sensors

MDPI

Article

Attention-Guided Personalised Federated Learning with Quality-Aware Aggregation for Wearable Physiological Signal Modelling

Maryam M. Al Dabel

Department of Computer Science and Engineering, College of Computer Science and Engineering, University of Hafr Al Batin, Hafr Al Batin 39524, Saudi Arabia; maldabel@uhb.edu.sa

Abstract

Building accurate machine-learning models for wearable health monitoring is challenging because raw physiological signals are private, users differ substantially, and edge devices have limited communication bandwidth. This paper presents *AGP-FED*, a personalised federated learning framework for stress recognition and heart-rate estimation. A shared ResidualCNN+Transformer backbone with cross-modal attention learns attended representations, while a private task head remains on-device for subject adaptation; an attention-derived quality summary enables optional server-side reweighting. Clients upload only backbone weights (not the private head), reducing total per-round upload from 26.25 MB (full-model FedAvg) to 10.57 MB across participating clients (59.8% reduction). Under LOSO evaluation on WESAD, *AGP-FED* achieves 0.812 accuracy and 0.950 AUC-ROC. On PPG-DaLiA, *AGP-FED* v4 attains 11.68 $\pm$ 2.79 BPM MAE, which is higher than that of FedAvg (7.66 BPM), and does not meet the 6.5 BPM deployment threshold under LOSO. Ablations show that quality-aware (non-uniform) aggregation does not consistently improve on equal-weight averaging in the moderate-heterogeneity WESAD setting; its benefit is expected mainly under strong client-quality shift. To avoid over-interpreting federated–centralised gaps, all performance claims are restricted to federated baselines evaluated under the same LOSO protocol; the centralised models are reported only as descriptive context. Federated SHAP attribution provides subject-level explanations without transmitting raw data off-device; a post-hoc audit of the model's cross-modal attention weights found them to be near-uniform across modalities rather than selectively weighted, so attention-based explanations are not reported as a reliable interpretability result. These derived explanations are treated as potentially sensitive information, and leakage risks and mitigations are discussed.

Keywords: federated learning; cross-modal attention; personalised federated learning; wearable health monitoring; physiological signal analysis; non-IID data heterogeneity; communication efficiency; explainable artificial intelligence; Internet of Medical Things; stress detection; heart-rate estimation

Check for updates

Academic Editor: Tao Liu

Received: 8 September 2026
Revised: 6 November 2026
Accepted: 12 November 2026
Published: 26 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Smartwatches and clinical-grade wearables continuously capture heart rate, electrodermal activity, skin temperature, and motion signals, enabling stress monitoring and cardiovascular assessment [1,2]. Training reliable models from these streams is challenging because physiology and signal quality vary across users and activities, and pooling raw physiological data centrally is often unacceptable in healthcare settings [3,4].

Federated learning (FL) trains models locally and exchanges only parameter updates [5], but wearable data are strongly non-IID across subjects, which degrades uniform aggregation such as FedAvg [6,7]. Personalised FL mitigates this by combining shared representations with user-specific adaptation, while communication constraints motivate reducing what is transmitted each round [8,9].

Applying standard federated optimisers to real wearable data exposes issues that synthetic benchmarks rarely capture simultaneously. Physiological signals reflect genetics, lifestyle, and immediate context, and data from 15 subjects can therefore follow markedly different distributions [10,11]. FedAvg and related methods implicitly assume that client distributions are broadly comparable; when this assumption fails, updates from one subject can reduce performance for another [5,12]. On WESAD [13], the Spearman rank correlation between per-subject KL divergence and accuracy is $\rho \approx -0.50$, indicating that biological heterogeneity limits uniform aggregation. Personalisation methods developed mainly for synthetic label-imbalance settings, including pFedMe and Per-FedAvg, perform poorly under this physiological non-IID structure: Per-FedAvg attains markedly lower accuracy, while pFedMe fails to make training progress at all and is excluded from quantitative comparison (Section 4.2). Both were run with their published default hyperparameters; this sensitivity is consistent with prior reports, although baseline-specific tuning may reduce the gap for Per-FedAvg [14,15].

Signal characteristics add further complexity. On WESAD, the amusement class is under-represented relative to baseline, biasing aggregation rules that weight clients by window count [6,16]. On PPG-DaLiA, PPG and accelerometer signals are sampled at different native rates (64 Hz and 32 Hz), so they must be aligned to a common temporal grid before windowing [17]. High-motion activities such as cycling and table soccer introduce systematic artefacts in PPG-derived features [17], increasing per-activity MAE and motivating mechanisms that down-weight compromised channels.

Most aggregation schemes make a further simplifying assumption: all clients contribute equally. In practice, signal quality varies with perspiration, sensor placement, and activity, so a noisy update from a single device can shift the global model in an unhelpful direction [16]. Communication is another constraint: full-model transmission in FedAvg and FedProx sends every parameter from each participating client in every round, and SCAFFOLD roughly doubles this cost by also transmitting control variates, so total upload grows quickly over many training rounds on bandwidth- and energy-limited wearable devices. Uploading only the shared backbone while keeping a private head on each device substantially reduces this per-round cost under the same accounting basis (Section 4).

Explainability is an additional gap. Deep models for physiological sensing are often opaque, limiting adoption in settings where clinicians need to interpret, validate, or override automated decisions [18,19]. Federated settings compound this difficulty because standard attribution methods require access to local activations. While *AGP-FED* avoids raw-data transfer, attention summaries and SHAP outputs are potentially sensitive derived information; therefore, this paper discusses leakage risks and mitigations rather than claiming no leakage.

Contributions: (i) a personalised FL design with a shared ResidualCNN+Transformer backbone and private on-device task heads; (ii) a temperature-scaled cross-modal attention module designed to perform modality selection and to produce a compact client-side quality summary for optional server-side reweighting ; an internal audit of the trained model's attention weights found them to be near-uniform across modalities rather than selectively weighted, and separately, non-uniform aggregation weighting does not consistently improve mean performance over equal-weight averaging in the moderate-heterogeneity WESAD setting; (iii) backbone-only communication with measured reporting of the to-

tal per-round upload across participating clients; and (iv) federated interpretability via SHAP , with an attention-based interpretability analysis also conducted but found not to show selective per-modality weighting in practice, with an explicit discussion of leakage risks from derived summaries. Evaluation uses LOSO cross-validation on WESAD (stress classification) and PPG-DaLiA (heart-rate regression); the term "strict LOSO" is avoided when a labelled calibration split from the held-out subject is used for head adaptation. On PPG-DaLiA, the current v4 configuration does not reach the 6.5 BPM deployment threshold under LOSO and is outperformed by FedAvg, and this limitation is discussed in Section 5.

The remainder of this paper is organised as follows. Section 2 reviews related work in federated learning and personalised wearable health monitoring. Section 3 describes *AGP-FED*, including the model architecture, cross-modal attention, the personalised training protocol, and the datasets. Section 4 details the experimental setup and reports the results and analysis. Section 5 discusses implications and limitations, including non-IID effects and federated explainability. Section 6 concludes the paper and outlines directions for future work.

2. Literature Review

Wearable physiological sensing has generated a substantial body of machine-learning research, driven by datasets covering heart rate, EDA, skin temperature, and PPG across diverse daily activities [20,21]. Early systems paired hand-crafted time- and frequency-domain features with classical classifiers, achieving reasonable accuracy in controlled laboratory settings but struggling to generalise across subjects and conditions [3,13]. Deep learning shifted the paradigm by learning representations directly from raw signals, reducing the need for manual feature engineering.

Convolutional networks capture local temporal patterns effectively and are widely used for stress detection and activity classification from wearable sensors [22,23]. Recurrent and Transformer-based models provide additional sequential context [24], and hybrid CNN–LSTM architectures have often outperformed either component alone [25,26]. Despite strong performance, these models typically rely on centralised training, which is increasingly problematic given the sensitivity of physiological data [10,11].

Privacy concerns have therefore motivated federated learning, in which devices train locally and exchange model updates rather than raw data [27]. FedHealth [10] provided an early demonstration that transfer learning can be federated for wearable health tasks; follow-up work [28] applied privacy-preserving federated deep learning to wearable IoT-based stress monitoring, and [21] presented a taxonomy of federated medical applications. A central challenge is data heterogeneity: physiological distributions are shaped by age, fitness, and behaviour in ways that violate the identical-distribution assumption underpinning FedAvg [11,29].

Addressing non-IID physiological data without sacrificing communication efficiency has proved difficult. Sattler et al. [12] showed that sparse gradient communication could improve convergence under distribution shift; DEEP-FEL [29] addressed decentralised edge settings with efficiency in mind. Personalised FL methods, including shared-backbone/private-head schemes such as FedPer and FedRep and normalisation-based variants such as FedBN, target subject adaptation by keeping part of the model local [30–32]. However, these methods generally do not address signal-quality-aware aggregation or provide federated interpretability mechanisms.

Attention mechanisms have recently moved from centralised physiological models into more complex architectures. AttX [33] demonstrated cross-modal attention for wearable emotion recognition, showing improved accuracy over unimodal and classical

multimodal fusion. Transformer-driven approaches [24] and cross-sensor fusion methods [34] report similar gains, but are typically evaluated centrally. In federated learning, attention-based aggregation has also been explored (e.g., FedAtt [35]), but not in combination with wearable-specific modality attention and backbone-only communication in a single pipeline.

On the communication side, gradient compression [36] and quantisation [37,38] reduce per-round cost but are not designed primarily to improve aggregation quality. Methods that weight clients non-uniformly based on data quality [16] address the aggregation problem but typically rely on hand-crafted importance metrics rather than learned attention.

Explainability is the least-explored dimension in federated wearable sensing. Jaber et al. [18] designed an explanatory report for stress prediction from wearable sensors, modelled on a standard blood-test report, that presents the physiological measurements contributing to each prediction together with reference intervals for stress and non-stress states; their explanations were consistent with ground truth and judged useful by psychiatrists, but their analysis requires full data access. A recent survey [39] confirmed that privacy-preserving explainability in federated healthcare remains largely unsolved. Taken together, no existing work integrates attention-based feature selection, non-uniform quality-aware aggregation, backbone-only communication, and federated explainability for wearable physiological tasks in a single framework. Table 1 summarises related work.

Table 1. Comparison of Representative Methods with the Proposed *AGP-FED* Framework.

Work	Federated Learning	Wearable Health Task	Personalised Client Model	Attention Mechanism	Non-IID Robustness	Comm. Efficiency	XAI Support
McMahan et al. [5]	✓	×	×	×	×	×	×
Li et al. [7]	✓	×	×	×	✓	×	×
Chen et al. [10]	✓	✓	×	×	×	×	×
Sattler et al. [12]	✓	×	×	×	✓	✓	×
Can and Ersoy [28]	✓	✓	×	×	×	×	×
Gil-Martín et al. [22]	×	✓	×	×	×	×	×
Lian et al. [29]	✓	×	×	×	✓	✓	×
Tan et al. [8]	✓	×	✓	×	✓	×	×
Tanwar et al. [25]	×	✓	×	✓	×	×	×
Bhatti et al. [33]	×	✓	×	✓	×	×	×
Almadhor et al. [40]	✓	✓	×	×	×	×	×
Lu et al. [41]	✓	✓	✓	×	✓	×	×
Singh et al. [26]	×	✓	×	×	×	×	×
Moser et al. [19]	×	✓	×	×	×	×	✓
Proposed (*AGP-FED*)	✓	✓	✓	✓	✓	✓	✓

3. Methodology

AGP-FED is illustrated in Figure 1. Clients align multi-rate signals and construct windows, train a shared ResidualCNN+Transformer backbone with cross-modal attention, and upload only backbone updates to the server. The server aggregates the received backbones (optionally using a learned quality summary), while private task heads remain on-device. Stress classification and heart-rate estimation share the same backbone; only the task head and loss function differ.

