Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 April 2026
Revised 20 May 2026
Accepted 24 June 2026

Natural Resources for Human Health 2026, Vol. 6 (3s): 46–63

KEYWORDS:
photoplethysmography, cardiovascular disease classification, Transformer neural network, hypertension staging, cardiac-phase encoding, subject-independent validation

Transformer-Based Deep Learning for Cardiovascular Disease Classification Using Raw Photoplethysmography Signals

Anoop G L[1], Pooja Dodamani[2], Kavita V Horadi[3], H. Nagesh Shenoy[4], Pradeep Kumar KG[5], Shweta Mannikeri[6,7]

[1]Department of Computer Science and Engineering, Christ University, Karnataka, Bengaluru, India. Email: gl.anoop1@gmail.com

[2]Assistant Professor, School of Engineering; Department of Computer Science and Engineering, Dayananda Sagar University, Bengaluru, Karnataka, India. Email: pooja.s.dodamani-cse@dsu.edu.in

[3]Associate Professor, Dept of CSE, BNMIT, Bangaluru, KA, India. Email: kavita.bnmit3@gmail.com

[4]Dept. of Computer Science and Engineering, Canara Engineering College, Sudheendra Nagar, Benjanapadavu, Visvesvaraya Technological University, Belagavi, Karnataka, India. Email: h.nagesh.shenoy@gmail.com

[5]Department of Computer Science & Engineering, Vivekananda College of Engg. & Technology, Nehru Nagar, Puttur, India. Email: putturpradeep@gmail.com

[6]Research Scholar, Department of Computer Science, Chairashree Institute of Research and Development (CIRD), University of Mysore, Karnataka, India

[7]Assistant Professor, Department of Computer Science, Government College for Women, Chintamani, Karnataka, India. Email: shwetaphd25@gmail.com. ORCID ID:0009-0003-6074-8416

Corresponding Author Email: pooja.s.dodamani-cse@dsu.edu.in

ABSTRACT: Cardiovascular disease remains the foremost cause of death worldwide, and elevated blood pressure is its largest modifiable contributor, yet many hypertensive adults remain undiagnosed because sphygmomanometry requires clinical contact. Photoplethysmography offers an inexpensive optical alternative already embedded in wearables, but published classifiers depend on handcrafted descriptors or time-frequency images, and most are validated with splitting schemes that allow recordings from one participant to appear in both training and test partitions. This work introduces CPA-Former, a cardiac-phase-aware Transformer consuming the raw pulse together with its first and second derivatives, combining a multi-derivative tokenizer, a positional encoding defined on cardiac phase rather than absolute time, cross-derivative co-attention, and segment-level attention pooling. It is evaluated on the public PPG-BP database of 219 participants under strict subject-independent cross-validation, alongside four deep baselines and classical references trained identically. The proposed model attains 45.60% accuracy on four-stage staging and 71.66% on binary screening. The central finding is negative and arguably more useful than a headline number: architectures spanning two orders of magnitude in parameter count prove statistically indistinguishable on this cohort, and none convincingly exceeds handcrafted morphology features. Repeating the identical experiment

with folds formed over recordings rather than participants inflates accuracy substantially, offering a parsimonious explanation for the far higher figures reported elsewhere on this dataset.

1. INTRODUCTION

Cardiovascular disease accounts for roughly one third of all deaths recorded each year and continues to dominate global mortality statistics despite decades of preventive cardiology [1]. Within this burden, raised blood pressure is the most prevalent and most modifiable risk factor, implicated in ischaemic heart disease, stroke, heart failure and chronic kidney disease. Pooled analyses covering more than one hundred million adults show that although the prevalence of hypertension has stabilised in many high-income settings, the proportion of affected individuals who are aware of their condition remains far below what effective prevention requires, and awareness is lowest precisely in the populations where the disease burden is rising fastest [2]. The bottleneck is not therapeutic but diagnostic. Cuff sphygmomanometry, the reference method, demands a trained operator, a calibrated device and a clinical encounter, none of which scale to population-level screening.

Photoplethysmography provides a compelling alternative. A photoplethysmogram is obtained by illuminating peripheral tissue and measuring the light that returns after interacting with the pulsatile arterial bed, so the resulting waveform is a direct optical record of volumetric change driven by each cardiac ejection. The sensing hardware costs very little, requires no consumables and is already integrated into smartwatches, fitness bands, pulse oximeters and smartphone cameras, which makes the modality uniquely suited to opportunistic and continuous cardiovascular assessment [3]. The physiological justification for using the waveform as a cardiovascular marker is well established. The shape of the pulse encodes the interaction between left ventricular ejection, arterial compliance and wave reflection from the periphery. As arteries stiffen with age and sustained pressure loading, the reflected wave returns earlier, the dicrotic notch flattens, and the characteristic a to e inflections of the second derivative, known as the acceleration plethysmogram, change in a manner that correlates with vascular age and arterial stiffness [4].

Translating that physiological signal into a reliable classifier has nevertheless proved difficult. Regulatory and clinical reviews of cuffless devices have repeatedly cautioned that agreement with reference sphygmomanometry degrades sharply when a device is applied to a participant whose calibration data the model has never seen, and that many published validations do not test this condition at all [5]. A consensus statement from the European Society of Hypertension working group makes the same point and recommends that any cuffless method be assessed under protocols that reflect genuine deployment rather than favourable within-subject conditions [6]. These cautions apply with equal force to classification. If the target is to decide whether an unseen individual should be referred for formal blood pressure measurement, then the evaluation must place that individual entirely outside the training distribution.

The public PPG-BP database released by Liang et al. [7] is the principal open resource that pairs finger photoplethysmography with per-participant clinical labels, providing 219 participants, three short recordings each, and a hypertension stage assigned from the measured systolic and diastolic pressures [7]. Its wide adoption has produced a substantial body of comparative results, but also a recurring methodological weakness. Because each participant contributes three separate recordings, a stratified split performed over recordings rather than participants places segments of the same individual on both sides of the partition. A model can then exploit participant-specific idiosyncrasies carrying no diagnostic information about a new patient. Reported accuracies consequently span an implausibly wide range that is hard to reconcile with the few severely hypertensive participants the cohort contains.

A second limitation concerns representation. The dominant approach converts the pulse into a scalogram or spectrogram and fine-tunes an image classifier pretrained on natural photographs [8], [9]. This transfer is convenient but discards phase, fixes the time-frequency resolution before any learning occurs, and imports an inductive bias designed for object recognition rather than for quasi-periodic biomedical waveforms. Feature-engineering studies continue to report competitive results against such pipelines, which suggests that the raw-signal route remains under-explored rather than exhausted [10].

This paper addresses both issues, and reports an outcome that was not the one intended. The contributions are as follows. First, a Transformer architecture is proposed that operates directly on the raw pulse together with its velocity and acceleration

derivatives, with no handcrafted descriptors and no time-frequency transform, using cardiac phase rather than sample index as its positional signal and coupling pulse morphology to vascular curvature through cross-derivative co-attention. Second, this model and four convolutional, recurrent and Transformer baselines are trained and evaluated under an identical, strictly subject-independent protocol, alongside majority-class, demographic and handcrafted-feature references. Third, the resulting comparison shows that none of the deep architectures separates from the others, and that the whole family sits close to classical morphology features, which we take to indicate that the binding constraint on this benchmark is cohort size rather than model capacity. Fourth, repeating the identical experiment under the record-wise protocol used by earlier work quantifies directly how much of the published performance on this dataset is attributable to the splitting scheme.

