

Efficient-Transformers Optimization for Continuous Blood Pressure Estimation Using Bio-Sensors and Wearable Sensors

Daulet Zhailau¹, Arystan Aldiyev¹, Aidana Kaziyeva¹, and Magzhan Karandashov¹ *

Abstract—Blood pressure is a key indicator for detecting and preventing cardiovascular disease, which remains one of the leading causes of mortality worldwide [1]. In this work, we focus on building a research pipeline to evaluate and compare deep learning models for cuffless BP estimation by using photoplethysmography (PPG) and electrocardiography (ECG) signals retrieved from open datasets such as MIMIC-III and MIMIC-IV. The methodology is based on a standardized and reproducible framework that includes signal preprocessing (filtering, segmentation, normalization) and inter-subject data splitting to ensure fair evaluation. The system supports PPG-only, ECG-only and multimodal inputs, and enables consistent training and benchmarking of models such as LSTM, BP-Net, and several Transformer variants, including optimized and hybrid architectures. In addition, transfer learning approaches are explored to improve model generalization and reduce dependence on large labeled datasets. Model performance is evaluated using standard metrics such as MAE, RMSE, mean error with standard deviation, and BHS grading. Overall, this project represents the design, implementation, and evaluation of a reproducible research framework for cuffless BP estimation, highlighting key challenges such as generalization, bias, and efficiency in wearable health applications.

Index Terms—Blood Pressure Estimation, Transfer Learning, Self-Supervised Learning, Transformer, Foundation Models, PPG, ECG, Wearable Sensors, MIMIC, Physiological Signal Processing, BHS Grading

I. INTRODUCTION

Early detection and prevention of cardiovascular diseases, which is the leading cause of mortality worldwide, is done by an important continuous blood pressure monitoring technique. While cuff-based measurements are the clinical standard, they can be disruptive and intermittent, which led to popularization of cuffless BP estimation using PPG and ECG.

Deep learning approaches ranging from LSTMs and CNNs to Transformers have shown promise in learning BP relevant features from raw waveforms. Although Transformers were successful in physiological signal analysis, the challenge that exists for the design of efficient Transformer models toward continuous BP estimation is linked to their immense computational load and memory requirements coupled with trade-offs between accuracy and efficiency.

In this paper, we propose a comprehensive research pipeline for cuffless BP estimation that evaluates and compares both supervised deep learning baselines and transfer learning approaches using foundation models. We reproduce several large-scale foundation architectures - REVE, UniCardio,

LUNA. In reproduction we used multiple fine-tuning strategies, including frozen backbones, partial layer unfreezing, and differential learning rates. We additionally trained Transformer baselines from scratch as a comparative reference.

The main contributions of this work are:

- Establishing a comprehensive evaluation covering supervised baselines and transfer learning approaches for cuffless BP estimation from PPG and ECG signals on MIMIC-III and MIMIC-IV clinical databases.
- A systematic comparison of cardiovascular foundation models and fine tuning strategies for cuffless BP regression.
- Demonstration that transfer learning provides a competitive alternative to training from scratch

The remainder of this report is organized as follows. Section II identifies constraints. Section III reviews background and related work. Section IV describes the project approach and system design. Section V covers project execution across both semesters. Section VI presents evaluation results. Section VII concludes with potential future enhancement discussion.

II. IDENTIFICATION OF CONSTRAINTS

A. Economic constraints

All cloud-based GPU computations were done exclusively using the free tier of resources offered by Kaggle, where users are restricted to around 30 hours per account per week. Such restrictions limited the amount of runs and parameters that could be experimented with during the course of the semester. Usage of MIMIC-III and MIMIC-IV requires credential-based registration through PhysioNet, creating an administrative burden. In the future, the commercial release of such a device for estimating blood pressure will require significant spending on medically certified equipment, adding financial constraints to the process.

B. Social constraints and political constraints

The BP measurement technology has numerous social ramifications. Models built based on patients' data like the models generated from the MIMIC dataset are unlikely to be transferable to healthy people, leading to inequity issues. Social issues that arise include data ownership and personal agency in the area of constant health monitoring. The use of the technology for estimating blood pressure in clinical practice would require regulatory approval in the form of a 510(k) clearance by the FDA in the USA and the CE mark in Europe, which were not within the purview of this study but nevertheless represent

This manuscript is currently under review.