Two publicly available benchmarks are used. WESAD [13] contains chest-worn recordings from 15 subjects labelled across three affective conditions (baseline, stress, and amusement) and exhibits pronounced class imbalance. PPG-DaLiA [17] contains 15 subjects performing eight structured activities, ranging from sitting to table soccer and cycling, with concurrent chest-worn ECG providing heart-rate reference values. Each subject forms one federated client; evaluation follows leave-one-subject-out (LOSO) cross-validation, ensuring that training and test subjects do not overlap.

All channels are resampled to 32 Hz and filtered per modality. On WESAD, class imbalance is handled within-client via oversampling: SMOTE [42] is attempted when available/feasible; otherwise, minority windows are duplicated with additive Gaussian jitter ($\sigma = 0.01$). On PPG-DaLiA, a per-window SNR score between PPG and accelerometer channels is used as a reliability weight in the regression loss. For WESAD, windows of 320 samples (10 s at 32 Hz) with a 160-sample step (50% overlap) are extracted; for PPG-DaLiA, windows of 256 samples (8 s) with a 64-sample step (75% overlap) are extracted. Time- and frequency-domain features accompany each raw segment, as described in Algorithm 1.

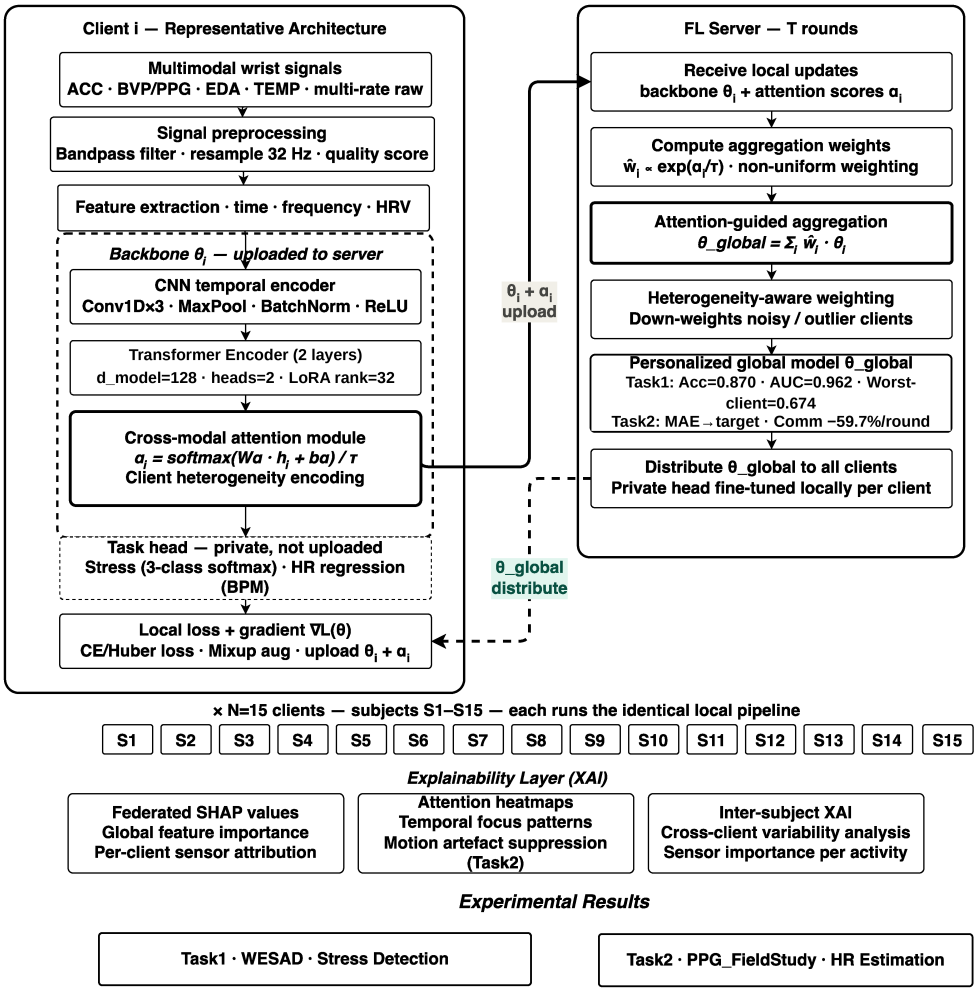


Figure 1. Overview of the proposed *AGP-FED* pipeline. Clients perform local preprocessing and backbone training with cross-modal attention, upload backbone updates (and a lightweight quality summary), and receive an aggregated backbone from the server; private task heads remain on-device.

Each client runs a model with 471,521 parameters in total: a 438,689-parameter shared backbone and a 32,832-parameter private head (Algorithm 2). The backbone comprises

a ResidualCNN encoder (filters $[32, 64, 128]$, kernel size 5, and skip connections) with sinusoidal positional encoding, followed by a two-layer Transformer encoder ($d_{\text{model}} = 128$ and two attention heads). For heart-rate estimation, LoRA adapters [43] (rank $r = 32$) are applied within the Transformer's attention and feed-forward projections, together with a per-client 16-dimensional subject embedding and a channel-gate mechanism. The attention module produces modality weights used to form an attended representation and a scalar quality summary $s_i = \mathbb{E}[\max_m a^{(m)}]$, which can be used for server-side weighting. Importantly, s_i is a heuristic proxy: its effectiveness depends on calibration (correlation with true update utility) and on the heterogeneity/noise regime; therefore, I treat quality-weighted aggregation as optional and evaluate it explicitly via ablations rather than assuming uniform gains. The private task head (classifier/regressor) is never transmitted. Training uses AdamW ($\eta = 10^{-3}$), cosine annealing ($\eta_{\min} = 10^{-5}$), FP16 mixed precision, and gradient clipping at norm 1.0.

On the server (Algorithm 3), approximately $\rho \approx 0.70$ of clients participate per round. Each participating client uploads only its updated backbone and scalar quality summary; the server aggregates backbones using softmax-normalised weights. Training runs for $T = 100$ rounds for stress classification and up to $T = 64$ rounds with per-fold early stopping for heart-rate estimation (patience $= 10$). Baselines follow the same protocol, using a common CNN–BiLSTM backbone (Section 4): FedAvg [5], FedProx [7], SCAFFOLD [44], FedNova [45], pFedMe [14], Per-FedAvg [15], and Ditto [46].

Explanations are derived primarily from per-client SHAP values [47,48] (Algorithm 4); the model's cross-modal attention weights were also examined as a candidate explanation source, but a post-hoc audit found them to be near-uniform across modalities rather than selectively weighted, so they are not treated as a reliable interpretability signal (see Section 4). Stress performance is assessed using accuracy, macro-F1, weighted F1, and AUC-ROC; heart-rate performance is assessed using MAE, RMSE, and Pearson r, with MAE < 6.5 BPM adopted as a deployment threshold. Worst-client accuracy is used to summarise federated fairness. On WESAD, *AGP-FED* reaches accuracy 0.812 and AUC-ROC 0.950. Communication is reported as total upload per round across all participating clients: 10.57 MB/round for *AGP-FED* vs. 26.25 MB/round for full-model FedAvg, a 59.8% per-round reduction.

Algorithm 1 describes the preprocessing pipeline. Heterogeneous sensor streams are aligned to 32 Hz, bandpass-filtered, balanced with within-client random oversampling (stress task), quality-scored (HR task), and segmented into fixed-length windows before federated client partitioning.

Algorithm 1 Physiological Signal Preprocessing and Federated Client Construction

Require: Subject set $\mathcal{S} = \{s_i\}_{i=1}^{N}$, sampling rates $\{f_k\}$, target rate f_t, window length W, stride S
Ensure: Per-client datasets $\{\mathcal{D}_i\}$ and client map $\mathcal{C}$

for each subject $s_i \in \mathcal{S}$ **do**
  for each modality k **do**
    Resample $\mathbf{r}_i^{(k)}$ to f_t and bandpass-filter: $\tilde{\mathbf{r}}_i^{(k)} \leftarrow \mathcal{F}(\mathbf{r}_i^{(k)})$
  end for
  Stack channels: $\mathbf{X}_i \leftarrow [\tilde{\mathbf{r}}_i^{(1)} \| \cdots \| \tilde{\mathbf{r}}_i^{(C)}]$
  if stress task **then**
    $(\mathbf{X}_i, \mathbf{y}_{\text{raw},i}) \leftarrow \text{Oversample}(\mathbf{X}_i, \mathbf{y}_{\text{raw},i})$ ▷ within-client; SMOTE if available, else duplication with Gaussian jitter ($\sigma = 0.01$)
  end if
  $\mathcal{D}_i \leftarrow \varnothing$
  for each window index j **do**
    $\mathbf{x}_{i,j} \leftarrow \mathbf{X}_i[jS : jS+W]$
    $y_{i,j} \leftarrow \text{mode}(\mathbf{y}_{\text{raw},i}[jS : jS+W])$ ▷ majority vote within window
    if HR task **then**
      $\rho_{i,j} \leftarrow \text{SNR}(\mathbf{x}_{i,j}^{(\text{PPG})}, \mathbf{x}_{i,j}^{(\text{ACC})})$
    else
      $\rho_{i,j} \leftarrow 1$
    end if
    $\phi_t \leftarrow [\mu, \sigma, \text{skew}, \text{kurt}, \text{RMSSD}](\mathbf{x}_{i,j})$
    $\phi_f \leftarrow \text{FFT}(\mathbf{x}_{i,j})$
    Store $(\mathbf{x}_{i,j}, [\phi_t \| \phi_f], y_{i,j}, \rho_{i,j}) \in \mathcal{D}_i$
  end for
end for
$\mathcal{C} \leftarrow \{c_i \equiv s_i\}_{i=1}^{N}$
return $\{\mathcal{D}_i\}, \mathcal{C}$

Algorithm 2 covers local training. The backbone is initialised from the global model; the private head is reinitialised each round. The cross-modal attention module produces both the attended representation for prediction and the quality summary $\boldsymbol{\alpha}_i$, which is uploaded to the server.

