlying doubt in a given prediction). For the model-based and post-hoc approaches considered, uncertainty reliability may be optimised using predictions from a hold-out dataset. E.g., model-based UQ techniques like Monte Carlo Dropout depend on model parameterisation, which can be optimised using a grid search against uncertainty reliability while post-hoc methods fit the predictions from a hold-out calibration set to optimise uncertainty reliability. With this approach, the reliability of the uncertainty estimates are optimised according to the model's behaviour on this hold-out dataset. There is no guarantee that the uncertainty reliability optimised for this dataset will generalise to other unseen data. Therefore, the evaluation of the generalisability of uncertainty reliability is an essential aspect of assessing the reliability of model performance.

Here, we investigate how the generalisability of uncertainty reliability may vary depending on the sources of uncertainty considered by the model (e.g. epistemic uncertainty alone, or both aleatoric uncertainty and epistemic uncertainty), as well as the type of UQ technique implemented (model-based approaches like Monte Carlo Dropout, or post-hoc approaches like temperature scaling). We also assess how post-hoc methods may influence generalisability by applying them to the uncertainties estimated with model-based approaches. These are assessed with splits of the datasets used for training (independent and identically distributed (IID) evaluation), as well as assessments on external out-of-distribution (OOD) datasets. We implement a large range of metrics (provided by the uncertainty evaluation framework) to provide a comprehensive assessment of uncertainty reliability.

We compare our results with other generalisbility studies performed on more generic tasks. E.g., [35] that reported how post-hoc methods exhibited poorer generalisability than model-based approaches on image classification tasks.

4.1 Atrial fibrillation detection using out-of-distribution data

4.1.1 Out-of-distribution data generation and parametric noise modelling

This task investigated the robustness of the atrial fibrillation detection performance and associated uncertainty estimates under OOD conditions in simulated PPG signals [36]. The OOD data were generated by augmenting the simulated artifact-free PPG signals with parametrised models of realistic PPG artifacts [37], including device displacement, forearm motion, hand motion and poor contact of different durations (4 s, 8 s, 12 s) as shown in Figure 11. This allowed a systematic assessment of how different types and durations of artifacts affect both the accuracy and uncertainty of the classification. Figure 12 shows the input signals, main processing stages, and classification outputs of the two AF detection models.

Each artifact type is generated by filtering white noise with a high-order FIR filter whose frequency response is controlled by a spectral slope sampled from a Gaussian distribution and scaling the amplitude using a Gamma-distributed normalised RMS [37]. Table 21 summarises the parameters used for artifact generation.

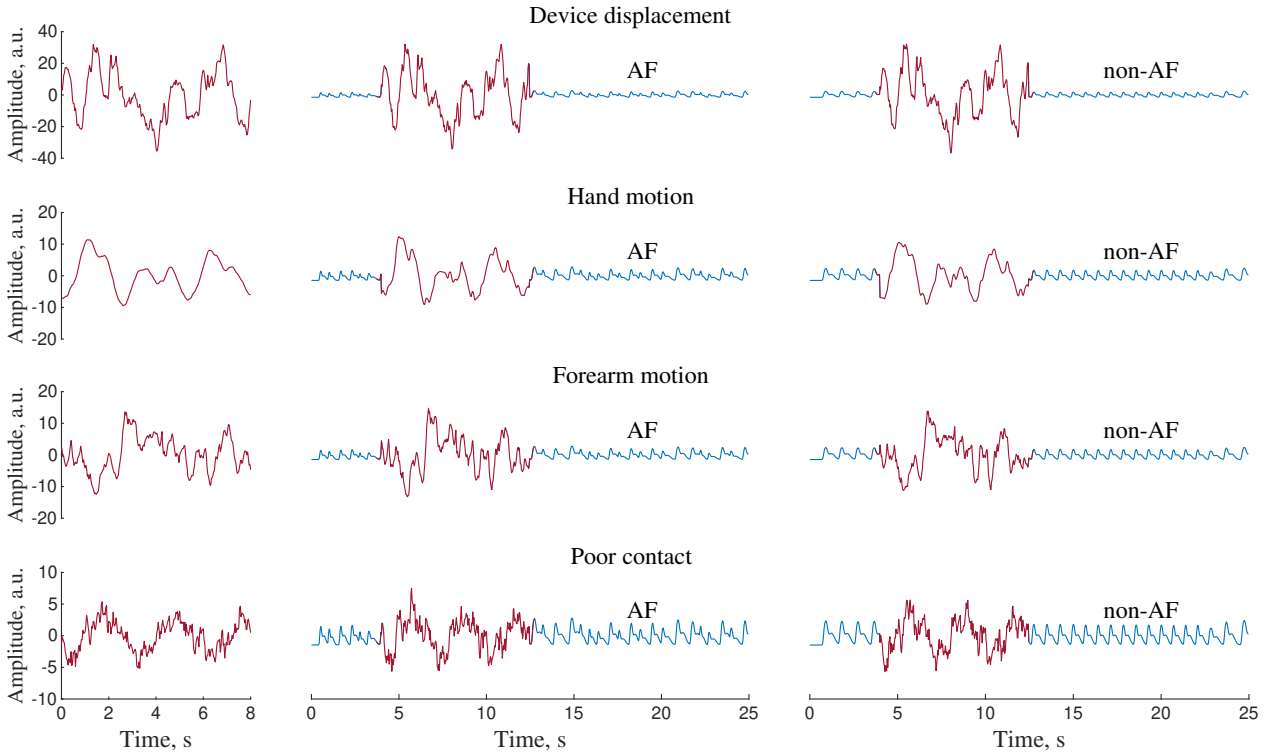


Figure 11: Different types of simulated artifacts in the investigation dataset and their representation in the PPG signal. An inserted 8 s artifact is shown in red.

Table 21: Parameters of spectral slope and normalised RMS for artifact types.

Artifact type	$\hat{\mu}_i$	$\hat{\sigma}_i$	$\hat{\alpha}_i$	$\hat{\beta}_i$	$\hat{\alpha}_i/\hat{\beta}_i$
Device displacement	-32.34	6.03	0.88	0.04	22.0
Forearm motion	-29.39	5.71	1.44	0.31	4.65
Hand motion	-25.45	4.13	1.40	0.45	3.11
Poor contact	-18.12	4.10	2.01	1.24	1.62

4.1.2 Atrial fibrillation detection models

Two AF detection models were evaluated. The first model was a convolutional neural network (CNN) operating directly on fixed-length PPG waveform segments ($\mathcal{D}_r$). The second model was a feed-forward neural network classifier operating on AF-related rhythm irregularity features together with a signal quality index reflecting the proportion of usable signal ($\mathcal{D}_f$) [38]. The generalised pipelines of both detectors are shown in Figure 12.

4.1.3 Datasets and preprocessing

The AF training cohort included 15 patients (age 72.9 ± 8.9 years; BMI $28.3 \pm 5.9\,\mathrm{kg\,m^{-2}}$; observation period 21.3 ± 3.8 h), while the non-AF cohort included 19 patients (age 67.5 ± 10 years; BMI $28.0 \pm 5.0\,\mathrm{kg\,m^{-2}}$; observation period 21.6 ± 3.1 h) [38]. After balancing by record (not by patient), the training set contained 37,102 AF and 37,102 non-AF 25 s segments, and the validation set contained 8,430 AF and 8,430 non-AF segments. Separate patients were used for the training and validation datasets. No signals were excluded based on signal quality.

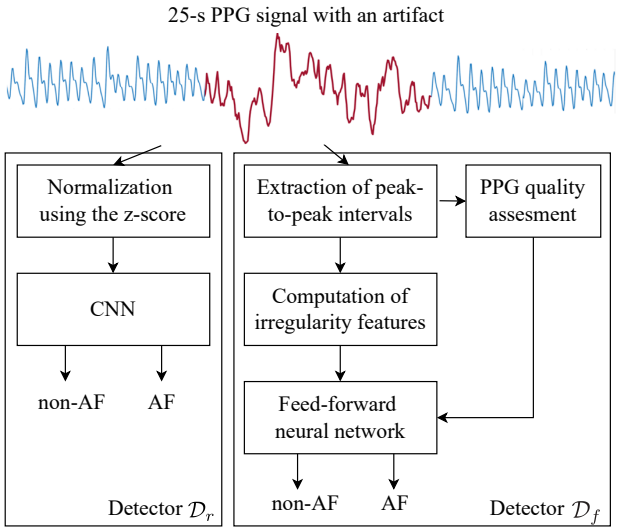


Figure 12: Block diagrams of the AF detectors under investigation: Detector $\mathcal{D}_r$ that uses the PPG signal as input, and detector $\mathcal{D}_f$ that uses AF-related rhythm irregularity features with a signal quality index.

A test set of 20,000 artifact-free simulated PPG signals was generated (10,000 AF and 10,000 non-AF), each 25 s long. For each artifact type, artifacts of 4 s, 8 s, and 12 s were added to all baseline signals, resulting in a total of $20{,}000 + 20{,}000 \times 4 \times 3 = 260{,}000$ simulated PPG signals.

PPG signals were preprocessed using a zero-phase Butterworth bandpass filter [39] with a passband of 0.5–5 Hz. For the waveform-based CNN detector, each 25 s PPG segment was resampled to 32 Hz and represented by 800 samples, which were normalised using z-score standardisation prior to inference.

4.1.4 Uncertainty evaluation methods

Uncertainty was quantified using three complementary approaches: threshold-based error rate analysis at multiple decision thresholds, conformal prediction to identify confident, ambiguous, or rejected predictions, and Monte Carlo Dropout, which estimates predictive entropy by generating multiple stochastic forward passes.

4.1.5 Out-of-distribution effects on prediction distributions

Figure 13 shows how the OOD artifacts affect the distribution of the model output. Although both detectors exhibit bimodal distributions for artifact-free signals, increasing artifact duration leads to flattened and shifted distributions, particularly for the feature-based detector. Device displacement causes the largest deviation from the in-distribution prediction patterns.

4.1.6 Out-of-distribution effects on detection performance

Figure 14 shows the effect of artifacts on sensitivity and specificity. Sensitivity decreases markedly with the duration of the artifact, especially with device displacement, while specificity trends differ between detectors. In general, $\mathcal{D}_f$ performs better under short artifacts, while $\mathcal{D}_r$ is less sensitive to artifact duration. Figure 15 presents threshold-based error rates, demonstrating how increasing decision thresholds amplifies performance differences under OOD conditions.

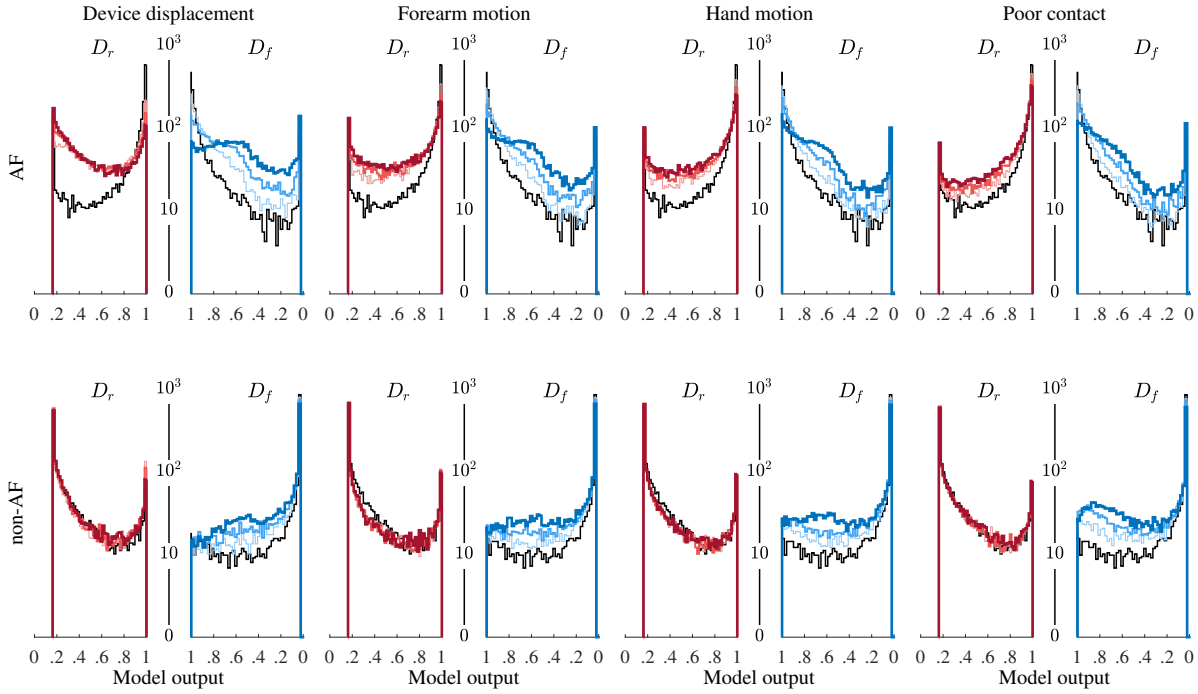


Figure 13: Effect of artifacts on prediction frequency of detectors $\mathcal{D}_r$ and $\mathcal{D}_f$ for PPG signals with AF and non-AF. Black represents the distribution for signals without added artifacts, and progressively darker colour shades indicate distributions for longer artifact durations, ranging from 4 s to 12 s in 4 s increments.

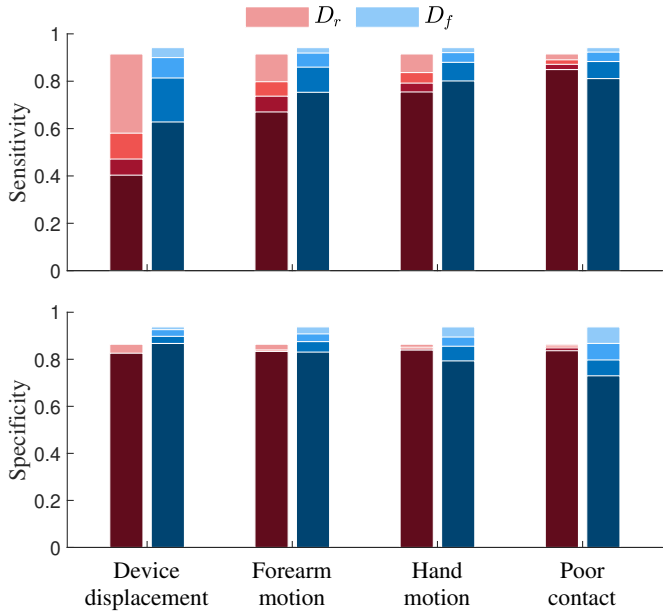


Figure 14: Effect of artifacts on sensitivity and specificity of $\mathcal{D}_r$ and $\mathcal{D}_f$. Sensitivity is defined as the proportion of correctly detected AF cases, while specificity is the proportion of correctly detected non-AF cases. Shades ranging from lighter to darker indicate increasing artifact durations, starting from no artifact and progressing up to 12 s in 4 s increments.

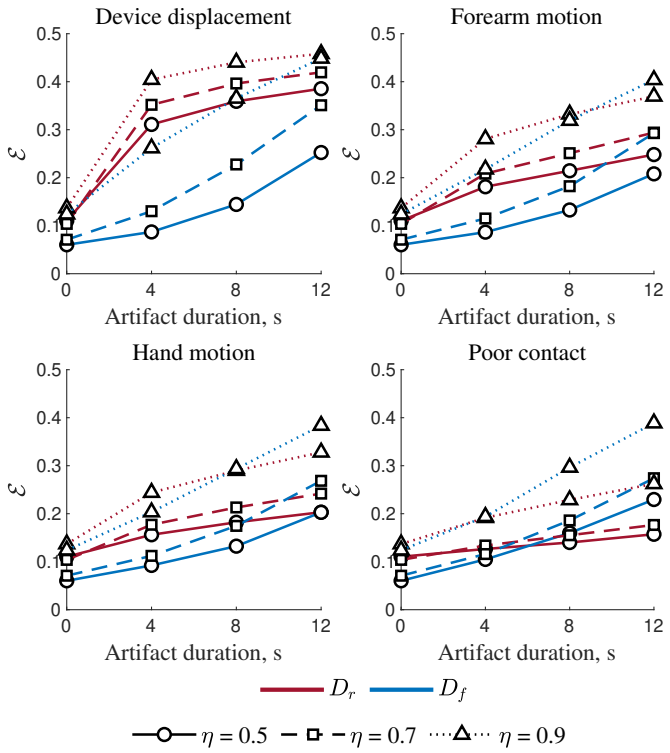


Figure 15: Threshold-based error rate $\mathcal{E}$ of $\mathcal{D}_r$ and $\mathcal{D}_f$ under different artifact types and durations.

4.1.7 Out-of-distribution effects on uncertainty

Figure 16 shows the proportion of uncertain predictions obtained using conformal prediction at different coverage levels. For $\mathcal{D}_r$, uncertainty remains relatively stable throughout the duration of the artifact, whereas for $\mathcal{D}_f$ uncertainty increases with the severity of the artifact. Table 22 summarises the composition of conformal prediction sets for artifact-free data at different coverage levels. Monte Carlo Dropout results in Figure 17 show that predictive entropy increases with dropout rate and artifact severity, with $\mathcal{D}_r$ consistently exhibiting higher entropy than $\mathcal{D}_f$.

4.1.8 Reduction of false positives

Uncertainty estimates can be used to suppress unreliable decisions under OOD conditions. Figure 18 demonstrates that discarding uncertain predictions using conformal prediction substantially reduces false positives,

Table 22: Composition of the conformal prediction sets and proportion of uncertain decisions $\mathcal{U}$.

Detector	γ	AF	non-AF	Both	Neither	$\mathcal{U}$
$\mathcal{D}_r$	0.99	2226	0	7774	0	0.78
	0.95	4517	2818	2665	0	0.27
	0.90	5131	4602	267	0	0.03
$\mathcal{D}_f$	0.99	1605	4089	4306	0	0.43
	0.95	4924	4908	168	0	0.02
	0.90	4584	4725	0	691	0.07

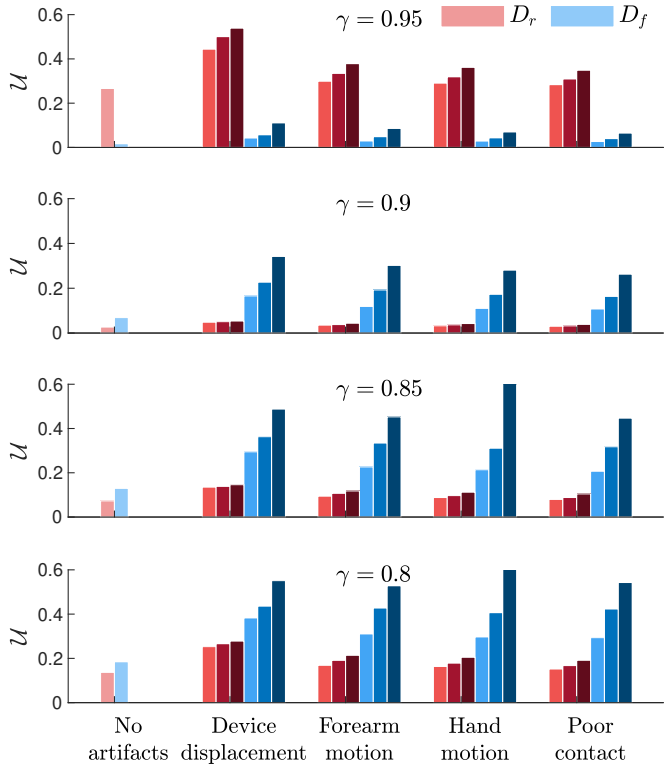


Figure 16: Proportion of uncertain predictions $\mathcal{U}$ for different artifact types and durations for different coverage levels γ. Shades ranging from lighter to darker indicate increasing artifact durations, starting from 4 s and progressing up to 12 s in 4 s increments.

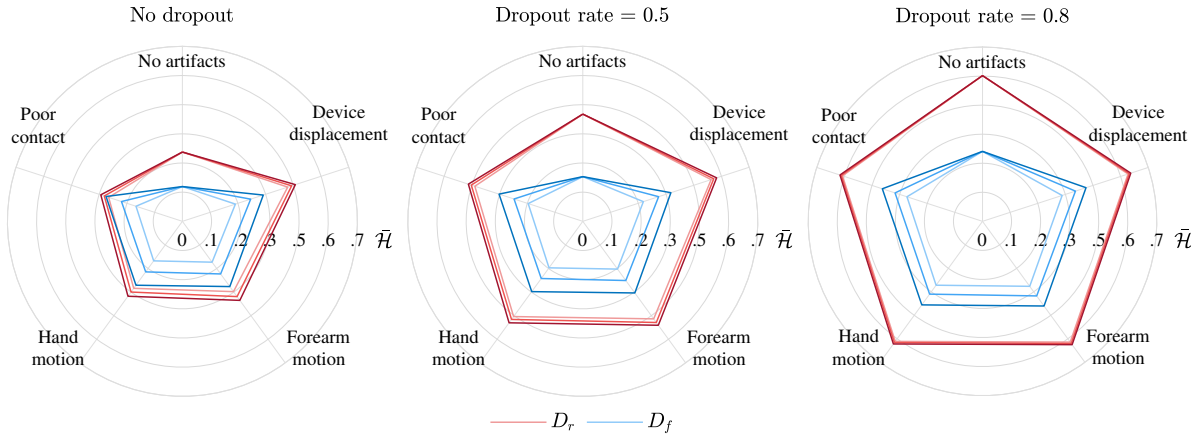


Figure 17: Effect of dropout on the predictive entropy $\bar{\mathcal{H}}$ of $\mathcal{D}_r$ (red) and $\mathcal{D}_f$ (blue). Shades ranging from lighter to darker indicate increasing artifact durations, starting from 4 s and progressing up to 12 s in 4 s increments.

particularly for severe artifacts.

4.1.9 Discussion

The OOD evaluation demonstrates that AF detection performance and uncertainty are strongly influenced by both the type and duration of PPG artifacts. While both detection models experience performance degra-

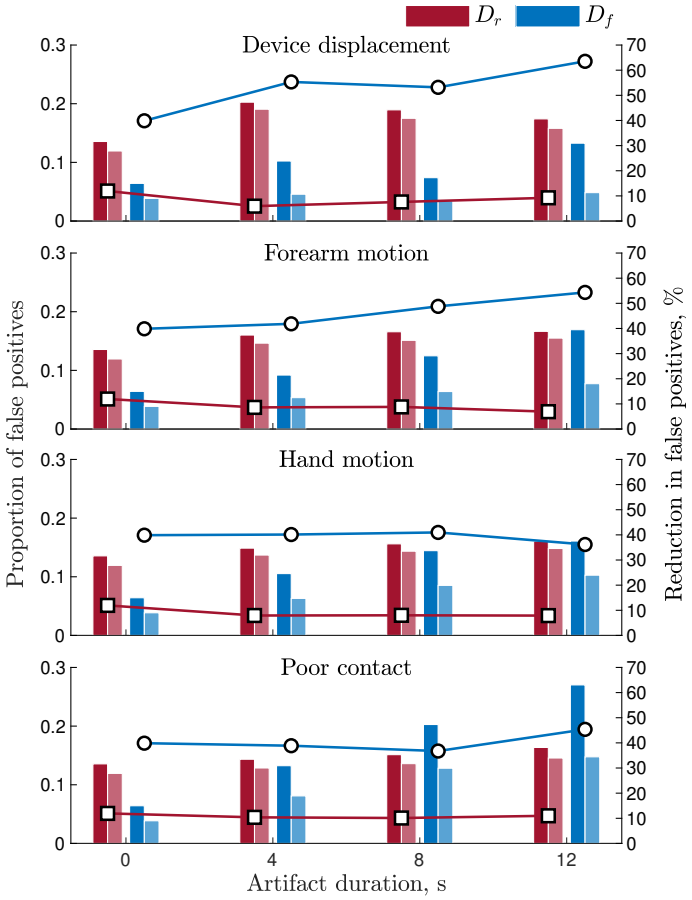


Figure 18: Reduction in false positives (lines) obtained by discarding decisions flagged as uncertain using conformal prediction with $\gamma = 0.9$. Dark and light shaded bars represent the proportion of false positives before and after applying conformal prediction, respectively.

dation under OOD conditions, their behaviour differs substantially, highlighting complementary strengths and limitations.