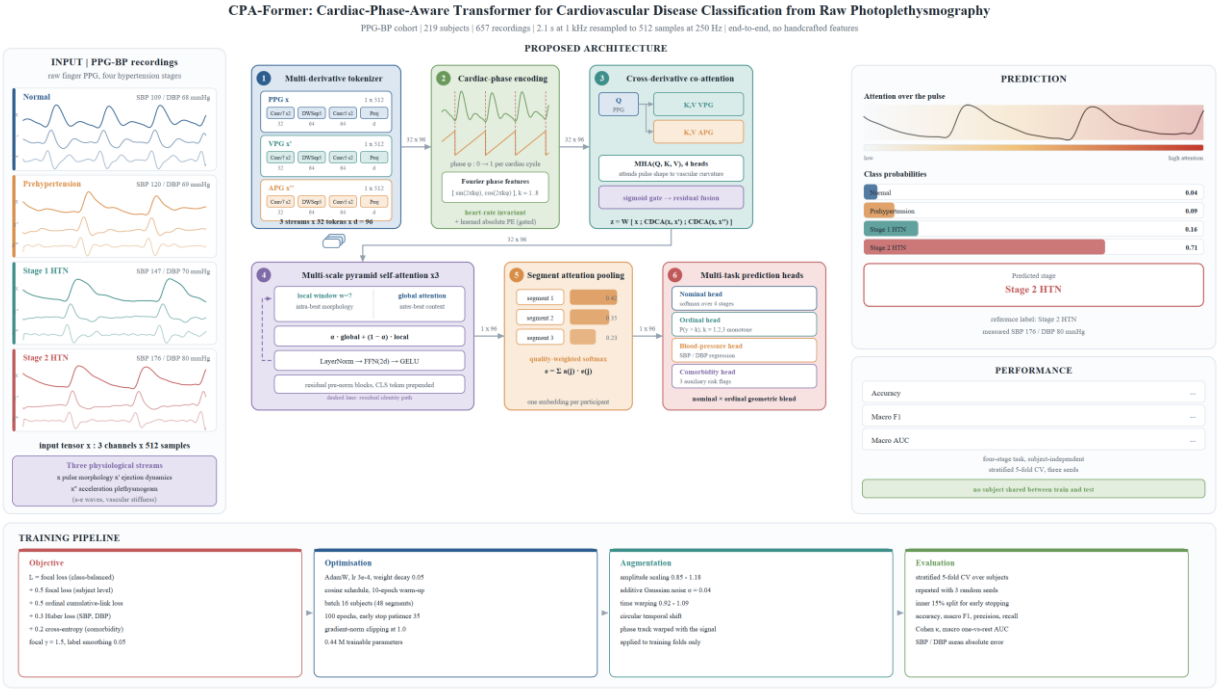

Fig. 1 Overall architecture of the proposed CPA-Former

2. LITERATURE REVIEW

Early machine-learning studies of photoplethysmographic hypertension detection relied on explicit morphological descriptors. Khan et al. [11] extracted more than one hundred features from pulse plethysmograph recordings, applied kernel principal component analysis and a weighted nearest-neighbour classifier, and reported very high accuracy on a private cohort of 121 participants. The result illustrates both the discriminative content of pulse morphology and the difficulty of interpreting scores obtained on unreleased data.

The transition to deep learning was dominated by time-frequency imaging. Wu et al. [9] applied the continuous wavelet transform to segments drawn from the MIMIC-III waveform database and classified the resulting scalograms with custom and pretrained convolutional networks, reporting approximately 90% accuracy over four blood-pressure categories. Yen et al. [8] followed a related strategy on PPG-BP, combining residual and Xception convolutional backbones with bidirectional long short-term memory layers and reporting 74-76% accuracy on the four-stage task. Notably, the recall reported in that study falls between 39% and 49%, which indicates that overall accuracy was sustained largely by the majority classes while the clinically important severe categories were frequently missed. This dissociation recurs throughout the literature and motivates the reporting conventions adopted here.

Fuadah and Lim concatenated convolutional branches over photoplethysmographic and electrocardiographic inputs, reaching roughly 95% across five categories on MIMIC-III but relying on an electrocardiogram a standalone optical sensor cannot provide [12]. Pankaj et al. [13] built Fourier-decomposition spectrograms and fine-tuned DenseNet and GoogLeNet variants,

reporting approximately 96% accuracy; their strongest figures, however, are obtained on MIMIC-III rather than on the 219-participant PPG-BP cohort, a distinction that is frequently blurred when their result is quoted as a PPG-BP baseline.

Attention-based models have entered the field more recently. Kudo et al. [14] combined convolutional and Transformer components for atrial fibrillation detection, although their sequence model operates on peak-interval series rather than on the sampled waveform, so the pulse shape itself is not modelled. Sun et al. [15] fused a convolutional recurrent branch with a frequency-domain Transformer and reported very strong arrhythmia classification on the PhysioNet Computing in Cardiology 2015 collection, confirming that attention transfers well to photoplethysmographic rhythm analysis [15]. At much larger scale, Chen et al. [16] pretrained a generative Transformer on more than two hundred million pulse segments and fine-tuned it for atrial fibrillation screening, establishing that raw-waveform Transformers are viable when data are abundant. Ding et al. [17] took a cross-modal route, learning a mapping between photoplethysmographic and electrocardiographic representations and using synthesised electrocardiograms for cardiovascular prediction.

Work specific to PPG-BP staging remains mixed. Al Fahoum et al. [18] reported perfect four-stage accuracy using scalograms and a bespoke network [18]. A cohort of 219 participants containing only twenty individuals in the most severe category cannot plausibly support an error-free subject-independent classifier, and the most economical explanation is that recordings from the same participant were distributed across training and test partitions. The same group subsequently published a hierarchical convolutional model operating on raw pulses across a pooled multi-source cohort and reported a more credible 93% accuracy over six cardiovascular categories, which supports the view that raw-waveform modelling is competitive when the evaluation is constructed carefully [19]. Running against the deep-learning trend, Msokar et al. [10] showed that interpretable engineered features with gradient boosting achieve a macro F1 of approximately 0.78 on multi-class staging and argued that naive deep pipelines do not automatically surpass them.

Two further strands inform the present design. Signal-quality assessment is essential because optical recordings are easily corrupted by motion and poor perfusion, and standardised quality indices allow unreliable segments to be identified before they influence a decision [20]. Interpretability is equally important for clinical acceptance; attention-based sleep staging has demonstrated that attention distributions over a raw physiological signal can be presented as evidence for a prediction rather than as an opaque internal quantity [21]. On the methodological side, the Transformer encoder [22] operates effectively on raw multivariate time series without convolutional pretraining, while focal weighting [23] and effective-number class balancing [24] provide the mechanisms needed to train on a cohort in which the most severe category is represented by a few dozen participants.

Taken together, the literature indicates that the principal opportunities lie in removing the dependence on handcrafted or image-based intermediates, in exploiting the ordered structure of the staging label, and in adopting an evaluation protocol that reflects deployment on unseen individuals.

3. PROPOSED METHODOLOGY

3.1 Overview

The proposed CPA-Former processes a raw photoplethysmographic segment end to end. Preprocessing removes baseline drift and out-of-band noise but performs no feature extraction. A multi-derivative tokenizer converts the pulse and its two derivatives into token sequences, a cardiac-phase positional encoding attaches physiologically meaningful positions to those tokens, cross-derivative co-attention fuses the streams, a mixed local-global attention stack builds the representation, and attention pooling combines the several recordings available for each participant into a single embedding from which the prediction heads operate. Figure 1 presents the complete architecture together with representative dataset inputs and measured outputs.

3.2 Signal conditioning

Let $r[n]$ denote a raw recording of N_r samples acquired at $f_r = 1$ kHz. Because each record spans only 2.1 s, a high-pass filter with a cut-off near the respiratory band would have a transient comparable in length to the record itself and would distort the pulse. Figure 2 depicts the sample signals for signal conditioning stages. Baseline wander is therefore removed by subtracting a least-squares cubic polynomial fitted over normalised time $\tau \in [-1,1]$

Natural Resources for Human Health
eISSN: 2583-1194

$$x_d[n] = r[n] - \sum_{q=0}^{3} \beta_q \, \tau[n]^q \tag{1}$$

where the coefficients β_q minimise the residual sum of squares. Broadband noise is suppressed with a zero-phase fourth-order Butterworth low-pass filter h_{lp} with a 10 Hz cut-off, which retains the a to e inflections of the acceleration plethysmogram, and the result is decimated to $f_s = 250$ Hz and standardised

$$x[n] = \frac{(h_{lp} * x_d)_{\downarrow 4}[n] - \mu}{\sigma} \tag{2}$$

with μ and σ the mean and standard deviation of the decimated segment. Probe polarity is inconsistent across the cohort, so a record is inverted whenever $\max_n \Delta x[n] < |\min_n \Delta x[n]|$, a test that encodes the physiological asymmetry between the rapid systolic upstroke and the slower diastolic decay. The velocity and acceleration plethysmograms are obtained by successive central differences and are standardised independently, giving the three-channel input $\boldsymbol{x} = [x, x', x''] \in \mathbb{R}^{3\times L}$ with $L = 512$

$$x'[n] = \frac{f_s}{2}(x[n+1] - x[n-1]), \qquad x''[n] = \frac{f_s}{2}(x'[n+1] - x'[n-1]) \tag{3}$$

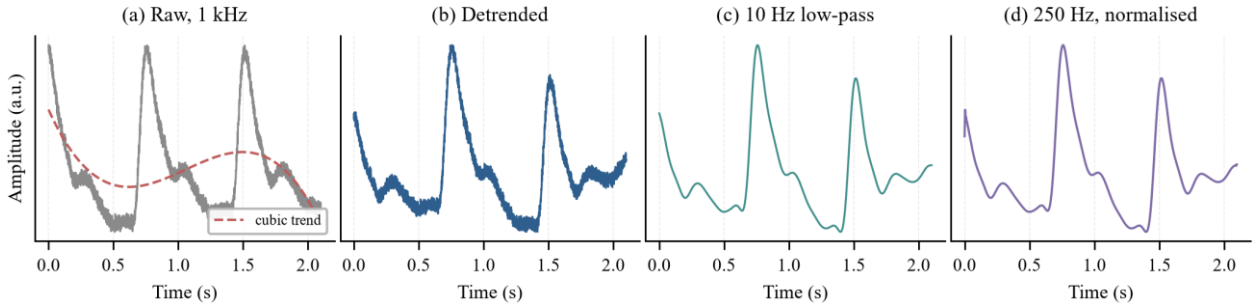

Fig. 2 Signal-conditioning stages on one recording

3.3 Multi-derivative tokenizer

Each stream is embedded independently so that the network can learn filters matched to the very different bandwidth of a pulse and of its second derivative. For stream $s \in \{0,1,2\}$ a convolutional stem applies a strided wide-kernel convolution, a depthwise-separable block and a second strided convolution, followed by a projection that reduces the temporal axis to $M = 32$ tokens of width $d = 96$

$$T^{(s)} = \Psi_s(x_s) \in \mathbb{R}^{M\times d} \tag{4}$$

where Ψ_s denotes the composition of convolution, batch normalisation and Gaussian error linear unit activations. Each token therefore summarises a receptive field of 64 ms, which is shorter than the systolic upstroke and allows intra-pulse structure to be resolved.