Algorithm 2 Local Client Training with ResidualCNN–Transformer and Cross-Modal Attention

Require: $\mathcal{D}_i$, global backbone parameters Θ_g^t
Ensure: Updated backbone Θ_i^{t+1}, attention summary α_i

$\Theta_{bb} \leftarrow \Theta_g^t$; reinitialise private head Θ_{hd}^i
if HR task **then**
  Attach LoRA adapters (rank $r = 32$)
end if
for epoch $\leftarrow 1$ to E_l **do**
  for each mini-batch (X, y, ρ) in $\mathcal{D}_i$ **do**
    if HR task **then**
      $\tilde{X} \leftarrow \lambda X + (1 - \lambda) X_\pi$ ▷ Mixup
    else
      $\tilde{X} \leftarrow X$; $\rho \leftarrow \mathbf{1}$
    end if
    $H \leftarrow f_{\text{ResidualCNN}}(\tilde{X})$
    $Z \leftarrow f_{\text{Transformer}}(H)$
    if HR task **then**
      $Z \leftarrow Z + \text{LoRA}(Z)$
    end if
    $\alpha \leftarrow \text{softmax}((W_a Z + b_a)/\tau)$
    $F \leftarrow \alpha \odot Z$
    $\hat{y} \leftarrow f_{\text{task}}(F)$
    if stress task **then**
      $\mathcal{L} \leftarrow -\sum_c w_c y_c \log \hat{y}_c$
    else
      $\mathcal{L} \leftarrow B^{-1} \sum_b \rho_b \, \text{Huber}(\hat{y}_b, y_b)$
    end if
    Update $(\Theta_{bb}, \Theta_{hd}^i)$ with AdamW; clip gradients to norm 1.0
  end for
  Validate; early stopping (patience 10)
end for
$\Theta_i^{t+1} \leftarrow \Theta_{bb}$
$\alpha_i \leftarrow \text{mean}(\alpha)$
return Θ_i^{t+1}, α_i

Algorithm 3 covers server aggregation. Non-uniform weights derived from uploaded attention scores are designed to let high-quality clients drive the global model more strongly than low-quality or artefact-heavy ones; the ablation study (Section 4) finds this does not consistently hold in practice, particularly in the moderate-heterogeneity WESAD setting.

Algorithm 3 *AGP-FED*: Attention-Guided Federated Aggregation

Require: Client set $\mathcal{C}$, initial global backbone Θ_g^0, rounds T, participation rate ρ, temperature τ_s, tolerance ε
Ensure: Final global backbone Θ_g^*

for $t \leftarrow 1$ to T **do**
  Sample participants $\mathcal{S}_t \subseteq \mathcal{C}$ with $|\mathcal{S}_t| = \lfloor \rho |\mathcal{C}| \rfloor$
  Broadcast Θ_g^{t-1} to $\mathcal{S}_t$
  for all $c_i \in \mathcal{S}_t$ in parallel **do**
    $(\Theta_i^{t+1}, \alpha_i) \leftarrow \textsc{LocalTrain}(\mathcal{D}_i, \Theta_g^{t-1})$
    $s_i \leftarrow \text{mean}(\alpha_i)$ ▷ scalar quality summary
  end for
  for all $c_i \in \mathcal{S}_t$ **do**
    $w_i \leftarrow \exp(s_i/\tau_s) / \sum_{j \in \mathcal{S}_t} \exp(s_j/\tau_s)$
  end for
  $\Theta_g^t \leftarrow \sum_{i \in \mathcal{S}_t} w_i \Theta_i^{t+1}$
  if $\|\Theta_g^t - \Theta_g^{t-1}\|_2 < \varepsilon$ **then**
    break
  end if
end for
$\Theta_g^* \leftarrow \Theta_g^t$
for all $c_i \in \mathcal{C}$ **do**
  Fine-tune private head Θ_{hd}^i with frozen Θ_g^*
end for
return Θ_g^*

Algorithm 4 specifies the LOSO evaluation. Each fold trains the federation on $N-1$ subjects, personalises the task head on a small validation split of the held-out subject, then evaluates on that subject's remaining windows.

Algorithm 4 LOSO Cross-Validation and Federated Evaluation

Require: Subject set $\mathcal{S} = \{s_i\}_{i=1}^N$, federated trainer Φ (*AGP-FED*)
Ensure: Per-fold metrics $\mathcal{M}$ and aggregated statistics

$\mathcal{M} \leftarrow \varnothing$
for $f \leftarrow 1$ to N **do**
  $s_{\text{test}} \leftarrow s_f$
  $\mathcal{S}_{\text{train}} \leftarrow \mathcal{S} \setminus \{s_f\}$
  $\Theta_g^{*,f} \leftarrow \Phi(\mathcal{S}_{\text{train}})$
  Personalise Θ_{hd}^f on calibration split $\mathcal{V}_f$
  Evaluate on $\mathcal{D}_{s_{\text{test}}} \setminus \mathcal{V}_f$
  if stress task **then**
    Compute $\text{Acc}_f, \text{MacroF1}_f, \text{AUC}_f$
  else
    Compute $\text{MAE}_f, \text{RMSE}_f, r_f$
  end if
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{\text{metrics}_f\}$
end for
Report mean $\pm$ std; fairness $\leftarrow \min_f(\text{Acc}_f)$
return $\mathcal{M}$

Dataset Collection and Preprocessing

Two publicly available benchmarks form the experimental basis. WESAD [13] was recorded in a controlled laboratory from 15 subjects (labelled S2–S17 in the dataset; S1 and S12 are absent from the public release) wearing an Empatica E4 wrist device and a RespiBAN chest device (ECG, EDA, Resp, EMG, Temp, ACC). Only chest modalities

(RespiBAN) are used here. Labels cover three affective conditions: a neutral baseline, the Trier Social Stress Test, and comedy-induced amusement with short stress and amusement periods, resulting in a markedly skewed class distribution. Recordings are serialised as Python pickle archives with synchronised sensor arrays and a reference label vector.

PPG-DaLiA [17] includes 15 subjects performing 8 daily-life activities defined by the dataset protocol. To avoid inconsistencies, I use the original dataset documentation for the exact activity list, and I use the provided activity labels only for analysis/visualisation (not as model inputs unless stated). A wrist device acquires PPG at 64 Hz and three-axis ACC at 32 Hz; a chest device captures ECG at 700 Hz, used only to derive heart-rate reference targets. Per-subject data are stored as pickle dictionaries containing multi-rate arrays, per-sample HR targets, and activity labels.

Preprocessing begins with polyphase anti-aliasing resampling of all channels to 32 Hz. Zero-phase Butterworth filters are then applied per modality: [0.5, 4.0] Hz bandpass for PPG/BVP to isolate the cardiac pulse band; [0.5, 40] Hz bandpass for ECG; 1.0 Hz low-pass for EDA; [0.1, 0.5] Hz bandpass for Respiration; and high-pass for ACC. On WESAD, within-client class balancing is performed via oversampling: SMOTE is attempted when available/feasible; otherwise, minority windows are duplicated with additive Gaussian jitter ($\sigma = 0.01$) to avoid exact copies. On PPG-DaLiA, each window receives a motion-artefact quality score $q_j = \mathrm{SNR}(\mathbf{r}^{(\mathrm{PPG})}, \mathbf{r}^{(\mathrm{ACC})})$ that feeds directly into the Huber regression loss as a reliability weight.

For WESAD, windows of $W = 320$ samples (10 s at 32 Hz) with step $S = 160$ (50% overlap) yield $n_w = \lfloor (T_i - W)/S \rfloor + 1$ segments per subject. For PPG-DaLiA, $W = 256$ samples (8 s at 32 Hz) with step $S = 64$ (75% overlap) are used. Time-domain summary statistics and spectral features accompany each raw window. Each subject maps to one federated client; the 15-fold LOSO split is disjoint between training and test participants. Mixup augmentation [49] ($\lambda \sim \mathrm{Beta}(0.2, 0.2)$) is applied only during HR training; stress training relies solely on the oversampling-balanced pool. Validation and test windows use the same normalisation as training, with no augmentation.

For manuscript verification, Claude (Claude Code and Claude, Anthropic) was used after all experiments were completed to cross-check reported numerical claims, table entries, and figure captions against the corresponding saved scripts, logs, and model checkpoints, and to flag inconsistencies for manual correction; it was not used to design experiments, generate results, or draw scientific conclusions.

4. Experimental Setup, Analysis and Results

All experiments use leave-one-subject-out cross-validation across $N = 15$ subjects for both WESAD stress classification and PPG-DaLiA heart-rate estimation. When private-head calibration on the held-out subject is enabled, this setting is reported explicitly as "LOSO with test-subject calibration" rather than strict LOSO. The proposed *AGP-FED* framework is trained for $T = 100$ communication rounds for stress and up to $T = 64$ rounds with per-fold early stopping (patience = 10) for heart-rate estimation, with a client participation rate of $\rho \approx 0.70$, i.e., approximately 70% of the training clients are sampled in each round. Local training runs for $E_l = 5$ epochs per round using AdamW with $\eta = 10^{-3}$, cosine annealing decay to $\eta_{\min} = 10^{-5}$, gradient clipping at norm 1.0, and FP16 mixed-precision training on a single NVIDIA GPU under the PyTorch framework.

Baselines: Comparisons are performed against FedAvg [5], FedProx [7], SCAFFOLD [44], FedNova [45], pFedMe [14], Per-FedAvg [15], and Ditto [46] (personalisation via a proximal regulariser towards the global model). Unless stated otherwise, baselines are evaluated under the same LOSO splits, client participation rate, and local epoch budget. All baselines use a common CNN–BiLSTM backbone, whereas *AGP-FED* uses the Residual-

CNN+Transformer backbone described in Section 3; differences between *AGP-FED* and the baselines therefore reflect both the training scheme and the architecture. FedProx, pFedMe, and Ditto are reported on WESAD only. For hyperparameter-sensitive methods (notably pFedMe and Per-FedAvg), I now report the exact hyperparameters used in Table 2, and I explicitly label any baseline run that diverged (or failed to converge) under the matched tuning/compute budget.

Additional personalised baselines (discussion): FedPer/FedRep/FedBN keep (different) subsets of parameters local to enable subject adaptation: FedPer keeps personalisation layers (typically heads) local, FedRep separates representation vs. head learning, and FedBN keeps batch-normalisation statistics local to mitigate feature-shift. These are strong, protocol-relevant comparators for the proposed shared-backbone/private-head design. However, FedPer/FedRep/FedBN are not included as executed baselines in this revision because (i) they require method-specific implementation details for the multimodal ResidualCNN+Transformer backbone (e.g., which blocks are treated as representation vs. *head*, and how BN is handled when some blocks lack BN), and (ii) a fair comparison would require a matched tuning budget (same number of trials/compute) across all methods. Quantitative claims are therefore restricted to the baselines that were fully implemented and tuned under the same training budget, and FedPer/FedRep/FedBN are listed as priority follow-up benchmarks under the same LOSO protocol.

Table 2. Hyperparameters used for sensitive personalised-FL baselines (reported for reproducibility).

Setting	**pFedMe Hyperparameters**
WESAD	$T = 100$ rounds; $E_l = 5$ local epochs; participation $\rho = 0.70$; batch size 64; local/outer LR $\eta = 10^{-3}$ (Adam); FedProx $\mu = 0.01$ (when applicable); Moreau regularisation $\lambda = 15.0$; inner SGD LR $\beta =$ **0.001**; $K = 5$ inner steps per epoch.
Setting	**Per-FedAvg Hyperparameters**
WESAD	Outer/meta LR $\alpha = 10^{-3}$; inner/fast LR $\beta = 10^{-3}$; $T = 100$ rounds; $E_l = 5$; participation $\rho = 0.70$; batch size 64.
PPG-DaLiA	Meta/support-set step size $\alpha = 0.01$; query-set optimiser Adam with LR 10^{-3}; $T = 100$ rounds; $E_l = 5$; participation in $[0.60, 0.80]$; batch size 64.

Stress classification is assessed using accuracy, macro-averaged F1, weighted F1, and AUC-ROC. Heart-rate estimation is assessed using mean absolute error (MAE, in BPM), RMSE, and Pearson correlation coefficient, with MAE < 6.5 BPM adopted as a deployment threshold. Where close variants are compared (e.g., Full vs. A2), 95% confidence intervals are reported and paired tests across LOSO folds are used to reduce the risk of over-interpreting small mean differences under substantial fold-wise variance.