Across all experiments, device displacement emerged as the most detrimental artifact type, leading to the largest reduction in sensitivity and the most pronounced changes in prediction distributions. This is consistent with the large amplitude and broadband nature of this artifact, which distorts the pulsatile component of the PPG signal. In contrast, forearm motion, hand motion, and poor contact generally caused more gradual degradation, with effects that depended on artifact duration and model type.

The waveform-based CNN detector ($\mathcal{D}_r$) exhibited an abrupt drop in performance when artifacts were introduced but showed relatively limited sensitivity to increasing artifact duration. This suggests that once the waveform structure deviates sufficiently from the training distribution, the model performance degrades rapidly, after which additional corruption has a smaller marginal effect. The feature-based detector ($\mathcal{D}_f$), on the other hand, demonstrated better robustness to short-duration artifacts but became increasingly unreliable as artifact duration increased, reflecting its dependence on accurate pulse detection and rhythm feature extraction.

Importantly, the results show that classification accuracy alone does not fully capture model reliability under OOD conditions. Changes in uncertainty estimates did not always mirror changes in sensitivity or specificity, particularly for longer artifacts. Conformal prediction proved effective in explicitly identifying signals for which reliable classification was not possible, enabling a clear separation between confident predictions and those affected by distributional shift. This behaviour is especially valuable in practical screening scenarios,

where suppressing unreliable outputs is preferable to producing potentially misleading classifications.

Monte Carlo Dropout further highlighted differences in internal model behavior. The consistently higher predictive entropy observed for the waveform-based model indicates greater model uncertainty under corrupted signal conditions, whereas the feature-based model produced lower entropy values, reflecting its lower-dimensional and more constrained representation. These findings underline that different uncertainty estimation methods capture complementary aspects of model reliability and should not be interpreted in isolation.

From an application perspective, the reduction in false positives achieved by discarding uncertain predictions demonstrates the practical benefit of uncertainty-aware decision-making. Under severe artifact conditions, particularly device displacement, uncertainty-based filtering substantially reduced false alarms without requiring changes to the underlying classifiers. This supports the integration of uncertainty estimation into wearable AF screening pipelines as a mechanism for improving reliability under real-world conditions.

Overall, the OOD evaluation shows that uncertainty-aware analysis provides essential information beyond conventional performance metrics. The complementary behaviour of waveform-based and feature-based detectors suggests that hybrid or adaptive strategies—selecting or weighting model outputs based on signal quality and uncertainty—could further improve robustness. These findings directly support the objectives of robust AF detection under realistic, artifact-prone conditions and provide a foundation for uncertainty-guided decision strategies in downstream tasks.

4.2 Blood pressure prediction using out of distribution data

4.2.1 Methods

We return to the problem of predicting blood pressure from PPG signals. Models predict both systolic blood pressure (SBP) and diastolic blood pressure (DBP) (see [1] for more details).

Table 23 summarises the datasets used in this study. Following [1], models are trained on PulseDB [40] (in-distribution (ID) datatset) and evaluated on external out-of-distribution datasets (OOD) [41]. Based on prior evidence that AAMI-Vital of PulseDB, improves OOD generalisation [42], we decided to use it as the in-distribution training set. The OOD datasets include Sensors, UCI, BCG, and PPGBP datasets. Further details are provided in [41].

Table 23: Summary of the datasets utilised in this study.

	ID dataset	**OOD dataset**			
Metric	**AAMI-VitalDB**	**Sensors**	**UCI**	**BCG**	**PPGBP**
Subjects	1,409	1,195	10,793	40	218
Total Duration (h)	∼1489	∼15	∼570	∼4	<1
Segments (count , length)	536,358 , 10s	11,102 , 5s	410,596 , 5s	3,063 , 5s	619 , 2.1s
Age (years, mean ± SD)	58.75 ± 14.91	57.1 ± 14.2	unknown	34.2 ± 14.5	56.9 ± 15.8
Gender (% female)	41.98	40.2	unknown	55.5	53.1
SBP (mmHg, mean ± SD)	115.65 ± 18.95	134.36 ± 21.78	131.57 ± 11.16	120.99 ± 15.29	128.02 ± 20.50
DBP (mmHg, mean ± SD)	63.05 ± 12.07	65.37 ± 10.51	66.79 ± 10.48	67.23 ± 9.30	71.91 ± 11.20

4.2.2 Results

Table 24 shows that total uncertainty is larger for OOD inputs. Modelling aleatoric uncertainty with a Gaussian negative log likelihood (GNLL) loss function provides better calibrated uncertainties (i.e. lower expected normalised calibration error (ENCE) scores) and mostly superior predictive performance than implementing Monte Carlo Dropout (MCD) with a mean squared error (MSE) loss, where the majority of the uncertainty for GNLL-optimised models is aleatoric.

Table 24: Comparison of prediction accuracy and uncertainty for GNLL and MSE losses using a dropout rate of 4%. All models are trained on the AAMI-Vital dataset.

Test Dataset	**MAE (mmHg) ↓**		**Uncertainty SBP/DBP (mmHg) ↓**						**ENCE ↓**	
			Epistemic		Aleatoric		Total			
	SBP	DBP	SBP	DBP	SBP	DBP	SBP	DBP	SBP	DBP
GNLL (Dropout = 4%)										
AAMI-Vital (ID)	18.26	12.53	0.83	0.00	9.80	7.22	9.83	7.22	1.28	1.10
BCG (OOD)	12.54	9.60	1.51	0.90	12.17	9.22	12.26	9.27	0.29	0.31
Sensors (OOD)	17.76	14.73	1.81	1.05	14.39	10.73	14.51	10.78	0.57	0.58
UCI (OOD)	18.91	13.81	1.99	1.14	15.01	11.15	15.14	11.21	0.62	0.46
PPGBP (OOD)	16.72	12.95	2.28	1.31	15.18	11.49	15.35	11.57	0.40	0.38
MSE (Dropout = 4%)										
AAMI-Vital (ID)	18.92	12.43	1.73	0.71	–	–	1.73	0.71	23.61	24.97
BCG (OOD)	16.74	9.20	1.21	0.71	–	–	1.21	0.71	17.26	17.10
Sensors (OOD)	17.90	13.10	1.81	0.69	–	–	1.81	0.69	18.95	22.43
UCI (OOD)	19.39	11.65	1.21	0.71	–	–	1.21	0.71	19.89	20.10
PPGBP (OOD)	20.94	10.26	1.53	0.94	–	–	1.53	0.94	15.79	12.83

Uncertainty reliability as assessed with the ENCE is better for OOD datasets for both modelling approaches. Predictive performance (as measured by the mean absolute error (MAE)) is also better for some OOD datasets than the ID data. The larger scale and duration of the ID test data likely provides a more diverse range of outputs compared to other datasets that may result in poor generalisability, where there appears to be a negative correlation between predictive performance and the number of segments considered across the datasets. Similarly, the ENCE for SBP has a positive correlation with the number of segments, which may suggest that variability across a larger number of samples results in poorer uncertainty reliability. However, this is not consistent with the trend observed for DBP.

There is a positive correlation between SBP mean, and SBP ENCE, and a negative correlation for DBP mean and DBP ENCE; this suggests that shifts in label distribution do not affect uncertainty reliability in a consistent manner.

While there are few consistent trends, it is apparent that uncertainty reliability does not generalise here for either modelling approach, and that there is considerable variability across the datasets. These results suggest that the composition/properties of previously unseen data should be carefully considered when using uncertainty quantification. It is not clear as to whether epistemic or aleatoric contributions alone are better indicators of OOD nature of the data.

4.3 Sleep apnea detection using out-of-distribution data

4.3.1 Introduction to apnea-related definitions

Apnea episode - a lack of an airflow for at least 10 s with an associated SpO2 decrease of at least 3–4 % and/or arousals.

Hypopnea episode - a 50 % or greater reduction in airflow for at least 10 s with an associated SpO2 decrease of at least 3–4 % and/or arousals.

Sleep apnea segment - 1 min signal segment containing at least 1 apnea/hypopnea episode.

Sleep apnea burden - a proportion of sleep apnea segments per subject signal.

4.3.2 Data: In-distribution subsets

The Multi-Ethnic Study of Atherosclerosis (**MESA**) dataset [43] comprises data from 2055 patients (53.63 % female), aged between 54 and 95 years. It contains a total of 16,300 hours of fully annotated overnight polysomnography (PSG) recordings collected during a sleep study (2010-2012), funded by the National Heart, Lung, and Blood Institute. The average age of subjects is 69.37±9.12 years. Participants included in the study had not used treatments for sleep apnea, such as continuous positive airway pressure or oxygen devices, for more than a month prior to the study, or only used it less than once a week. The dataset represents a cohort, featuring participants from a variety of racial/ethnic groups, including White, African American, Hispanic, and Chinese American individuals.

PSG recordings were obtained by the patients at home using the Compumedics Somte monitoring system (Compumedics Ltd., Abbotsville, Australia). The dataset includes signals such as PPG, ECG, electrooculography (EOG), electroencephalography (EEG), electromyography (EMG), SpO2, nasal airflow, snoring, abdominal and thoracic movements. The sampling rate of PPG and ECG signals is 256 Hz. PPG signals were recorded from the finger using the Nonin 8000 sensor. In addition, time labels and durations of apnea and hypopnea episodes in the respiratory flow signal are annotated. However, this database is not open-access and is only available on request. The MESA dataset has already been used in the following state-of-the-art studies [44, 45].

4.3.3 Data: Out-of-distribution subsets

The open-access Obstructive Sleep Apnea Stroke Unit Dataset (**OSASUD**) polysomnography dataset [6] involves 30 patients (36.67 % female) admitted to the stroke unit at the Clinical Neurology Unit of Udine University Hospital between 2019-2020 due to suspected cerebrovascular events, such as ischemic and hemorrhagic stroke, or transient ischemic attack. The dataset provides information of subject age (68.97 ± 11.22 years), the body mass index (29.17 ± 5.74 kg/m^2), the apnea-hypopnea index (25.43 ± 21.31 counts/h), and signal duration (8.91 ± 1.47 h). Patients of age <18 years, those unable to comply with standard requirements for PSG monitoring, individuals with severe aphasia impacting understanding or consent, and those at high risk of alcohol or drug withdrawal syndrome were excluded from the study. Notably, conditions such as obesity, diabetes mellitus, atrial fibrillation and other cardiac diseases did not serve as exclusion criteria.

The dataset includes PSG signals such as finger photoplethysmography (PPG), electrocardiography (ECG), oxygen saturation (SpO2), respiratory rate, perfusion index, heart rate, nasal airflow, snoring, 3-axis accelerometer, abdominal and thoracic movements. The sampling rate of PPG and ECG signals is 80 Hz, while respiratory rate and SpO2 time series are sampled at 1 Hz. In addition, SpO2 values < 50 % or > 100 % were identified as artifacts and marked as null. This method was similarly applied to heart rate measurements < 20 bpm or > 200 bpm, as well as respiratory rate values < 5 bpm or > 40 bpm.

Recorded PSG data underwent thorough analysis using the Embla RemLogic software, version 3.4.1.2371 (Natus Medical Inc., Pleasanton, CA, USA), which facilitates signal processing, detailed examination, and annotation procedure. The OSASUD data were annotated by a trained sleep medicine physician based on the sleep scoring guidelines established by the American Academy of Sleep Medicine, identifying occurrences of central, obstructive, mixed apnea, and hypopnea events (1 s granularity). The OSASUD dataset has already been used in the following state-of-the-art studies [46, 47, 48, 49]. A comparison of the two datasets used is provided in Table 25.

In our experiments, the MESA dataset was used for training, validation, and for in-distribution (ID) testing. As already mentioned, this dataset consists of 2055 subjects, but we only used those with the highest polysomnography study quality grade and the apnea-hypopnea index AHI > 25 (due to class imbalance issues between true positives and true negatives). For the out-of-distribution (OOD) case, the OSASUD dataset was used. Moreover, PPG features were standardised for each subject individually in order to handle OOD data. Apnea burden distributions in ID and OOD datasets are provided in Figure 19.

Table 25: Dataset differences between MESA and OSASUD.

Dataset differences	#1 MESA	#2 OSASUD
Population shift	Diverse population: White, African American, Hispanic, and Chinese American participants	Racial/ethnic groups not considered
Population shift	General population	After-stroke patients
Environmental conditions	Home: full overnight unattended polysomnography, 6 hospitals	Hospital: the Clinical Neurology Unit, Udine, Italy
Device drift	Compumedics Somte monitoring system (Compumedics Ltd., Australia)	Embla RemLogic (Natus Medical Inc., Pleasanton, USA)
PPG sensor drift	Nonin 8000 sensor (finger PPG, sampling rate – 256 Hz)	Embletta polysomnograph + Mindray monitoring system (finger PPG, sampling rate – 80 Hz)

4.3.4 Methods: PPG feature estimation and labelling

As model inputs, five PPG pulse wave features were used (see Figure 20), namely pulse wave interval, T_p; peak-to-peak amplitude, A_{pp}; maximum slope of PPG, S_{max}; systolic time, ST; diastolic time, DT. In addition, oxygen saturation SpO2 time series were used separately as a model input, and also together with PPG pulse wave features. An illustration of signal segment labelling is showed in Figure 21.

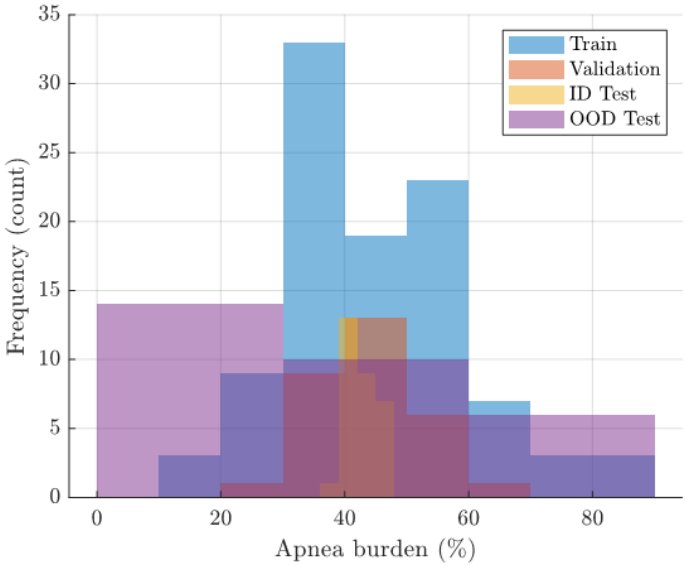


Figure 19: Apnea burden distributions in ID and OOD datasets. Training, validation and ID testing datasets show a symmetric distribution of data according to apnea burden, while the OOD testing dataset shows an asymmetric distribution.

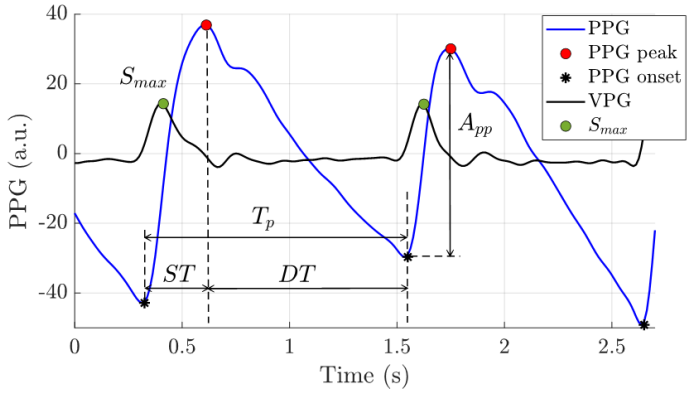


Figure 20: PPG features estimated from PPG pulse waveforms.

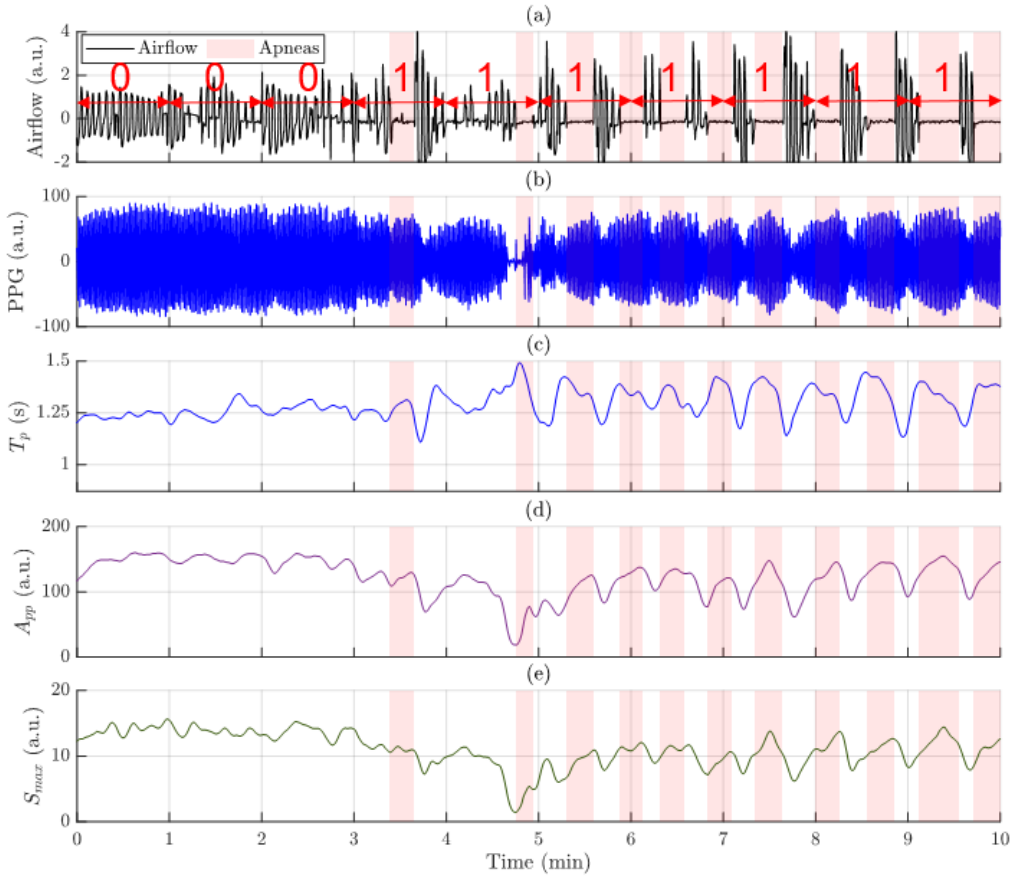


Figure 21: An illustration of signal segment labelling: A sleep apnea segment is defined as a 1 min signal segment containing at least 1 apnea/hypopnea episode (the pink colour shows apnea/hypopnea episodes); true positives (apnea segments) – 1, true negatives – 0.

4.3.5 Methods: Model architectures and performance metrics

In our experiments, we used two artificial neural network architectures - CNN-GRU (see Figure 22) and Inception1d (see Figure 23).

In order to evaluate model performance, we used the metrics sensitivity, specificity, balanced accuracy, and F1-score. Additionally, the mean absolute error (MAE) of the estimated apnea burden and the proportion of absolute differences (less than 10%) between the true and estimated apnea burden values were calculated.

4.3.6 Results: ID and OOD testing

The combination of PPG features and SpO2 demonstrated the highest sensitivity, accuracy and F1-score in the ID testing case (see Tables 26 and 28). However, in the OOD testing case (see Tables 27 and 29), SpO2 input demonstrated the highest specificity metric and a significantly lower MAE of the estimated apnea burden in most cases. During OOD testing, no significant differences were observed in accuracy and F1-score, i.e. the difference is less than 2%, when comparing PPG features + SpO2 vs. only SpO2. The lowest MAE of the apnea burden was obtained when only SpO2 was used as the model input.

When comparing OOD and ID testing, as expected, the ID case (see Tables 26 and 28) showed higher performance metrics. When comparing models, the CNN-GRU model (see Tables 26 and 27) performed significantly better than the Inception1d model, especially on OOD data, thus suggesting better generalisation.

Additional experiments were performed by resampling PPG signals with different sampling rates (see Figure 24). The results showed that reducing the PPG sampling frequency did not significantly affect the performance metrics when using PPG feature time series for classification. However, we can see that changing the PPG sampling rate had a more significant impact (albeit small) on OOD generalisation.

4.3.7 Discussion

The results highlight a clear distinction between model behaviour in ID and OOD settings. When evaluated on ID data, both architectures — CNN-GRU and Inception1d — benefit from combining PPG features with SpO2, with the CNN-GRU model performing particularly well in sensitivity, accuracy, and F1-score metrics. This suggests that PPG and SpO2 together provide complementary information, helping the models detect apnea events more reliably when the test data match the training distribution. However, the trend changes substantially in the OOD testing. SpO2 alone consistently delivers better or more stable performance than the combined feature set, especially for the CNN-GRU model, which outperforms Inception1d in nearly all OOD metrics. The degradation observed when PPG features are included suggests that these features may encode domain-specific patterns that do not generalise well to new patient populations or recording conditions. This leads to a form of overfitting where the model becomes sensitive to subtle PPG characteristics that are not preserved across datasets. The OOD robustness of SpO2-only models, although initially counterintuitive, indicates that SpO2 carries more invariant and transferable information across domains. In general, the findings emphasise that while multimodal features improve ID performance, generalisation requires careful feature selection, and future work should investigate how to make PPG features less domain-dependent.

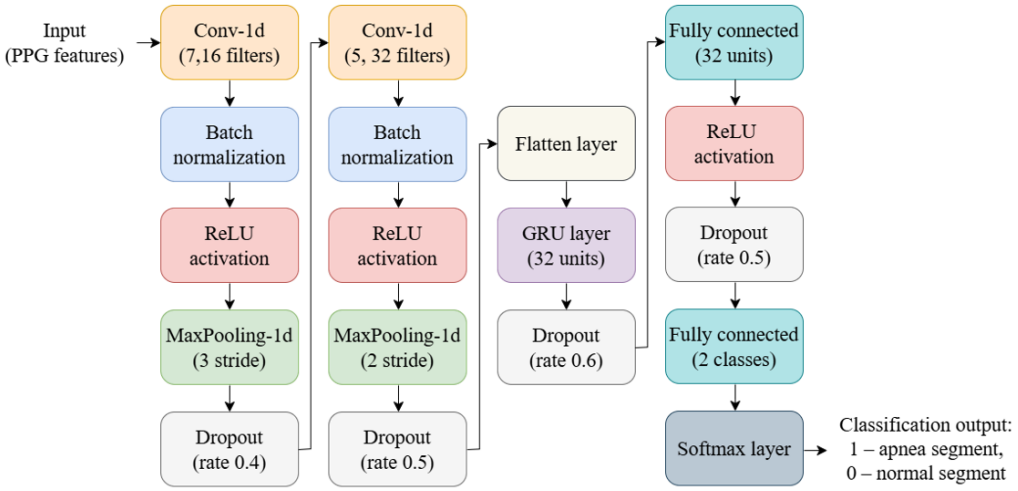


Figure 22: The model architecture structure of CNN-GRU.