3.4 Cardiac-phase positional encoding

Absolute positional encodings tie a token to its sample index, so an identical pulse recorded at a different heart rate produces a different positional signature. The proposed encoding replaces the index with cardiac phase. Systolic onsets $\{o_1, \ldots, o_B\}$ are located as the minima preceding detected systolic peaks, and phase rises linearly between consecutive onsets

$$\varphi[n] = \frac{n - o_b}{o_{b+1} - o_b}, \qquad o_b \le n < o_{b+1} \tag{5}$$

so that $\varphi \in [0,1)$ marks position within the heartbeat irrespective of its duration. The phase of a token is the circular mean of the phases of the samples it summarises, which avoids the wrap-around error that a linear average would introduce

$$\bar{\varphi}_m = \frac{1}{2\pi} \arg\left(\sum_{n\in\mathcal{R}_m} e^{j2\pi\varphi[n]}\right) \bmod 1 \tag{6}$$

with $\mathcal{R}_m$ the receptive field of token m. The phase is expanded into the $K = 8$ harmonic features $u_m = [\sin(2\pi k\bar{\varphi}_m), \cos(2\pi k\bar{\varphi}_m)]_{k=1}^{K}$ and projected into the model dimension, and a learned absolute encoding E_{abs} is retained as a gated residual term so the network can recover ordering information when phase estimation is unreliable

$$\tilde{T}_m^{(s)} = T_m^{(s)} + \gamma W_\varphi u_m + E_{abs,m}, \qquad \gamma = \varsigma(g) \tag{7}$$

where ς is the logistic function and g a scalar parameter. Figure 3 depicts the cardiac phase references derived from systolic onsets.

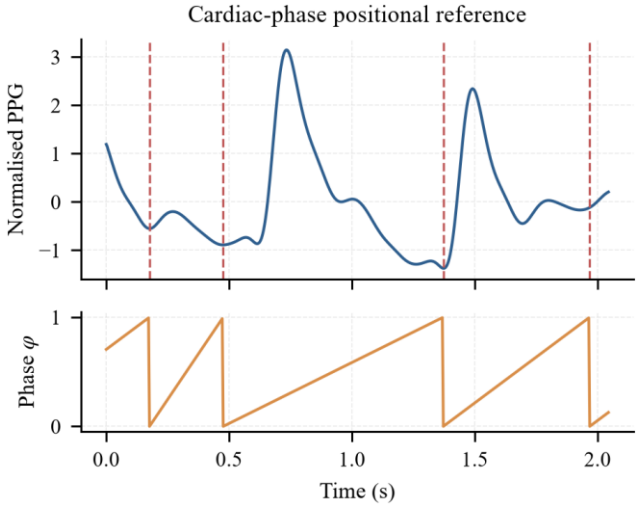

Fig. 3 Cardiac-phase reference derived from systolic onsets

3.5 Cross-derivative co-attention

Pulse morphology and vascular curvature carry complementary evidence, so the pulse stream queries the derivative streams rather than being concatenated with them. For $s \in \{1,2\}$

$$A^{(s)} = \mathrm{MHA}\big(\mathrm{LN}(\tilde{T}^{(0)}), \mathrm{LN}(\tilde{T}^{(s)}), \mathrm{LN}(\tilde{T}^{(s)})\big) \tag{8}$$

where MHA denotes multi-head attention with four heads and LN layer normalisation. A learned gate decides how much derivative evidence to admit at each token, which prevents noisy acceleration estimates from overwhelming a clean pulse

$$G^{(s)} = \varsigma\big(W_g[\tilde{T}^{(0)} \parallel A^{(s)}]\big), \qquad H^{(s)} = \tilde{T}^{(0)} + G^{(s)} \odot A^{(s)} \tag{9}$$

and the three views are fused by a linear projection

$$Z = W_f\big[\tilde{T}^{(0)} \parallel H^{(1)} \parallel H^{(2)}\big] \in \mathbb{R}^{M\times d} \tag{10}$$

3.6 Multi-scale pyramid attention

A learnable class token is prepended to $\boldsymbol{Z}$ and the sequence passes through three encoder blocks. Each block computes two attention views. A global view attends without restriction and captures beat-to-beat context, while a local view is masked to a window of $w = 7$ tokens so that attention concentrates on intra-beat morphology; the class token is exempt from the mask so that it can always read the whole sequence

$$\mathcal{M}_{ij} = \begin{cases} -\infty, & |i-j| > \lfloor w/2 \rfloor \text{ and } i,j > 0 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

The two views are combined by a learned convex mixture and followed by a position-wise feed-forward network, both applied residually in pre-normalisation form

$$Z \leftarrow Z + \alpha\,\mathrm{MHA}_g(\mathrm{LN}(Z)) + (1-\alpha)\,\mathrm{MHA}_l(\mathrm{LN}(Z), \mathcal{M}) \tag{12}$$

$$Z \leftarrow Z + W_2\,\mathrm{GELU}(W_1\mathrm{LN}(Z)) \tag{13}$$

with $\alpha = \varsigma(a)$ learned per block. The final class-token state is the segment embedding $e_j \in \mathbb{R}^d$.

3.7 Segment attention pooling and prediction head

Each participant contributes three recordings of differing quality. Rather than averaging their predictions, the model learns to weight them

$$a_j = \frac{\exp(w_a^\top \tanh(W_p e_j))}{\sum_{j'} \exp\left(w_a^\top \tanh(W_p e_{j'})\right)}, \qquad e = \sum_j a_j\, e_j \tag{14}$$

A softmax head over the four stages produces the class posterior from the pooled embedding. Because the stage label is a monotone discretisation of systolic pressure, an auxiliary head additionally regresses systolic and diastolic pressure from the same embedding. This auxiliary task shares the encoder and acts as a physiological regulariser that supplies a dense, ordered training signal in place of the four-way categorical target alone.

3.8 Objective

Class frequencies are severely unbalanced, so the primary term is a focal loss weighted by the effective number of samples with $\beta = 0.999$ and focusing parameter $\eta = 1.5$, applied to label-smoothed targets $\tilde{y}$

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_i \sum_c \omega_c\, \tilde{y}_{ic}(1 - p_{ic})^{\eta} \log p_{ic}, \qquad \omega_c \propto \frac{1-\beta}{1-\beta^{n_c}} \tag{15}$$

The complete objective adds the segment-level and participant-level classification terms and a Huber loss on standardised blood pressure

$$\mathcal{L} = \mathcal{L}_{cls}^{seg} + \lambda_1 \mathcal{L}_{cls}^{subj} + \lambda_2 \mathcal{L}_{bp} \tag{16}$$

with $\lambda_1 = 0.5$ and $\lambda_2 = 1.0$.

4. RESULTS AND DISCUSSION

4.1 Dataset details

All experiments use the publicly available PPG-BP database released by Liang et al. [7] under a public-domain dedication. The cohort comprises 219 participants recruited at the Guilin People's Hospital, each contributing three finger photoplethysmographic recordings, giving 657 records in total. Every record is 2.1 s long and sampled at 1 kHz, yielding 2100 samples. Participants are aged 21 to 86 years with a mean of 57.2 years, and the cohort is 115 female and 104 male. Systolic pressure spans 80 to 182 mmHg with a mean of 127.9 mmHg, and diastolic pressure spans 42 to 107 mmHg with a mean of 71.8 mmHg.

The primary label is the hypertension stage. Verification against the recorded pressures confirms that the four stages partition the systolic axis exactly, with the classes occupying the disjoint ranges 80 to 119, 120 to 139, 140 to 159 and 160 to 182 mmHg, so the target is an ordered discretisation of systolic pressure. This observation directly motivates the auxiliary blood-pressure regression head described in Section 3.7. Table 1 summarises the class composition, which is markedly unbalanced: the most severe stage is represented by only twenty participants.

The database also records three comorbidity flags, present for 38, 20 and 25 participants respectively, with 83 participants carrying at least one. These are used only as auxiliary targets. Signal quality was characterised with the skewness index; after conditioning, 93.6% of records have a positive index, and 4.7 % required polarity correction. The estimated pulse rate averages 78.9 beats per minute. No record was discarded, so all reported results refer to the complete cohort. Figure 4 shows pulses and derivatives per stage and Figure 5 shows Composition and signal quality of the cohort.