¹The School of Engineering and Digital Sciences, Nazarbayev University, Astana, Kazakhstan (✉corresponding author: jurn.park@nu.edu.kz). [†]Work started as a MS thesis at Nazarbayev University.

significant limitations in terms of practical implementation. The regulatory process itself is another form of dependency on politics.

C. Ethical constraints

Future clinical use will require transparent means for quantifying and explaining uncertainty, allowing clinicians and patients to appropriately set the level of confidence in automated blood pressure estimation.

D. Environmental constraints

Training large neural networks uses a lot of energy. We reduced computation costs by fine-tuning transfer learning rather than pre-training from scratch. However, still the environmental issue remains here.

E. Health, Safety, and Manufacturability Constraints

The model is not suitable for making decisions in clinical practice. Mistakes in estimating blood pressure (BP) might cause misdiagnosis when applying the model in practice without proper caution. Models can be assessed based on the BHS safety standards, with those scoring Grade C and below being classified as unsafe for clinical use.

III. BACKGROUND AND RELATED WORK

A. Cuffless Blood Pressure Estimation

Continuous, non-invasive BP monitoring has been an active research area for over two decades. The initial cuffless techniques were based upon the selection of physiological markers by hand, with the primary focus being the pulse transit time (PTT), which is the duration from the peak of the ECG R-wave to the PPG wave's arrival [2]. While PTT-based methods are computationally lightweight, they assume a linear PTT-to-pressure relationship that fails to generalize across individuals and physiological states [3]. More recent work has shifted toward end-to-end deep learning approaches that learn directly from raw waveform segments, treating BP estimation as a regression problem over fixed-length signal windows.

This body of work has evolved along three distinct modality branches: PPG-only estimation, multimodal ECG+PPG, and ECG-only analysis. While multimodal methods generally outperform unimodal baselines, all three branches share a common dependence on training from random initialization with limited labeled data, leaving their generalization to unseen subjects fundamentally bounded.

B. Deep Learning Baselines

LSTM networks were among the first deep learning models applied to cuffless BP estimation [4], capturing short-range temporal dependencies in PPG signals. Their sequential processing, however, limits parallelization and scalability.

Among the first models in deep learning applied to cuffless BP estimation [4], LSTM networks were found to extract short-range temporal correlations in the PPG signal. Their sequential computation, however, restrict the parallelism and scale. Transformer networks can bridge the gap through self-attention, that relates two positions in a sequence independently of their distance.

C. Self-Supervised Learning for Physiological Signals

Mirroring the success of BERT [5] and GPT [6] on natural language processing tasks, self-supervised learning (SSL) methods on biomedical time-series have gained much attention. To deal with the diversity of signals, BioT [7] tokenizes every channel and uses segment, channel, and position embedding, combined together, to pretrain on contrastive masked reconstruction objective. GPT style autoregressive pretraining [6] trains a transformer to predict next segment given prior segments and do not requires labels for the training. The above SSL framework can learn generalizable representation on abundant but unlabeled biosignal corpora, to overcome the scarcity and noisiness of BP labels which constraint the supervised methods.

D. Foundation models for cardiovascular signals

There have been several large-scale pretrained foundation models proposed for cardiovascular and general biosignal data. They vary in their pretraining objective, architectural inductive bias, and how they handle the multimodalities of the input signals:

HeartGPT [8] is a GPT-style transformer pretrained autoregressively on large scale PPG and ECG dataset. Its causal next token prediction objective makes it inherently sensitive to temporal dynamics over sequences, which can potentially benefit the temporal dynamics between successive cardiac cycles.

UniCardio [9] is a diffusion-based multimodal foundation model pretrained jointly on ECG, PPG and ABP signals. Its reconstruction objective of predicting a masked modality using other modalities structurally suits well for the cuffless BP problem, especially when using multimodality information for signal fusion.

LUNA [10] is a topology-agnostic transformer that incorporates FFE to conduct simultaneous temporal and frequency domain analysis. Such a design is structurally suitable for cardiovascular signals as they contain information across different frequency ranges.

BENDR [11] builds upon BERT's masked signal modeling principle, but applies it to biosignals to obtain temporal representations through unsupervised contrastive pretraining on ECG and EEG corpora. The masked reconstruction objective employed by BENDR is conceptually very different from an autoregressive or a diffusion model as the local information needs to be reconstructed from the global signal structure.

