Transfer Learning for Continuous Blood Pressure Change: Personalizing for SCI Individuals during Autonomic Dysregulation Episodes

Itilekha Podder, Udo Bub, Robert Riener, Jürgen Pannek, Inge Eriks-Hoogland and Diego Paez-Granados, Member, IEEE

S1. Extended Evaluation Tables (Aurora BP)

This section presents a detailed evaluation of the proposed multimodal framework for systolic blood pressure (SBP) change prediction on the Aurora BP test set (n=93). Performance is stratified by blood pressure category, including Normal, Elevated, Hypertension Stage 1, and Hypertension Stage 2, as defined by the American Heart Association (AHA). We evaluate four input configurations: PPG only, ECG only, dual-modal PPG+ECG, and a full multimodal variant incorporating demographic features (PPG+ECG+Demo). Metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Deviation (MD), each reported with their corresponding standard deviations (SD). These metrics reflect both the accuracy and consistency of the model across physiologically distinct BP categories. In addition to SBP change (ΔSBP), we also compare performance on static SBP prediction across all modalities.

Table 1: Modality wise performance comparison between SBP and ΔSBP prediction with MAE (SD) on Aurora BP test set (hold-out) (n = 93)

Modalities	Target	MAE (SD) mmHg
PPG, ECG, Demographic	ΔSBP	9.05 (9.02)
	SBP	9.91 (14.06)
ECG, PPG	ΔSBP	9.67 (9.52)
	SBP	10.74 (24.09)
PPG	ΔSBP	11.24 (10.53)
	SBP	10.008(13.812)
ECG	ΔSBP	10.45 (10.7)
	SBP	11.02 (15.34)

Table 1 presents a modality-wise comparison of mean absolute error (MAE) and standard deviation (SD) for both SBP and ΔSBP prediction on the Aurora BP hold-out set (n = 93). Across all configurations, ΔSBP estimation yielded lower or comparable error margins relative to static SBP, particularly when using multimodal input. The full model integrating PPG, ECG, and demographic features achieved the lowest MAE for ΔSBP, outperforming models using single or dual modalities. These results support the advantage of modelling dynamic BP changes using deep spatiotemporal features and personalized inputs. Thus, further analysis is focussed on ΔSBP results.

Category-Wise Results:

Table 2: Category-wise performance of ΔSBP prediction with RMSE, MAE, and MD (mean (std)) on Aurora BP test set (hold-out) (n = 93)

BP Group	PPG	ECG	PPG+ECG	PPG+ECG+Demo
RMSE				
Normal	13.56	16.50	13.03	12.28
Elevated	10.84	13.81	9.77	8.93
Hypertension S1	16.89	12.97	13.75	12.71
Hypertension S2	20.89	21.16	20.48	20.21
MAE ± SD (mmHg)				
Normal	9.77 ± 9.40	11.97 ± 11.35	9.52 ± 8.90	8.62 ± 8.75
Elevated	8.04 ± 7.28	9.38 ± 10.14	7.11 ± 6.71	6.65 ± 5.97
Hypertension S1	9.46 ± 9.17	11.01 ± 10.41	9.85 ± 9.59	9.30 ± 8.73

Hypertension S2	14.93 ± 15.10	24.27 ± 17.86	15.19 ± 13.74	14.80 ± 13.76
MD ± SD (mmHg)				
Normal	2.44 ± 13.34	9.43 ± 13.54	4.85 ± 12.09	3.53 ± 11.77
Elevated	-4.80 ± 9.72	6.54 ± 12.17	1.00 ± 9.72	0.10 ± 8.93
Hypertension S1	-8.39 ± 14.66	3.81 ± 12.39	-3.98 ± 13.16	-1.24 ± 12.70
Hypertension S2	-10.48 ± 18.07	3.56 ± 20.86	-1.14 ± 20.45	4.15 ± 19.86