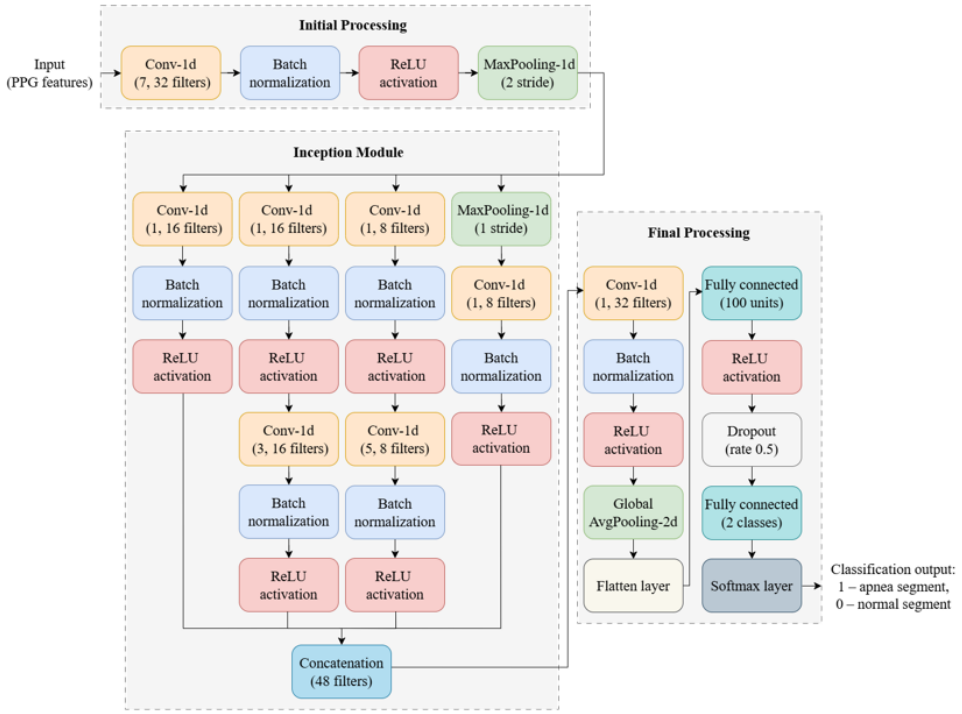


Figure 23: The model architecture structure of Inception1d.

Table 26: Performance metrics obtained using the CNN-GRU model. A difference of > 2 % is shown in bold (PPG features + SpO2 vs. SpO2). PPG sampling rate = 32 Hz, ID testing (batch size = 1024, learning rate = 0.001).

fs = 32 Hz	**PPG 5 features**	**SpO2**	**PPG 5 features + SpO2**
Sensitivity	49.56	68.16	**76.61**
Specificity	78.64	73.12	72.08
Balanced accuracy	64.10	70.64	**74.34**
F1-score	55.55	66.65	**71.47**
MAE (apnea burden)	16.59	7.81	9.77
Proportion (MAE < 10)	36.67	**70.00**	43.33

Table 27: Performance metrics obtained using the CNN-GRU model. A difference of > 2 % is shown in bold (PPG features + SpO2 vs. SpO2). PPG sampling rate = 32 Hz, OOD testing (batch size = 1024, learning rate = 0.001).

fs = 32 Hz	**PPG 5 features**	**SpO2**	**PPG 5 features + SpO2**
Sensitivity	61.63	72.10	73.31
Specificity	56.75	**65.92**	61.18
Balanced accuracy	59.19	69.01	67.25
F1-score	48.91	59.27	57.63
MAE (apnea burden)	27.81	**17.25**	19.61
Proportion (MAE < 10)	10.00	33.33	33.33

Table 28: Performance metrics obtained using the Inception1d model. A difference of > 2 % is shown in bold (PPG features + SpO2 vs. SpO2). PPG sampling rate = 32 Hz, ID testing (batch size = 1024, learning rate = 0.001).

fs = 32 Hz	**PPG 5 features**	**SpO2**	**PPG 5 features + SpO2**
Sensitivity	59.24	62.26	**72.21**
Specificity	70.45	**79.41**	75.53
Balanced accuracy	64.84	70.84	**73.87**
F1-score	59.47	65.50	**70.37**
MAE (apnea burden)	13.72	6.50	8.00
Proportion (MAE < 10)	36.67	**70.00**	66.67

Table 29: Performance metrics obtained using the Inception1d model. A difference of > 2 % is shown in bold (PPG features + SpO2 vs. SpO2). PPG sampling rate = 32 Hz, OOD testing (batch size = 1024, learning rate = 0.001).

fs = 32 Hz	**PPG 5 features**	**SpO2**	**PPG 5 features + SpO2**
Sensitivity	59.77	63.55	**67.14**
Specificity	55.42	**70.39**	64.26
Balanced accuracy	57.60	66.97	65.70
F1-score	47.27	56.39	55.54
MAE (apnea burden)	25.90	**16.05**	19.38
Proportion (MAE < 10)	20.00	26.67	26.67

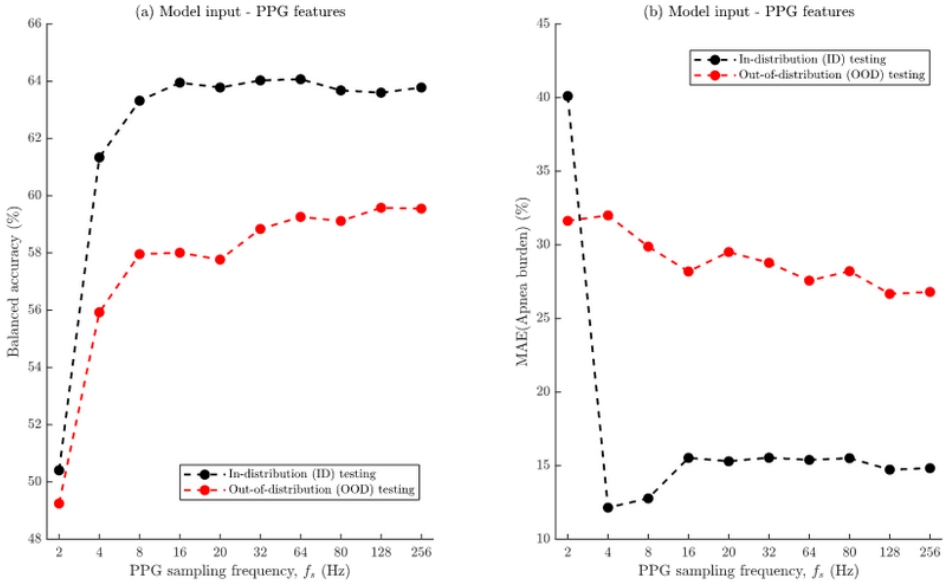


Figure 24: Balanced accuracy and MAE dependence on PPG sampling rate.

5 A1.3.4 – Dependence of results on different factors

PTB, FVB, KTU and THM will investigate the dependence of the results from A1.3.1 and A1.3.2 on the number of samples, sampling rate, pre-processing (smoothing, filtering, coherent averaging, periodification, additional transformations, etc.) and further PPG-specific dataset characteristics such as baseline wander, signal to noise ratio (modifiable for synthetic datasets or through adding additional parametric noise as in A1.3.3), length of signals, label noise, real vs. synthetic data.

We study how predictive performance and uncertainty calibration of machine learning models depend on several dataset characteristics. In particular, we examine the effects of additive Gaussian noise, high-pass and low-pass Butterworth filtering, and the sizes of the training and validation sets on two state-of-the-art 1D convolutional architectures, PP-Net [50] and XResNet1d50 [51].

Our experiments cover one classification dataset, Deepbeat [52], where the task is to detect atrial fibrillation, and one regression dataset, VitalDB [53], where the task is to predict a patient's blood pressure. We use the same preprocessed versions of the datasets as described in [1]. The VitalDB dataset is further subdivided into two datasets, calib and calibfree. In calib, the same subjects appear in the training, validation and test splits, so that the models can calibrate to the individual subjects. In calibfree, the dataset splits were created with samples coming from different subjects, meaning that the subjects in the test and validation splits did not appear in the training split.

We observe that, although these dataset transformations substantially affect model accuracy, the uncertainty calibration — as measured by ACE (adaptive calibration error) for classification and ENCE (expected normalised calibration error) for regression — remains comparatively stable, with notable deviations only in a few specific cases.

5.1 Number of samples

To investigate the relevance of the dataset size, we train the models only on a fraction of the training and validation set and then evaluate it on the full test set. To generate the smaller sets, we draw the samples randomly from the original dataset. The original datasets have the following sizes: calib has 419k samples for the training and 41k for the validation dataset, calibfree 417k for training and 32k for validation, and Deepbeat 106k for training and 15k for validation.

As shown in Figure 26 for the calib dataset, the MAE decreases approximately linearly when plotted against the logarithm of the number of samples for both models. For the calibfree dataset (Figure 25 centre), the MAE decreases initially, but then plateaus when around 6% of the samples are included in the training and validation set, thus with around 25k samples the models seem to have learned all they can learn from this dataset.

The fact that the models plateau at some MAE value and cannot learn further can be explained with blood pressure measurement noise. Measurement devices used in the VitalDB dataset [53] include the Solar 8000M Patient Monitor [54] that has an indicated clinical standard deviation of blood pressure readings of 8 mmHg. Then, an evaluation of various blood pressure measurement studies [55] showed that finger-based pressure measurements have a standard deviation of 11.7 mmHg for systolic and 7.7 mmHg for diastolic blood pressure compared to a reference device. Combining the error sources, from the reference measurement device σ_{device}, and finger based versus reference measurement variability $\sigma_{\text{reference}}$, and assuming that they are independent, we can calculate the standard deviation of finger-based pressure measurement as $\sigma_{\text{finger}} = \sqrt{\sigma^2_{\text{device}} + \sigma^2_{\text{reference}}}$. The mean absolute error, making the simplifying assumption that the blood pressure errors are normally distributed, can then be calculated as MAE $\approx 0.8\sigma$. This leads to MAE(SBP) $\approx 0.8\sqrt{11.7^2 + 8^2}$ mmHg ≈ 11.3 mmHg and MAE(DBP) $\approx 0.8\sqrt{7.7^2 + 8^2}$ mmHg ≈ 8.9 mmHg. Since PPG curves are measured by finger-tip devices, they presumably contain information about finger blood pressure, and the theoretical lower bound on the MAE for models trained on such a dataset should be around these values. Indeed, the models achieve a minimal MAE of 12.5 mmHg for systolic and 7.8 mmHg for diastolic

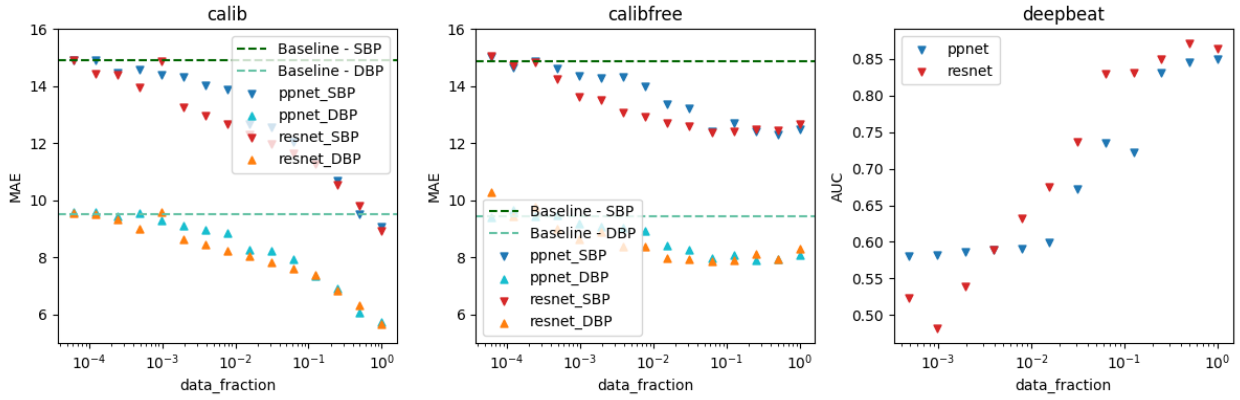


Figure 25: Influence of the number of samples in the training/validation set on prediction performance as measured by MAE and AUC. Data fraction is the fraction of the training dataset used for training.

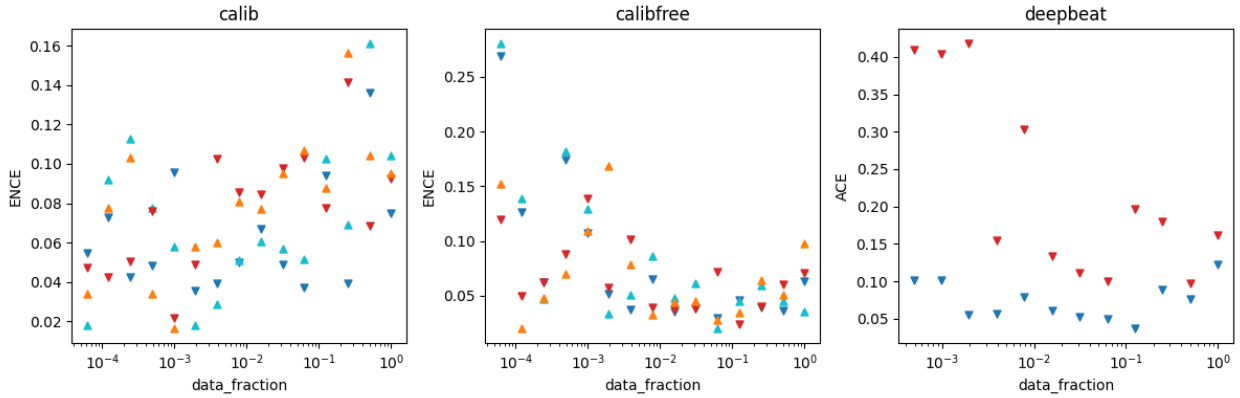


Figure 26: Influence of the number of samples in the training/validation set on prediction uncertainty calibration as measured by ENCE and ACE. Data fraction is the fraction of the training dataset used for training.

blood pressure, which is close to the expected MAE for such a measurement.

In contrast, the reason that in the calib case the MAE can be further decreased by increasing the number of samples is that there is more information that the models can learn to improve the predictions for unseen samples. In calib, subjects are present in both training and test sets, so the models can exploit subject-specific PPG features; in calibfree, train and test subjects are disjoint, and such subject-specific fitting harms generalisation. To learn the subject-specific features the networks need more samples.

For Deepbeat (Figure 25 right), both models exhibit substantial AUC gains with increasing dataset size, plateauing around 0.85. ResNet starts from a lower AUC and improves more smoothly with sample size, whereas PP-Net begins higher but shows a delayed, sharper increase once around $1-2$ % of the data is used. With a larger dataset the models would still be able to further increase the accuracy.

The calibration metrics do not show a consistent trend across datasets. For the calib case (Figure 26 left), fewer samples seem to lead to a slightly smaller ENCE, while for the calibfree case (Figure 26 center) fewer samples lead to a higher ENCE with a higher variability. For Deepbeat (Figure 26 right), the ACE of the PP-Net model remains in the range of $4-12$ %, whereas the ResNet model has a high calibration error of around 40 % when using only 0.1 % of the samples, and decreases when using more samples.

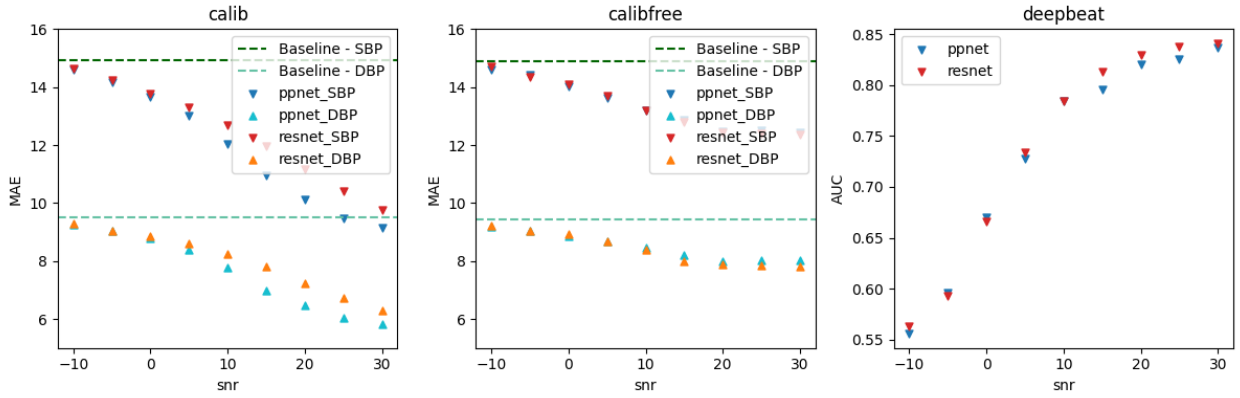


Figure 27: Influence of Gaussian noise perturbations parametrised by the signal to noise ratio (SNR) on prediction performance as measured by MAE and AUC.

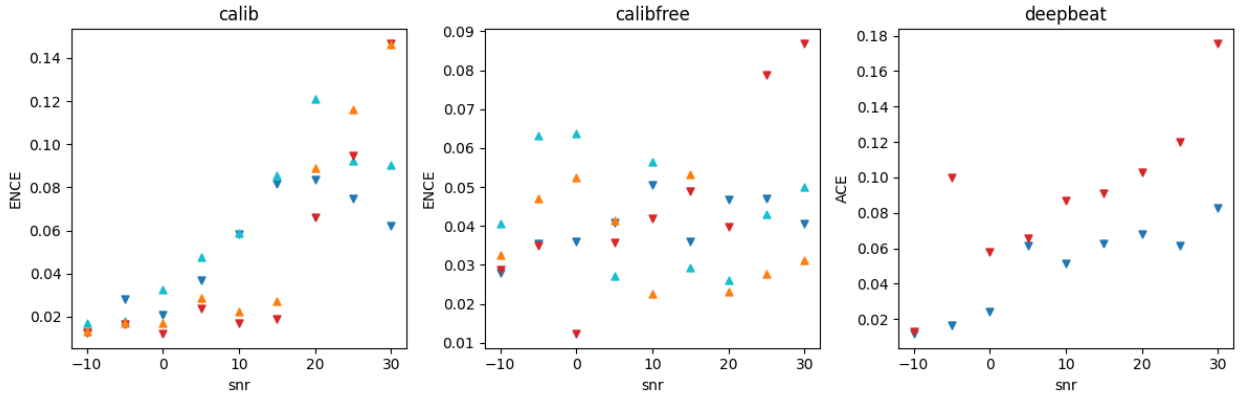


Figure 28: Influence of Gaussian noise perturbations parametrised by the signal to noise ratio (SNR) on prediction uncertainty calibration as measured by ENCE and ACE.

5.2 Additive Gaussian noise

We perturb the signals by adding white Gaussian noise at a specified signal-to-noise ratio (SNR). For each signal, we first generate a noise time series of equal length by sampling from a standard normal distribution. We then rescale this noise so that the ratio between the signal energy and the noise energy—where energy is defined as the variance of the time series—matches a target SNR expressed in decibels. The rescaled noise is finally added to the original signal.

As expected, higher SNRs yield better performance. For the calib dataset (Figure 27 left), the MAE decreases approximately linearly with SNR. At an SNR of 30 dB, the MAE nearly reaches the level of the clean signals, at about 9.0 mmHg for systolic and 5.7 mmHg for diastolic blood pressure (cf. Section 5.1). The fact that even very small noise (SNR=30 dB) affects the MAE shows that the model is learning very fine-grained-subject related features.

For the calibfree dataset (Figure 27 center), performance improves only up to an SNR of 20 dB; beyond this point, increasing the SNR does not lead to further gains. Similar to the discussion in Section 5.1, the models achieve MAE values that are due to blood pressure measurement variability and the noise in the signal becomes smaller than the noise in the finger-based blood pressure measurement.

For Deepbeat (Figure 27 right), the models increase in AUC linearly with an increase of SNR up to a SNR of around 10 dB and then approach more slowly the noiseless AUC value of almost 0.85.

The calibration metrics for Deepbeat and calib show a clear reduction of the calibration error when the SNR

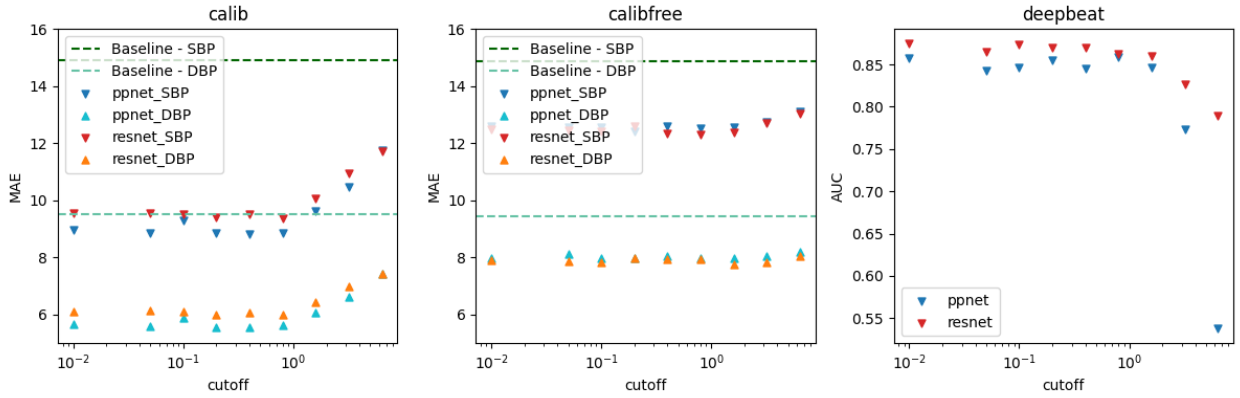


Figure 29: Influence of filtering through a 4th order Butterworth high-pass filter with given cutoff frequency on prediction performance as measured by MAE and AUC.

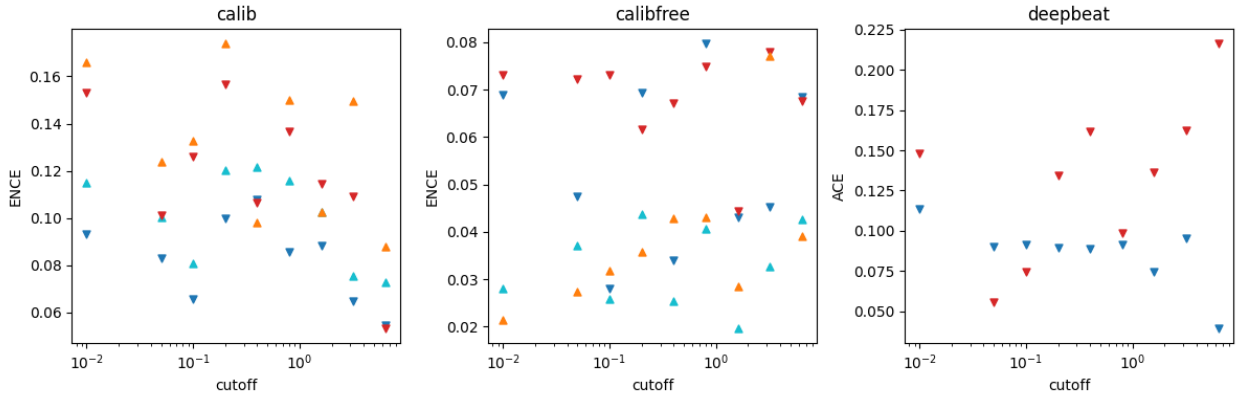


Figure 30: Influence of filtering through a 4th order Butterworth high-pass filter with given cutoff frequency on prediction uncertainty calibration as measured by ENCE and ACE.

is low (Figure 28 left and right), which is to be expected. When the noise dominates, the signal contain no information about the blood pressure, and the models can only learn global statistics of the dataset, such as the mean and standard deviation for blood pressure regression and mean class probability for classification. A forecaster that predicts exactly these global statistics is, in principle, perfectly calibrated under the assumed data-generating distribution. This explicitly illustrates that calibration is not related to accuracy and can be "good" even when predictions are uninformative. Interestingly, this does not happen in the calibfree case (Figure 28 center), where the calibration error stays in the same range for all SNR values.