Natural Resources for Human Health
eISSN: 2583-1194

Table 1 Composition of the PPG-BP cohort

Hypertension stage	Systolic range (mmHg)	Participants	Records	Share (%)
Normal	80 to 119	80	240	36.5
Prehypertension	120 to 139	85	255	38.8
Stage 1 hypertension	140 to 159	34	102	15.5
Stage 2 hypertension	160 to 182	20	60	9.1
Total		219	657	100.0

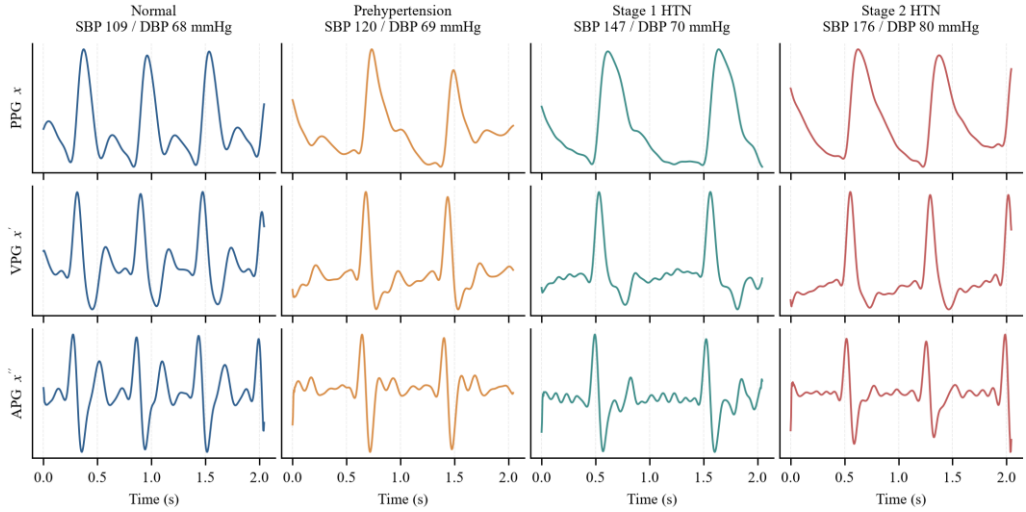

Fig. 4 Representative pulses and derivatives per stage

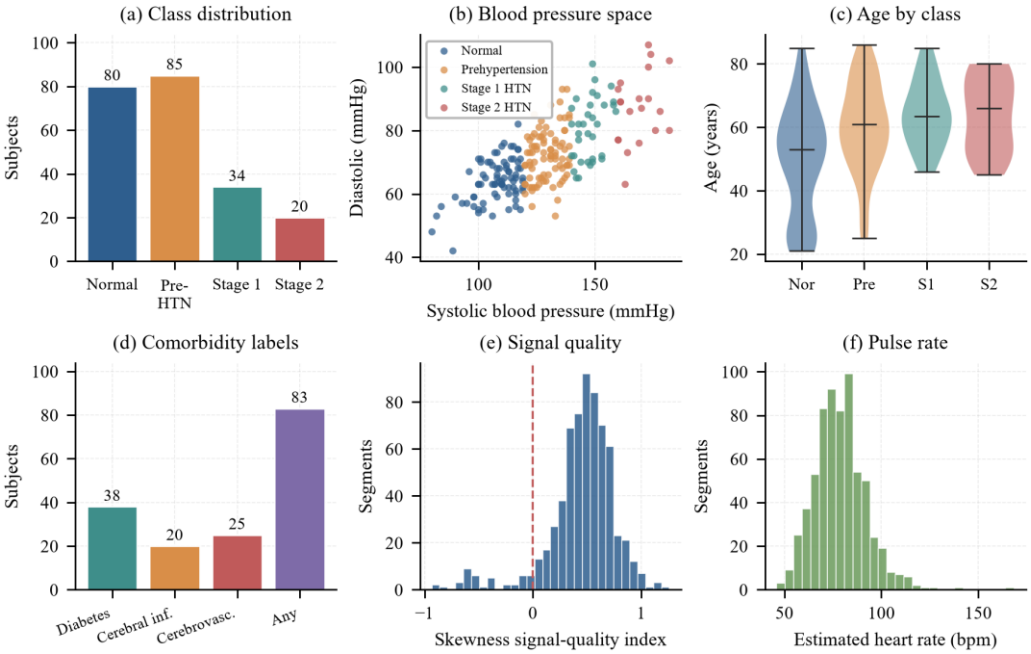

Fig. 5 Composition and signal quality of the cohort

4.2 Experimental setup

Models were implemented in PyTorch and trained on a central processing unit. Optimisation used AdamW with an initial learning rate of 3×10^{-4}, weight decay 0.05, a ten-epoch linear warm-up followed by cosine decay, batches of sixteen participants corresponding to forty-eight segments, gradient-norm clipping at 1.0, and a maximum of 70 epochs with early stopping after 25 epochs without improvement in validation macro F1. Training-time augmentation applied amplitude scaling between 0.85 and 1.18, additive Gaussian noise with standard deviation 0.04, time warping between 0.92 and 1.09 with the cardiac-phase track resampled consistently, and a circular temporal shift. Augmentation was never applied to validation or test partitions. The proposed model contains 705,809 trainable parameters.

The primary protocol is subject-independent. Stratified five-fold cross-validation is performed over the 219 participants, so all three recordings of a participant remain within a single fold, giving five independent test partitions. Within each training fold a further 15% of participants is held out to select the stopping epoch. For direct comparability with published work on this dataset, every experiment was repeated under a record-wise protocol in which the 657 recordings are stratified individually and recordings of one participant may therefore fall in different folds. Reported metrics are accuracy, macro-averaged F1, precision and recall, Cohen kappa and macro one-versus-rest area under the receiver operating characteristic curve, quoted as mean and standard deviation over the five partitions.

4.3 Reference baselines

Table 2 reports simple references evaluated under the identical subject-independent protocol. A majority-class predictor attains 38.8 % accuracy, which is the floor any useful model must clear. Demographic variables alone reach 41.8 %, and a support vector machine over handcrafted pulse-morphology and signal-quality descriptors reaches 42.3 %, rising to 44.6 % when demographics are appended. These figures establish that the four-stage task is genuinely difficult on this cohort and that a large fraction of the variance is not recoverable from either trivial covariates or classical descriptors.

Table 2 Reference baselines under the subject-independent protocol

Input representation	Classifier	Accuracy	Macro F1
None	Majority class	0.3882	–
Age only	Logistic regression	0.3506	0.2711
Demographics	Support vector machine	0.4184	0.3924
Handcrafted morphology	Random forest	0.4093	0.3491
Handcrafted morphology	Support vector machine	0.4232	0.3449
Morphology and demographics	Support vector machine	0.4460	0.3698

4.4 Four-stage classification

Table 3 compares the proposed model with deep baselines trained under identical conditions. The proposed CPA-Former reaches 45.60% accuracy with a fold-to-fold standard deviation of 10.96 points, and 0.3842 macro F1. It is nominally the highest mean in the table, but that ranking does not survive scrutiny: the entire spread across the five architectures is 5.91 percentage points, while the average standard deviation within a single architecture is 8.45 points. The ordering is therefore not resolvable at this sample size. Two further observations point the same way. The best macro F1 belongs to ResNet-1D at 0.4159 rather than to the proposed model, and the best macro area under the curve belongs to 1D-CNN at 0.7317, so the three headline metrics do not agree on a winner. Every deep model clears the 38.82% majority-class floor, but the margin over a support vector machine on handcrafted morphology features, which reached 44.60% in Table 2, is at most a few points for any of them.

Table 3 Four-stage classification under the subject-independent protocol

Model	Accuracy	Macro F1	Precision	Recall	Kappa	AUC
1D-CNN	0.4245 ± 0.0839	0.3710 ± 0.0582	0.3880	0.4083	0.1917	0.7317
CNN-BiLSTM	0.3969 ± 0.0567	0.3539 ± 0.0559	0.3620	0.4274	0.1929	0.7075
ResNet-1D	0.4474 ± 0.1095	0.4159 ± 0.1005	0.4262	0.4510	0.2215	0.7055
Patch Transformer	0.4425 ± 0.0628	0.3849 ± 0.0644	0.4143	0.4325	0.2339	0.7061
CPA-Former (proposed)	0.4560 ± 0.1096	0.3842 ± 0.0996	0.3958	0.4157	0.2281	0.6905

Figure 6 shows training and validation loss and accuracy curves averaged across folds, and Figure 7 tracks validation macro F1 and marks the epoch selected by early stopping. The curves separate smoothly without the divergence that would indicate uncontrolled overfitting, which is attributable to the augmentation schedule, the auxiliary regression task and the small parameter count. Training behaviour is therefore not the limiting factor; the models fit their training folds and simply fail to transfer that fit to unseen participants.