REVE [12] presents a cardiovascular transformer-based foundation model to obtain learned cardiovascular representations that can then be successfully transferred to the task of BP regression via fine-tuning.

Compared to supervised methods, they provide models trained on magnitudes more data than any available labeled BP dataset, though the exact pre-training differs in that HeartGPT targetscausal temporal prediction, UniCardio Cross-modal diffusion reconstruction, Lunatime-frequency duality, and BEHDRmasked contrastive learning. This is the first effort at a systematic benchmark of these different foundation models under the same inter-subject BP regression protocol to discover

what paradigm has the best cross-over to cuffless estimation, and provides the direct motivation for this comparative framework.

E. Architectural Efficiency in Physiological Transformers

In particular, standard Transformer self-attention has a time and space complexity that is quadratic to the sequence length, which makes it a bottleneck for large biosignal window lengths. TS-Fastformer [13] aims to tackle this issue with three components: 1) A Sub-Window Tokenizer that shortens the sequence and thus allows attention calculation over shorter sequence length; 2) A Pretrained Time-series Encoder that efficiently learns trending patterns; and 3) A Past Attention Decoder to model long-range dependencies without computationally intensive decoding. Across various datasets, TS-Fastformer shows superior performance to other Transformer baselines on both accuracy and efficiency. Such architectural efficiency are important for physiological monitoring, as longer temporal windows are beneficial for clinical interpretations while simultaneously overwhelming limited resources during training and potential for deployment in wearables. This efficiency consideration leads to exploring some highly optimized Transformer variants against a foundation model.

Compared to supervised methods, the presented models obtain representation learned from many more data than any labeled BP dataset, thus addressing limitation in generality of prior methods.

IV. PROJECT APPROACH

A. System Overview

The main part of the study centers around transfer learning with repetition and evaluation of baseline cardiovascular models across various fine-tuning methods. A Transformer model (randomly initialized) was added as a control to highlight the benefit of using pretrained models. This two part approach was designed so that we first knew the limitation of supervised learning.

B. Datasets

Two publicly available clinical databases are used: MIMIC-III and MIMIC-IV [14]. These are databases with simultaneously measured PPG, ECG, and invasive arterial blood pressure (ABP) waveforms from ICU patients at 125 Hz. Ground-truth labels for SBP and DBP are derived from the ABP waveforms as maximum and minimum beat-to-beat values. The data are grouped into two testing protocols: intra-subject splits, which demonstrate personalization, and inter-subject splits, which demonstrate generalizability to unseen patients.

C. Preprocessing Pipeline

All signals are preprocessed using the same pipeline. All raw waveforms were filtered using bandpass filters (0.5–8 Hz for PPG; 0.5–40 Hz for ECG) to eliminate baseline wander and high frequency noise. For each beat-to-beat window, beats are segmented into fixed-length 625 sample windows

and then min-max scaled in $[0, 1]$. All fixed-length non-overlapping windows are chunked into tokens mapped into learnable embeddings encoding channel, positional and segment information if necessary.

D. Third-Party Tools and Integration

The following external tools and libraries were integrated into the pipeline:

- **PyTorch** primary framework for all model implementation, training, and evaluation.
- **Hugging Face Hub** loading pretrained checkpoints of foundation models.
- **Kaggle** cloud GPU environment (P100/T4 GPUs) for all training experiments.
- **WFDB / PhysioNet Toolkit** reading and processing MIMIC waveform data.
- **NumPy / SciPy** signal processing and numerical computation for preprocessing.
- **Scikit-learn** data splitting, scaling, and metric computation.
- **GitHub** collaborative code management.
- **Google Drive / Docs** collaborative documentation and data sharing.
- **Telegram** daily team coordination and updates.

We did not design any foundation model architecture from scratch; we take those released models (REVE, UniCardio, LUNA, BENDR), load their pretrained weights, and fine-tune them by changing the original output heads to the regression heads specialized for our tasks.