5.3 High pass filter

We used a 4th order Butterworth high-pass filter [39] to filter out low frequencies under a given cutoff. Filtering out low frequencies results in removing artefacts unrelated to blood pressure such as baseline wander.

For the calib case (Figure 29 left), removing frequencies up to around 0.8 Hz had no significant impact on the MAE, and setting a higher cutoff then lead to an increase of the MAE. The calibfree case was similar (Figure 29 center), except that the increase in the MAE started at a higher cutoff and was less steep.

Models trained on the Deepbeat dataset also showed similar results (Figure 29 right); up to a cutoff value of 0.8 Hz there was no significant decrease in AUC. At a cutoff of 6.4 Hz, the AUC of ResNet decreased only slightly, whereas PP-Net dropped sharply to 0.54, close to the AUC value of a random forecaster 0.5. This shows that the ResNet can still learn, while too much information has been filtered out for PP-Net.

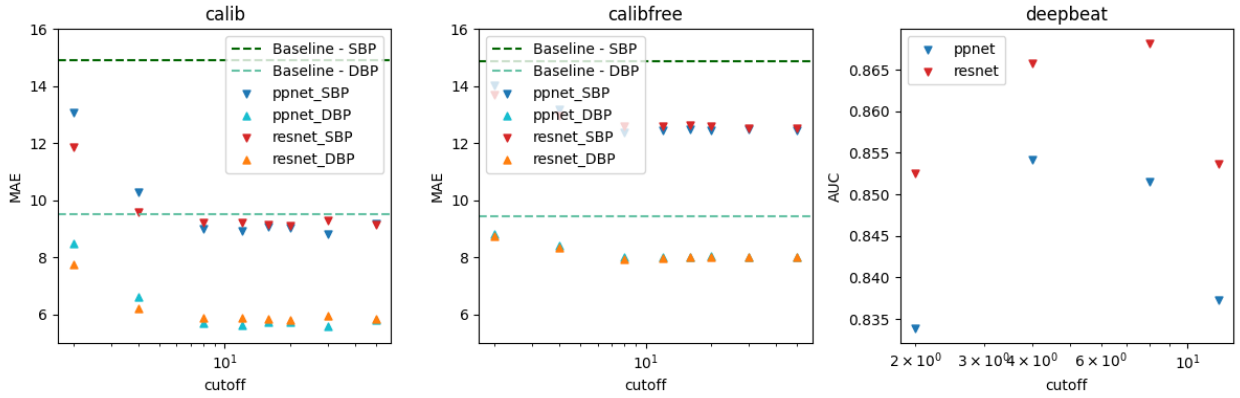


Figure 31: Influence of filtering through a 4th order Butterworth low-pass filter with given cutoff frequency on prediction performance as measured by MAE and AUC.

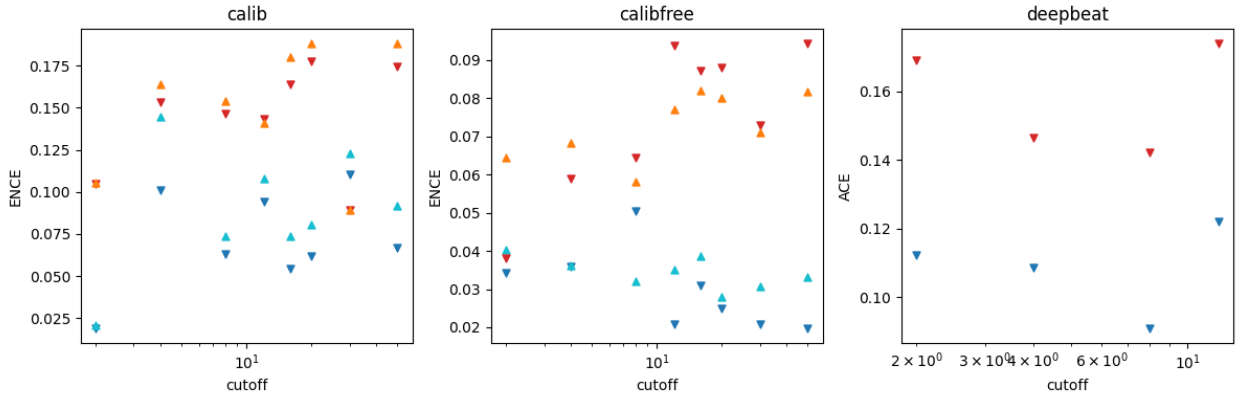


Figure 32: Influence of filtering through a 4th order Butterworth low-pass filter with given cutoff frequency on prediction uncertainty calibration as measured by ENCE and ACE.

The calibration metrics do not show any clear relationship with the high-pass filter frequency cutoff (Figure 30).

5.4 Low pass filter

We used a 4th order Butterworth low-pass filter [39] to filter out high frequencies over a given cutoff. The models did not show any significant increase or decrease of the MAE for the calib and calibfree dataset down to a cutoff frequency of 4Hz, as shown in Figure 31 (left and center). A cutoff under that frequency lead to less precise predictions and to a higher MAE for both models and both systolic and diastolic blood pressure predictions. This implies that signals with a sampling rate of around 16 Hz (double frequency is needed due to aliasing) appear to contain all the relevant information of the PPG curve for a model to predict the blood pressure.

For the Deepbeat dataset (Figure 31 right), the cutoffs 4 Hz and 8 Hz produced slightly better AUC values than higher cutoff or lower cutoff frequencies. However, there were few data points, so this may have been due to random variation, and more experiments are needed to confirm it.

The calibration metrics did not exhibit any clear dependence on the cutoff frequency. The uncertainties obtained by PP-Net are better calibrated, as shown in Figure 32 (right).

5.5 Discussion

Across all experiments, dataset transformations had a pronounced effect on predictive accuracy, whereas uncertainty calibration, as quantified by ACE and ENCE, was comparatively stable. For blood pressure regression, the observed performance plateau is consistent with known variability of finger-based blood pressure measurements, indicating a practical lower bound on MAE inherent to the acquisition modality rather than the models. The relevant information of the PPG curves for blood pressure regression is contained mostly in the frequencies up to 8 Hz, meaning that low sampling rates of 16 Hz are a sufficient data quality. The two architectures performed comparatively well, where PP-Net obtained more accurate results in some cases and ResNet in others. These findings suggest that, within realistic PPG acquisition conditions, moderate variations in pre-processing and sampling primarily affect accuracy, while calibration remains relatively robust. Further work should extend this analysis to additional artefacts, sampling schemes, and calibration approaches.

6 A1.3.5 – Disentangling aleatoric and epistemic uncertainty

NPL, LNE, IPQ and PTB will investigate aleatoric vs. epistemic for those uncertainty quantification methods where the two can be disentangled (A1.2.1, A1.2.2, A1.2.5), while also varying the data quality and model size. A1.3.1 and A1.3.2 will be partly reconsidered in this context using selected ML/DL models from A1.1.2-A1.1.5 on suitable datasets of A2.1.4, but also the relative scales of epistemic vs. aleatoric uncertainty will be investigated. LNE will investigate the relation between aleatoric uncertainty and accuracy for at least one of the benchmark problems of A2.1.4. In addition, PTB and NPL will consider whether epistemic uncertainty is the most appropriate for detecting out-of-distribution samples for at least one of the benchmark problems from A2.1.4, which were assessed in A1.3.2.

It is possible to model several sources of uncertainty [33], where estimates of a given source may assist with model prototyping or dataset curation. Therefore, it is of interest to evaluate how the composition of uncertainties may change with respect to variations in dataset and architecture. Epistemic (reducible model uncertainty), and aleatoric uncertainty (irreducible data uncertainty) are the most significant sources of uncertainty in most practical measurement scenarios. The former can be reduced with a more effective choice of model architecture, or more comprehensive dataset, while the latter is irreducible, and intrinsic to the task or data (e.g. noise or ill-posedness). Most UQ techniques capable of modelling these sources of uncertainty output a total uncertainty, where the contribution from either source must be disentangled from the model's estimate. Here, following the protocol described in [33] we model aleatoric uncertainty using a Gaussian Negative Log-Likelihood for regression, and by learning logit variances for classification. The Law of Total Variances is used for disentangling uncertainties output from regression models [56], while we implement a custom approach for disentangling aleatoric from total uncertainty for classification (Algorithm 2). Epistemic uncertainty is modelled using Monte Carlo Dropout.

We aim to observe whether changes in the composition of predicted uncertainties may align with expectations concerning choice of dataset or architecture. We trained to perform two tasks: the BP regression task as described in [2, 1] and the sex classification task as described in Section 3.2. For the former, we consider a 1D variant of AlexNet, and an XResNet, with dropout applied throughout each architecture. We considered two custom splits of the VitalDB datset for BP regression: calibfree (no patient overlap in the splits) and calib datasets (some patient overlap in the splits) as described in [2, 1]. For sex classification, we considered two custom versions of the AuroraBP dataset and used them to train the same AlexNet architecture: model *s1* was trained with high-quality signals (optical quality ≥ 0.65) and other details as described in Section 3.2.6, while model *s2* was trained on a dataset that retains signals of all qualities (no quality thresholding) and that extracted only one instance from the raw PPG signals. The latter model, therefore, was trained on a lower-quality and smaller dataset. Commentary on whether epistemic uncertainty is the most appropriate for OOD detection is given in Section 4.2.2.

For classification, the model predicted means $\mu_{i,c,t}$ and variances $\sigma^2_{i,c,t}$, which parametrise a Gaussian distribution over the logits $z_{i,c,t}$ for each class c, input i, and stochastic forward pass t. These parameters were computed independently across T iterations. For each iteration t, we drew $K = 100$ samples from the distribution

$$z_{i,c,t,k} \sim \mathcal{N}(\mu_{i,c,t}, \sigma^2_{i,c,t})$$

for all classes, then applied a softmax to each sampled logits vector $\mathbf{z}_{i,t,k}$ to obtain class probabilities $\mathbf{p}_{i,t,k}$ (see Algorithm 1). Aleatoric and total uncertainty were disentangled using a custom averaging procedure inspired by Kendall et al. [33] (see Algorithm 2). The aggregated results reflected uncertainty as entropy over the distributions derived from the sampling procedure, where prediction variance broadens the distribution.

For regression, the model $f_{\boldsymbol{\theta},\omega}(x_i)$ output the mean and variance of a Gaussian distribution for each predicted quantity. Using MAP estimation, it predicted

$$\mu_{SBP,i,t}, \mu_{DBP,i,t}, \quad \text{and variances} \quad \sigma^2_{SBP,i,t}, \sigma^2_{DBP,i,t}$$

for input x_i (e.g., a raw PPG time series), parametrising independent Gaussian distributions across T stochastic forward passes. We then estimated the mean and both aleatoric and epistemic uncertainties using the Law of Total Variance [56], as detailed in Algorithm 3.

Algorithm 1 Monte Carlo Dropout Classification Training Loop

for each batch **do**
  Compute predicted mean $\mu_{i,c}$ and variance $\sigma^2_{i,c}$ for each class c for each input i in the batch
  for $k = 1$ to K **do**
    for each class c **do**
      Sample $\epsilon_{i,c,k} \sim \mathcal{N}(0, 1)$
      Compute sampled logit $z_{i,c,k} = \mu_{i,c} + \sigma_{i,c}\epsilon_{i,c,k}$
    end for
    Compute $\mathbf{p}_{i,k} = \text{softmax}(\mathbf{z}_{i,k})$
  end for
  Compute average $\bar{\mathbf{p}}_i = \frac{1}{K}\sum_{k=1}^{K} \mathbf{p}_{i,k}$
  Evaluate NLL on P where P is the batch of $\bar{\mathbf{p}}_i$ values.
end for

Algorithm 2 Monte Carlo Dropout Classification Evaluation Loop

for each input **do**
  for $t = 1$ to T **do**
    Acquire the predicted mean $\mu_{c,t}$ and variance $\sigma^2_{c,t}$ for each class c
    for $k = 1$ to K **do**
      for each class c **do**
        Sample $\epsilon_{c,t,k} \sim \mathcal{N}(0, 1)$
        Compute sampled logit $z_{c,t,k} = \mu_{c,t} + \sigma_{c,t}\epsilon_{c,t,k}$
      end for
      Compute $\mathbf{p}_{t,k} = \text{softmax}(\mathbf{x}_{t,k})$
    end for
    Compute average $\bar{\mathbf{p}}_t = \frac{1}{K}\sum_{k=1}^{K} \mathbf{p}_{t,k}$
  end for
  $H_{\text{ale}} = \frac{1}{T}\sum_{t=1}^{T} H(\bar{\mathbf{p}}_t)$
  $H_{\text{total}} = H(\frac{1}{T}\sum_{t=1}^{T} \bar{\mathbf{p}}_t)$
  Where H_{ale} is the aleatoric uncertainty expressed as an entropy.
end for

6.1 Results

6.1.1 Blood pressure estimation

Figures 33, 34, 35 and 36 show the disentangled uncertainty estimates for each BP model/dataset, while Table 30 shows the mean ratio of aleatoric and total uncertainty (variances). We found that the ratio is consistent for SBP and DBP for a given model and task, where the models trained on the calibfree data have total uncertainties composed of a larger proportion of aleatoric uncertainty compared to models trained on the calib dataset. For calib, the uncertainty estimates from the larger XResNet model produced total uncertainties with smaller proportions of aleatoric uncertainty compared to AlexNet, while little difference was observed for the models trained on calibfree. A more significant proportion of aleatoric uncertainty for the calibfree is in line with prior expectations given the lack of patient overlap and potential ill-posedness of the prediction task [1]. The lower proportion of aleatoric uncertainty for the calib case for the larger XResNet model compared to the AlexNet model contrasts with expectations that more capacity should lower epistemic uncertainty. It is possible that some of this could be attributed to differences in how well-optimised each model may be, but could also indicate the poor effectiveness of the disentanglement strategy.

Here, aleatoric uncertainty appears to broadly correlate with model performance (MAE) for the calib task; i.e. a larger proportion of aleatoric uncertainty corresponds to lower predictive performance. Similar predictive performance across the two models for calibfree is reflected in the comparable proportions of aleatoric uncertainty for each model. This suggests that aleatoric uncertainty as disentangled here may be an effective

Algorithm 3 Monte Carlo Dropout Regression Evaluation Loop

for each input i **do**
  for $t = 1$ to T **do**
    $\mu_t, \sigma_t^2 = f(\boldsymbol{\theta})$
  end for
  $\mu_i = \frac{1}{T}\sum_{t=1}^{T} \mu_{i,t}$
  $\sigma^2_{\text{epi},i} = \frac{1}{T}\sum_{t=1}^{T}(\mu_{i,t} - \frac{1}{T}\sum_{t=1}^{T}\mu_{i,t})^2$
  $\sigma^2_{\text{ale},i} = \frac{1}{T}\sum_{t=1}^{T}\sigma^2_{i,k}$
end for

proxy metric for model performance.

Table 30: Aleatoric ratio and MAE by task and dataset, and for both models.

Task	**Ale ratio SBP (%)**	**Ale ratio DBP (%)**	**SBP MAE**	**DBP MAE**
calib AlexNet	97	97	10.24	6.64
calib XResNet	93	93	8.38	5.31
calibfree AlexNet	100	100	12.42	7.98
calibfree XResNet	100	100	12.53	7.89

6.1.2 Sex classification

Table 31 reports the testing performance of two AlexNet sex classifiers: *s1* (optimal) and *s2* (sub-optimal: trained on a smaller, lower-quality dataset). Because *s1* uses a dataset configuration that extracts multiple instances from each raw PPG signal, it is trained on substantially more examples despite quality thresholding. As expected, *s1* achieved higher testing performance. Nevertheless, the two models remain broadly comparable for the purposes of this analysis.

Aleatoric and total uncertainty estimates for *s1* and *s2* are illustrated in Figures 37 and 38, respectively. The average aleatoric and total uncertainties and their ratio are also given in Table 32. As explained at the beginning of this section, classification uncertainty is expressed as *entropies*. The following patterns are observed:

- The sub-optimal model *s2* estimated higher aleatoric and total uncertainties, leading to larger average values. The higher aleatoric uncertainty in *s2* may be attributed to the absence of quality thresholding in the training data, which increases data-related uncertainty.
- The plot of total versus aleatoric uncertainty qualitatively shows that some samples have higher proportions of epistemic uncertainty for *s2*, consistent with expectations for models trained on smaller, lower-quality datasets. The small drop in the ratio of aleatoric to total uncertainty (see Table 32) also suggests that the relative increase in epistemic uncertainty magnitude is more significant than seen for the aleatoric uncertainty in *s2*. This is because for this ratio to decrease, epistemic uncertainty must go higher by a larger percentage than the aleatoric uncertainty, given their smaller values. That said,

Table 31: Sex classification performance metrics of the *s1* and *s2* AlexNets for the testing set.

Model	Data count	Loss	Accuracy	Precision	Recall (sensitivity)	F1 score	Specificity	AUC
s1	64,401	0.5787	0.7101	0.7290	0.7349	0.7287	0.6817	0.7714
s2	35,606	0.6264	0.6483	0.6343	0.6280	0.6271	0.6684	0.7060

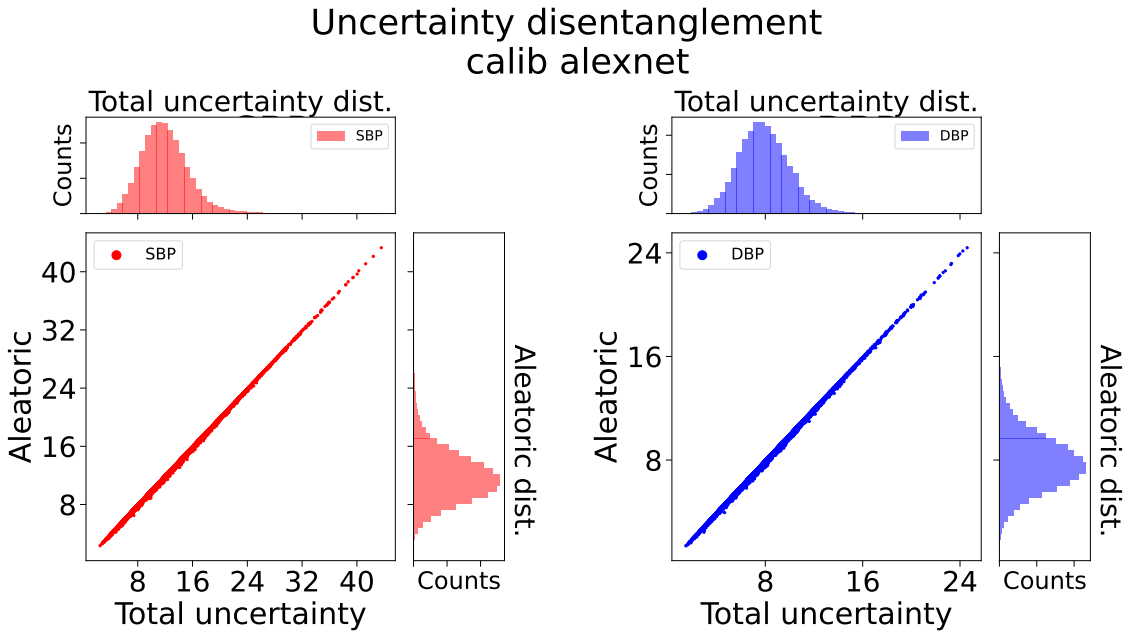


Figure 33: Disentangled uncertainty estimates for the BP regression task on the calib dataset with the AlexNet architecture.

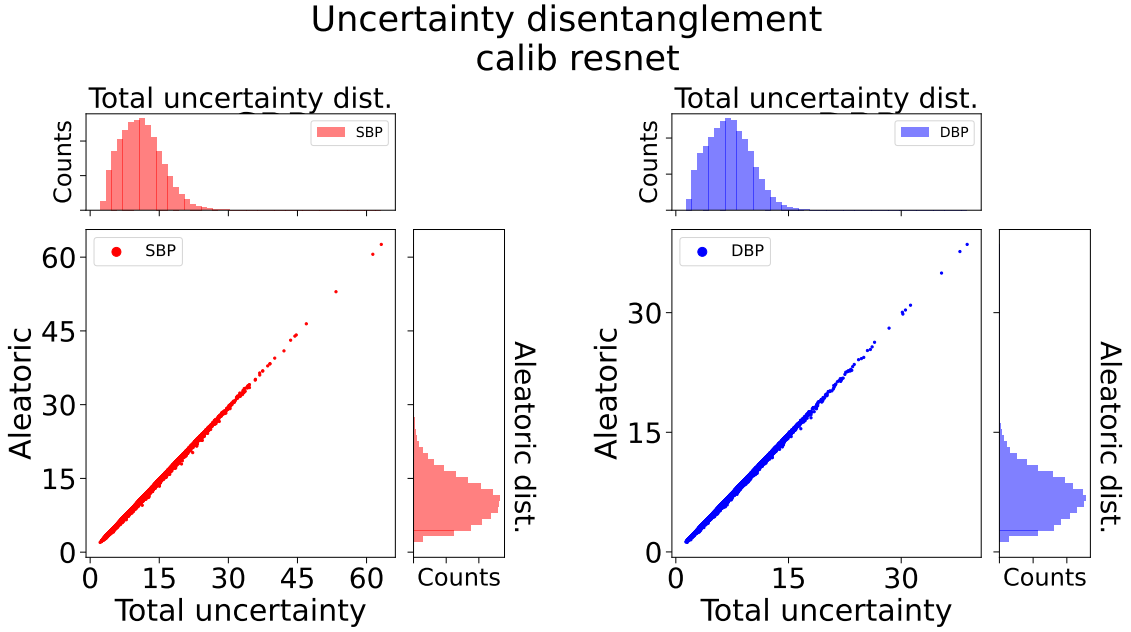


Figure 34: Disentangled uncertainty estimates for the BP regression task on the calib dataset with the XResNet architecture.

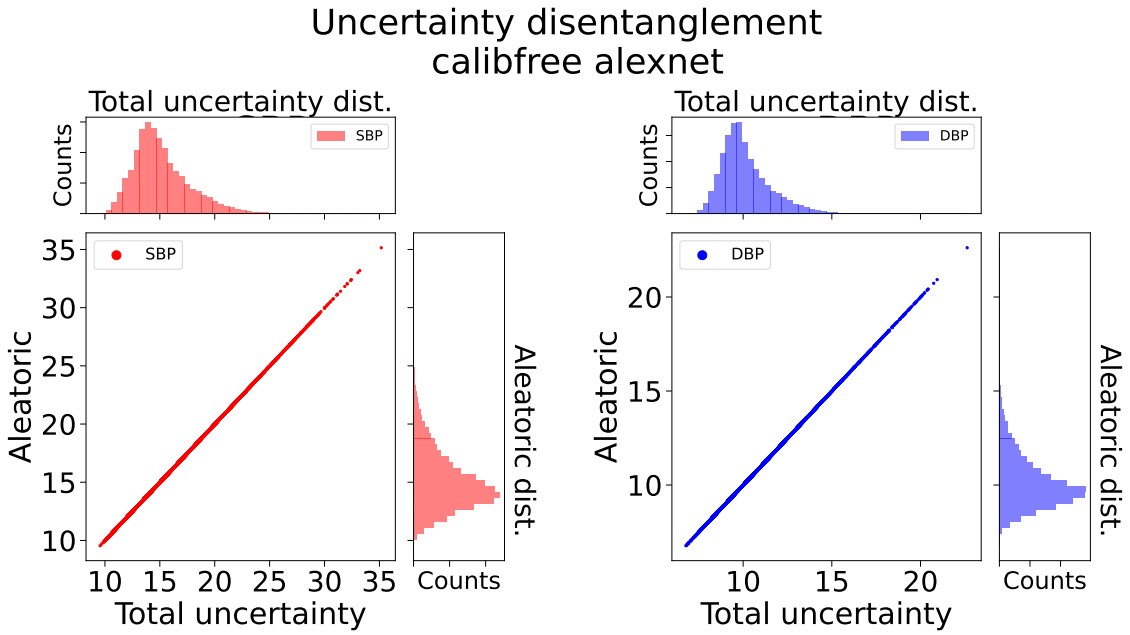


Figure 35: Disentangled uncertainty estimates for the BP regression task on the calibfree dataset with the AlexNet architecture.