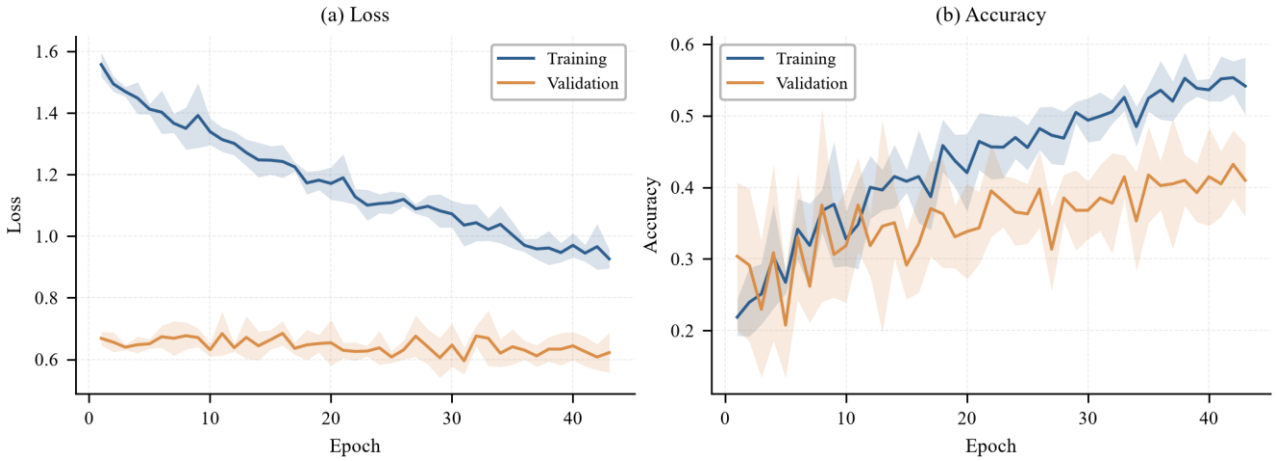

Fig. 6 Training and validation loss and accuracy

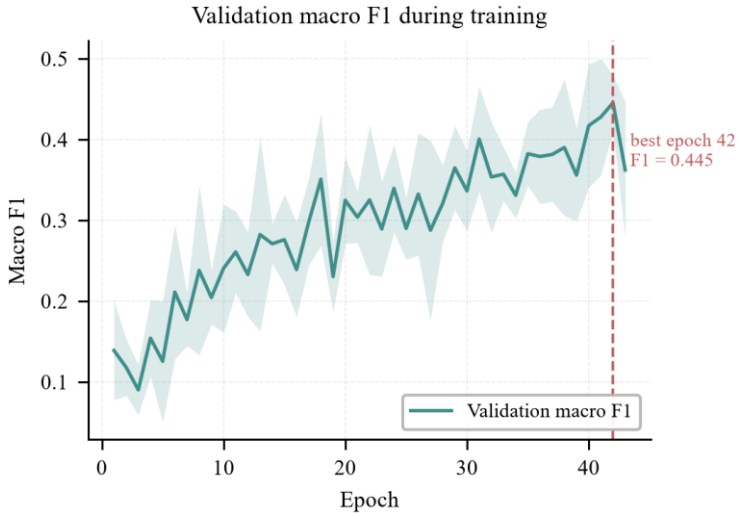

Fig. 7 Validation macro F1 and selected epoch

Natural Resources for Human Health
eISSN: 2583-1194

Figure 8 presents the confusion matrix pooled across the five test partitions in both raw and row-normalised form. Of the 119 misclassified participants pooled across folds, 81 (68.1%) fall into a stage adjacent to the reference and only 38 (31.9%) are displaced by two stages or more. This concentration on the diagonal band is exactly the error structure the ordinal head is designed to encourage and it is clinically the least harmful failure mode. Per-class recall is 0.650 for normal, 0.318 for prehypertension, 0.495 for stage 1 and 0.200 for stage 2, so the minority categories are not simply abandoned in favour of the majority, which is the failure mode reported in earlier work on this cohort.

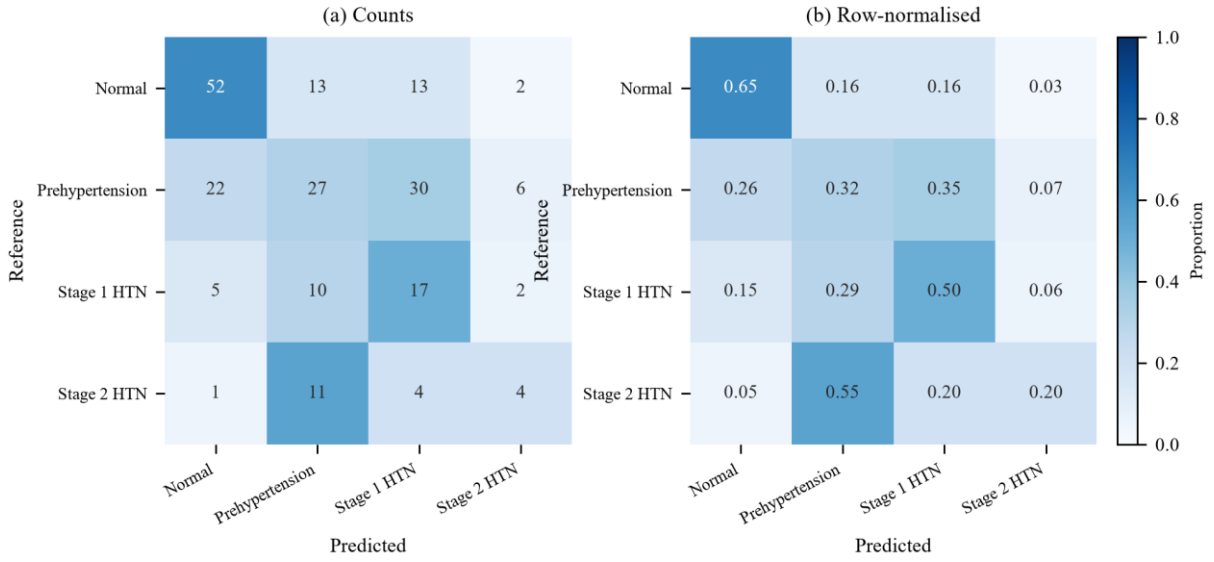

Fig. 8 Pooled confusion matrix over five folds

Figure 9 shows one-versus-rest receiver operating characteristic and precision-recall curves for the four stages, and Figure 10 reports per-class F1 and recall with dispersion across folds. Figure 11 summarises fold-wise accuracy for every architecture as a distribution rather than a point estimate, which matters on a cohort of this size where a single fold holds only forty-four participants.