E. Transfer Learning Framework

Models were reproduced from random initialization; we load the weights of the backbone part from public foundation models checkpoints. These fine-tuning methods we reproduced are:

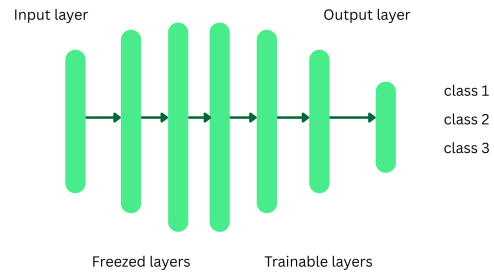


Fig. 1. Transfer learning

- **Frozen backbone** all backbone parameters frozen; only the regression head trained. Tests direct feature transferability.
- **Partial layer unfreezing** early layers frozen to preserve low-level features; later layers fine-tuned for task adaptation.

- **Differential learning rates** layerwise rates (high for head, low for backbone core) enabling full adaptation without catastrophic forgetting.
- **Full fine-tuning** all parameters jointly trained, maximizing adaptation at the risk of overwriting pretrained representations.

1) *REVE*: REVE [12] is a cardiovascular foundation model based on a Transformer. The pretrained weights were loaded and the original classification output layer replaced with a regression MLP. We use partial layer unfreezing and learning rate scheduling in training.

2) *HeartGPT*: HeartGPT [8] is a GPT-style autoregressive Transformer model that consists of 8 backbone blocks. The classification head is replaced by a three-layer MLP (Embedding $\rightarrow$ 128 $\rightarrow$ 64 $\rightarrow$ 1). We freeze the first 2 blocks and embeddings ($\sim 30\%$ of parameters); fine-tune the last 6 blocks ($\sim 70\%$). The Adam optimizer ($\text{lr} = 1 \times 10^{-4}$), weight decay 0.001, dropout 0.3, gradient clipping 1.0, were used for training, and we implemented early stopping over 50 epochs using MSE loss.

3) *UniCardio*: UniCardio [9] is a multimodal, diffusion-based foundation model pretrained on ECG, PPG, and ABP together. We compare two configurations: (1) *the backbone layers are frozen* with 3-layer MLP head, AdamW, L1 loss, 10 epochs; and (2) *differential LR* the model is fine-tuned using differential LR, where the head, projections and core layers have learning rates of (1×10^{-3} head, 1×10^{-4} projections, 5×10^{-5} core layers).

4) *LUNA*: LUNA [10] is a topology-free Transformer utilizing Fourier Feature Embeddings to jointly study in time and frequency domains. Inputs were padded from 625 to 640 samples in zero to match the native patch geometry. A two-stage training strategy consisting of initial base training, followed by a OneCycleLR fine-tuning process converges the model in 20 epochs, with SBP R^2 0.2053 and DBP R^2 0.4174.

5) *BENDR*: BENDR [11] is a 12-layer Transformer encoder that was pretrained using contrastive masked signal modeling on large unlabeled ECG and EEG corpora. We load the checkpoint from the official release, discard the original classification head, and add a task-specific regression MLP. To maintain the low-level temporal feature extractors, layers 1–4 are frozen. Layers 5–12 are then fine-tuned using AdamW at a lower learning rate (5×10^{-5}) with AdamW and a linear warmup period..

F. Regression head design

All models share the same regression head design. We discard the original classification head, and add a new task-specific MLP that maps backbone representations to continuous BP values. For models that output a regression value for every time step, we take the average of predictions over the window. For models that output a single representation, we perform global average pooling over all the output tokens, and then apply a linear projection.

G. Evaluation protocol

Models are evaluated using: MAE, RMSE, ME $\pm$ SD, BHS grading (Grade A–D by percentage of predictions within 5, 10, and 15 mmHg of ground truth).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

H. Team Workflow

The team of four people worked on things at the same time. Each person was in charge of one foundation model. They all worked together on some things like the preprocessing pipeline and the evaluation scripts. They also shared a GitHub repository. When it came to deciding how to do things the team of four made these decisions together.

V. PROJECT EXECUTION

This project was executed across two consecutive phases. Phase 1 established the supervised learning baselines; Phase 2 adopted transfer learning with pretrained foundation models. The structure of two phases was deliberate-establishing the performance cap and generalization limitations of supervised models provided direct, empirical justification for the foundation model approach

A. Phase 1: Evaluation of deep learning models

1) *Data acquisition and preprocessing*: The raw waveform data was downloaded from MIMIC-III and MIMIC-IV. A shared preprocessing pipeline-bandpass filtering, beat segmentation, and min-max normalization-was developed and utilized by all members of the team. Strict consistency in the preprocessing pipeline ensures that differences in performance are solely attributable to the tested model, not variability in data setup.