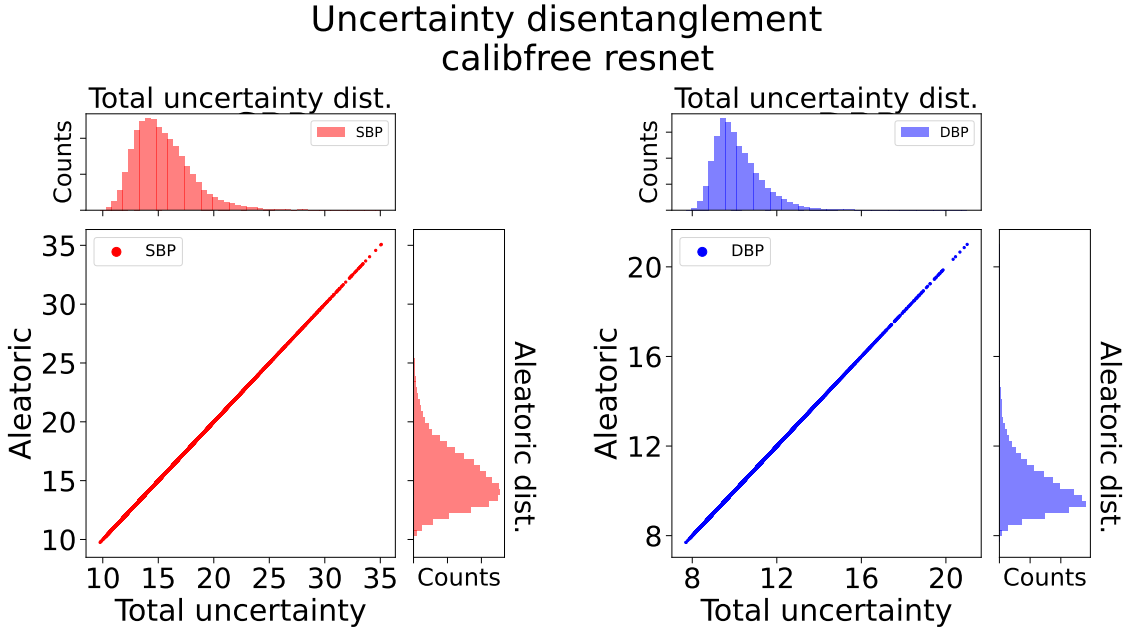


Figure 36: Disentangled uncertainty estimates for the BP regression task on the calibfree dataset with the XResNet architecture.

Table 32: Average sex-classification uncertainty results (entropies).

Model	Mean ale.	Mean total	Mean ale./total
s1	0.8794	0.8897	0.9883
s2	0.9214	0.9345	0.9859

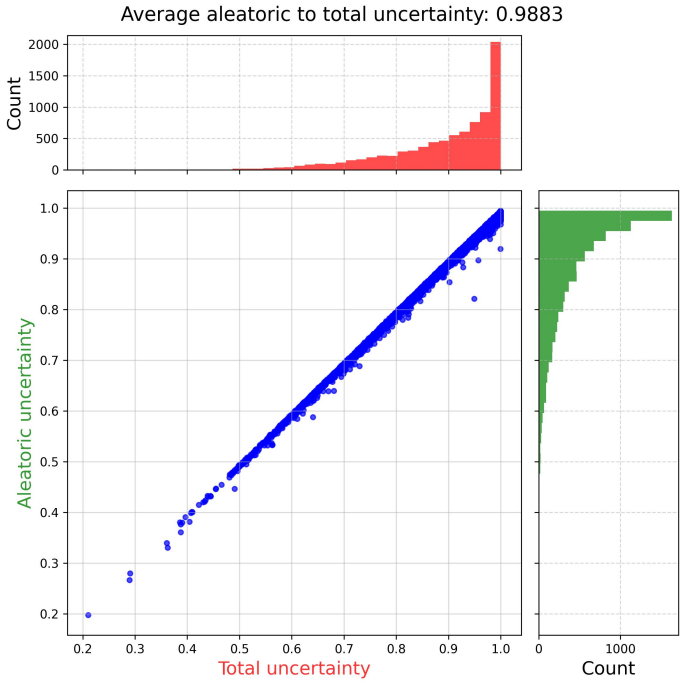


Figure 37: Disentangled uncertainty estimates for the optimal sex classifier (model *s1*).

the ratios are still quite similar, which indicates that in both cases the bulk of the total uncertainty comes from the aleatoric uncertainty.

The relative magnitudes of aleatoric uncertainty are in line with expectations with respect to changes in data quality. This may provide some evidence suggesting there could be some practical utility in using diesntangled estimates to acquire information about datset/curation/modelling choices. However, several other factors (e.g. differences in the extent to which each model is optimised), may also affect uncertainty estimates which makes it challenging to assess the quality of disentangled estimates.

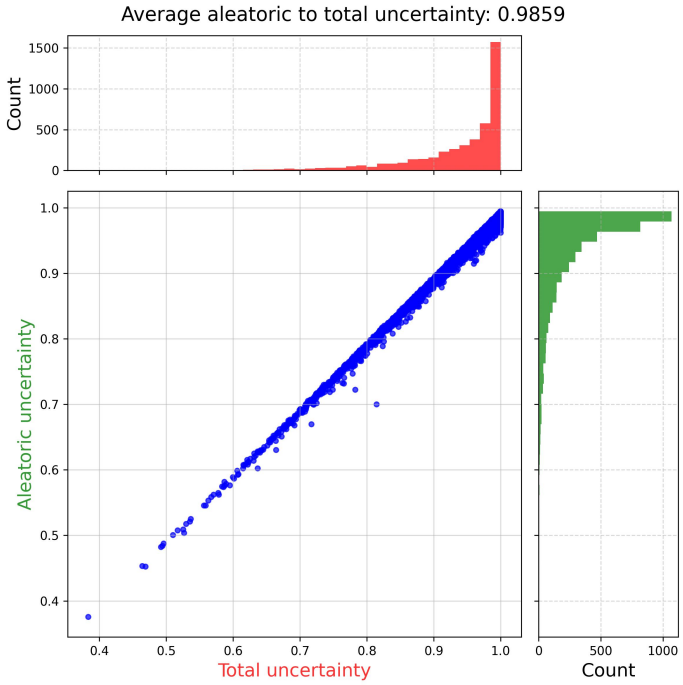


Figure 38: Disentangled uncertainty estimates for the sub-optimal sex classifier (model *s2*).

7 A1.3.6 – Classification of skin tone

THM and NPL supported by FC and IMBiH will consider classification of PPG signals by skin tone, using the six classes of skin tone of the Fitzpatrick scale as labels, using at least one of the models developed in A1.1.2-A1.1.4., evaluating both performance and uncertainty, in order to determine whether or not machine learning can accurately distinguish between signals from different skin tones. The performance metrics from A1.1.6 will be analysed with respect to sex to check for any differences. In addition, an unsupervised clustering analysis will be performed by THM and NPL to see whether there are any natural clusters in this data.

Photoplethysmography (PPG) signals are collected by many wearable devices including smart watches and pulse oximeters. A light is shone onto the skin and the light either reflected or transmitted is measured to generate the signal. It is a reasonable hypothesis that skin tone will have an effect on these signals. This was confirmed during the covid-19 pandemic where it was found that blood oxygenation readings for patients with darker skin tones were inaccurate, potentially resulting in sub-optimal treatment for some ethnic groups [57]. An independent report was commissioned in the UK entitled *Equity in Medical Devices: Independent Review* [58] which found "extensive evidence of poorer performance of pulse oximeters for patients with darker skin tones". Additionally, real world studies have recently been published that suggest "worse patient outcomes among patients with darker skin pigmentation" [59].

PPG signals are of much current interest as they can be collected non-invasively from unobtrusive wearable devices and, as such, are ideal for long term health monitoring. Some health-related quantities can be derived directly from the signal, such as heart rate, but for others, it is not obvious how they can be extracted from the PPG signals. Machine learning is therefore widely used to infer other quantities of interest. Examples include monitoring of blood pressure [60] and detection of atrial fibrillation [61].

The aim of this work is to consider machine learning prediction of skin tone itself, using the six-point Fitzpatrick skin tone scale [62] for a six-class classification. Machine learning is useful for discerning the mapping from signal to classification. Thus, if machine learning accuracy is high for this task, then it indicates that there is a distinguishable difference between PPG signals associated with different skin tones. On the contrary, if the accuracy is low, it implies that machine learning is struggling to distinguish between signals from different skin tones, which would imply that skin tone has little effect on the signals. In addition, we also consider binary classifications of light (classes 1, 2, 3) vs. dark (classes 4, 5, 6) skin tones.

One problem associated with this task is the inherent inaccuracy in the labelling of records according to skin tone. Currently, the assessment of skin tone is subjective. A person looks at the skin and compares it to the representative shades for each of the six skin tone classes and then decides which class the skin belongs to. There are no clearly defined boundaries for each of the skin tone classes, so the "correct" class could easily be one different from the chosen class. We take this inaccuracy in labelling of skin tone into account by adopting a "fuzzy accuracy" which is a modified version of the standard accuracy. Instead of requiring an exact match between predicted and true labels, fuzzy accuracy considers a prediction to be correct if it is within a certain tolerance range of the recorded label. In this case, the tolerance range is ± 1.

The effect of skin tone on machine learning predictions has recently been considered in [63] where classification of high blood pressure using PPG signals was considered. If a machine learning model was trained on only Fitzpatrick skin tone class 1 records (the lightest skin tone) and subsequently evaluated on data from the other skin tone classes, classification accuracy generally declined markedly for those classes in comparison with models which had been trained on data which included all skin tone classes. Many PPG datasets do not record skin tone, and for those that do, there is often a strong bias towards the lighter skin tones with only a few records from darker skin tones. This applies for example to the Aurora-BP dataset that is used in this study [64]. As a result, models trained primarily on data from lighter-skinned subjects may not generalise well to individuals with darker skin tones, raising concerns about potential performance disparities.

7.1 Fitzpatrick skin type

The Fitzpatrick skin type scale consists of six classes, ranging from light (class 1) to dark (class 6). It was originally developed to classify people with white skin in order to select the correct UV dose for treating skin disease, not for characterising skin tone [62]. However, its use as a scale for skin tone is now "nearly ubiquitous" [65] and there are a few datasets that record the Fitzpatrick skin tone for subjects, along with other metadata.

Alternative skin tone scales have been proposed, including the 10-point Monk skin tone scale which was specifically designed for skin tone classification in a diverse population and has been adopted by Google [65]. However, we are not aware of any datasets of physiological signals to date that include the Monk skin tone class.

A severe limitation of the Fitzpatrick skin tone assessment (together with assessments using other skin tone scales) is that it is purely subjective. Ideally, the skin tone at the site of data collection (i.e. the wrist if using a smartwatch) should be carefully measured using for example a spectrophotometer. However, this is not current practice. Consequently, there is often uncertainty about which class a particular subject should be in, since there are no defined boundaries between classes. This is a potential problem when using the Fitzpatrick skin tone class as a label for use with machine learning, as the designated class could easily be one class different from the predicted class. Indeed, in an assessment of manual grading, there were many one class differences between the classes designated by the graders for a particular facial image, and differences of two or more classes also occurred [66]. Such subjective assessment can result in a large number of misclassifications which correspond to inaccurate labels for our problem of skin tone classification. We deal with this problem by adapting the accuracy metric to a fuzzy accuracy, that we consider in Section 7.7.1.

7.2 Data

For this study, we used the AuroraBP dataset [64] which includes PPG waveforms and Fitzpatrick skin tone labels for 823 subjects (51.6 % female). (Note that there are 824 subjects with Fitzpatrick skin tone labels, but the PPG waveforms are missing in one case.) The PPG waveforms were collected at 250 Hz using a wrist-worn optical sensing device and the signals were subsequently resampled to 500 Hz. Two distinct protocols were used, namely auscultatory and oscillometric, and we combined both of these into a single dataset giving a combined total of 19,017 records. The signals were of varying length but were generally around 15 s in length.

The distribution of subjects by skin tone class is shown in Figure 39 (left) which shows that the dataset is highly imbalanced with the majority of subjects having lighter skin tones (90.3% in classes 1, 2 and 3). Overall, the gender balance is quite good, apart from skin tone class 6 which has only male subjects. It is a similar story for the distribution of records by skin tone class, as shown in Figure 39 (right) (88.2% in classes 1, 2 and 3).

7.3 Data analysis

In this section, we consider two aspects of the PPG signals that show a distinctive trend for different skin tone classes, namely noise and amplitude.

In Figure 39, we show the number of records for each skin tone class that have optical quality ≥ 0.65 (Dataset 1). Of more interest is the percentage of records in each skin tone class that satisfy this quality criterion, which is shown in Figure 40 for males and females separately. This shows that there is a steady and rapid decline in the percentage of records meeting this quality threshold as the skin tone class increases. Consequently, there are very few records in the darker skin classes, with only 35 (all male) in class 6. Moreover, where there is a reasonable amount of data, namely the first 5 classes for males and the first 4 classes for females, the rate of decline is remarkably consistent. Also, for the first 4 classes, there is little difference in this percentage for males and females. This observation suggests that PPG signals are increasingly noisy for darker skin tones which is not dependent on sex.

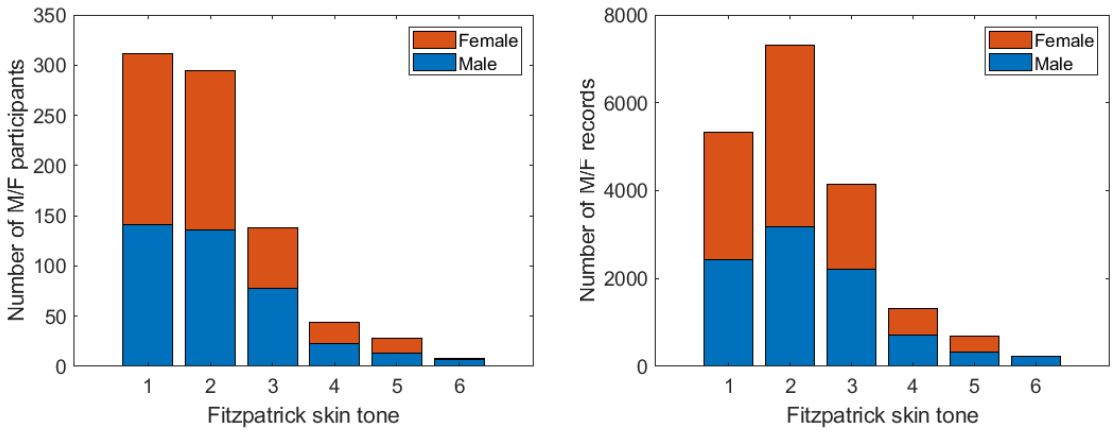


Figure 39: The number of subjects (left) and records (right) in each skin tone class.

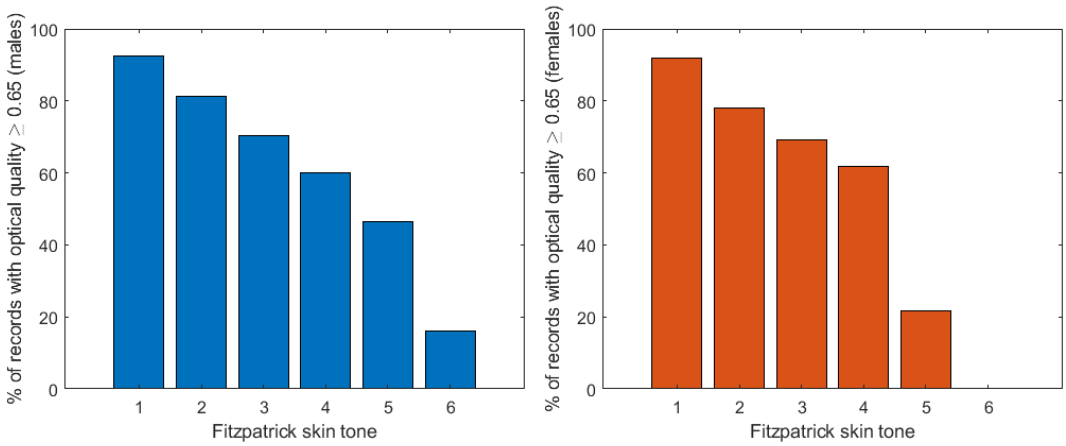


Figure 40: The percentage of records in each skin tone class that have optical quality ≥ 0.65 for males (left) and females (right).

The optical quality measure provided with the AuroraBP dataset penalises windows of data that have "substantial artifacts, poor signal-to-noise ratio, and lack of consistency between the pulses within a window" [64]. We investigated just the signal to noise ratio to see how this varies with skin tone. The median signal to noise ratio for all signals and signals with optical quality ≥ 0.65 for each skin tone class are shown in Figure 41 and the more informative box plots of signal to noise ratio for these two datasets are shown in Figure 42. We note that there is a steady decline in the median signal to noise ratio for all signals, indicating noisier signals for higher skin tone classes, but that decline does not start until class 4 for the better quality signals. As expected, the summary statistics shown in the box plots move in the positive direction for the better quality signals, although there are still some of these signals with a negative signal to noise ratio. This is likely to be due to baseline wander on some signals.

Next we consider the amplitude of the PPG signals by skin tone class. We find the amplitude for a signal by first filtering using a fourth order band pass Chebyshev II filter [67] with pass frequencies of 0.5–15 Hz [68] which filters out the baseline wander in particular. Some signals contain noise over a small part of the signal. We therefore split each signal in half and assessed the quality of each half by finding the height of the first peak of the autocorrelation function. For the half that had the higher peak (and thus was of higher quality), we found the amplitude as the difference between the 99^{th} and 1^{st} percentiles of the signal.

We found the amplitudes for all the signals, and for signals with optical quality ≥ 0.65. The median amplitude for each skin tone class is shown in Figure 43 for both datasets, with the median for all the signals having

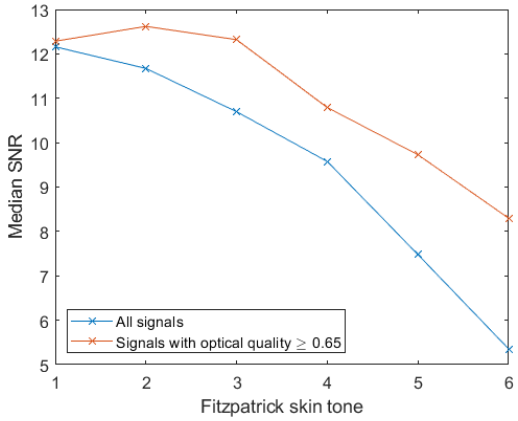


Figure 41: The median signal to noise ratio for all signals and for signals with optical quality ≥ 0.65.

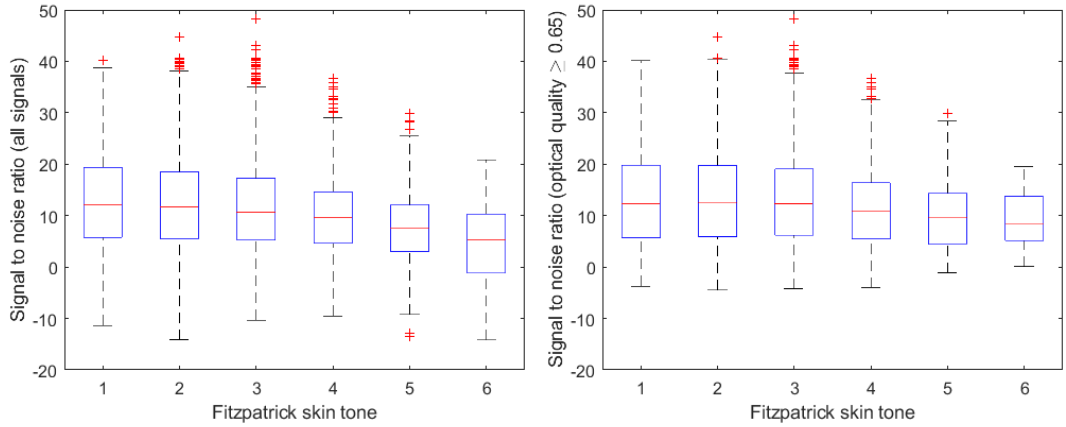


Figure 42: A box plot of the signal to noise ratio for all signals (left) and for signals with optical quality ≥ 0.65 (right).

the higher value for all classes. Again, we note that there is a steady and consistent decline in the median amplitude for increasing skin tone class. Of course the median only provides limited information, so we also show the box plots for both datasets in Figure 44 where some outliers (up to 459,971 in class 2 for all signals and 4,207 in class 1 for signals with optical quality ≥ 0.65) have been excluded from the plots. These plots show that there is a wide range of amplitudes for each skin tone class with overlapping ranges. However, the trend towards lower amplitudes for higher skin tone classes is clear. This is not altogether surprising, since darker skin will generally absorb more light and thus reflect less back, resulting in smaller amplitudes.

7.4 Clustering analysis

We performed a clustering analysis to see whether there were any clearly defined clusters in the data and, if so, whether they were aligned with the skin tone in any way. For this analysis, we used Symmetric Projection Attractor Reconstruction (SPAR) features. The SPAR method converts an approximately periodic signal into a compact image which encapsulates the morphology and variability of the signal [31].

To generate the SPAR images, we first filtered the signals using a fourth order high pass Chebyshev II filter with cutoff frequency of 0.5 Hz [68]. This removed any baseline wander from the signals which can affect the SPAR attractors. The size of the attractor is determined by the amplitude of the signals [31] and we have seen in Figure 44 that there is a huge range in the amplitudes in this dataset. Thus, we scaled the signals in the vertical direction by the amplitude, which was found as described in Section 7.3, so that all signals had

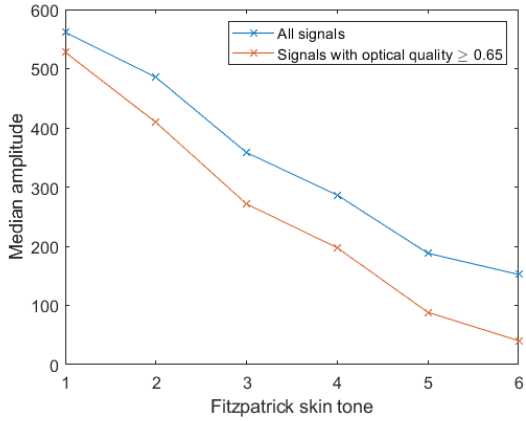


Figure 43: The median amplitude for all signals and for signals with optical quality ≥ 0.65.

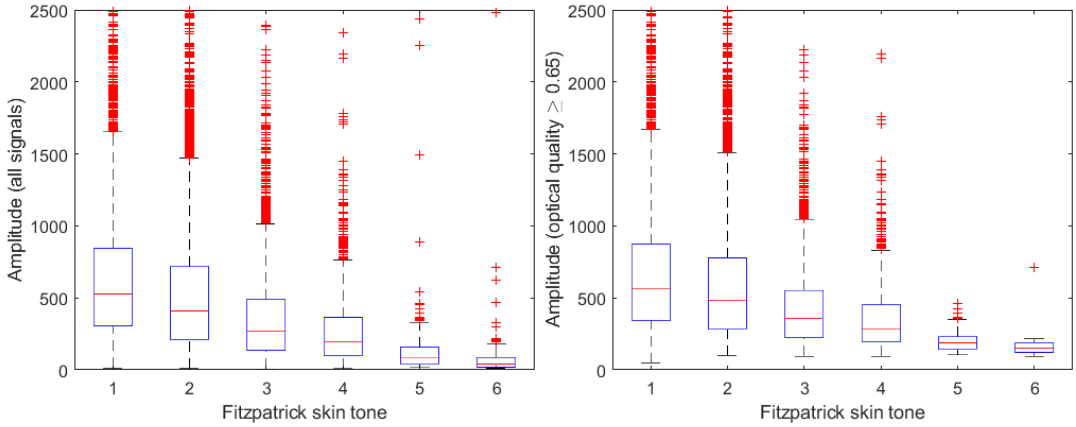


Figure 44: A box plot of amplitudes for all signals (left) and for signals with optical quality ≥ 0.65 (right). Some outliers in both cases have been excluded from the plot.