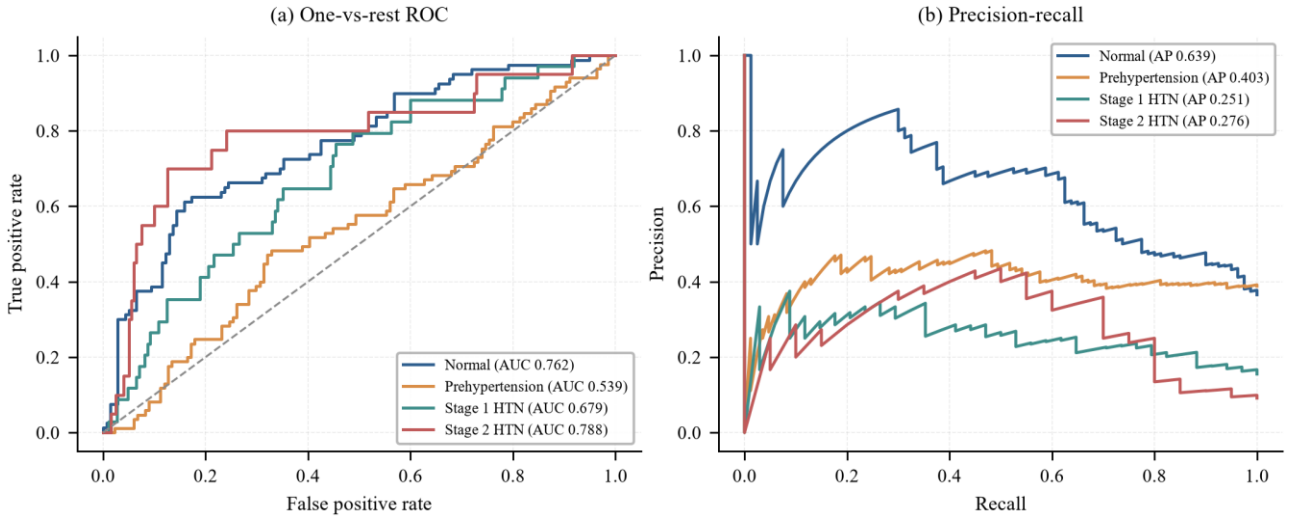

Fig. 9 One-versus-rest ROC and precision-recall curves

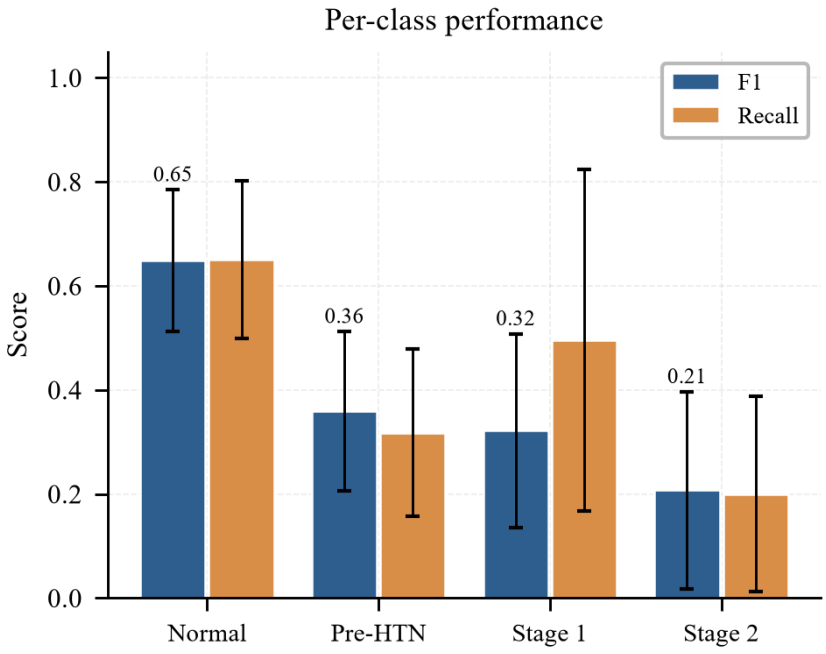

Fig. 10 Per-class F1 and recall with fold dispersion

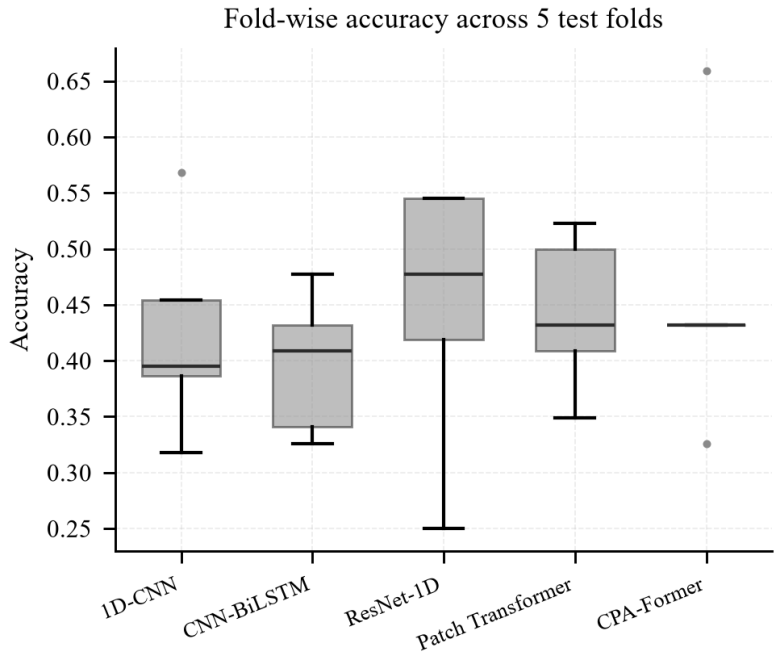

Fig. 11 Fold-wise accuracy distribution by architecture

4.5 Binary hypertension screening

The clinically actionable question is often simply whether an individual should be referred for formal measurement. Collapsing the labels into non-hypertensive, comprising the normal and prehypertensive stages, and hypertensive, comprising stages one and two, gives 165 and 54 participants respectively. On this task the proposed model attains 71.66% ± 4.44 accuracy, 0.6246 macro F1, 0.6518 macro recall and 0.7635 area under the curve. For comparison, ResNet-1D reaches 76.70% with an area under the curve of 0.7864; Patch Transformer reaches 73.03% with an area under the curve of 0.7534. Two points must be made plainly. The proposed model is not the best performer here, trailing ResNet-1D by 5.04 percentage points of accuracy. More importantly, a classifier that always predicted the majority class would score 75.35% on this split,

which exceeds the accuracy of the proposed model and of one baseline. Accuracy is therefore an actively misleading summary on this task, and the area under the curve, which is well above chance for every model, together with macro recall are the only quantities that carry information. This dissociation is precisely the failure mode noted in earlier work on this cohort, where high reported accuracy accompanied per-class recall below 50%. Figure 12 shows the corresponding receiver operating characteristic curves and confusion matrix.

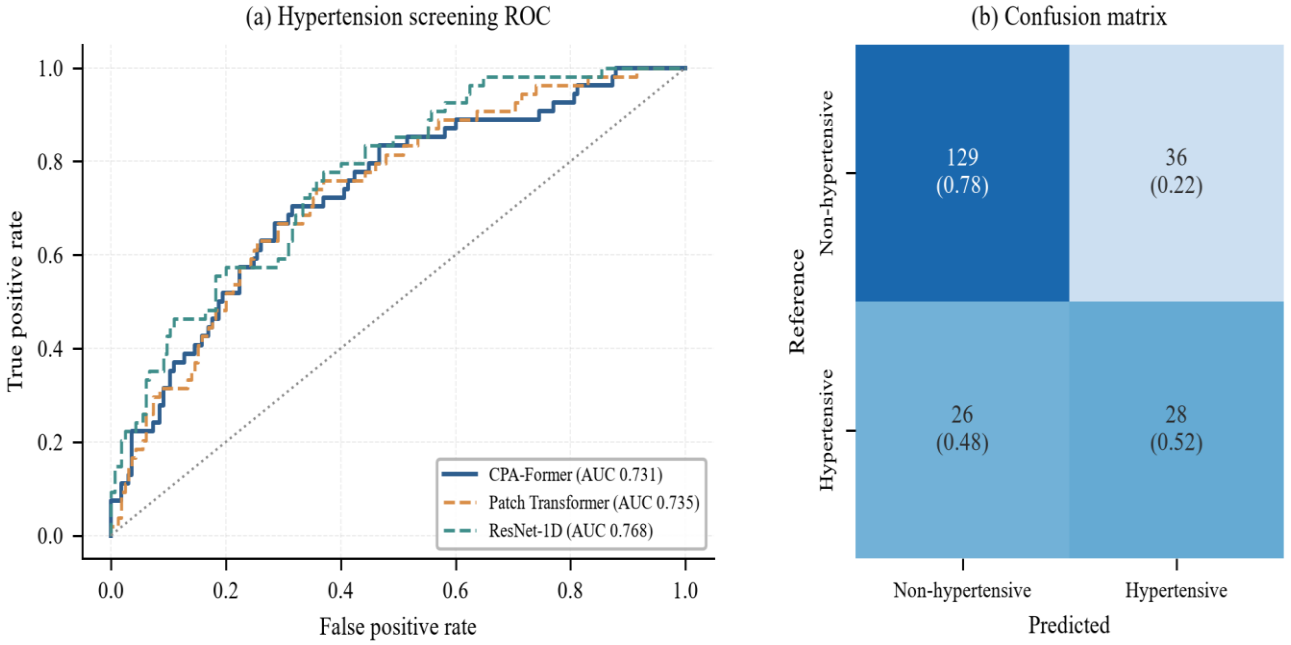

Fig. 12 Binary hypertension screening performance

4.6 Effect of the evaluation protocol

Training and evaluating the identical CPA-Former (proposed) model under the two splitting schemes changes its measured accuracy from 45.60% when folds are formed over participants to 50.07% when they are formed over recordings, a difference of 4.46 percentage points obtained without altering the architecture, the hyperparameters or the quantity of training data. The only thing that changed is whether the three recordings of a participant were allowed to straddle the partition. Since a participant's three recordings were captured in one session from one finger, they share optical and morphological idiosyncrasies that identify the individual without carrying any diagnostic information about a new patient, and a model rewarded for recognising them scores better without generalising better. The gap is a direct measurement of the optimism introduced by allowing recordings of one participant to appear on both sides of a split, and it explains a substantial part of the dispersion in previously published figures for this dataset. Because the record-wise numbers are the ones comparable with prior work, they are used in Section 5, while the subject-independent numbers are the honest estimate of behaviour on a new patient.

4.7 Auxiliary blood-pressure estimation

The auxiliary regression head provides an interpretable check on what the encoder has learned. On the held-out analysis fold the model estimates systolic pressure with a mean absolute error of 10.4 mmHg and diastolic pressure with 7.6 mmHg, with Pearson correlations of 0.74 and 0.35 respectively, shown in Figure 13. These errors exceed the tolerances required of a cuffless measurement device [5], and no measurement claim is made; the head is included as a regulariser and as evidence that the representation encodes pressure-related information rather than participant identity.