2) *Model Development*: Four standard baseline architectures were implemented and tested: LSTM-standard, LSTM-advanced, BP-Net, and a baseline Transformer. The three Transformer models evaluated by subsequent experiments (FastFormer, cBP-Tnet, and CTRhythm) were then implemented to determine how varying efficient attention mechanisms affect the accuracy-efficiency space. All models were trained from random initialization on MIMIC BP labeled data.

3) *Challenges and Solutions*: During Phase 1, two significant challenges were encountered. The LSTM-based models encountered memory overflow issues and were prohibitively slow due to the sequential nature of processing of long biosignal windows. This was resolved by implementing early stopping on the validation loss, thereby mitigating both memory depletion and unnecessary training duration. Secondly, intense training runs would consistently exhaust the Kaggle weekly GPU allocation. This issue was dealt with by waiting and careful resource management.

4) *Key Observations:* Under within-subject analysis, the Transformer yielded a BHS Grade A on DBP and Grade B on SBP. The optimized FastFormer trained in less than one hour while retaining competitive accuracy levels, indicating that efficient attention significantly improves training efficiency within resource-limited environments. Regardless, all models encountered reduced performance levels under within-subject analysis, reaffirming that generalization and long training times remained fundamental outstanding limitations of Phase 1.

B. Phase 2: Transfer Learning

1) *Motivation:* Phase 1 clearly highlighted a fundamental limitation: while models trained well on known individuals, they showed significant performance degradation on unknown subjects. Instead of augmenting or regularizing a supervised solution, a paradigm shift was undertaken by incorporating transfer learning and starting model training with representations learned from much larger and more diverse biosignal datasets. The emerging field of biosignal foundation models [7], [8] informed this decision.

2) *Model Selection:* Foundation models were selected based on a thorough review of available literature and public pretrained checkpoint accessibility. Domain proximity to cardiovascular signals (ECG, PPG, or ABP) was a primary criterion; enhanced transferability was anticipated to be achieved from domains closer to the cardiovascular domain.

3) Challenges and Solutions:

- **Input format mismatches.** LUNA expected input sequences of 640 samples while our preprocessed windows contained 625 samples; this was resolved through dynamic zero-padding. Several models required inputs projected to a specific hidden dimension before the backbone; this was resolved by inserting a linear projection layer prior to the backbone.
- **Catastrophic forgetting.** Full fine-tuning of UniCardio initially caused rapid overwriting of pretrained cardiovascular representations, resulting in worse performance than the frozen backbone baseline. This was resolved by adopting the differential learning rate strategy, assigning learning rates orders of magnitude lower to backbone core layers than to the task-specific regression head.
- **GPU memory constraints.** Fine-tuning large foundation models exceeded available Kaggle GPU memory when storing all intermediate activations during backpropagation. This was resolved through gradient checkpointing, which recomputes forward-pass activations during the backward pass rather than storing them, trading additional compute for reduced memory usage.
- **Domain mismatch.** There were models with which our database has a domain mismatch, for example, TinyMyo’s EMG-domain pretraining proved incompatible with cardiovascular BP regression.

4) *Design Decisions:* We evaluated multiple foundation models rather than optimizing a single one, because it allowed us to compare different architectural families. Testing each model in both frozen and fine-tuned states, allowed separation

of the value of the original pretraining from the value of adapting specifically to BP data. We kept the preprocessing pipeline identical across all tests, ensuring that any performance gaps were truly due to the modeling choices, not differences in the input data.

5) *Team Collaboration:* When individual team members encountered technical blockers such as input dimension mismatches or memory overflow during fine-tuning solutions were discussed. This collaborative approach reduced individual resolution time and ensured that solutions such as gradient checkpointing were documented and available to the whole team. Regular progress check-ins allowed maintaining overall momentum toward the project deadline.

VI. EVALUATION

A. Evaluation Protocol

Each of the models is tested on an inter-subject split, where different patients are placed in training and test sets — the only evaluation setting that is relevant to a deployable BP monitoring system. DWT-based preprocessing (PP3), applying wavelet denoising before bandpass filtering and segmentation, is also used in supervised baselines. Performance is measured using MAE, RMSE, $ME \pm SD$, and BHS grading, where Grade A indicates clinically acceptable accuracy and Fail means the model does not meet minimum clinical standards.