4.1. Centralised Reference Performance

Figure 2 presents the training and validation convergence curves of the centralised CNN–BiLSTM model on the WESAD stress classification task. Across all 15 LOSO folds, the mean cross-entropy loss decreases steeply over the first five epochs and falls below 0.05 by approximately epoch 10. Validation accuracy, computed on the held-out subject, varies widely across subjects, and the mean curve stabilises at approximately 0.76; folds end at different epochs because of early stopping (patience 10). Final per-fold results use the checkpoint with the highest held-out-subject accuracy, giving a LOSO mean accuracy of 0.855, macro-F1 of 0.794, and AUC-ROC of 0.935. Because checkpoint selection used the held-out subject, these centralised figures are an optimistic reference rather than an unbiased upper bound. The reference model also differs from the federated pipeline in architecture (CNN–

BiLSTM vs. ResidualCNN+Transformer) and does not use Mixup or LoRA adapters; it is therefore not used to attribute gains specifically to federated personalisation.

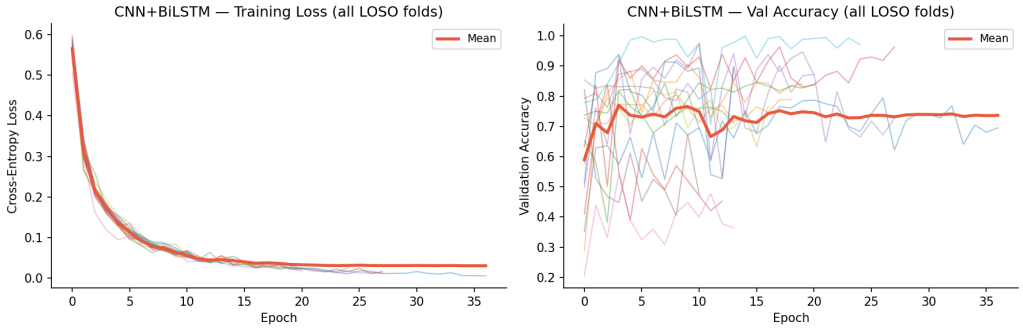


Figure 2. Centralised CNN–BiLSTM convergence on WESAD (15-fold LOSO). (**Left**): training loss. (**Right**): validation accuracy. Thin coloured lines: individual LOSO folds; bold red line: mean across folds.

Figure 3 shows the centralised CNN-BiLSTM convergence on the PPG-DaLiA heart-rate estimation task. Each subplot corresponds to one LOSO fold, with both training (MSE) and validation loss converging consistently across subjects. The LOSO-averaged MAE is 7.082 BPM with Pearson $r = 0.861$.

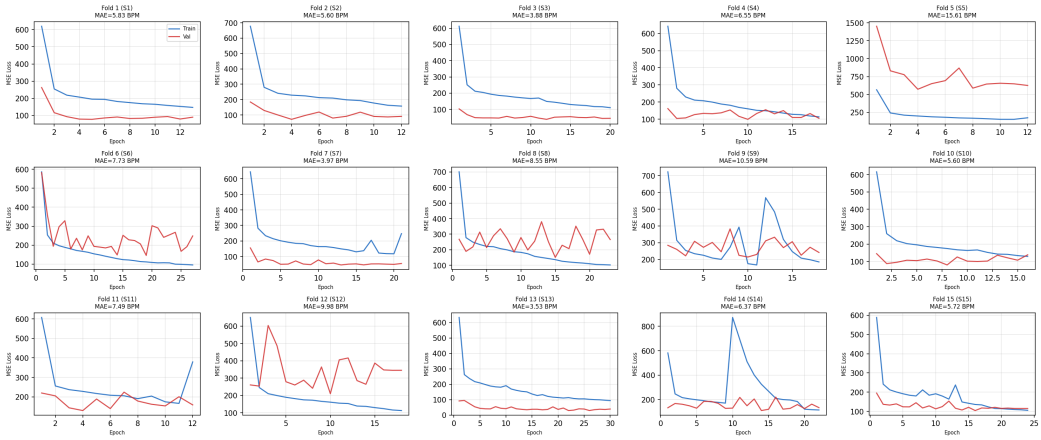


Figure 3. Centralised CNN–BiLSTM convergence on PPG-DaLiA heart-rate estimation (15-fold LOSO). Per-fold training/validation loss; LOSO mean MAE = 7.082 BPM, $r = 0.861$.

The final LOSO results are summarised in Table 3 (WESAD) and Table 4 (PPG-DaLiA).

Table 3. LOSO performance comparison on the WESAD stress classification benchmark. Worst-Client Acc. is the lowest per-fold accuracy across the 15 LOSO folds. † pFedMe's personalised parameters and per-round accuracy history showed no meaningful change in 14 of the 15 LOSO folds (collapsing to one of two values in the remaining fold); see Section 4.2 and Section 5 for diagnostics. "–" indicates that the value was not recorded in the experimental logs.

Method	Accuracy	Macro F1	AUC-ROC	Worst-Client Acc.
CNN-BiLSTM (Centralised)	0.855	0.794	0.935	0.543
FedAvg [5]	0.740	0.639	0.833	0.341
FedProx [7]	0.762	0.663	0.868	0.361
SCAFFOLD [44]	0.809	0.664	0.838	0.482
FedNova [45]	0.790	0.709	0.891	0.246
pFedMe [14]†	Excluded – frozen in 14/15 LOSO folds (Section 4.2)			
Per-FedAvg [15]	0.514	0.493	0.838	0.127
Ditto [46]	0.828	0.777	0.970	0.454
AGP-FED (Proposed)	0.812	0.729	0.950	0.596

Table 4. LOSO performance comparison on PPG-DaLiA heart-rate estimation. Reported values are means across $N = 15$ LOSO folds (with std where available); "Fail" counts folds with MAE ≥ 6.5 BPM. † Per-FedAvg failed to train in every LOSO fold: predictions were non-finite in 6 of the 15 folds, and in the remaining 9 the model collapsed to a single constant output regardless of input, in several cases reaching physiologically impossible magnitudes. The reported MAE ($\sim 2.7 \times 10^{10}$ BPM) is the mean over the 9 finite folds, dominated by the most extreme cases. It is excluded from quantitative comparison. "–" indicates that the value was not recorded in the experimental logs.

Method	MAE (BPM)	RMSE (BPM)	r	Fail (#)
CNN-BiLSTM (Centralised)	7.082	11.767	0.861	–
FedAvg [5]	7.662	12.011	0.767	–
FedNova [45]	14.168	17.467	0.718	–
SCAFFOLD [44]	8.842	13.329	0.747	–
Per-FedAvg† [15]	diverged (MAE$\sim 2.7 \times 10^{10}$ BPM)			
AGP-FED v4 (Proposed)	11.676 $\pm$ 2.787	15.620	0.525	15

4.2. Federated Learning Convergence Analysis

Figure 4 presents the round-by-round accuracy and macro-F1 trajectories of all federated baselines on the WESAD task over 100 communication rounds. Final LOSO results for each method are reported in Table 3. Both metrics are shown with $\pm$ 1 std shading across LOSO folds, making fold-to-fold variance visible. Round-level diagnostics show that pFedMe's personalised parameters did not move materially away from the global model during local training, and the aggregated global model showed no meaningful parameter change across communication rounds. A separate audit of the per-round accuracy history confirms this: it is frozen at a single value for the entire 100-round run in 14 of the 15 LOSO folds (and collapses to one of two values in the remaining fold), with the model predicting only one of the three classes throughout in those 14 folds. This matches the invalid-run flag already raised by our own Phase 6 analysis script, which identifies this configuration's mean accuracy (0.302) as falling below the chance-level threshold for this 3-class task. It is therefore excluded from quantitative comparison (Table 3) and is shown here only to illustrate this non-training behaviour, as discussed in Section 5. These results suggest that personalisation schemes validated primarily on synthetic label-imbalance benchmarks may not transfer directly to real wearable physiological data.

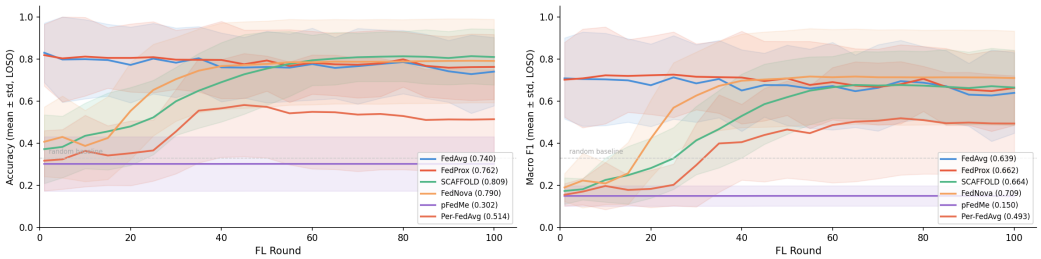


Figure 4. Baseline FL convergence on WESAD (100 rounds, 15-fold LOSO). Mean accuracy and macro-F1 ($\pm$ 1 std). pFedMe's flat trajectory reflects a non-converging run under this configuration (excluded from Table 3; see Section 5). Final LOSO results are reported in Table 3.

Figure 5 shows per-fold MAE convergence on the PPG-DaLiA task over 100 rounds for the first six LOSO folds. FedAvg and SCAFFOLD reduce MAE within the first 20–30 rounds in most folds, whereas FedNova remains consistently higher throughout training. Difficulty varies markedly across held-out subjects: all methods show substantially higher MAE on S5 and S6. Over all 15 folds, FedAvg achieves the lowest LOSO-mean MAE (7.662 BPM), followed by SCAFFOLD (8.842 BPM) and FedNova (14.168 BPM) (Table 4). Per-FedAvg failed to train in every LOSO fold: predictions were non-finite in 6 of the 15 folds, and in the remaining 9 the model collapsed to a single constant prediction regardless of input, in several cases reaching physiologically impossible magnitudes (up to $\pm 2 \times 10^{11}$ BPM). It is therefore excluded from quantitative comparison.

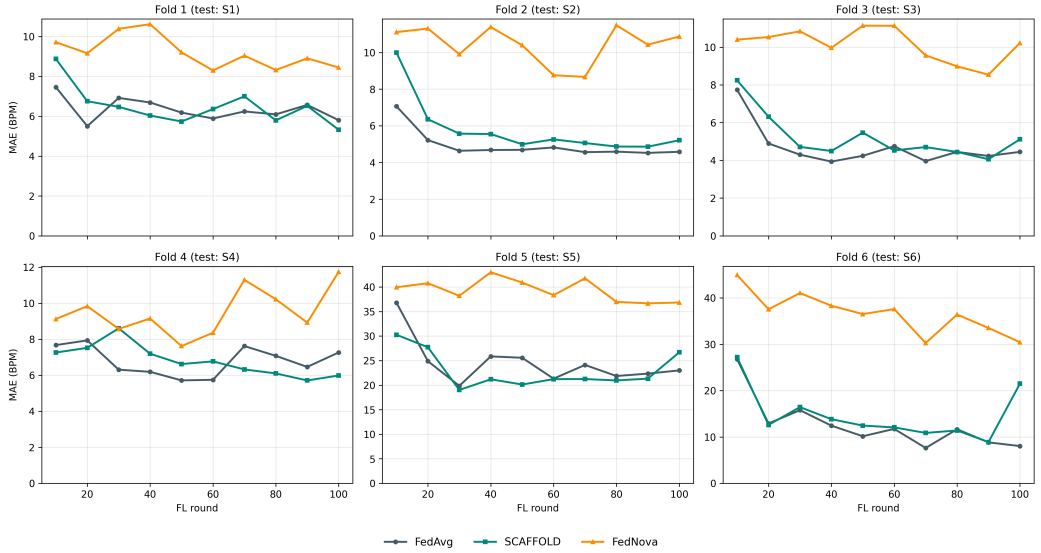


Figure 5. Baseline FL convergence on PPG-DaLiA heart-rate estimation (first 6 LOSO folds shown). MAE (BPM) vs. round. Per-FedAvg is omitted because it failed to train in every LOSO fold: predictions were NaN in 6 of the 15 folds, and in the remaining 9 the model collapsed to a single constant prediction for every window regardless of input, in several folds reaching physiologically impossible magnitudes (up to $\pm 2 \times 10^{11}$ BPM). LOSO-mean MAE values are summarised in Table 4 (FedAvg = 7.662; SCAFFOLD = 8.842; FedNova = 14.168 BPM).