Table 2 presents category-wise evaluation of ΔSBP prediction across four input configurations. The full multimodal model (PPG+ECG+Demographics) consistently achieved the lowest RMSE and MAE in all BP groups, with particularly strong gains in the 'Normal' and 'Elevated' categories. It also yielded the smallest mean deviation (MD) values, indicating reduced prediction bias and improved calibration across clinically relevant BP ranges. Single-sensor models exhibited higher error and variability, especially in hypertensive groups, reflecting limited robustness under physiologically unstable conditions. While demographic fusion enhanced performance in stable BP profiles, its impact diminished in more variable states (e.g., Hypertension Stage 2), likely due to waveform degradation and increased signal noise. Nonetheless, the full model demonstrated consistent bias reduction and lower dispersion (MD ± SD), supporting its role as a reliable initialization point for transfer learning in small, high-variance datasets such as SCI. These results underscore the effectiveness of multimodal integration in improving both accuracy and generalization for beat-to-beat BP estimation.

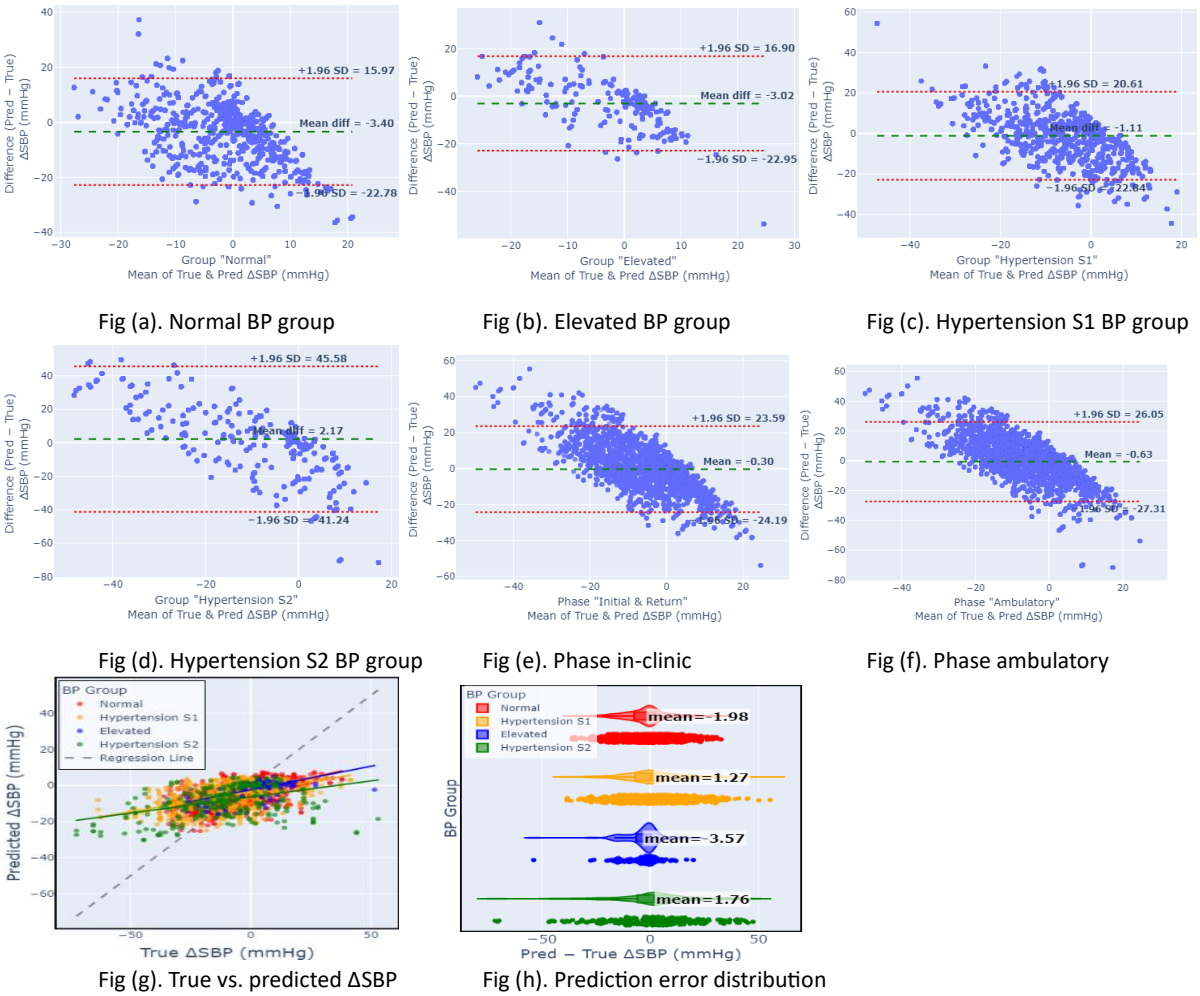

Fig (a). Normal BP group

Fig (b). Elevated BP group

Fig (c). Hypertension S1 BP group

Fig (d). Hypertension S2 BP group

Fig (e). Phase in-clinic

Fig (f). Phase ambulatory

Fig (g). True vs. predicted ΔSBP

Fig (h). Prediction error distribution

Figure 1. Bland Altman plot for various BP subgroups and phases

To complement the numerical metrics Figure 1 shows the Bland–Altman plots stratified by BP categories. Subgroup-specific analyses reveal important differences. In the Normal (1a) and Elevated (1b) BP groups, the SD are relatively narrower, and residuals appear symmetrically distributed, indicating more consistent model

behavior. Conversely, the Hypertension Stage 1 (1c) and Stage 2 (1d) groups exhibit wider dispersion and increased bias, reflecting diminished reliability in more dysregulated physiological states. Measurement conditions also affect model performance. During in-clinic assessments (1e), prediction errors are centered with reduced spread, suggesting