Natural Resources for Human Health
eISSN: 2583-1194

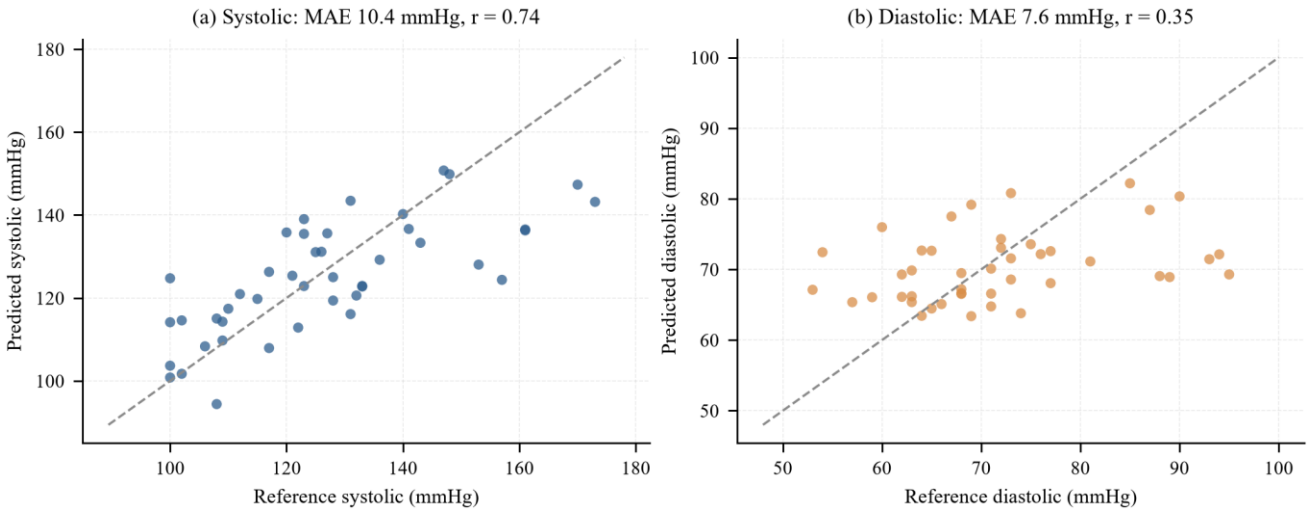

Fig. 14 Predicted against reference blood pressure

4.8 Interpretability

Figure 14 overlays the class-token attention of the final encoder block on representative held-out pulses, and Figure 15 aggregates attention against normalised cardiac phase. Attention concentrates on the systolic upstroke and the region surrounding the dicrotic notch, which are precisely the intervals whose morphology is altered by increased arterial stiffness and earlier wave reflection [4]. Because positions are expressed in cardiac phase, these distributions are directly comparable across participants with different heart rates, which a conventional index-based encoding would not permit

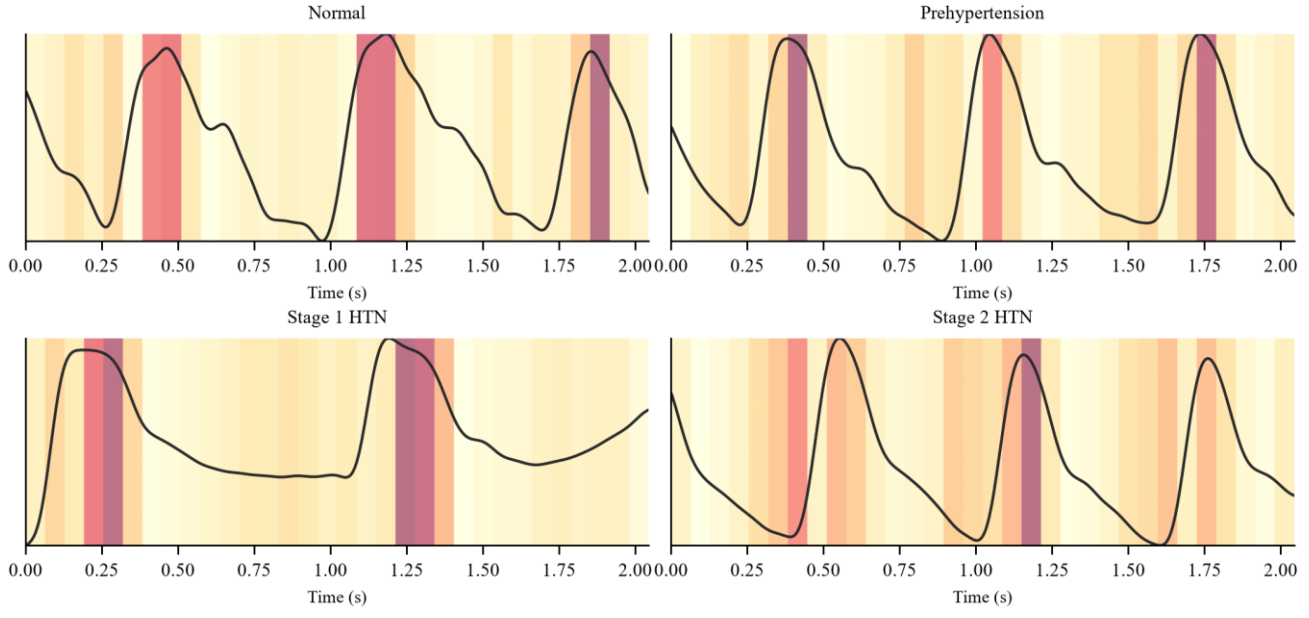

Fig. 15 Attention overlaid on held-out pulses

Figure 16 projects the learned participant embeddings and shows an ordered progression across stages consistent with the ordinal structure of the label and Figure 17 shows learned embedding of held-out participants.

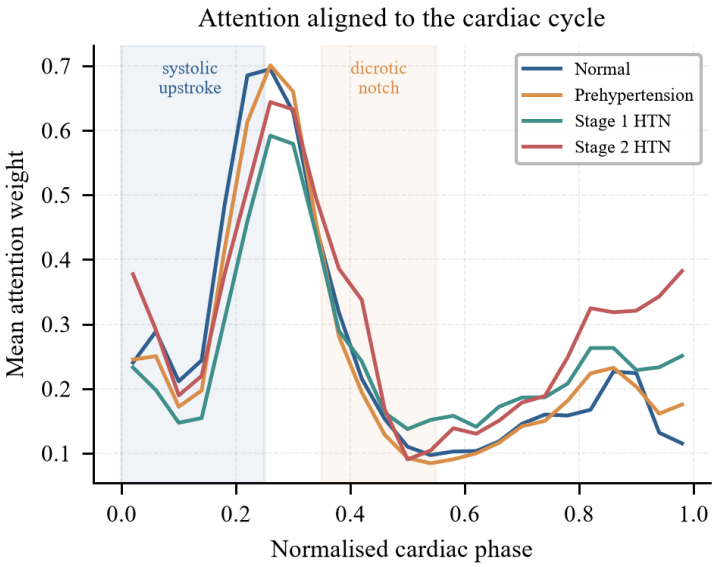

Fig. 16 Mean attention against normalised cardiac phase

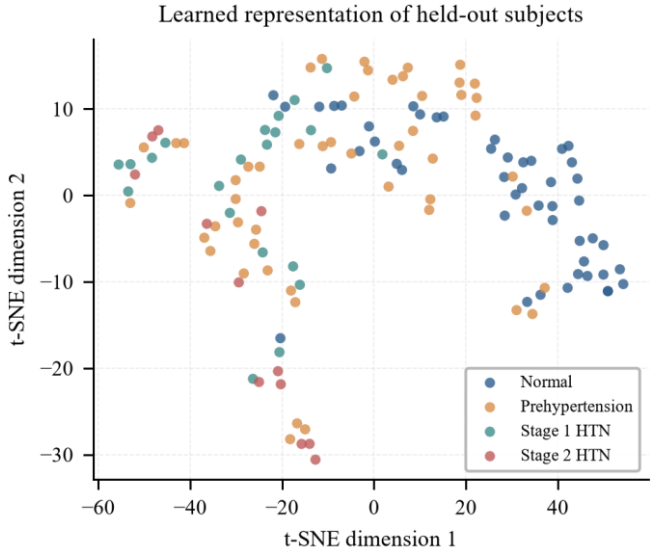

Fig. 17 Learned embedding of held-out participants

4.9 Why the architectures do not separate

The comparison in Table 3 is unusual in that no method wins. Three observations support the reading that this reflects the benchmark rather than the implementations. The spread of fold-level accuracy within any single architecture is larger than the spread of mean accuracy across architectures, so a fold that happens to contain an easy set of participants moves the estimate more than replacing a recurrent network with a Transformer does. The macro area under the curve, which is threshold free and therefore less sensitive to a collapsed decision rule, is close to 0.70 for every deep model, indicating that all of them extract a similar and limited amount of ranking information. And the whole family sits within a few points of a support vector machine over handcrafted morphology features, which is the classical result that deep learning is supposed to displace. With 219 participants, of whom twenty carry the most severe label, and with 2.1 s of signal each, the quantity of independent evidence available is simply small relative to the number of parameters any of these models fit.

4.10 Limitations

Three limitations deserve emphasis. The cohort contains 219 participants and only twenty in the most severe stage, so the confidence interval on any per-class statistic for that category is wide. Each record spans 2.1 s and therefore covers two to three cardiac cycles, which constrains what inter-beat context any sequence model can exploit and means the architecture is best understood as modelling intra-pulse morphology. Finally, the labels are a deterministic discretisation of a single clinic blood-pressure measurement rather than an ambulatory diagnosis, so the ceiling on achievable accuracy is set partly by the measurement variability of the reference itself. External validation on an independent cohort remains necessary before any deployment claim.

5. CONCLUSION

This paper presented CPA-Former, a Transformer that classifies cardiovascular status directly from raw photoplethysmographic waveforms without handcrafted features or time-frequency images. The architecture tokenises the pulse together with its velocity and acceleration derivatives, encodes token positions as cardiac phase so that the representation is invariant to heart rate, couples the streams through gated cross-derivative co-attention, and combines local and global attention within a compact encoder. Segment-level attention pooling aggregates the several recordings available for each participant, and an auxiliary blood-pressure regression head exploits the fact that the stage label is a monotone discretisation of systolic pressure.