4.3. Proposed AGP-FED Performance

Figure 6 compares *AGP-FED* with the federated baselines and the centralised reference model on WESAD stress classification across LOSO evaluation metrics. *AGP-FED* achieves 0.812 accuracy, exceeding FedAvg (0.740) and most conventional baselines. Because the centralised reference model was trained under a different recipe (different architecture, no Mixup or LoRA, and checkpoint selection on the held-out subject), the centralised-vs-federated gap is treated as descriptive context rather than evidence attributable to

federated personalisation alone; correspondingly, claims are restricted to comparisons among methods trained under the same federated recipe and evaluation protocol.

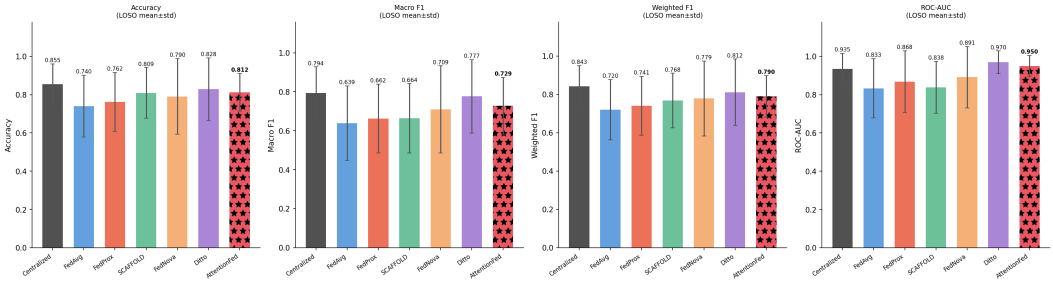


Figure 6. WESAD stress classification summary (15-fold LOSO). Accuracy, macro-F1, weighted F1, and AUC-ROC (mean ± std). Final *AGP-FED* values: acc = 0.812, AUC = 0.950 (Table 3). Bold values indicate the best result for each metric.

Figure 7 compares per-subject MAE between *AGP-FED* v4 and FedAvg on PPG-DaLiA across the 15 LOSO folds. FedAvg achieves lower MAE than *AGP-FED* v4 in all 15 folds (Figure 7). Fold-level aggregates and the number of folds above the 6.5 BPM threshold are reported in Table 4.

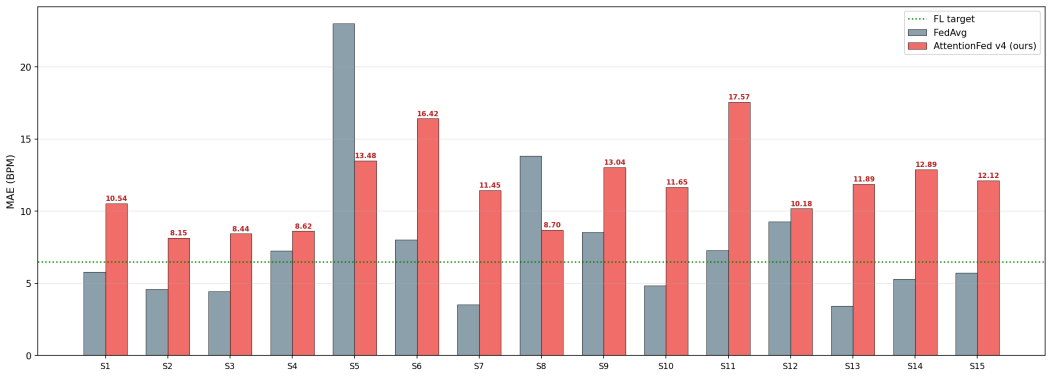


Figure 7. PPG-DaLiA per-fold MAE under LOSO. FedAvg vs. *AGP-FED* v4; dashed line: 6.5 BPM threshold. LOSO means: FedAvg = 7.662 BPM; *AGP-FED* v4 = 11.676 ± 2.787 BPM; *AGP-FED* fails 15/15 folds (Table 4).

4.4. Fairness and Non-IID Analysis

Figure 8 presents the worst-client accuracy across all algorithms on the WESAD task, providing a federated fairness diagnostic. *AGP-FED* achieves worst-client accuracy of 0.596, exceeding FedNova (0.246) and FedAvg (0.341). Ditto achieves a lower worst-client accuracy of 0.454 despite its higher mean accuracy of 0.828. The ablation in Table 5 shows that replacing quality-weighted aggregation with equal-weight averaging (A2) can improve worst-client accuracy relative to the full model (0.642 vs. 0.596). To check whether this accuracy difference reflects a genuine effect of the aggregation scheme or fold-to-fold variance, Full *AGP-FED* is compared against A2 across the three ablation folds (S2, S9, S13): the mean difference is small (-0.06, 95% CI $[-0.50, +0.37]$; paired t-test $p = 0.60$; Wilcoxon $p = 0.75$). This indicates that aggregation choice has a modest, fold-dependent effect on mean accuracy rather than a uniform one, and a larger-fold ablation would sharpen this estimate (noted as future work).

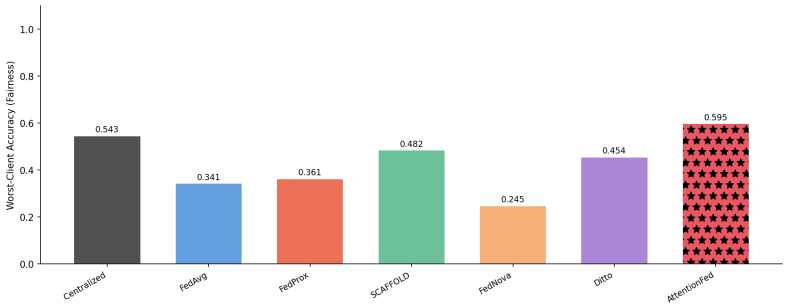


Figure 8. WESAD worst-client accuracy under LOSO: minimum per-fold accuracy across 15 folds. *AGP-FED* = 0.596; FedAvg = 0.341; FedNova = 0.246; Ditto = 0.454 (Table 3).

Table 5. Ablation and comparison study of the proposed *AGP-FED* framework on the WESAD stress classification benchmark under LOSO cross-validation ($N = 15$ folds). Accuracy is reported as the mean $\pm$ std across folds. Comm. denotes total upload per round across all participating clients in megabytes. Communication reduction is computed relative to FedAvg (26.25 MB/r). A1/A2 ablation rows are evaluated on three representative LOSO folds (S2, S9, S13); all other rows use the full 15-fold protocol.

Method	Configuration	Accuracy (Mean $\pm$ std)	Macro F1	Worst-Client Accuracy	Comm. (MB/r)	Reduction vs. FedAvg
CNN+BiLSTM (Centralised)	Single centralised model; no FL; no privacy constraint	0.8549 ± 0.1055	0.7944	0.5432	∞	N/A
FedAvg [5]	Uniform equal-weight aggregation; no personalisation; full-model upload	0.7399 ± 0.1622	0.6389	0.3409	26.25	baseline
Ditto [46]	Per-client model copy; personalised local fine-tuning; full-model upload	0.8282 ± 0.1642	0.7767	0.4537	10.88	−58.6%
A1: No personal head	Attention aggregation retained; no private task head; full backbone exposed	0.8362 ± 0.1166	0.7399	0.8333	10.57	−59.8%
A2: No attention aggregation	Private head retained; equal-weight aggregation replaces attention weighting	0.8534 ± 0.0561	0.8405	0.6415	10.57	−59.8%
Full *AGP-FED* (Proposed)	Attention aggregation + private head; backbone-only upload; quality weighting	0.8124 ± 0.0986	0.7286	0.5955	10.57	−59.8%

Figure 9 summarises dual non-IID effects in PPG-DaLiA, capturing both heterogeneity in heart-rate distributions and activity covariate shift across the 15 federated clients. Across subjects, the mean pairwise HR EMD is 17.415 BPM and the mean activity KL divergence is 0.872 nats (computed from the reported client-level distributions; median 0.014 nats, as the mean is dominated by pairs involving S6). Subject S5 exhibits the highest combined non-IID severity (normalised EMD + KL = 1.93), driven by the largest mean pairwise HR EMD (38.9 BPM) together with an above-average activity KL divergence (0.916 nats). Subject S6 has the second-largest HR EMD (31.3 BPM) but the lowest outgoing activity KL (0.572 nats); its activity shift is one-directional, because its shorter recording contains no lunch or walking windows, so the divergence of other subjects toward S6 averages 12.3 nats. These results indicate simultaneous inter-subject HR variability and differential

activity participation, constituting a two-dimensional non-IID structure stronger than the synthetic Dirichlet partitions considered [6]. This dual non-IID structure contributes to the higher and more variable MAE observed across baselines on S5 and S6, and helps explain why *AGP-FED* v4 does not achieve deployable MAE on this dataset despite early-round improvements; this is discussed as a limitation in Section 5.

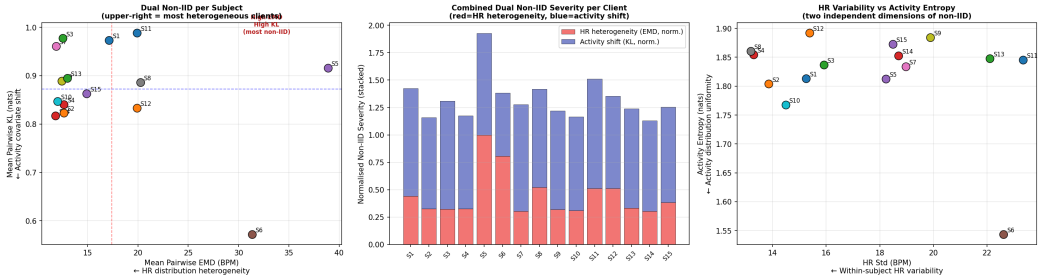


Figure 9. PPG-DaLiA dual non-IID analysis. HR-distribution heterogeneity vs. activity covariate shift across 15 clients, illustrating the two-dimensional non-IID structure driving unstable FL convergence.