higher stability under controlled conditions. In contrast, ambulatory phases (1f) show greater variance and mild skew, likely due to posture changes, motion artifacts, and environmental noise impacting waveform quality. Figure 1g illustrates the relationship between true and predicted ΔSBP values stratified by BP group. While overall predictions follow the diagonal trend, regression lines reveal varying slopes and offsets across subgroups, suggesting group-specific differences in prediction scaling and calibration. This is further supported by the error distributions in 1h, where hypertensive groups—particularly Stage 2—exhibit larger residual spread and a negative mean bias, indicating systematic underestimation. In contrast, the elevated group shows a mild overestimation bias, while normotensive predictions remain relatively centered. Collectively, these results confirm that while the model maintains low overall bias, its reliability varies with both clinical context and BP category—highlighting the importance of stratified evaluation.

Regulatory Evaluation of ΔSBP Prediction

Table 3: Evaluation of ΔSBP prediction on the Aurora BP test set (hold-out) under standard protocols

Standard	Metric	Value
IEEE 1708a-2019	Static ME (mmHg)	–0.78
	Static SD (mmHg)	4.44
	Dynamic ME (mmHg)	–0.38
	Dynamic SD (mmHg)	14.99
	Combined ME (mmHg)	0.80
	Combined SD (mmHg)	12.75
	Combined MAD (mmHg)	9.05
ANSI/AAMI SP10	ME (mmHg)	–0.49
	SD (mmHg)	13.02
	Meets Criteria (SD ≤ 8 mmHg)	No
BHS	≤ 5 mmHg (%)	44.6%
	≤ 10 mmHg (%)	64.7%
	≤ 15 mmHg (%)	78.9%
	Grade	D

Table 3 presents the evaluation of ΔSBP prediction under IEEE 1708a-2019, AAMI, and BHS protocols. The model performed notably better under static conditions, with substantially lower variability compared to dynamic segments, reflecting the impact of motion artifacts and physiological noise in ambulatory settings. While overall bias remained low, the standard deviation exceeded the AAMI threshold, and the BHS grading corresponded to a Grade D. Although formal regulatory criteria were not met, these outcomes underscore the difficulty of beat-to-beat BP estimation in real-world scenarios and reinforce the need for state-aware evaluation strategies.