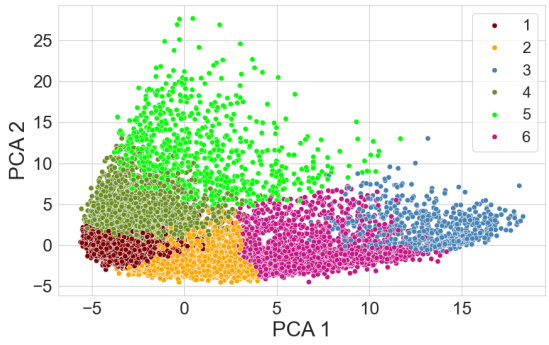
amplitude 1 resulting in consistently sized attractors. Note that this implies that the variation in amplitude for the different skin tone classes (see Figure 44) has been discarded. We then generated SPAR attractor images for each signal using a 64×64 grid and three delay coordinates, resulting in an approximate threefold rotational symmetry in the attractors.

To generate feature vectors from the SPAR attractor images, each point in the image was expressed in polar coordinates (r, θ) and then a density was generated for the r values and for the θ values. The r density vectors and θ density vectors were used as input for k-means clustering, with $k = 6$ to match the number of Fitzpatrick skin tone classes.

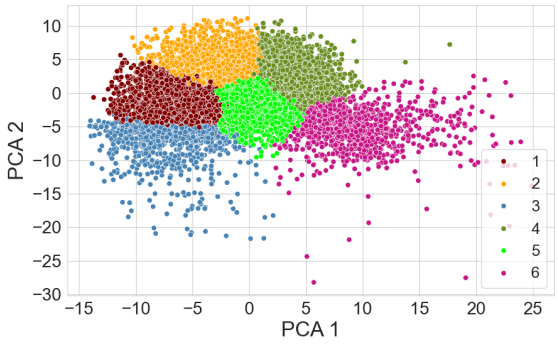
The results from the clustering analysis are shown in Figure 45 in a projection using the first two vectors from Principle Component Analysis (PCA), which reveals six clusters in both cases. In this projection, the clusters are quite distinct when using the θ density features, but there is some overlap when using the r density features. However, the underlying rationale for these groupings remains unclear.

To gain deeper insight into the clustering structure, we conducted a further clustering analysis by plotting the distribution of skin tone classes within each cluster. If the data was separated based on skin tone class, we would expect each cluster to have a dominant skin tone.

From Figure 46 we observe that some clusters have a bias towards either lighter skin tones (e.g. r density features, cluster 1) or darker skin tones (e.g. r density features, clusters 3 and 6) but many have a fairly uniform pattern across the skin tone classes. This implies that the clustering is based partly on skin tone

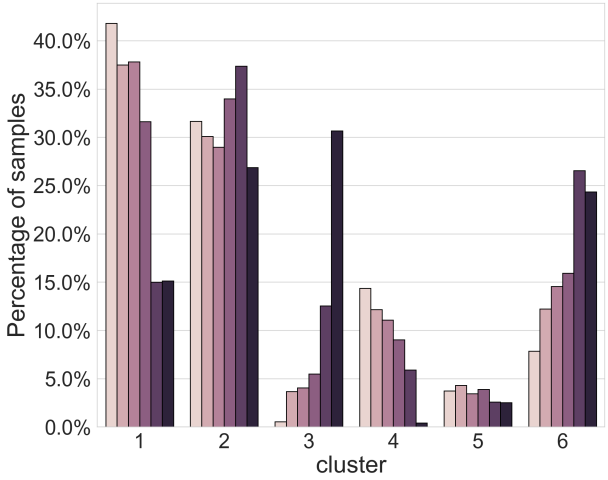


(a) k-means clustering results with r density features.

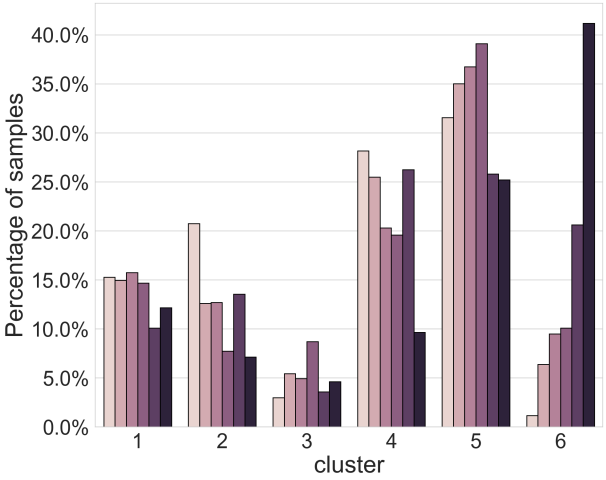


(b) k-means clustering results with θ density features.

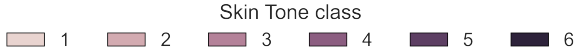
Figure 45: Results from k-means clustering when applied to (a) r density SPAR attractor features and (b) θ density SPAR attractor features.



(a) Clustering analysis for clusters based on r density features.



(b) Clustering analysis for clusters based on θ density features.



(c) Legend

Figure 46: Results from the cluster analysis on clusters found using (a) r density SPAR attractor features and (b) θ density SPAR attractor features.

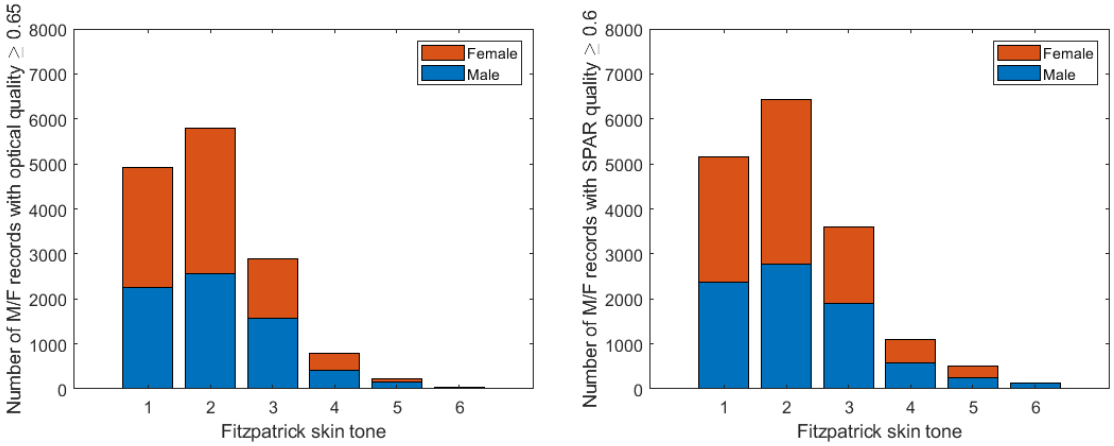


Figure 47: The number of male and female records in each skin tone class for Dataset 1 (optical quality ≥ 0.65) (left) and Dataset 2 (autocorrelation quality ≥ 0.6) (right).

but involves other, unknown factors as well.

7.5 Datasets

For classification of PPG signals by skin tone, we used four different datasets. The first two were used for six class classification and other two were used for a binary classification of "light" vs. "dark" skin. The binary classification also helps to address the dataset imbalance which can strongly influence model performance as we chose an approximately equal number of records in each of the two classes.

- **Dataset 1**

  The signals are of varying quality and a measure of "optical quality" is included in the dataset. This measure takes into account artifacts, signal to noise ratio and consistency between pulses [64]. Records with optical quality ≥ 0.65 were used in [64] so we used the same threshold. There are a total of 14,667 records satisfying this condition and the number of records in each class satisfying this condition are shown in Figure 47. Note that there are only 35 records for Fitzpatrick skin tone class 6.

- **Dataset 2**

  We constructed a second dataset using a different quality measure. The first step of the Symmetric Projection Attractor Reconstruction method [31] (which we later use) is to find the average cycle length of the signal. This is done by filtering the signal to remove any baseline wander and then constructing the graph of the autocorrelation function which finds the correlation between the signal and a time shifted copy of the signal. This function has the value 1 when the time shift is zero, since the signal perfectly correlates with itself. If the signal is exactly periodic with period T_0, then the autocorrelation function also has the value 1 whenever the time shift is any multiple of T_0. When the function is approximately periodic with average cycle length T, the autocorrelation function has peaks around the multiples of T. The average cycle length is then defined to be the shift corresponding to the first peak in the autocorrelation function (see Figure 48). However, the height of the peak can also be used as a measure of quality (or periodicity) of the signal since it achieves its maximum possible value of 1 when the signal is exactly periodic. Decreasing values of the peak indicate signals where the signal and its shift by the average cycle length correlate more poorly, as shown in Figure 48.

  This quality measure is very different from the measure of optical quality provided with the dataset, Indeed, a signal for one subject (o364) has the lowest possible value of optical quality, namely 0, but a high value of autocorrelation quality of 0.9422. The likely reason for this is that there is a lot of baseline wander on the signal, which results in a poor optical quality. However, the signals are filtered

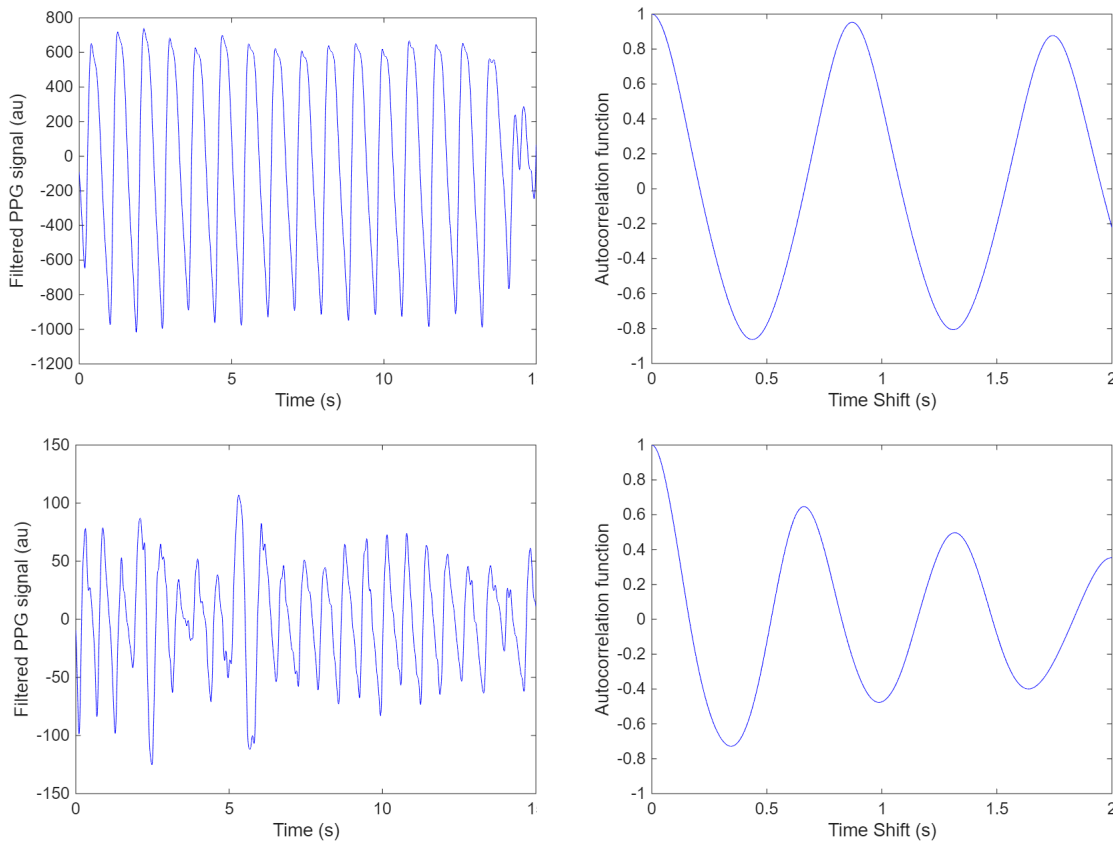


Figure 48: The top signal gives rise to an autocorrelation curve with the height of the first peak being 0.9533. The bottom signal gives rise to an autocorrelation curve with the height of the first peak being 0.6469, and is clearly of lower quality than the top signal.

to remove baseline wander before the construction of the autocorrelation function and once the baseline wander has been removed, it is quite a good quality signal.

For the second dataset, also summarised in Figure 47, we took all signals with autocorrelation quality ≥ 0.6. This gave a total of 16,898 records, which is larger than Dataset 1. Moreover, the number of Fitzpatrick skin tone class 6 records is 128 which is significantly larger (although still small) than the 35 in Dataset 1.

- **Dataset 3** We created a subset of the data with two classes, namely light (Fitzpatrick skin tone classes 1, 2, 3) and dark (Fitzpatrick skin tone classes 4, 5, 6). We selected records from Dataset 2 as the number of available records in class 6 is significantly higher than for Dataset 1. We first grouped classes 4, 5 and 6 representing dark skin tones. We then accounted for the number of unique patient IDs and records in this class and used this as a threshold to sample an equal (or as close to equal as possible) number of records from classes 1, 2 and 3 that make up the lighter skin tone class. Patient IDs were taken into account when sampling records to ensure a consistent amount of variability for both the classes. This process resulted in a balanced dataset with 1,718 records in each of the two classes.

- **Dataset 4** Due to the subjective nature of the Fitzpatrick labels, there will inevitably be some records labelled as class 3 that should actually be class 4 and vice versa, and so we refer to classes 3 and 4 as boundary classes. In Dataset 3, these records would be in the wrong binary subset. We thus eliminated this possibility by using only classes 1 and 2 for light skin, and classes 5 and 6 for dark skin. In this case, again using Dataset 2, the dark skin tone class was first created by grouping classes 5 and 6 and then we sampled records from classes 1 and 2 to match the number of records and patient IDs in the

dark skin tone class. This gave an approximately balanced dataset with 735 light records and 628 dark records.

7.6 Machine learning classification of skin tone

We consider several machine learning methods for this task, employing different inputs, namely machine learning using interpretable PPG features, deep learning or random forest using raw or filtered PPG signals, and deep learning using Symmetric Projection Attractor Reconstruction (SPAR) images [31].

7.6.1 Interpretable PPG features

We used the PulseAnalyse code [69] to extract PPG features from the raw signals.Further details about these PPG features are provided in [70]. This function is designed to analyse arterial pulse waves, including blood pressure and PPG signals.

A fourth-order Chebyshev Type II band-pass filter (0.5–15 Hz) was applied to remove low-frequency drift and high-frequency noise before passing the signals to the PulseAnalyse function. In total, 66 features were extracted, including 39 fiducial point–based features, 15 amplitude-based features, and 12 cardiovascular indices, and organised into a feature table of dimension $14{,}667 \times 66$. All missing values were replaced with zeros. Approximately 0.3 % (41 signals) could not be processed using PulseAnalyse, and additional effort will be devoted to investigating this issue.

Three supervised machine learning methods were applied to these features: a Neural Network, Classification Trees, and a Support Vector Machine. The Neural Network included two fully connected layers. The first layer had 20 nodes and was followed by a ReLU activation function.

For dataset 3 in the binary skin tone classification task, the same preprocessing and feature extraction procedure was used. A dataset with dimension $2{,}905 \times 66$ was created, and missing values were replaced with zeros. The same three supervised machine learning methods were applied. The Neural Network again had two fully connected layers. The first layer contained 50 nodes and was followed by a ReLU activation function.

For dataset 4, the same preprocessing, feature extraction, and modelling steps were performed. A feature table with dimension $1{,}176 \times 66$ was created, and missing values were replaced with zeros. The same three supervised machine learning methods were applied. The Neural Network again had two fully connected layers. The first layer contained 50 nodes and was followed by a ReLU activation function.

7.6.2 Raw or filtered signals

Multiclass classification problem

The signals in the various datasets have different lengths with minimum length 5,877 samples and so we considered the first 5,877 samples of each signal to ensure that every signal had uniform length.

In the next step, the signals are processed further in two different ways:

i) Machine learning and deep learning methods were applied directly to these raw signals.

ii) The signals were filtered with a Chebyshev type II bandpass filter of order 4 with 0.5 Hz and 15 Hz as cutoff frequencies and stopband attenuation of 20 dB. Machine learning and deep learning methods were then applied to these filtered signals.

The following machine learning and deep learning classifiers were evaluated:

- Random forest classifier from the scikit-learn package [71]

- Gradient boosted decision tree classifier from the yggdrasil decision forests package (ydf) [72]
- LeNet1d [1]
- ResNet1d50 [1]
- ResNet1d101 [1]
- VGG16 [1]

A stratified 5-fold cross-validation by subject was employed in all cases to ensure that no signals from a subject were in both the training and test datasets. For Gradient boosted decision trees, multiple loss functions were tested, including the multinomial log-likelihood, a cross entropy loss function and two self-defined variants of multinomial log-likelihood and cross entropy respectively reflecting fuzzy accuracy. This was done using label smoothing and an indicator matrix which treats adjacent-class predictions as correct.

Moreover, in the case of the tree-based models, accuracy and fuzzy accuracy were calculated also for the case that the classification problem was restricted to either signals from male or female subjects while using dataset 1.

Binary skin tone classification problems

For both binary datasets 3 and 4 that were described in 7.5, the raw signals were processed in the same way as in case of the multiclass classification problem and the same classifiers were evaluated. In case of the Gradient boosted decision trees, the multiclass specific loss functions were replaced by a binomial log likelihood loss function and a binary focal loss function. Again, as in the multiclass scenario, a stratified 5-fold cross-validation by subject was employed in all cases to ensure that no signals from a subject were in both the training and test datasets.

7.6.3 Deep learning with SPAR attractor images

SPAR attractor images were generated from each signal as described in Section 7.4, using only the first 15 s for any longer signals. We then treated the classification of skin tone as purely an image classification exercise using the ResNet50, ResNet152, and VGG16 models. All of the models were executed using PyTorch and were pretrained on IMAGENET weights. We used 5-fold cross validation by subject. Before training models, we performed augmentations on the dataset consisting of random horizontal and vertical flips, rotations and resized crop. This improves the diversity of the image dataset allowing the model to generalise better. Augmentations also help reduce overfitting and improve robustness of the model by expanding the amount of data available for model training [73, 74, 75, 76]. To account for the inaccuracies associated with the Fitzpatrick labels, we adopted the fuzzy accuracy for multi-class tasks, coupled with the Earth Mover's Distance (EMD) loss [77]. EMD loss accounts for the distance between probability bins allowing for misclassifications between closer labels to be penalised less, making it better suited to use with fuzzy accuracy when addressing label inaccuracy. Each of the models was trained using two loss metrics: a standard cross entropy loss when using strict accuracies and an EMD loss when using fuzzy accuracies.

7.7 Results

For all the results, we used 5-fold cross validation by subject.

7.7.1 Metrics

In this work, we consider the usual accuracy metric, which is the sum of the diagonal entries in the confusion matrix divided by the total number of records. However, we discussed the problem with subjective labelling

Table 33: Accuracy and fuzzy accuracy values for different classifiers and loss functions applied to the PPG features (dataset 1). The best results are shown in bold.

Classifier	**Loss Function**	**Accuracy**	**Fuzzy Accuracy**
Trees	Classification error	0.3473	0.8100
Neural Network	Cross entropy	**0.4219**	**0.8705**
Support Vector Machine	Classification error	0.4024	0.8638

Table 34: Accuracy values for different classifiers and loss functions applied to the PPG features from the binary dataset 3 (123 vs. 456). The best result is shown in bold.

Classifier	**Loss Function**	**Accuracy**
Trees	Classification error	0.5852
Neural Network	Cross entropy	**0.6723**
Support Vector Machine	Classification error	0.6282

Table 35: Accuracy values for different classifiers and loss functions applied to the PPG features from the binary dataset 4 (12 vs. 56). The best result is shown in bold.

Classifier	**Loss Function**	**Accuracy**
Trees	Classification error	0.6845
Neural Network	Cross entropy	**0.7253**
Support Vector Machine	Classification error	0.7049

of skin tone in Section 7.1 such that it is quite likely that some (and possibly many) skin tone labels are incorrect by one class. Thus, we also consider a new metric that we call *fuzzy accuracy* which counts a predicted skin tone class as correct if it differs from the skin tone class label by at most one class. So for example a record predicted to be in class 3 would be counted as correct if the skin tone label was class 2, 3 or 4. In terms of the confusion matrix, this equates to considering the diagonal together with one superdiagonal and one subdiagonal as being correct classifications, so clearly, by definition the fuzzy accuracy will be greater than or equal to the accuracy.

7.7.2 Interpretable PPG features

The results obtained for the accuracy and fuzzy accuracy for dataset 1 and for the different classifiers using PPG features are shown in Table 33.

Figure 49 shows an additional confusion matrix combined across all folds for the Neural Network model trained with the cross-entropy loss function. The evaluation was performed using 5-fold cross-validation by subject on raw signals from dataset 1. The numbers represent classification results per record, not per subject.

The results obtained for different classifiers using PPG features applied to the two binary classifications using datasets 3 and 4 are shown in Tables 34 and 35. The corresponding confusion matrices are shown in Figure 50. In all cases, the Neural Network model gave the best results.

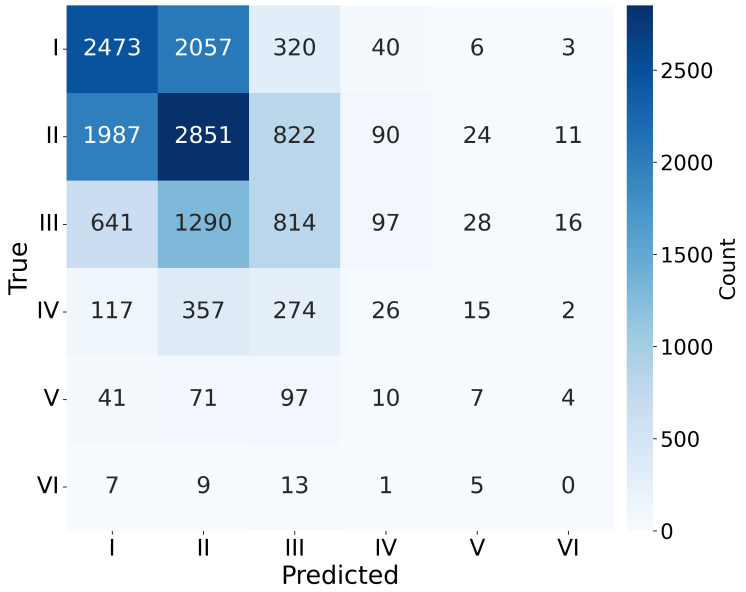
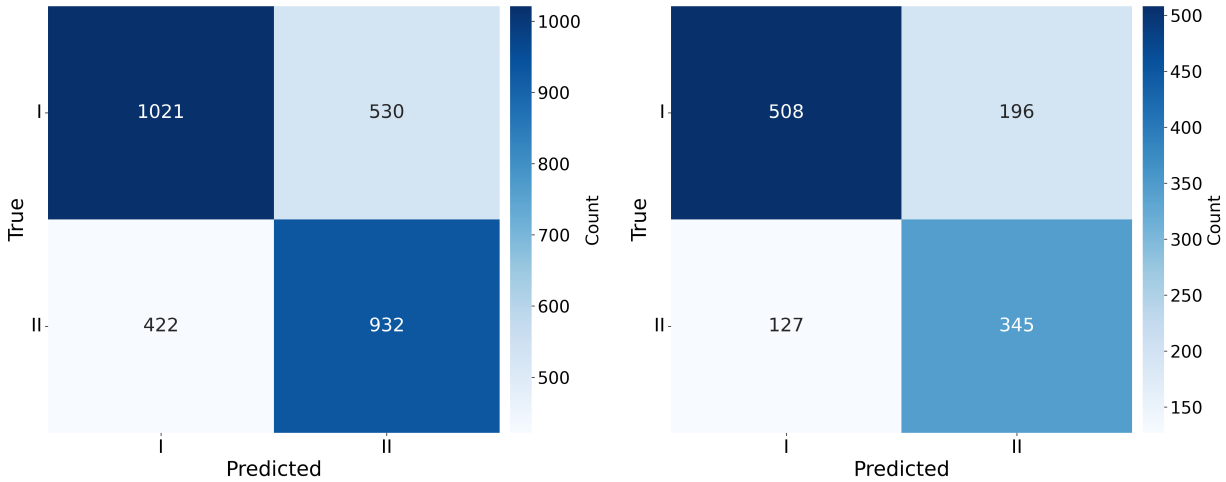


Figure 49: Confusion matrix over 5-fold cross validation for the Neural Network model using PPG features with cross entropy loss function trained on raw signals from dataset 1.