4.5. Explainability Analysis

Figure 10 presents the federated SHAP sensor-modality importance map for WESAD. The interpretation matches the plotted attributions and the sensor set used in training (chest modalities only: ECG/HRV, EDA, respiration, ACC, EMG, and temperature from RespiBAN). Temperature shows the highest global mean |SHAP| (0.00644), followed by ACC (0.00438) and respiration (0.00353). Rankings are reported as shown in the figure; over-interpretation of any single modality as dominant is avoided because attributions vary across folds and subjects.

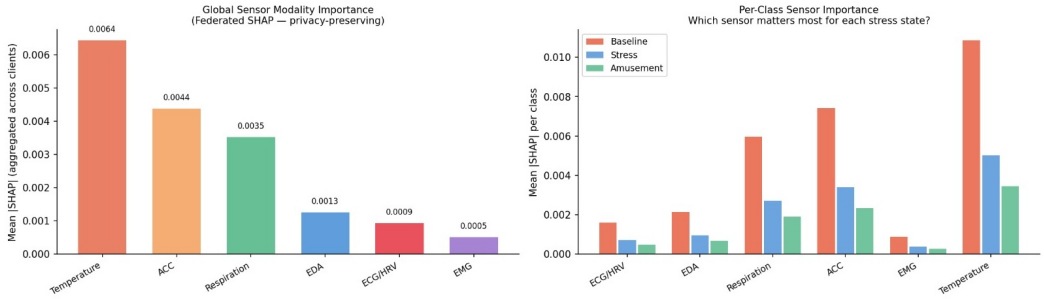


Figure 10. Federated SHAP modality importance on WESAD. Global mean |SHAP| (**left**) and per-class breakdown (**right**) using chest RespiBAN modalities; values are derived summaries (not a formal privacy guarantee).

Federated SHAP on PPG-DaLiA indicates that ACC and PPG dominate heart-rate estimates, contributing 40.2% and 39.9% of the total mean |SHAP|, respectively, followed by EDA (11.9%) and temperature (7.9%).

4.6. Communication Efficiency

Figure 11 presents the accuracy–communication trade-off on WESAD. Communication is reported consistently as *total upload per round across all participating clients*, as measured in the experimental codebase. Under this accounting, *AGP-FED* uploads 10.57 MB/round (backbone only, summed over all participating clients), while FedAvg/FedProx upload 26.25 MB/round (full model), a 59.8% per-round reduction. The reported MB figures follow the accounting used in the baseline codepath (serialised FP32 tensors plus protocol overhead). Over $T = 100$ rounds, this corresponds to 1.057 GB vs. 2.625 GB total upload across all participating clients. Error bars show ± 1 std of per-fold accuracy across LOSO folds.

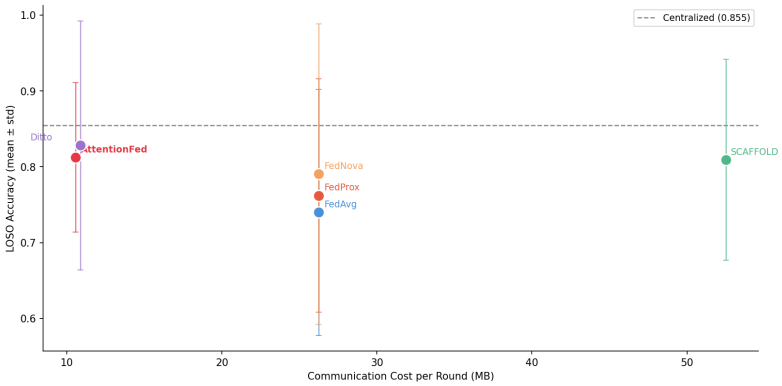


Figure 11. WESAD accuracy vs. communication (15-fold LOSO). Upload is total per round across participating clients. *AGP-FED*: 0.812 accuracy at 10.57 MB/round vs. FedAvg/FedProx 26.25 MB/round (Table 6).

Figure 12 reports the effect of full-model quantisation on model size for the proposed *AGP-FED* v4 client model (471,521 parameters), and the corresponding CPU inference latency. Model size scales directly with numerical precision: FP32 (1.886 MB) → FP16 (0.943 MB, 2×) → INT8 (0.472 MB, 4×) → INT4 (0.236 MB, 8×). Inference latency for the full v4 model was measured on a desktop-class CPU (Intel Core i9-14900HX, 24 threads, Windows 11, PyTorch 2.5.1) as a proxy for edge deployment, using single-sample (batch size 1) forward passes (20 warm-up, 200 timed runs). P95 latency was 3.61 ms at FP32 and 5.66 ms at INT8 (dynamic quantisation of all 17 linear layers; the CNN encoder, LayerNorm, and LoRA/gating parameters remain FP32); both are well below the 50 ms target adopted in this work for IoT deployment and well within the 2000 ms budget set by the 2-s window stride (64 samples at 32 Hz). Notably, INT8 is consistently slower than FP32 at this batch size (by a factor of approximately 1.5–1.6× across repeated runs): dynamic quantisation converts activations on every call, and this per-call overhead outweighs the cheaper INT8 arithmetic at batch size 1. INT8 quantisation in this work therefore reduces communication cost, not inference latency; a genuine edge/embedded processor with native low-precision support may behave differently. I do not report a partial (Transformer-only) quantisation configuration or an accuracy (MAE) comparison across quantisation levels: on internal audit, the previously reported Transformer-only communication figures ("3.2×" and "3.46×") were computed on a smaller, never-trained proxy model (approx. 364k parameters) rather than the 471,521-parameter v4 model evaluated elsewhere in this paper, and the quantised-model MAE values were the full-precision v4 MAE plus fixed offsets added by a simulation routine. A true per-fold accuracy evaluation of quantised v4 weights, and a parameter-level Transformer-only breakdown, are left as future work.

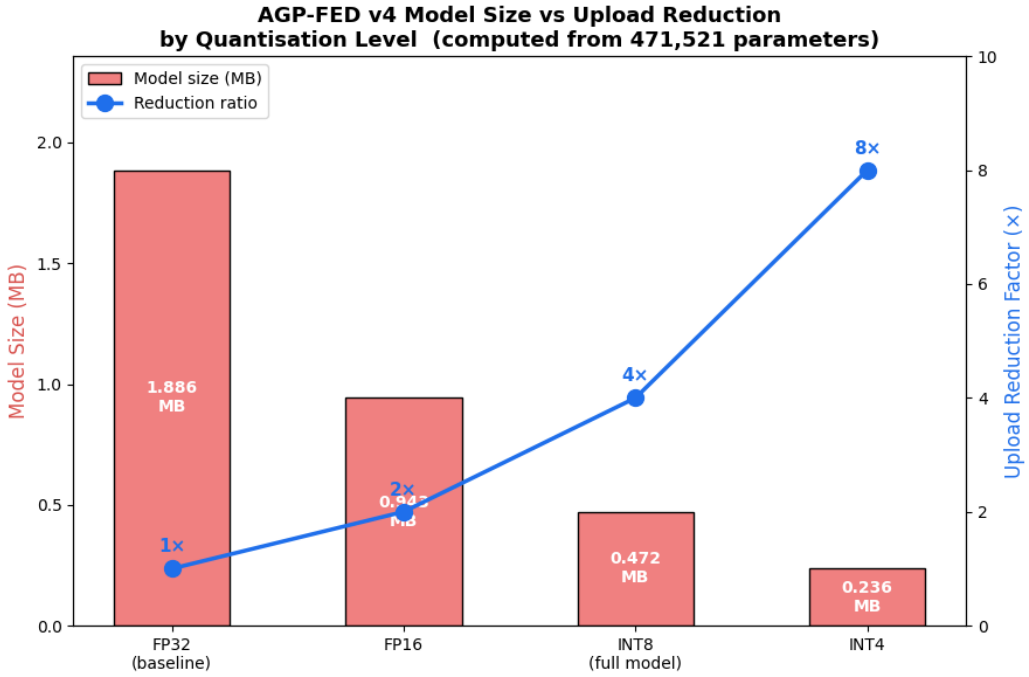


Figure 12. *AGP-FED* v4 model-size reduction by full-model quantisation level (FP32/FP16/INT8/INT4), computed from the 471,521-parameter v4 model. MAE under quantisation and partial (Transformer-only) quantisation are not reported (see text); inference latency (3.61 ms FP32 / 5.66 ms INT8, measured on the real v4 model on a desktop-class CPU) is reported in the text, not plotted.

4.7. Ablation Study

Per-fold convergence on PPG-DaLiA varies substantially: most folds converge within rounds 10–20 (S1, S2, S3, S4, S5, S6, S7, S8, S12, S14, S15), while S9 and S13 run to the maximum 64 rounds. This per-fold variability underscores the importance of the per-fold early-stopping criterion, which prevents overfitting on small local validation splits while allowing each test subject to benefit from the appropriate number of global aggregation rounds.

Table 5 evaluates each component's contribution under LOSO cross-validation. Removing the private task head (A1) yields a mean accuracy of 0.836 (vs. 0.812 Full), while worst-client accuracy improves to 0.833 (vs. 0.596 Full). Removing quality-weighted aggregation (A2) yields a slightly higher mean accuracy (0.853) and improved worst-client accuracy (0.642 vs. 0.596). Because A1 and A2 are evaluated on three representative folds whereas the full model uses all 15 folds, these differences are indicative rather than conclusive (see the paired comparison in Section 4.4). These results suggest that the private head supports subject-level adaptation, whereas the current quality-weighted aggregation does not consistently improve mean accuracy on WESAD. This motivates reporting aggregation effects conditionally (e.g., heterogeneity-stratified LOSO folds and, for PPG-DaLiA, activity-stratified errors) rather than as a uniform average-gain claim. The aggregation module is retained because it is motivated by signal-quality shifts (e.g., motion artefacts on PPG-DaLiA) and is evaluated transparently via ablations.

4.8. Quantitative Results

Table 3 summarises the LOSO-averaged performance of all evaluated methods on the WESAD stress classification task. *AGP-FED* achieves 0.812 accuracy and 0.950 AUC-ROC, surpassing FedAvg, FedProx, SCAFFOLD, FedNova, and Per-FedAvg among conventional federated baselines, and approaching Ditto (0.828 accuracy, 0.970 AUC-ROC). SCAFFOLD

is the strongest conventional baseline by accuracy (0.809), whereas FedNova attains the highest conventional macro-F1 (0.709). AGP-FED offers a favourable trade-off: comparable accuracy to SCAFFOLD and FedNova with substantially reduced communication overhead, personalised private heads, and federated explainability, features absent from all conventional baselines. pFedMe is excluded from this comparison because it failed to make training progress under this configuration (Section 4.2 and Section 5); its raw per-round accuracy history averages to 0.302, below chance level for this 3-class task, consistent with the frozen, single-class predictions documented there.

Table 4 summarises heart-rate estimation on PPG-DaLiA. To make the deployment claim auditable, this paper reports the LOSO-mean MAE, RMSE, and Pearson r, together with the number of folds exceeding the 6.5 BPM threshold.

Table 6 compares communication cost over $T = 100$ rounds on WESAD using a single accounting convention: *total upload per round across all participating clients*. Under this convention, *AGP-FED* reduces total upload from 26.25 MB/round (FedAvg/FedProx) to 10.57 MB/round, a 59.8% per-round reduction. SCAFFOLD incurs higher upload (52.51 MB/round) due to its control-variate transmission. Ditto uploads 10.88 MB/round, similar to *AGP-FED* but without quality-weighted aggregation or federated explainability. Table 7 lists the serialised parameter counts logged for each component. The full-model count differs from the 471,521-parameter v4 client model described in Section 3, so these per-client sizes are indicative only; the measured per-round totals in Table 6 are the reference values for all communication comparisons.