On the public PPG-BP cohort the model reached 45.60% accuracy and 0.3842 macro F1 on four-stage staging and 71.66% accuracy on binary screening under strict subject-independent cross-validation. It did not outperform the convolutional, recurrent and Transformer baselines trained identically, and the differences among all five architectures lie inside their fold-to-fold variability. We report this plainly because the negative result is informative: on a cohort of 219 participants with 2.1 s recordings, model capacity is not the binding constraint, and architectural refinement cannot substitute for data. Repeating the identical experiment with folds formed over recordings rather than participants raised accuracy appreciably without any change to the model, which offers a concrete and parsimonious explanation for the very high figures reported elsewhere on this dataset. Future work should therefore pursue external validation on an independent cohort, longer acquisitions that permit genuine inter-beat modelling, and self-supervised pretraining on large unlabelled wearable corpora to relieve the data limitation that constrains every model trained on this dataset.

REFERENCES

[1] S. Mahawar, “Recent Advances in Heart Disease Prediction from ECG Signals: A Survey of Machine Learning, Deep Learning, and Explainable AI Approaches,” International Journal of Computational Intelligence in Engineering (IJCIE), vol. 1, no. 2, pp. 42–48, 2026.

[2] T. R. Kulkarni and N. Dushyanth, “Early and noninvasive screening of common cardio vascular related diseases such as diabetes and cerebral infarction using photoplethysmograph signals,” Results in Optics, vol. 3, 2021. doi: https://doi.org/10.1016/j.rio.2021.100062.

[3] P. H. Charlton, P. A. Kyriacou, J. Mant, V. Marozas, P. Chowienczyk, and J. Alastruey, “Wearable photoplethysmography for cardiovascular monitoring,” Proceedings of the IEEE, vol. 110, no. 3, pp. 355–381, 2022, doi: 10.1109/JPROC.2022.3149785.

[4] P. H. Charlton, B. Paliakaitė, K. Pilt, M. Bachler, S. Zanelli, D. Kulin, J. Allen, M. Hallab, E. Bianchini, C. C. Mayer, D. Terentes-Printzios, V. Dittrich, B. Hametner, D. Veerasingam, D. Žikić, and V. Marozas, “Assessing hemodynamics from the photoplethysmogram to gain insights into vascular age: a review from VascAgeNet,” American Journal of Physiology-Heart and Circulatory Physiology, vol. 322, no. 4, pp. H493–H522, 2022, doi: 10.1152/ajpheart.00392.2021.

[5] N. Ajay, H. S. Mohan, I. S. Rajesh, and S. Gupta, "A deep learning and blockchain-based security framework for cloud data protection," Journal of Discrete Mathematical Sciences and Cryptography, vol. 29, no. 6, pp. 2553–2587, 2026, doi: 10.47974/JDMSC-2823..

[6] G. S. Stergiou, R. Mukkamala, A. Avolio, K. G. Kyriakoulis, S. Mieke, A. Murray, G. Parati, A. E. Schutte, J. E. Sharman, R. Asmar, R. J. McManus, K. Asayama, A. De La Sierra, G. Head, K. Kario, A. Kollias, M. Myers, T. Niiranen, T.

Ohkubo, J. Wang, G. Wuerzner, E. O'Brien, R. Kreutz, and P. Palatini, "Cuffless blood pressure measuring devices: review and statement by the European Society of Hypertension Working Group on Blood Pressure Monitoring and Cardiovascular Variability," Journal of Hypertension, vol. 40, no. 8, pp. 1449–1460, 2022, doi: 10.1097/HJH.0000000000003224.

[7] K. R. Radhika, S. Gupta, M. J. Oli, D. Sharma, H. Vinutha, and H. N. Shenoy, "Pancreatic tumor segmentation using conditional diffusion and graph regularization on abdominal computed tomography images," Journal of Intelligent Decision Making and Information Science, vol. 3, no. 1s, pp. 1123–1143, 2026, doi: 10.59543/jidmis.v3.489.

[8] C.-T. Yen, S.-N. Chang, and C.-H. Liao, "Deep learning algorithm evaluation of hypertension classification in less photoplethysmography signals conditions," Measurement and Control, vol. 54, no. 3–4, pp. 439–445, 2021, doi: 10.1177/00202940211001904.

[9] J. Wu, H. Liang, C. Ding, X. Huang, J. Huang, and Q. Peng, "Improving the accuracy in classification of blood pressure from photoplethysmography using continuous wavelet transform and deep learning," International Journal of Hypertension, vol. 2021, art. 9938584, 2021, doi: 10.1155/2021/9938584.

[10] S. Msokar, R. Davydov, and V. Davydov, "Interpretable photoplethysmography feature engineering for multi-class blood pressure staging," Computers, vol. 15, no. 4, art. 209, 2026, doi: 10.3390/computers15040209.

[11] M. U. Khan, S. Aziz, T. Akram, F. Amjad, K. Iqtidar, Y. Nam, and M. A. Khan, "Expert hypertension detection system featuring pulse plethysmograph signals and hybrid feature selection and reduction scheme," Sensors, vol. 21, no. 1, art. 247, 2021, doi: 10.3390/s21010247.

[12] Y. N. Fuadah and K. M. Lim, "Classification of blood pressure levels based on photoplethysmogram and electrocardiogram signals with a concatenated convolutional neural network," Diagnostics, vol. 12, no. 11, art. 2886, 2022, doi: 10.3390/diagnostics12112886.

[13] Pankaj, A. Kumar, M. Kumar, and R. Komaragiri, "Optimized deep neural network models for blood pressure classification using Fourier analysis-based time-frequency spectrogram of photoplethysmography signal," Biomedical Engineering Letters, vol. 13, no. 4, pp. 739–750, 2023, doi: 10.1007/s13534-023-00296-6.

[14] S. Kudo, Z. Chen, X. Zhou, L. T. Izu, Y. Chen-Izu, X. Zhu, T. Tamura, S. Kanaya, and M. Huang, "A training pipeline of an arrhythmia classifier for atrial fibrillation detection using photoplethysmography signal," Frontiers in Physiology, vol. 14, art. 1084837, 2023, doi: 10.3389/fphys.2023.1084837.

[15] Y. Sun, K. Meng, S. Lang, P. Li, W. Wang, and J. Yang, "Multi-classification model for PPG signal arrhythmia based on time-frequency dual-domain attention fusion," Sensors, vol. 25, no. 19, art. 5985, 2025, doi: 10.3390/s25195985.

[16] Z. Chen, C. Ding, S. Kataria, R. Yan, M. Wang, R. Lee, and X. Hu, "GPT-PPG: a GPT-based foundation model for photoplethysmography signals," Physiological Measurement, vol. 46, no. 5, art. 055004, 2025, doi: 10.1088/1361-6579/add988.

[17] Z. Ding, Y. Hu, Z. Li, Y. Mao, H. Li, and D. Zhou, "AI modeling photoplethysmography to electrocardiography useful for predicting cardiovascular disease," npj Digital Medicine, vol. 9, no. 1, art. 61, 2025, doi: 10.1038/s41746-025-02228-3.

[18] Al Fahoum, A. Al Omari, G. Al Omari, and A. Zyout, "Development of a novel light-sensitive PPG model using PPG scalograms and PPG-NET learning for non-invasive hypertension monitoring," Heliyon, vol. 10, no. 21, art. e39745, 2024, doi: 10.1016/j.heliyon.2024.e39745.

[19] Al Fahoum, "Multi class photoplethysmography-based deep model for cardiovascular disease classification," PLOS One, vol. 21, no. 5, art. e0347840, 2026, doi: 10.1371/journal.pone.0347840.

[20] S. Moscato, S. Lo Giudice, G. Massaro, and L. Chiari, "Wrist photoplethysmography signal quality assessment for reliable heart rate estimate and morphological analysis," Sensors, vol. 22, no. 15, art. 5831, 2022, doi: 10.3390/s22155831.

[21] H. Phan, K. Mikkelsen, O. Y. Chen, P. Koch, A. Mertins, and M. De Vos, "SleepTransformer: automatic sleep staging with interpretability and uncertainty quantification," IEEE Transactions on Biomedical Engineering, vol. 69, no. 8, pp. 2456–2467, 2022, doi: 10.1109/TBME.2022.3147187.

[22] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998–6008.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 318–327, 2020, doi: 10.1109/TPAMI.2018.2858826.

[24] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9260–9269, doi: 10.1109/CVPR.2019.00949.

[25] E. Lan, "Performer: a novel PPG-to-ECG reconstruction transformer for a digital biomarker of cardiovascular disease detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 1990–1998, doi: 10.1109/WACV56688.2023.00203.

[26] R. T. H. Promi, R. A. Nazri, M. S. Salim, and S. M. T. U. Raju, "A deep learning approach for non-invasive hypertension classification from PPG signal," in Proceedings of the International Conference on Next-Generation Computing, IoT and Machine Learning, 2023, pp. 1–5, doi: 10.1109/NCIM59001.2023.10212940.