(a) Neural network model trained on dataset 3.

(b) Neural network model trained on dataset 4.

Figure 50: Confusion matrices depicting model predictions when using the Neural Network model using PPG features trained on binary datasets 3 and 4 when predicting light (class I) and dark (class II) skin tones.

Table 36: Accuracy and fuzzy accuracy values for different classifiers and loss functions applied to the raw signals (dataset 1). The best results are shown in bold. GBT = gradient boosted decision tree, RF = random forest classifier.

Classifier	**Loss Function**	**Accuracy**	**Fuzzy Accuracy**
GBT	Cross entropy	0.5502	0.9561
GBT	Fuzzy cross entropy	0.5336	**0.9594**
GBT	Multinomial log likelihood	**0.5503**	0.9561
GBT	Fuzzy multinomial log likelihood	0.4585	0.9339
RF	Standard	0.5332	0.9531
LeNet1d	Cross Entropy	0.3748	0.8433
ResNet1d50	Cross entropy	0.3499	0.8454
ResNet1d101	Cross entropy	0.2990	0.7204
VGG16	Cross entropy	0.3954	0.9278

Table 37: Accuracy and fuzzy accuracy values for different classifiers and loss functions applied to the filtered signals (dataset 1). The best results are shown in bold. GBT = gradient boosted decision tree, RF = random forest classifier.

Classifier	**Loss Function**	**Accuracy**	**Fuzzy Accuracy**
GBT	Cross entropy	0.5427	0.9563
GBT	Fuzzy cross entropy	0.5369	0.9554
GBT	Multinomial log likelihood	0.5427	0.9563
GBT	Fuzzy multinomial log likelihood	0.4440	0.9264
RF	Standard	**0.5431**	**0.9564**
LeNet1d	Cross entropy	0.3837	0.8413
ResNet1d50	Cross entropy	0.2677	0.6300
ResNet1d101	Cross entropy	0.2609	0.6611
VGG16	Cross entropy	0.3815	0.8904

7.7.3 Raw or filtered signals

Multiclass skin tone classification

The results obtained for the accuracy and fuzzy accuracy for dataset 1 and for the different classifiers and loss functions are shown in Table 36. Table 37 shows similar results for the filtered signals in dataset 1.

An exemplary confusion matrix combined over all folds for the case of Gradient boosted decision trees with fuzzy cross entropy loss function with 5-fold cross validation by subject on raw signals from dataset 1 is shown in Figure 51. (Note that the numbers indicate classification by record not by subject.)

The results obtained for the accuracy and fuzzy accuracy for the slightly larger dataset 2 and for the different classifiers and loss functions are shown in Table 39 while the results using the filtered signals are shown in Table 40.

Clearly, the results are very similar when using either the raw or the filtered signals from dataset 1 and also

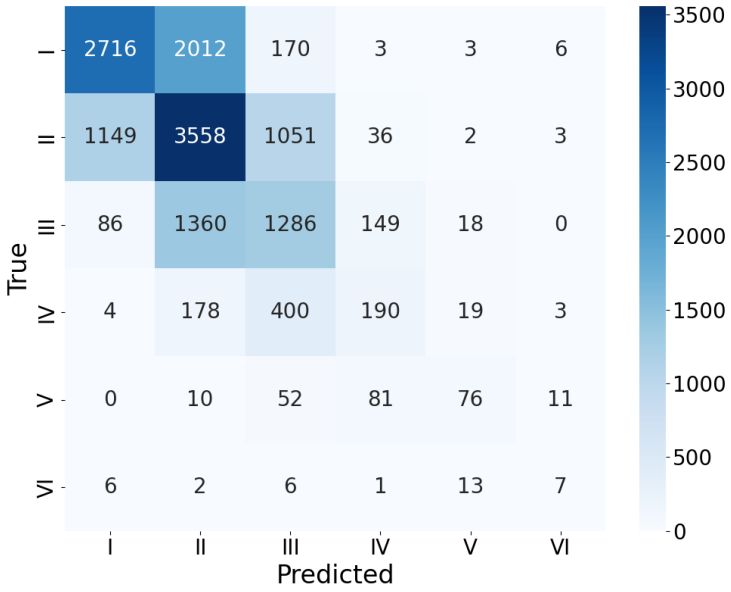


Figure 51: Confusion matrix over 5-fold cross validation for the Gradient boosted decision tree model with fuzzy cross entropy loss function trained on raw signals from dataset 1.

Table 38: Accuracy and fuzzy accuracy values for different classifiers and loss functions applied to the raw signals (dataset 1) in different scenarios. The best results are shown in bold. GBT = gradient boosted decision tree, RF = random forest classifier.

		Accuracy			**Fuzzy accuracy**		
Classifier	**Loss Function**	male only	female only	all	male only	female only	all
GBT	Fuzzy cross entropy	0.5202	0.5458	0.5336	**0.9553**	0.9639	**0.9594**
GBT	Multinomial log likelihood	**0.5284**	**0.5663**	**0.5503**	0.9488	0.9643	0.9561
RF	Standard	0.5149	0.5495	0.5332	0.9360	**0.9651**	0.9531

very similar when using either the raw or the filtered signals from dataset 2. Overall, the results for dataset 1 are slightly better. For dataset 1, the best fuzzy accuracy for the raw signals was obtained by the Gradient boosted tree classifier with the fuzzy cross entropy loss function, while for the filtered signals, the Random forest classifier gave the best results. For dataset 2, however, the best fuzzy accuracy was obtained by the Gradient boosted tree classifier.

Finally, we note that in all scenarios, tree-based methods gave better results than deep learning methods. The evaluation of tree-based models (i.e. Random forest and Gradient boosted decision trees) on female or male only subsets of dataset 1 delivered similar values for accuracy and fuzzy accuracy with a difference that was always less than 0.04 as shown in Table 38.

Binary skin tone classification

The results for the binary classification of light vs. dark skin tone including boundary classes (dataset 3) are shown in Table 41. Those for the binary classification of light vs. dark skin tone without boundary classes (dataset 4) are shown in Table 42.

For dataset 3, i.e. with boundary classes, high accuracy values of up to nearly 90 % were obtained for tree-based models. Furthermore, if we omit these boundary classes in dataset 4, an accuracy of more than 99 % accuracy was achieved, and so the records were classified nearly perfectly.

Table 39: Accuracy and fuzzy accuracy values for different classifiers and loss functions applied to the raw signals (dataset 2). The best results are shown in bold. GBT = gradient boosted decision tree, RF = random forest classifier.

Classifier	**Loss Function**	**Accuracy**	**Fuzzy Accuracy**
GBT	Cross entropy	**0.5303**	**0.9464**
GBT	Fuzzy cross entropy	0.5156	**0.9464**
GBT	Multinomial log likelihood	**0.5303**	**0.9464**
GBT	Fuzzy multinomial log likelihood	0.4438	0.9143
RF	Standard	0.5140	0.9403
LeNet1d	Cross Entropy	0.3534	0.8142
ResNet1d50	Cross entropy	0.2850	0.6898
ResNet1d101	Cross entropy	0.2895	0.7027
VGG16	Cross entropy	0.3277	0.7673

Table 40: Accuracy and fuzzy accuracy values for different classifiers and loss functions applied to the filtered signals (dataset 2). The best results are shown in bold. GBT = gradient boosted decision tree, RF = random forest classifier.

Classifier	**Loss Function**	**Accuracy**	**Fuzzy Accuracy**
GBT	Cross entropy	**0.5302**	0.9427
GBT	Fuzzy cross entropy	0.5202	**0.9466**
GBT	Multinomial log likelihood	**0.5302**	0.9427
GBT	Fuzzy multinomial log likelihood	0.4311	0.9094
RF	Standard	0.5246	0.9447
LeNet1d	Cross Entropy	0.3807	0.8209
ResNet1d50	Cross entropy	0.3084	0.7658
ResNet1d101	Cross entropy	0.2011	0.5047
VGG16	Cross entropy	0.3497	0.8068

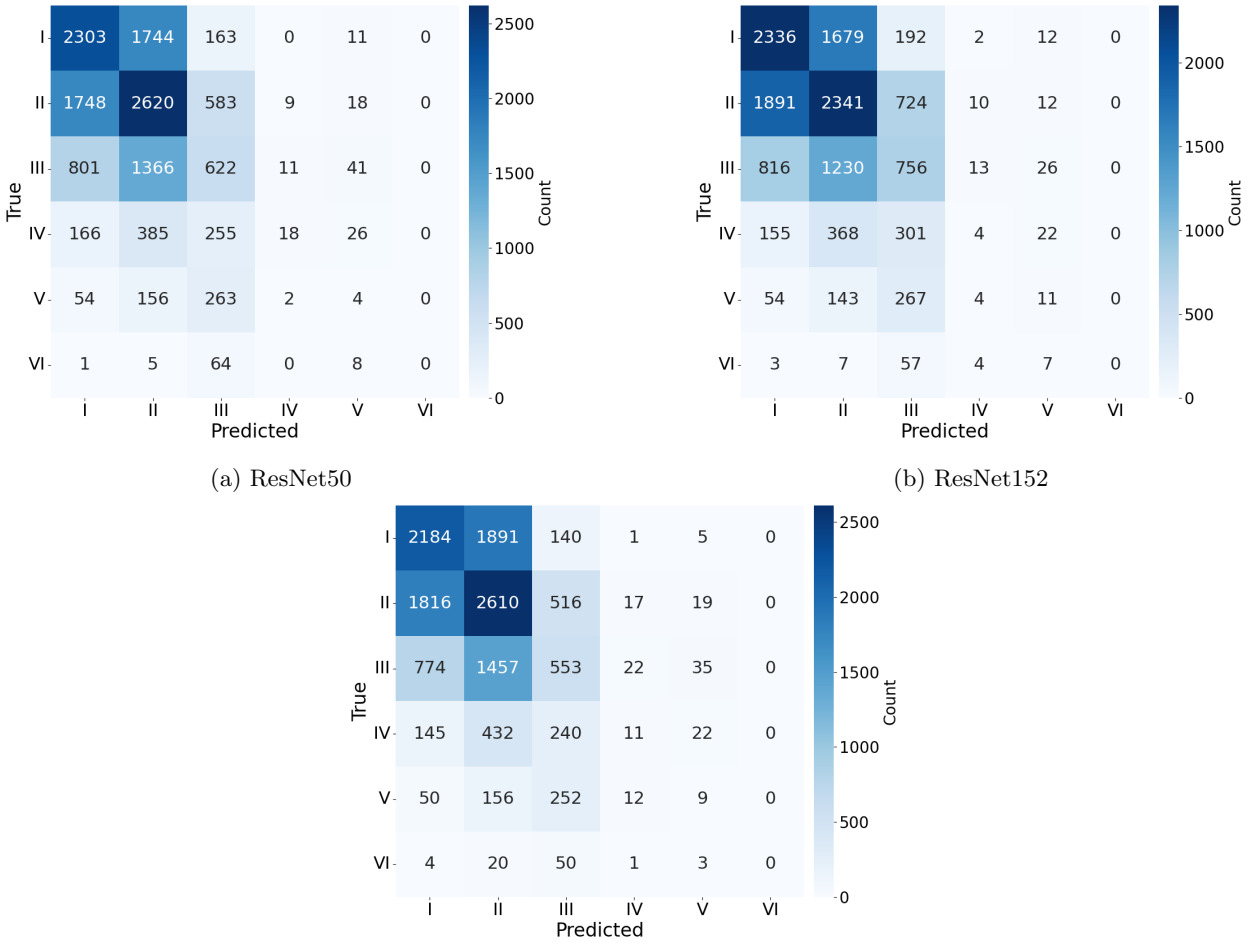
7.7.4 Deep learning with SPAR images

In this section we outline the results from using deep learning models for multiclass and binary classification of skin tones using SPAR images.

Multiclass skin tone classification

We assessed ResNet50, ResNet152 and VGG16 models using accuracy and fuzzy accuracy, implemented with 5 fold cross validation by subject and the results are shown in Table 43. These results indicate that VGG16 was the best model, achieving a fuzzy accuracy of 0.8588 and strict accuracy of 0.4603. Table 44 lists classwise performance of each of the models.

Model predictions were illustrated using confusion matrices depicted in Figure 52. We observe that the models are able to classify lighter skin tones to some extent but fail when it comes to darker skin tones.



(a) ResNet50

(b) ResNet152

(c) VGG16

Figure 52: Confusion matrices for ResNet50, ResNet152, and VGG16 trained on SPAR images with 5 fold cross validation.

Table 41: Accuracy values for different classifiers and loss functions applied to the raw signals from the binary dataset 3 (123 vs. 456). The best result is shown in bold. GBT = gradient boosted decision tree, RF = random forest classifier.

Classifier	**Loss Function**	**Accuracy**
GBT	Binomial log likelihood	0.8373
GBT	Binary focal loss	**0.8414**
RF	Standard	0.8225
LeNet1d	Cross Entropy	0.6714
ResNet1d 50	Cross Entropy	0.5445
ResNet1d 101	Cross Entropy	0.5489
VGG16	Cross Entropy	0.4837

Table 42: Accuracy values for different classifiers and loss functions applied to the raw signals from the binary dataset 4 (12 vs. 56). The best result is shown in bold. GBT = gradient boosted decision tree, RF = random forest classifier.

Classifier	**Loss Function**	**Accuracy**
GBT	Binomial log likelihood	0.9596
GBT	Binary focal loss	**0.9611**
RF	Standard	0.9567
LeNet 1d	Standard	0.6875
ResNet1d50	Cross entropy	0.5767
ResNet1d101	Cross entropy	0.5745
VGG16	Cross entropy	0.5363

Binary skin tone classification

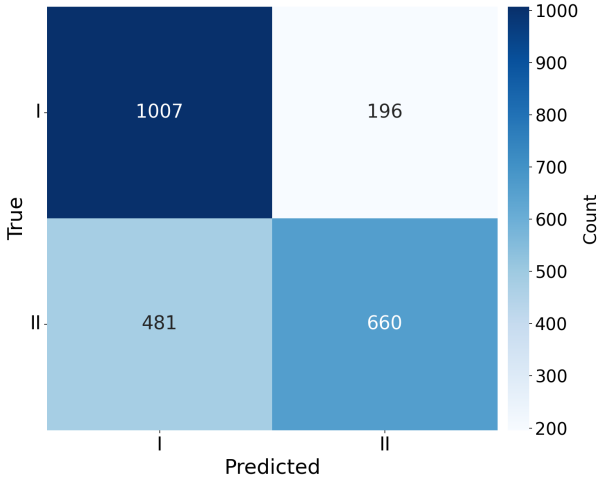
The performance comparison of ResNet50, Resnet152 and VGG16 for the two binary classification problems using datasets 3 and 4 is shown in Tables 45 and 46. The corresponding confusion matrices are shown in Figure 53.

We noticed a massive improvement in standard accuracy after adopting a binary classification approach with ResNet50 achieving an accuracy of 0.8097 when trained on dataset 4 without the boundary classes. When trained on dataset 3 with boundary classes, VGG16 achieved the best accuracy of 0.7125.

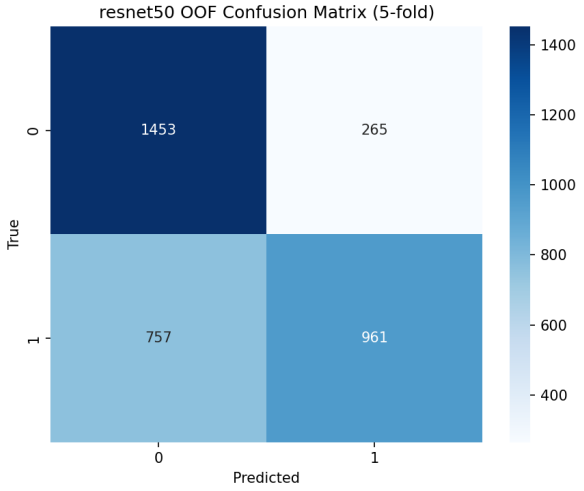
7.8 Uncertainty quantification

The underlying doubt in a given prediction was further analysed using the deep ensembles technique and was measured via the Uncertainty Calibration Error (UCE) [2, 78]. This UCE metric visualises the "mean predicted uncertainty (expressed as entropy) vs. mean observed error in each uncertainty bin." In particular, it is "the weighted average of the squared difference between the average entropy of the predicted distribution and the corresponding mean inaccuracy of predictions binned by magnitude of the entropies." [2, Table4]. An implementation of the UCE can be found in [79]. We implement deep ensembles with the best performing models. With respect to the SPAR image dataset, this was Resnet50 (marginally) with dataset 4. The model reported a UCE value of 0.4749 and an ECE value of 0.0250.

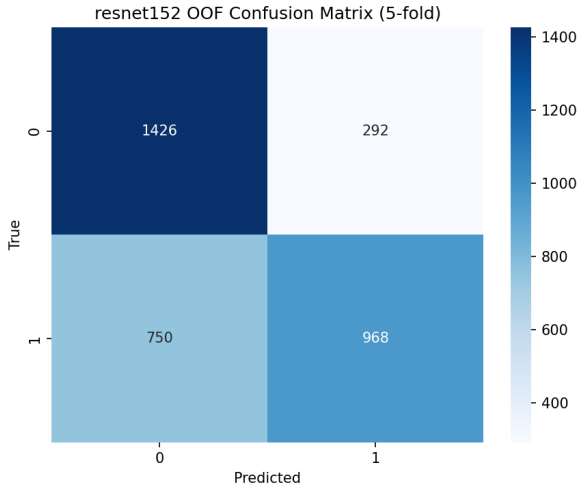
With respect to raw signals from dataset 3, this was GBT with binomial log-likelihood loss function. The model reported a UCE value of 0.6097. With respect to raw signals from dataset 4, Random forest and both



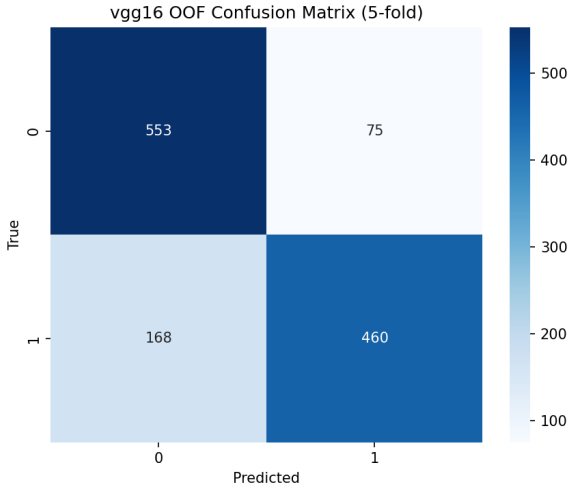
(a) VGG16 trained on dataset 3.



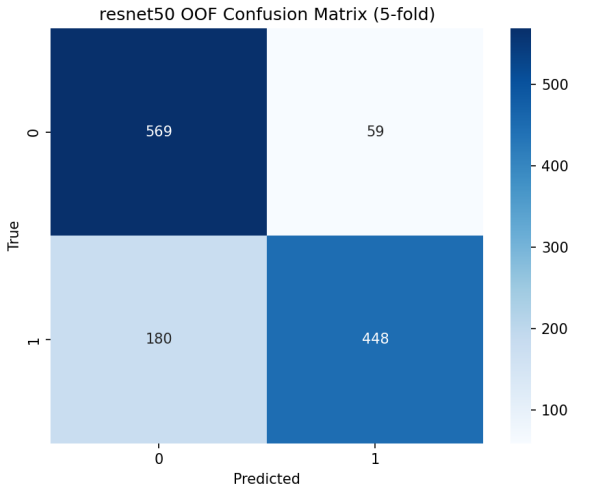
(b) ResNet50 trained on dataset 3.



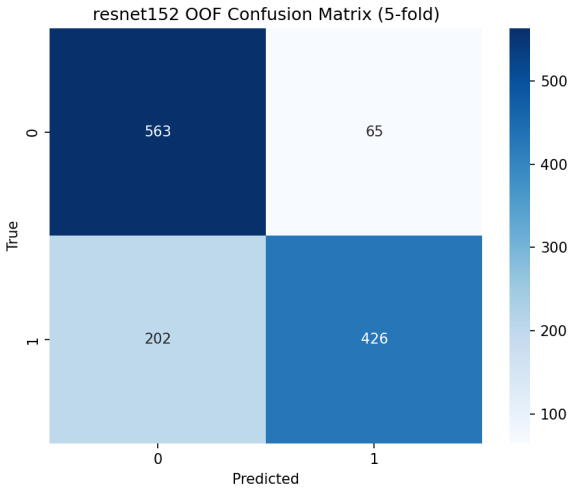
(c) ResNet152 trained on dataset 3.



(d) VGG16 trained on dataset 4.



(e) ResNet50 trained on dataset 4.



(f) ResNet152 trained on dataset 4.

Figure 53: Confusion matrices depicting model predictions when using VGG16, ResNet50 and ResNet152 models trained on binary datasets 3 and 4 when predicting light (class I) and dark (class II) skin tones.

Table 43: Performance metrics for ResNet50, ResNet152, and VGG16 trained on multi class SPAR images using cross entropy loss with strict accuracy and EMD loss with fuzzy accuracy. The best results are shown in bold.

Classifier	Loss function	Loss Score	Accuracy	Fuzzy Accuracy
ResNet50	Cross entropy	1.2138	0.4096	0.8454
ResNet152	Cross entropy	1.2071	**0.4439**	0.8450
VGG16	Cross entropy	1.2111	0.4018	0.8295
ResNet50	Earth Mover Distance	0.0761	0.4097	0.8514
ResNet152	Earth Mover Distance	0.0783	0.4339	0.8504
VGG16	Earth Mover Distance	0.076	0.4047	**0.8600**

Gradient boosted decision tree variants gave similar values for accuracy, but different UCE values. Table 47 show very good UCE results for Random Forest and much worse for both Gradient boosted decision tree variants.

Figure 54 illustrates a reliability diagram to visualise model calibration. The diagram plots model confidence (maximum probability) against empirical accuracy. We observe that in higher confidence bins, the model achieves near perfect calibration. The rapid increase in empirical accuracy near mid range bins shows the model is moderately confident. The model rarely outputs predictions of low confidence. In confidence 0.5 - 0.7 we observe some degree of overconfidence which quickly decreases as confidence increases. These characteristics highlight that the model is well calibrated and produces reliable predictions with moderate to high confidence.