B. Supervised Baseline Results

Table I presents inter-subject results for all five supervised baselines across ECG-only, PPG-only, and PPG+ECG inputs. The same pattern appears across all models: SBP is substantially harder to estimate than DBP under inter-subject conditions. No model reaches better than Grade Fail for SBP in any modality, with MAE ranging from 14.48 mmHg (LSTM-standard, ECG-only) to 26.71 mmHg (Transformer, ECG-only). DBP results are more encouraging — several configurations reach BHS Grade A, including U-Net (PPG+ECG) at 4.47 mmHg and Transformer (ECG-only) at 5.26 mmHg.

Among the baselines, LSTM-standard achieves the lowest SBP MAE of 14.48 mmHg using ECG-only input, suggesting that sequential models pick up some subject-invariant temporal structure from ECG waveforms. Still, the failure of all models on inter-subject SBP points to the core problem this work addresses: training from scratch on limited labeled data is not enough to generalize SBP-relevant morphological features across subjects.

C. Transfer Learning Results

Table II shows results for the evaluated foundation models. REVE achieves the best overall performance, with SBP MAE of 9.12 mmHg and DBP MAE of 4.71 mmHg — a 37% reduction in SBP error compared to the best supervised baseline. UniCardio with differential learning rates reaches DBP BHS Grade A (MAE 4.67 mmHg), though its SBP performance is limited at 12.70 mmHg (Grade Fail), with a systematic negative bias of -8.99 mmHg pointing to consistent underestimation. HeartGPT gives the weakest results among the

TABLE I
SUPERVISED BASELINE RESULTS — INTER-SUBJECT SPLIT, DWT PREPROCESSING (PP3)

Model	Modality	SBP MAE	SBP RMSE	SBP ME $\pm$ SD	SBP BHS	DBP MAE	DBP RMSE	DBP ME $\pm$ SD	DBP BHS
U-Net	ECG-only	20.78	24.31	11.12 $\pm$ 21.58	Fail	5.12	7.21	1.58 $\pm$ 7.87	A
U-Net	PPG-only	22.76	26.81	15.69 $\pm$ 21.73	Fail	4.98	6.38	1.43 $\pm$ 7.36	B
U-Net	PPG+ECG	20.34	23.21	13.44 $\pm$ 18.92	Fail	4.47	6.04	0.45 $\pm$ 7.27	A
BP-Net	ECG-only	17.09	21.34	2.36 $\pm$ 21.21	Fail	5.67	7.53	-1.13 $\pm$ 15.23	Fail
BP-Net	PPG-only	17.39	20.93	1.55 $\pm$ 20.87	Fail	4.56	5.91	-0.37 $\pm$ 18.93	Fail
BP-Net	PPG+ECG	19.57	22.73	-0.01 $\pm$ 18.63	Fail	6.71	9.11	4.47 $\pm$ 19.38	Fail
LSTM-stan.	ECG-only	14.48	18.90	7.63 $\pm$ 17.29	Fail	8.01	9.24	12.44 $\pm$ 10.71	Fail
LSTM-stan.	PPG-only	18.18	21.53	11.03 $\pm$ 17.61	Fail	8.36	9.46	2.49 $\pm$ 13.84	Fail
LSTM-stan.	PPG+ECG	18.96	20.97	9.89 $\pm$ 19.26	Fail	8.87	10.10	2.60 $\pm$ 8.01	Fail
LSTM-adv.	ECG-only	18.33	20.97	11.32 $\pm$ 17.65	Fail	7.22	8.35	3.78 $\pm$ 12.14	Fail
LSTM-adv.	PPG-only	25.42	27.96	20.83 $\pm$ 18.66	Fail	5.18	8.32	-0.96 $\pm$ 18.26	B
LSTM-adv.	PPG+ECG	16.75	20.40	4.09 $\pm$ 19.99	Fail	6.63	8.96	3.22 $\pm$ 8.36	Fail
Transformer	ECG-only	26.71	29.66	20.35 $\pm$ 21.58	Fail	5.26	8.26	-0.12 $\pm$ 8.26	A
Transformer	PPG-only	16.57	20.52	4.09 $\pm$ 19.99	Fail	6.63	8.96	3.22 $\pm$ 8.36	Fail
Transformer	PPG+ECG	18.98	21.61	20.35 $\pm$ 21.58	Fail	12.41	14.52	6.70 $\pm$ 10.79	A