Table 6. Communication cost comparison over $T = 100$ FL rounds (WESAD), reported as *total upload per round across all participating clients*. ‡ The INT8 Acc/MB figure assumes WESAD classification accuracy is unchanged under INT8 quantisation; this was not separately measured (see Figure 12 and Section 4) and should be read as an upper-bound estimate, not a verified value.

Method	Upload/Round (Total)	Reduction	Acc/MB
FedAvg [5]	26.25 MB	baseline	0.028
FedProx [7]	26.25 MB	0.0%	0.029
SCAFFOLD [44]	52.51 MB	+100%	0.015
Ditto [46]	10.88 MB	−58.6%	0.076
AGP-FED (FP32)	10.57 MB	−59.8%	0.077
AGP-FED (INT8) ‡	**2.64 MB**	**−89.9%**	**0.308**

Table 7. Parameter and communication breakdown per participating client per round. "Size" counts serialised parameters only. Total-round upload in Table 6 is measured empirically (including serialisation and protocol overhead); per-client parameter sizes are therefore indicative and are not intended to reproduce the measured totals.

Component	Params (#)	Precision	Per-Client Upload (MB)	Notes
Backbone (v4)	438,689	FP32	1.75	transmitted each round (AGP-FED, per client)
Head	32,832	FP32	0	kept local (AGP-FED)
Full model	712,353	FP32	2.85	transmitted each round (FedAvg/FedProx, per client)

4.9. Comparison with State-of-the-Art

Table 8 benchmarks *AGP-FED* against representative state-of-the-art methods on the two tasks. To avoid incomplete or protocol-mismatched reporting, the entries are marked explicitly when performance is not reported under the same (3-class, LOSO) evaluation protocol, rather than leaving cells blank.

Table 8. Comparison with state-of-the-art physiological sensing methods. FL = federated learning; XAI = explainability support. ‡ Evaluated on a different dataset and protocol; not directly comparable with the PPG-DaLiA LOSO results reported here.

Method	Task	FL	XAI	Performance (Protocol Notes)
Gil-Martín et al. [22]	Stress	×	×	Centralised protocol; not directly comparable with the federated 3-class LOSO setting used here.
Tanwar et al. [25]	Stress	×	×	Centralised protocol; 3-class LOSO number not reported in the source in a directly comparable form.
Singh et al. [26]	Stress	×	×	Centralised protocol; 3-class LOSO number not reported in a directly comparable form.
Moser et al. [19]	Stress	×	✓	Centralised protocol; comparable LOSO metric not reported.
Almadhor et al. [40]	Stress	✓	×	FL setting; protocol differs from the 3-class LOSO used here (number not protocol-matched).
AGP-FED (Proposed)	Stress	✓	✓	Acc 0.812 (3-class LOSO; this work)
Fan [50]	HR est.	×	×	MAE 5.2 BPM ‡ (large-scale running-workout dataset, >250,000 records; not PPG-DaLiA LOSO)
Reiss et al. [17]	HR est.	×	×	MAE 7.65 BPM (centralised CNN ensemble on PPG-DaLiA; not federated)
AGP-FED v4 (Proposed)	HR est.	✓	✓	Mean MAE 11.68 ± 2.79 BPM (15-fold LOSO; Table 4)

5. Discussion

AGP-FED achieved 0.812 accuracy and 0.950 AUC-ROC on WESAD stress classification and reduced total per-round communication by 59.8% (10.57 MB vs. 26.25 MB for FedAvg, total across participating clients) via backbone-only upload. For PPG-DaLiA, per-fold HR metrics are reported in Table 4; *AGP-FED* v4 achieves a mean MAE of 11.68 BPM, which does not surpass FedAvg (7.66 BPM) under the LOSO protocol. This is attributed to the extreme dual non-IID structure of PPG-DaLiA (simultaneous inter-subject HR variability and activity covariate shift), which causes the shared backbone to diverge past its best checkpoint after rounds 10–30, and is reported as a primary limitation of the current framework on noisy continuous-output regression tasks. Across tasks, the main contributions are (i) local private-head adaptation, (ii) communication reduction from backbone-only transfer, and (iii) federated explainability via SHAP ; an additional attention-based interpretability analysis was conducted but found the model's cross-modal attention to be near-uniform (Section 5) and is therefore not listed as a contribution; quality-weighted aggregation is evaluated via ablations and is sensitive to the score definition and task setting.

Quality-aware aggregation as a heterogeneity-conditioned mechanism (classification vs. regression): The observed behaviour is consistent with the expected operating regime of non-uniform aggregation once the analysis is conditioned on the heterogeneity regime of each dataset. On WESAD (LOSO classification), the non-IID analysis indicates a moderate inter-subject shift: in this regime, simple equal-weight aggregation already approximates a good consensus update, while attention-derived reweighting can introduce extra variance without a sufficiently strong signal to justify it. In addition, under LOSO, the aggregation weights are estimated entirely from training clients and therefore cannot encode any notion of predictive quality *with respect to the held-out test subject*; consequently, the equal-weight ablation (A2) becomes a competitive upper bound for the aggregation component in this moderate-heterogeneity setting. On PPG-DaLiA (LOSO heart-rate regression), the regression objective has a smoother loss surface, and FedAvg-style equal averaging provides a strong implicit ensembling inductive bias that tends to generalise well across subjects. In contrast, attention-based weighting can bias the shared backbone toward high-

weight clients, which may reduce generalisation to unseen subjects when the weighting proxy is not calibrated on a representative meta-validation distribution.

Boundary-condition contribution: Accordingly, the role of the quality-aware aggregation module is reframed as follows: quality-aware aggregation is expected to improve performance primarily in high-heterogeneity regimes where uniform averaging would over-represent atypical/noisy clients, whereas in moderate-heterogeneity regimes the dominant generalisation gains come from the personalised components (private head / task-specific adaptation and, for regression, the LoRA-enhanced backbone variants), with quality-aware aggregation acting as a complementary mechanism.

Additional "no-new-training" ablations extracted from existing artefacts: Although no additional training runs were performed, the current outputs already enable two informative, protocol-matched analyses: (A3) *heterogeneity-stratified WESAD performance*: rank LOSO folds by the held-out subject's mean pairwise divergence to training subjects (using the already-computed divergence matrix), and report Full vs. A2 accuracy on the most-heterogeneous vs. least-heterogeneous fold subsets; (A4) *activity-stratified PPG-DaLiA error*: report FedAvg vs. *AGP-FED* MAE/RMSE broken down by activity label (low-motion vs. high-motion activities), using the dataset metadata already loaded for the non-IID analysis/visualisation. These analyses directly test the heterogeneity-conditioning hypothesis using results already produced by the pipeline.

Privacy and leakage considerations: Although *AGP-FED* does not transmit raw physiological signals, attention summaries and SHAP values are derived from private data and may encode subject-specific information. The method is therefore positioned as reducing raw-data exposure rather than providing formal privacy guarantees, and a threat model and mitigation options (e.g., secure aggregation, clipping/noising summaries) are outlined as future work.

The ablation study separates the effects of personalisation and aggregation. On WESAD, equal-weight aggregation (A2) slightly increases mean accuracy (0.853 vs. 0.812) and improves worst-client accuracy (0.642 vs. 0.596) and macro-F1 (0.841 vs. 0.729); because A2 is evaluated on three folds and the full model on 15, and the paired three-fold comparison is not significant ($p = 0.60$), this indicates that the benefit of quality-weighted aggregation on WESAD is not established and depends on the aggregation rule and the quality-score definition. Removing the private task head (A1) yields a mean accuracy of 0.836 and worst-client accuracy of 0.833, suggesting the private head introduces some variance in exchange for subject-level adaptation. For PPG-DaLiA, the v4 backbone with LoRA-32 adapters, ResidualCNN skip connections, subject embedding, and channel gate converges to a local minimum in rounds 10–30 before validation MAE diverges; per-fold early stopping captures the best checkpoint, but mean MAE remains at 11.68 BPM. Addressing this limitation likely requires architectural changes beyond attention weighting, such as local fine-tuning with more data or Ditto-style proximal regularisation. pFedMe made no meaningful training progress on WESAD under this configuration: round-level diagnostics show its personalised parameters never moved materially away from the global model, and its per-round accuracy history is frozen at a single value for the entire 100-round run in 14 of the 15 LOSO folds (collapsing to one of two values in the remaining fold), with the model predicting only one of the three classes throughout in those 14 folds. It is therefore excluded from quantitative comparison (Table 3). Per-FedAvg exhibited catastrophic divergence on PPG-DaLiA. The Spearman correlation of $\rho \approx -0.50$ between per-subject KL divergence and FedAvg accuracy formalises the impact of biological non-IID. On WESAD specifically, heterogeneity is signal-space rather than label-space: label distributions are near-identical by protocol design, while per-client feature cosine distances yield a mean of 0.3528, confirming physiological signal heterogeneity as the dominant non-IID axis.

Several limitations remain. *AGP-FED* v4 does not achieve deployable MAE on PPG-DaLiA, and the dataset's dual non-IID structure may require architecture changes beyond attention weighting. With only 15 subjects per benchmark, generalisation across demographics and device types needs validation on larger, multi-centre cohorts. On-device federated training has not been profiled under strict battery or memory budgets, so the FP32 P95 latency of 3.61 ms (INT8 P95: 5.66 ms) reported in Figure 12 should not be taken as a proxy for training cost; note also that INT8 quantisation increased inference latency relative to FP32 in this measurement, so it should be understood as a communication-cost optimisation rather than a latency optimisation. Future directions include larger temporal backbones for richer temporal modelling, formal differential privacy guarantees [28] layered on top of the existing task-head isolation and adaptive round scheduling driven by client battery state. Separately, an audit of the trained model's cross-modal attention weights (Section 4) found them to be near-uniform across modalities rather than selectively weighted, suggesting that channel-level interpretability claims should not be drawn from this architecture's attention layer in its current form; improving this is left as future work.

6. Conclusions and Future Work

This paper presented *AGP-FED*, a personalised federated learning framework for multimodal physiological signal analysis that targets privacy, biological non-IID heterogeneity, communication overhead, and interpretability in a single architecture. A shared Residual-CNN+Transformer backbone learns temporal representations across subjects; a cross-modal attention module is designed to reweight sensor modalities and yields a client-side quality summary, although a post-hoc audit of the trained model's attention weights found them to be near-uniform across modalities rather than selectively weighted (Section 4); a private task head retains subject-specific prediction parameters locally. Only backbone weights and a lightweight summary travel to the server each round, reducing total per-round upload by 59.8% relative to full-model FedAvg on WESAD (10.57 MB vs. 26.25 MB across participating clients). Under LOSO evaluation, *AGP-FED* achieves 0.812 accuracy and 0.950 AUC-ROC on WESAD; ablations show that quality-aware aggregation does not consistently improve on equal-weight averaging in this moderate-heterogeneity setting. On PPG-DaLiA, *AGP-FED* v4 reaches 11.68 ± 2.79 BPM mean MAE (Table 4), which is higher than that of FedAvg (7.66 BPM) and does not meet the 6.5 BPM deployment threshold; the dataset's dual non-IID structure (inter-subject HR variability and activity covariate shift) is the likely primary limitation for continuous regression. blackCentralised results are reported only for context (different recipe, and an optimistic reference because checkpoints were selected on the held-out subject) and are not used to claim a federated personalisation advantage. Finally, "strict LOSO" claims are avoided when a labelled calibration split from the held-out subject is used for private-head adaptation.