Figure 55 depicts model predictions and their associated variance across model ensembles against the maximum predicted probability. We observe that misclassifications are often associated with higher variance and mid range predicted probabilities, while correct classifications are associated with high predicted probabilities with mostly lower variance. We observe some prediction outliers associated with mid range predicted probability and very high variance. We observe relatively clear separation of correct predictions in high confidence regions. We once again have no predictions of low predictive probability, indicating the model only outputs when confidence is reasonably high and variance is low.

7.9 Conclusions

If skin tone had no influence on the PPG signals, then machine learning would not be able to classify signals according to Fitzpatrick skin tone class with any accuracy. In this scenario, the labels would essentially be allocated randomly and so we would expect approximately 1 in 6 (16.67 %) to be correct. The accuracies we obtained were up to around 55 % which is clearly much higher, but still poor in classification terms. However, when we take into account the subjective nature of the Fitzpatrick skin tone labels, and define a corresponding fuzzy accuracy which counts a prediction as correct if the prediction and the label differ by at most one class, then we obtain much higher accuracies of around 95 %, thus demonstrating a clear dependence of the PPG signals on skin tone.

The best results for this classification problem when considering fuzzy accuracy were obtained by deep learning models using raw signals as input. In particular, the best fuzzy accuracy was obtained using a Gradient boosted decision tree model with a fuzzy cross entropy loss when working with the raw signals, or a Random forest model when working with filtered raw signals. There was very little difference between the results using raw and filtered signals and predictions of the underrepresented darker skin tones was reasonable. The SPAR and PPG features methods gave fuzzy accuracies approximately 10 % lower, and with poor prediction of the darker skin tones, with the dominant lighter skin tone classes being favoured. However, we note that the SPAR method also normalised the amplitudes of the signals in order to give comparably

Table 44: Classwise performance metrics for ResNet50, ResNet152 and VGG16 implemented with 5 fold cross validation when using strict and fuzzy accuracies for predicting skin tone labels from SPAR images.

Model	Accuracy Type	Class	Precision	Recall	F1-score
ResNet152	Strict	Class 1	0.4352	0.5394	0.4818
		Class 2	0.4153	0.4831	0.4467
		Class 3	0.3065	0.2503	0.2755
		Class 4	0.2000	0.0059	0.0114
		Class 5	0.0641	0.0104	0.0180
		Class 6	0.0000	0.0000	0.0000
ResNet50	Strict	Class 1	0.4468	0.5229	0.4818
		Class 2	0.4013	0.5625	0.4684
		Class 3	0.3358	0.1749	0.2300
		Class 4	0.4545	0.0059	0.0116
		Class 5	0.0000	0.0000	0.0000
		Class 6	0.0000	0.0000	0.0000
VGG16	Strict	Class 1	0.4297	0.6330	0.5119
		Class 2	0.3900	0.4221	0.4054
		Class 3	0.3116	0.1816	0.2295
		Class 4	0.2133	0.0188	0.0346
		Class 5	0.0753	0.0146	0.0245
		Class 6	0.0000	0.0000	0.0000
ResNet152	Fuzzy	Class 1	0.7823	0.9270	0.8485
		Class 2	0.9057	0.9950	0.9482
		Class 3	0.7593	0.6864	0.7210
		Class 4	0.9586	0.3812	0.5455
		Class 5	0.1000	0.0125	0.0223
		Class 6	0.9000	0.1154	0.2045
ResNet50	Fuzzy	Class 1	0.8083	0.9747	0.8837
		Class 2	0.8709	0.9988	0.9305
		Class 3	0.8558	0.7163	0.7798
		Class 4	0.9959	0.2824	0.4400
		Class 5	0.0000	0.0000	0.0000
		Class 6	1.0000	0.1026	0.1860
VGG16	Fuzzy	Class 1	0.7710	0.9690	0.8582
		Class 2	0.8947	0.9916	0.9406
		Class 3	0.8290	0.6434	0.7245
		Class 4	0.9048	0.3353	0.4893
		Class 5	0.1667	0.0271	0.0467
		Class 6	0.2381	0.0641	0.1010

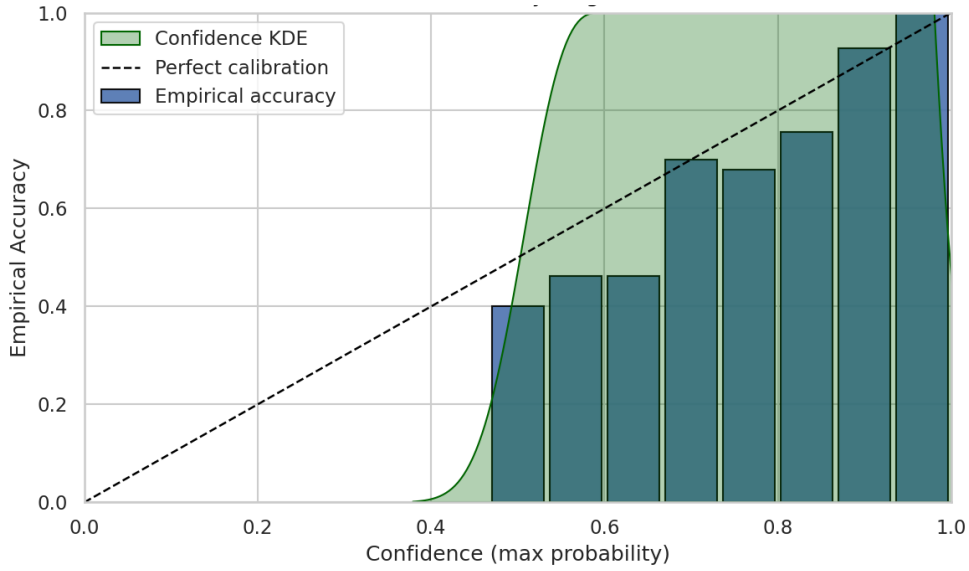


Figure 54: Reliability diagram representing empirical accuracy vs. model confidence when using the best performing model - VGG16 with dataset 4 and SPAR images.

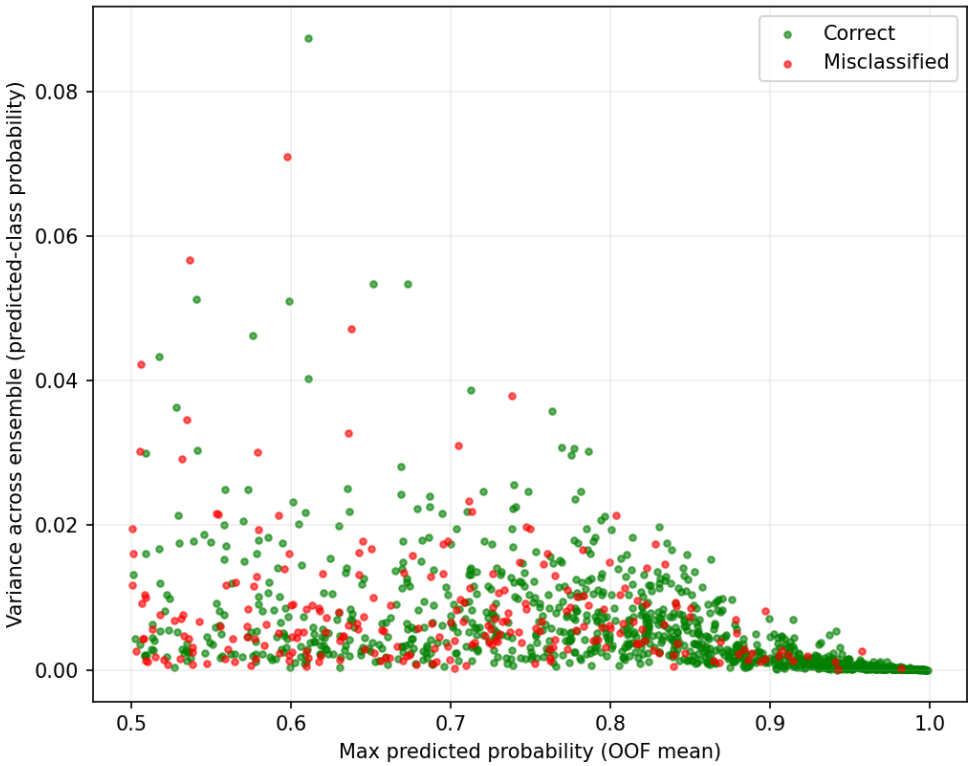


Figure 55: Visualising model prediction variance when using the best performing model - VGG16 with dataset 4 and SPAR images.

Table 45: Accuracies for ResNet50, ResNet152 and VGG16 trained on SPAR images for the two binary classifications. The best results are shown in bold.

Model	Dataset	Accuracy
ResNet50	3	0.7026
ResNet 152	3	0.6967
VGG16	3	**0.7125**
ResNet50	4	**0.8097**
ResNet152	4	0.7874
VGG16	4	0.8065

Table 46: Classwise performance metrics for ResNet50, ResNet152 and VGG16 implemented with 5 fold cross validation when predicting light (class I) and dark (class II) skin tones from SPAR images.

Model	Dataset	Class	Precision	Recall	F1-score
ResNet152	3	Class I	0.6553	0.8300	0.7324
		Class II	0.7683	0.5634	0.6501
ResNet50	3	Class I	0.6575	0.84587	0.7398
		Class II	0.7838	0.5594	0.6529
VGG16	3	Class I	0.6753	0.8184	0.7400
		Class II	0.7696	0.6065	0.6784
ResNet152	4	Class I	0.7359	0.8965	0.8083
		Class II	0.8676	0.6783	0.7641
ResNet50	4	Class I	0.7597	0.9061	0.8264
		Class II	0.8836	0.7134	0.7894
VGG16	4	Class I	0.7670	0.8806	0.8199
		Class II	0.8598	0.7325	0.7911

Table 47: Accuracy and UCE values for different classifiers and loss functions applied to the raw signals from the binary dataset 4 (12 vs. 56). GBT = gradient boosted decision tree, RF = random forest classifier.

Classifier	**Loss Function**	**Accuracy**	**UCE**
GBT	Binomial log likelihood	0.9596	0.8966
GBT	Binary focal loss	0.9611	0.6689
RF	Standard	0.9567	0.0039

sized attractors and so the variation in amplitude with skin tone is lost in this case. We showed in Section 7.3 that the amplitude of the signals is skin tone dependent and so clearly this useful information is not available for the SPAR method, resulting in lower classification accuracies.

Similarly, for the binary classification of light vs. dark skin tones, we obtained an accuracy of around 70 % using the PPG features and for the SPAR method or around 90 % using the raw signals with the boundary skin tone classes 3 and 4 included. Clearly, given the subjective nature of the labels, some subjects who were classified as class 3 should have been class 4 and vice versa, and these subjects were therefore mislabelled for the binary classification. Removing these boundary classes gives a more accurately labelled (but smaller) dataset on which we achieved a classification accuracy of around 80 % using SPAR or an impressive 96 % using the raw signals, indicating near perfect classification. Thus, we conclude that there is a very distinct and discernable difference between PPG signals for light and dark skin.

When the data was separated out into male and female subsets, the accuracy and fuzzy accuracy were both a little higher for females than for males. We conjecture that one possible reason for this is the fact that there are no female records in class 6, which is also the smallest class and so is more likely to be misclassified.

The uncertainty quantification results for the binary case and raw signals illustrate that the accuracy values might be quite similar, but the UCE metrics could be quite different. While the UCE for Gradient boosted decision trees is relatively high, the results for Random forest are very low and hence very good.

We note that the best performing models show near perfect calibration at higher confidence bins and do not report any predictions in lower confidence bins. Models report some under confidence at mid range bins, indicating they are more right with predictions than expected. Models report lower misclassification with a reduction in variance in predicted class probabilities. Results from the uncertainty quantification metrics indicate models with balanced records and suitable separation between skin tones are able to provide reliable and trustworthy predictions that strongly differentiate between the light and dark skin tones.

We acknowledge a limitation of this study in that we used a dataset which is highly imbalanced with regard to skin tone for the six class classification. In future, it would be interesting to repeat this exercise using a more skin tone balanced dataset.

A consequence of this work is that skin tone should be taken into consideration in the calibration and use of wearable devices and pulse oximeters. Failure to do so could result in biased results and, potentially, consequent harm to the wearer.

References

[1] M. Moulaeifard, L. Coquelin, M. Rinkevičius, et al. Machine learning for photoplethysmography analysis: benchmarking feature, image, and signal-based approaches arXiv:2502.19949, 2025.

[2] C. Bench, O. Pfeffer, V. Desai, et al. A systematic evaluation of uncertainty quantification techniques in deep learning: A case study in photoplethysmography signal analysis. *Machine Learning: Health*, 2026.

[3] 22HLT01 QUMPHY D4 deliverable report, 2025.

[4] R. Mieloszyk, H. Twede, J. Lester, et al. A comparison of wearable tonometry, photoplethysmography, and electrocardiography for cuffless measurement of blood pressure in an ambulatory setting. *IEEE Journal of Biomedical and Health Informatics* 26(7), 2864–2875, 2022.

[5] MIMIC PERform AF dataset. `https://ppg-beats.readthedocs.io/en/latest/datasets/mimic_perform_af/`.

[6] A. Bernardini, A. Brunello, G. L. Gigli, et al. OSASUD: A dataset of stroke unit recordings for the detection of Obstructive Sleep Apnea Syndrome. *Scientific Data* 9(1), 177, 2022.

[7] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics 1992.

[8] C. Guo, G. Pleiss, Y. Sun, et al. On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330, 2017.

[9] D. Levi, L. Gispan, N. Giladi, et al. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors* 22(15), 5540, 2022.

[10] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Advances in Large Margin Classifiers*. Cambridge, MA, 2000, 61–74.

[11] M. Kull, T. Silva Filho, and P. Flach. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In: *Artificial Intelligence and Statistics*, 623–631, 2017.

[12] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International Conference on Machine Learning*, 625–632, 2005.

[13] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In: *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, 609–616, 2001.

[14] M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 291, 2015.

[15] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press 2006.

[16] A. Rahimi, T. Mensink, K. Gupta, et al. Post-hoc calibration of neural networks by g-layers. *arXiv:2006.12807*, 2020.

[17] H. Song, T. Diethe, M. Kull, et al. Distribution calibration for regression. In: *International Conference on Machine Learning*, 5897–5906, 2019.

[18] J. V. Lyle, M. Nandi, and P. J. Aston. Symmetric Projection Attractor Reconstruction: Sex differences in the ECG. *Frontiers in Cardiovascular Medicine* 8, 709457, 2021.

[19] T. Aydemir, M. Şahin, and Ö. Aydemir. Gender recognition from PPG signals of people with different blood pressure. In: *2021 29th Signal Processing and Communications Applications Conference (SIU)*, 1–4, 2021.

[20] S. Dehghanojamahalleh and M. Kaya. Sex-related differences in photoplethysmography signals measured from finger and toe. *IEEE Journal of Translational Engineering in Health and Medicine* 7, 1900607, 2019.

[21] The Aurora-BP study and dataset. GitHub repository. Accessed: 2025-09-01.

[22] Aurora - eliminating cardiovascular surprises. `https://www.microsoft.com/en-us/research/project/aurora/`.

[23] Y. Liang, M. Elgendi, Z. Chen, et al. An optimal filter for short photoplethysmogram signals. *Scientific Data* 5, 2018.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[25] T. He, Z. Zhang, H. Zhang, et al. Bag of tricks for image classification with convolutional neural networks. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 558–567, June 2019.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations*, 2021.

[27] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[28] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In: *International Conference on Learning Representations*, 2017.

[29] I. Loshchilov and F. Hutter. SGDR: stochastic gradient descent with warm restarts. In: *International Conference on Learning Representations*, 2017.

[30] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, 1050–1059, 2016.

[31] P. Aston, M. Christie, Y. Huang, et al. Beyond HRV: Attractor reconstruction using the entire cardiovascular waveform data for novel feature extraction. *Physiological Measurement* 39, 024001, 2018.

[32] Z. J. Wang, R. Turko, O. Shaikh, et al. CNN explainer: Learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics* 27(2), 1396–1406, 2020.

[33] A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in neural information processing systems* 30, 2017.

[34] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems* 30, 2017.

[35] Y. Ovadia, E. Fertig, J. Ren, et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems* 32, 2019.

[36] A. Sološenko, A. Petrėnas, V. Marozas, et al. Modeling of the photoplethysmogram during atrial fibrillation. *Comput. Biol. Med.* 81, 130–138, 2017.

[37] B. Paliakaitė, A. Petrėnas, A. Sološenko, et al. Modeling of artifacts in the wrist photoplethysmogram: Application to the detection of life-threatening arrhythmias. *Biomed. Signal Process. Control.* 66, 102421, 2021.

[38] A. Sološenko, A. Petrėnas, B. Paliakaitė, et al. Detection of atrial fibrillation using a wrist-worn device. *Physiol. Meas.* 40(2), 025003, 2019.

[39] L. Paarmann. Butterworth filters. In: *Design and Analysis of Analog Filters: A Signal Processing Perspective*. Boston, MA: Springer US, 2001, 113–130.

[40] W. Wang, P. Mohseni, K. L. Kilgore, et al. PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health* 4, 1090854, 2023.

[41] S. González, W.-T. Hsieh, and T. P.-C. Chen. A benchmark for machine-learning based non-invasive blood pressure estimation using photoplethysmogram. *Scientific Data* 10(1), 149, 2023.

[42] M. Moulaeifard, P. H. Charlton, and N. Strodthoff. Generalizable deep learning for photoplethysmography-based blood pressure estimation: A benchmarking study. *arXiv:2502.19167*, 2025.

[43] X. Chen, R. Wang, P. Zee, et al. Racial/ethnic differences in sleep disturbances: The multi-ethnic study of atherosclerosis (mesa). *Sleep* 38(6), 877–888, 2015.

[44] A. S. Alarcón, N. M. Madrid, R. Seepold, et al. Obstructive sleep apnea event detection using explainable deep learning models for a portable monitor. *Frontiers in Neuroscience* 17, 1155900, 2023.

[45] T. Choksatchawathi et al. ApSense: Data-driven algorithm in PPG-based sleep apnea sensing. *IEEE Internet of Things Journal* 11(20), 33915–33926, 2024.

[46] M. A. Almarshad, S. Al-Ahmadi, M. S. Islam, et al. Adoption of transformer neural network to improve the diagnostic performance of oximetry for obstructive sleep apnea. *Sensors* 23(18), 7924, 2023.

[47] A. Bernardini, A. Brunello, G. L. Gigli, et al. AIOSA: An approach to the automatic identification of obstructive sleep apnea events based on deep learning. *Artificial Intelligence in Medicine* 118, 102133, 2021.

[48] Y. Ji, D. Chen, Y. Zuo, et al. Accurate apnea and hypopnea localization in PSG with multi-scale object detection via dual-modal feature learning. *Biomedical Signal Processing and Control* 89, 105717, 2024.

[49] P. Kulkarni et al. Obstructive sleep apnea syndrome identification using CNN-LSTM hybrid model. *Journal of Electrical Systems* 20(2), 2386–2394, 2024.

[50] M. Panwar, A. Gautam, D. Biswas, et al. PP-Net: A deep learning framework for PPG-based blood pressure and heart rate estimation. *IEEE Sensors Journal* 20(17), 10000–10011, 2020.

[51] N. Strodthoff, P. Wagner, T. Schaeffter, et al. Deep learning for ecg analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics* 25(5), 1519–1528, 2020.

[52] J. Torres-Soto and E. A. Ashley. Multi-task deep learning for cardiac rhythm detection in wearable devices. *npj Digital Medicine* 3(1), 116, 2020.

[53] H.-C. Lee, Y. Park, S. B. Yoon, et al. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data* 9(1), 279, 2022.

[54] *Solar 8000M Patient Monitor: Service Manual*. Revision C. Document number 2000701-123, dated 17 February 2003. GE Medical Systems Information Technologies. 2003.

[55] B. P. Imholz, W. Wieling, G. A. van Montfrans, et al. Fifteen years experience with finger arterial pressure monitoring: Assessment of the technology. *Cardiovascular research* 38(3), 605–616, 1998.

[56] M. Valdenegro-Toro and D. S. Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1508–1516, 2022.

[57] D. Martin. The impact of skin tone on performance of pulse oximeters used by NHS England COVID Oximetry @home scheme: Measurement and diagnostic accuracy study. *BMJ* 392, e085535, 2026.

[58] Equity in medical devices: Independent review. `https://www.gov.uk/government/publications/equity-in-medical-devices-independent-review-final-report`. Accessed: 22/07/2025. 2024.

[59] Performance evaluation of pulse oximeters taking into consideration skin pigmentation, race and ethnicity. `https://www.fda.gov/media/175828/download`. Accessed: 09/02/2026. 2024.

[60] R. Mukkamala, S. Shroff, K. Kyriakoulis, et al. Cuffless blood pressure measurement: Where do we actually stand? *Hypertension* 82, 957–970, 2025.

[61] A. Sološenko, A. Petrėnas, B. Paliakaitė, et al. Detection of atrial fibrillation using a wrist-worn device. *Physiological Measurement* 40, 025003, 2019.

[62] T. Fitzpatrick. The validity and practicality of sun-reactive skin types I through VI. *Arch. Dermatol.* 124, 869–871, 1988.

[63] P. Aston. Does skin tone affect machine learning classification accuracy applied to photoplethysmography signals? *Computing in Cardiology* 51, 038, 2024.

[64] B. Mieloszyk, H. Twede, J. Lester, et al. A comparison of wearable tonometry, photoplethysmography, and electrocardiography for cuffless measurement of blood pressure in an ambulatory setting. *IEEE Journal of Biomedical and Health Informatics* 26, 2864–2875, 2022.

[65] E. Monk. The Monk skin tone scale (MST). *SocArXiv* https://doi.org/10.31235/osf.io/pdf4c, 2023.

[66] K. Krishnapriya, G. Pangelinan, M. King, et al. Analysis of manual and automated skin tone assignments. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 429–438, 2022.

[67] L. Paarmann. Chebyshev type II filters. In: *Design and Analysis of Analog Filters: A Signal Processing Perspective*. Boston, MA: Springer US, 2001, 155–176.

[68] Y. Liang, M. Elgendi, Z. Chen, et al. An optimal filter for short photoplethysmogram signals. *Sci. Data* 5, 180076, 2018.

[69] P. H. Charlton. Pulse-Analyse. `https://github.com/peterhcharlton/pulse-analyse`. GitHub repository, accessed Jan. 20, 2026. 2023.

[70] P. H. Charlton, P. Celka, B. Farukh, et al. Assessing mental stress from the photoplethysmogram: A numerical study. *Phys. Meas.* 39(5), 054001, 2018.

[71] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12, 2825–2830, 2011.

[72] M. Guillame-Bert, S. Bruch, R. Stotz, et al. Yggdrasil decision forests: A fast and extensible decision forests library. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, 4068–4077, 2023.

[73] P. Krishnadas, K. Chadaga, N. Sampathila, et al. Classification of malaria using object detection models. *Informatics* 9(4), 76, 2022.

[74] A. Mikolajczyk and M. Grochowski. Data augmentation for improving deep learning in image classification problem. In: *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 117–122, 2018.

[75] M. Elgendi, M. Nasir, Q. Tang, et al. The effectiveness of image augmentation in deep learning networks for detecting COVID-19: A geometric transformation perspective. 2021.

[76] S. Yang, W. Xiao, M. Zhang, et al. Image data augmentation for deep learning: A survey. *arXiv:2204.08610*, 2023.

[77] L. Hou, C.-P. Yu, and D. Samaras. Squared earth mover's distance-based loss for training deep neural networks. *arXiv:1611.05916*, 2016.

[78] M.-H. Laves, S. Ihler, K.-P. Kortmann, et al. Uncertainty calibration error: A new metric for multi-class classification. 2021.

[79] Uncertainty quantification metrics. `https://gitlab.com/qumphy/d2-code/-/blob/dev/uq_eval/uq_metrics.py?ref_type=heads`. 2025.