TABLE II
TRANSFER LEARNING RESULTS — INTER-SUBJECT SPLIT

Model	Strategy	SBP MAE	SBP RMSE	SBP R ²	SBP BHS	DBP MAE	DBP BHS
REVE	Partial unfreeze	9.12	12.34	0.223	—	4.71	—
UniCardio	Differential LR	12.70	17.12	—	Fail	4.67	A
HeartGPT	Partial unfreeze (70%)	15.84	19.95	—	D	7.61	D
BENDR	Partial unfreeze (5–12)	20.01	—	-1.743	Fail	7.96	Fail
LUNA*	Two-stage OneCycleLR	0.578 [†]	—	0.205	—	0.220 [†]	—

* Evaluated on intra-subject split; results not directly comparable.

[†] Normalized units (StandardScaler), not mmHg.

comparable models, with SBP MAE of 15.84 mmHg and DBP MAE of 7.61 mmHg, both at BHS Grade D.

BENDR, pretrained via contrastive masked learning on EEG and ECG data, achieves SBP MAE of 20.01 mmHg and DBP MAE of 7.96 mmHg. Both SBP R² (-1.74) and DBP R² (-1.00) are negative, meaning the model performs worse than simply predicting the mean. This suggests that BENDR’s representations, shaped by a different pretraining objective and a different signal domain, do not transfer to BP regression without considerably more task-specific adaptation.

LUNA was evaluated under a different setup — a random intra-subject split rather than inter-subject — and its outputs are in StandardScaler-normalized units rather than mmHg, since SBP and DBP were extracted as the max and min of normalized ABP waveform predictions. The normalized SBP MAE of 0.578 and DBP MAE of 0.220, with R² of 0.205 and 0.417 respectively, indicate moderate predictive capacity, particularly for DBP. Due to the protocol and unit mismatch, LUNA results cannot be directly compared to the other models and are excluded from the cross-model ranking.

D. Comparative Analysis and Discussion

Transfer learning consistently improves SBP estimation over training from scratch. The 37% SBP reduction of REVE is consistent with the hypothesis that cardiovascular foundation models learn subject-invariant representations that supervised models trained on limited labeled data simply cannot acquire.

DBP follows a different pattern. Several supervised baselines already reach Grade A, so the benefit of transfer learning is smaller. This likely reflects the physiology of DBP —

diastolic minima are less sensitive to inter-subject vascular differences and are more learnable from waveform morphology alone.

The ordering among foundation models — REVE > UniCardio > HeartGPT > BENDR — tracks closely with how close each model’s pretraining domain is to the BP regression task. REVE and UniCardio were pretrained on cardiovascular signals including ABP, directly aligning their representations with the target. HeartGPT’s causal autoregressive objective, while trained on PPG and ECG, is oriented toward sequential prediction rather than regression over a full cardiac cycle. BENDR’s contrastive masked pretraining on EEG and ECG, combined with the domain gap between neural signals and arterial pressure, likely explains why it fails to generalize. Overall, domain proximity of pretraining data appears to be a critical factor in whether transfer learning works for cuffless BP estimation.

Even with these improvements, no model in either phase reaches clinically acceptable SBP estimation under inter-subject evaluation. REVE’s SBP MAE of 9.12 mmHg is still above the BHS Grade A threshold of 5 mmHg, and the wide limits of agreement (LoA [-26.59, 19.05] mmHg) show substantial variance at the individual level. Cuffless BP estimation from ICU waveforms remains an open problem, especially for SBP generalization across subjects.

VII. CONCLUSION AND FUTURE WORK

In this work, we built a reproducible research pipeline for cuffless blood pressure estimation from PPG and ECG signals, covering both supervised deep learning baselines and transfer learning with cardiovascular foundation models, evaluated on MIMIC-III and MIMIC-IV.

The supervised baseline phase showed a clear performance ceiling under inter-subject evaluation. All five models — U-Net, BP-Net, LSTM-standard, LSTM-advanced, and Transformer — failed to achieve acceptable SBP estimation, with MAE ranging from 14.48 to 26.71 mmHg. DBP was more tractable, with several configurations reaching BHS Grade A.