Future work: (i) benchmark against additional personalised FL baselines (e.g., FedPer/FedRep/FedBN) under matched tuning budgets; (ii) evaluate robust aggregation under stronger client-quality shifts and partial participation; (iii) add formal privacy mechanisms (secure aggregation and/or calibrated noise for transmitted summaries) and quantify the privacy–utility trade-off; (iv) validate on larger cohorts and additional wearable devices to test demographic and device generalisation; and (v) investigate Ditto-style proximal regularisation or local fine-tuning strategies to address convergence instability on continuous HR regression under strong non-IID.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The WESAD and PPG-DaLiA datasets analysed in this study are publicly available from their original authors [13,17].

Acknowledgments: During the preparation of this manuscript, the author used Claude (Claude Code and Claude, Anthropic) to cross-check the numerical results, figures, and tables reported in this manuscript against the underlying analysis code, scripts, and saved experimental outputs, and to help identify and correct inconsistencies between the manuscript and these artefacts after the revision. The author has reviewed and verified all resulting changes and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest regarding the publication of this manuscript.

References

1. Chen, X.; Liang, D.; Zhang, J.; Zhang, Y. Deploying emotion recognition on resource-constrained device: Utilising lightweight deep learning models with multi-physiological signals. In *Proceedings of the 3rd International Conference on Signal Processing, Computer Networks and Communications (SPCNC '24)*; New York, NY, USA, 22 - 24 December 2024; pp. 519–524.
2. Ghadi, Y.Y.; Shah, S.F.A.; Waheed, W.; Mazhar, T.; Ahmad, W.; Saeed, M.M.; Hamam, H. Integration of wearable technology and artificial intelligence in digital health for remote patient care. *J. Cloud Comput.* **2025**, *14*, 39.
3. Bolpagni, M.; Pardini, S.; Dianti, M.; Gabrielli, S. Personalised stress detection using biosignals from wearables: A scoping review. *Sensors* **2024**, *24*, 3221.
4. Marrah, S.A.; Wang, J.; Koroma, A.B.; Kamara, G.D.; Babatunde, O.S.; Marrah, M.; Stamber, R.T.; Saidu, S.E. Federated deep learning for privacy-preserving disease detection in IoT-enabled healthcare systems. *Front. Comput. Sci.* **2026**, *8*, 1725597.
5. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralised data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*; Fort Lauderdale, FL, USA, 20-22 April 2017; pp. 1273–1282.
6. Jimenez G., D.M.; Solans, D.; Heikkilä, M.; Vitaletti, A.; Kourtellis, N.; Anagnostopoulos, A.; Chatzigiannakis, I. Non-IID data in federated learning: A survey with taxonomy, metrics, methods, frameworks and future directions. *arXiv* **2024**, arXiv:2411.12377.
7. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimisation in heterogeneous networks. *Proc. Mach. Learn. Syst.* **2020**, *2*, 429–450.
8. Tan, A.Z.; Yu, H.; Cui, L.; Yang, Q. Towards personalised federated learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 9587–9603.
9. Johnson, F.T.; Oladapo, E.; Olukumoro, O.; Victor,A.; Akande, A. Communication-efficient federated learning: A systematic review of model compression and aggregation techniques. *Artif. Intell. Appl.* **2025**.
10. Chen, Y.; Qin, X.; Wang, J.; Yu, C.; Gao, W. FedHealth: A federated transfer learning framework for wearable healthcare. *IEEE Intell. Syst.* **2020**, *35*, 83–93.
11. Ali, M.; Naeem, F.; Tariq, M.; Kaddoum, G. Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 778–789.
12. Sattler, F.; Wiedemann, S.; Müller, K.-R.; Samek, W. Robust and communication-efficient federated learning from non-IID data. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 3400–3413.
13. Schmidt, P.; Reiss, A.; Dürichen, R.; Marberger, C.; Van Laerhoven, K. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*; ACM: Boulder, CO, USA, 2018; pp. 400–408.
14. Dinh, C.T.; Tran, N.H.; Nguyen, T.D. Personalised federated learning with Moreau envelopes. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Virtual, 6–12 December 2020; Volume 33, pp. 21394–21405.
15. Fallah, A.; Mokhtari, A.; Ozdaglar, A. Personalised federated learning: A meta-learning approach. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Virtual, 6–12 December 2020.
16. Kundroo, M.; Singh, T.; Kim, T. FedLBW: A loss-based weighting strategy for federated learning on non-IID data in wireless networks. *Expert Syst. Appl.* **2025**, 130487.
17. Reiss, A.; Indlekofer, I.; Schmidt, P.; Van Laerhoven, K. Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors* **2019**, *19*, 3079.
18. Jaber, D.; Hajj, H.; Maalouf, F.; El-Hajj, W. Medically oriented design for explainable AI for stress prediction from physiological measurements. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 38.
19. Moser, M.K.; Ehrhart, M.; Resch, B. An explainable deep learning approach for stress detection in wearable sensor measurements. *Sensors* **2024**, *24*, 5085.

20. Li, F.; Zhang, D. Multimodal physiological signals from wearable sensors for affective computing: A systematic review. *Intell. Sport. Health* **2025**, *1*, 210–222.
21. Rauniyar, A.; Hagos, D.; Jha, D.; Håkegård, J.E.; Bagci, U.; Rawat, D.B.; Vlassov, V. Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet Things J.* **2024**, *11*, 7374–7398.
22. Gil-Martín, M.; San-Segundo, R.; Mateos, A.; Ferreiros-López, J. Human stress detection with wearable sensors using convolutional neural networks. *IEEE Aerosp. Electron. Syst. Mag.* **2022**, *37*, 60–70.
23. Singh, G.; Phukan, O.C.; Kumar, R. Stress recognition with multimodal sensing using bootstrapped ensemble deep learning model. *Expert Syst.* **2023**, *40*, e13239.
24. Li, F.; Zhang, D. Transformer-driven affective state recognition from wearable physiological data in everyday contexts. *Sensors* **2025**, *25*, 761.
25. Tanwar, R.; Phukan, O.C.; Singh, G.; Pal, P.K.; Tiwari, S. Attention-based hybrid deep learning model for wearable-based stress recognition. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107391.
26. Singh, G.; Chetia Phukan, O.; Gupta, R.; Nayyar, A. Hybrid deep learning model for wearable sensor-based stress recognition for Internet of Medical Things (IoMT) system. *Int. J. Commun. Syst.* **2024**, *37*, e5657.
27. Thilakarathne, N.N.; Muneeswari, G.; Parthasarathy, V.; Alassery, F.; Hamam, H.; Mahendran, R.K.; Shafiq, M. Federated learning for privacy-preserved medical Internet of Things. *Intell. Autom. Soft Comput.* **2022**, *33*, 157–172.
28. Can, Y.S.; Ersoy, C. Privacy-preserving federated deep learning for wearable IoT-based biomedical monitoring. *ACM Trans. Internet Technol.* **2021**, *21*, 21.
29. Lian, Z.; Yang, Q.; Wang, W.; Zeng, Q.; Alazab, M.; Zhao, H.; Su, C. DEEP-FEL: Decentralised, efficient and privacy-enhanced federated edge learning for healthcare cyber-physical systems. *IEEE Trans. Netw. Sci. Eng.* **2022**, *9*, 3558–3569.
30. Arivazhagan, M.G.; Aggarwal, V.; Singh, A.K.; Choudhary, S. Federated learning with personalisation layers. *arXiv* **2019**, arXiv:1912.00818.
31. Collins, L.; Hassani, H.; Mokhtari, A.; Shakkottai, S. Exploiting shared representations for personalised federated learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*; PMLR: Virtual Event, 18–24 July 2021; pp. 2089–2099.
32. Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; Dou, Q. FedBN: Federated learning on non-IID features via local batch normalisation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 3–7 May 2021.
33. Bhatti, A.; Behinaein, B.; Hungler, P.; Etemad, A. AttX: Attentive cross-connections for fusion of wearable signals in emotion recognition. *ACM Trans. Comput. Healthc.* **2024**, *5*, 1–24.
34. Pallapothula, V.K.N.; Anand, S.; Bhattacharjee, S.D.; Yuan, J. Generalised multisensor wearable signal fusion for emotion recognition from noisy and incomplete data. *Smart Health* **2025**, *36*, 100571.
35. Ji, S.; Pan, S.; Long, G.; Li, X.; Jiang, J.; Huang, Z. Learning private neural language modelling with attentive aggregation. In *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*; Budapest, Hungary, 14–19 July 2019; pp. 1–8.
36. Xu, Y.; Jiang, Z.; Xu, H.; Wang, Z.; Qian, C.; Qiao, C. Federated learning with client selection and gradient compression in heterogeneous edge systems. *IEEE Trans. Mob. Comput.* **2024**, *23*, 5446–5461.
37. Ji, Y.; Chen, L. FedQNN: A computation-communication-efficient federated learning framework for IoT with low-bitwidth neural network quantisation. *IEEE Internet Things J.* **2023**, *10*, 2494–2507.
38. Mao, Y.; Zhao, Z.; Yan, G.; Liu, Y.; Lan, T.; Song, L.; Ding, W. Communication-efficient federated learning with adaptive quantisation. *ACM Trans. Intell. Syst. Technol.* **2022**, *13*, 67.
39. Madathil, N.T.; Dankar, F.K.; Gergely, M.; Belkacem, A.N.; Alrabaee, S. Revolutionising healthcare data analytics with federated learning: A comprehensive survey of applications, systems, and future directions. *Comput. Struct. Biotechnol. J.* **2025**, *28*, 217–238.
40. Almadhor, A.; Sampedro, G.A.; Abisado, M.; Abbas, S.; Kim, Y.-J.; Khan, M.A.; Baili, J.; Cha, J.-H. Wrist-based electrodermal activity monitoring for stress detection using federated learning. *Sensors* **2023**, *23*, 3984.
41. Lu, W.; Wang, J.; Chen, Y.; Qin, X.; Xu, R.; Dimitriadis, D.; Qin, T. Personalised federated learning with adaptive BatchNorm for healthcare. *IEEE Trans. Big Data* **2024**, *10*, 915–925.
42. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.
43. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685.
44. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.U.; Suresh, A.T. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*; PMLR: Virtual Event, 13–18 July 2020; pp. 5132–5143.
45. Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; Poor, H.V. Tackling the objective inconsistency problem in heterogeneous federated optimisation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Virtual, 6–12 December 2020; Volume 33, pp. 7611–7623.

46. Li, T.; Hu, S.; Beirami, A.; Smith, V. Ditto: Fair and robust federated learning through personalisation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*; PMLR: Virtual Event, 18–24 July 2021; pp. 6357–6368.
47. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 4–9 December 2017.
48. Ducange, P.; Marcelloni, F.; Miglionico, G.C.; Renda, A.; Ruffini, F. Federated SHAP: Privacy-preserving and consistent post-hoc explainability in federated learning. *Mach. Learn.* **2026**, *115*, 24.
49. Zhang, H.; Cissé, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond empirical risk minimisation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 30 April–3 May 2018.
50. Fan, Y. Wearable-based personalised exercise heart rate estimation and distribution analysis using dual-context LSTM model. *Internet Technol. Lett.* **2025**, *8*, e627.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.