This gap between SBP and DBP performance, observed consistently across all architectures and modalities, confirms that SBP generalization to unseen subjects is the main unresolved challenge in supervised cuffless BP estimation.

The transfer learning phase showed that starting from pre-trained cardiovascular foundation models improves over training from scratch. REVE achieved the best results overall (SBP MAE 9.12 mmHg, DBP MAE 4.71 mmHg), cutting SBP error by 37% relative to the best supervised baseline. UniCardio with differential learning rates reached BHS Grade A for DBP. The ordering across foundation models — REVE > UniCardio > HeartGPT > BENDR — correlated strongly with how close each model’s pretraining domain is to arterial BP signals, which has practical implications for model selection in future biosignal transfer learning work.

Despite these gains, no model reached clinically acceptable SBP estimation under inter-subject evaluation, and REVE’s wide limits of agreement (LoA [−26.59, 19.05] mmHg) reflect persistent individual-level variance. Cuffless BP estimation remains an open problem, and foundation model fine-tuning alone is not yet enough for clinical deployment.

Several directions are worth exploring going forward. Subject-adaptive calibration — using a small number of labeled samples per subject to personalize a pretrained backbone — could reduce inter-subject variance without full retraining. Pretraining on larger and more varied corpora, including ambulatory and wearable-derived PPG beyond ICU data, may also improve generalization to real-world deployment conditions. Adding uncertainty quantification to the regression head would provide confidence intervals alongside BP estimates, which is a basic requirement for clinical use. Finally, model compression and quantization are needed before any of these approaches can run on wearable hardware, since current foundation model backbones are too large for real-time edge inference.

REFERENCES

- [1] World Health Organization, “Cardiovascular diseases (CVDs),” *WHO Fact Sheets*, 2023.
- [2] R. Makkamala, J.-O. Hahn, O. T. Inan, L. K. Mestha, C.-S. Sung, C.-S. Kim, and J. Sola, “Toward ubiquitous blood pressure monitoring via pulse transit time: Theory and practice,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 8, pp. 1879–1901, 2015.
- [3] R. Makkamala and J. O. Hahn, “Pulse transit time-based blood pressure estimation: Is it a sustainable technique?,” *Journal of Hypertension*, vol. 33, p. e1, 2015.
- [4] P. Su, X.-R. Ding, Y.-T. Zhang, J. Liu, F. Miao, and N. Zhao, “Long-term blood pressure prediction with deep recurrent neural networks,” pp. 323–328, 2018.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [6] S. A. T. P. Y. Z. W. Y. Y. L. Defu Cao, Furong Jia, “Tempo: Prompt-based generative pre-trained transformer for time series forecasting,” 2023.
- [7] C. Yang, M. B. Westover, and J. Sun, “Biot: Biosignal transformer for cross-data learning in the wild,” *NeurIPS*, 2023.
- [8] H. J. Davies, J. Monsen, and D. P. Mandic, “Interpretable pre-trained transformers for heart time-series data,” *arXiv preprint arXiv:2407.20775*, 2024.
- [9] Z. Chen, Y. Miao, L. Wang, L. Fan, D. P. Mandic, and J. Zhu, “Versatile cardiovascular signal generation with a unified diffusion transformer,” 2025. *arXiv:2505.22306v2 [cs.LG]*.
- [10] B. Döner, T. M. Ingolfsson, L. Benini, and Y. Li, “LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis,” *arXiv preprint arXiv:2510.22257*, 2025.
- [11] D. Kostas, S. Aroca-Ouellette, and A. J. Casson, “Bendr: using transformers and a masked autoencoder to learn from raw eeg-based brain-computer interface data,” *Frontiers in Human Neuroscience*, vol. 15, p. 67, 2021.
- [12] Y. El Ouahidi, G. Cisotto, R. Nahon, J.-E. Le Douget, and G. Neri, “REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects,” *arXiv preprint arXiv:2510.21585*, 2025.
- [13] W. Zuo and et al., “Ts-fastformer: A fast transformer for time-series analysis,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [14] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “Mimic-iii, a freely accessible critical care database,” *Scientific Data*, vol. 3, p. 160035, 2016.