Privacy-Preserving ECG-HRV Emotion Classification for Mood Swing Detection using Federated Learning on Constrained Edge Devices

Chidike Vincent, Muhammad Aslam Jarwar, Najam Ul Hasan
Laurence Hirsch, Jonathan Zasada-James, Hem Raj Dadral
School of Computing and Digital Technologies, Sheffield Hallam University, Sheffield, United Kingdom
{v.vincent, a.jarwar, n.ul-hasan, l.hirsch j.zasada-james, hemraj.r.dadral}@shu.ac.uk

arXiv:2611.01853v1 [cs.LG] 1 Nov 2026

***Abstract*—This paper presents a privacy-preserving mood monitoring framework that classifies emotional states from ECG-derived heart rate variability (HRV) features. The proposed framework is evaluated on three datasets comprising 74 participants and validated on an unseen individual using a wearable in a pilot study. Leave-One-Participant-Out cross-validation LOPO-CV, which evaluates the model on each participant without exposing their data during training, serves as the primary generalization benchmark. A multi-phase experimental design examines how temporal windowing, HRV feature selection, and class equalisation affects four-class valence-arousal and binary valence classification. Class imbalance is identified as the dominant suppressor of low-valence; and equalisation improves minority class recall by 48.5 % on the DREAMER dataset. Binary valence classification achieves 53.17% LOPO-CV accuracy using standard HRV features. Under identical conditions, Logistic Regression outperforms the other five classifiers and exceeding comparable HRV-based baselines. Pilot validation using a Polar H10 wearable achieves 75% binary valence accuracy on an unseen individual without calibrations, including 100% precision for negative valence predictions. Despite being limited to a single participant, the finding demonstrates the framework's feasibility for validation on a larger cohort. Finally, the proposed framework is deployed on a six-node Raspberry Pi federated learning cluster to evaluate the privacy-preserving distributed training. The results show that federated Logistic Regression achieves accuracy comparable to centralized Logistic Regression across all three datasets, without sharing raw physiological data, thereby preserving user privacy. These findings establish the proposed framework as an interpretable, and privacy preserving solution for continuous mood monitoring in real-world clinical and community settings.**

***Index Terms*—Federated Learning, Emotion Classification, HRV Analysis, ECG Signals, Privacy Preservation**

I. Introduction

Mood disorders in young people often go undetected due to reliance on self-reporting and scheduled assessments. Millions of young people experience symptoms of depression, anxiety, and mood disorders without a diagnosis. Because individuals rarely present to a clinic that fits traditional diagnostic frameworks, the average delay between symptom onset and treatment stretches to eight to twelve years [1]. This gap is not because effective treatments are unavailable, but because tools for detecting mood changes early enough to intervene are still largely absent from everyday life. Depression is the leading global cause of disability, affecting an estimated 322 million people, with particularly high prevalence among adolescents and young adults [2].

The human body does not wait for a clinical appointment to signal distress. The autonomic nervous system responds to emotional states continuously, and those responses are measurable through the heart. Heart rate variability (HRV), defined as the variation in time intervals between successive R-peaks in the electrocardiogram (ECG), reflects the dynamic balance between the sympathetic and parasympathetic nervous systems. Emotions can be described along two dimensions: valence, which captures whether a feeling is positive or negative, and arousal, which captures how activated or calm a person feels [3]. A person feeling joyful scores high on both; someone feeling sad scores low on valence and low on arousal; someone anxious scores high on arousal but low on valence. HRV is an established physiological correlate of both dimensions [4], [5]. Wearable devices such as chest straps and smartwatches are now widely available and capable of recording cardiac electrical activity with sufficient fidelity to extract meaningful physiological information. This research evaluates whether these signals, processed through intelligent privacy-preserving frameworks, can serve as reliable longitudinal markers of an individual's emotional state.

Despite an increased volume of research applying machine learning to ECG-HRV signals, three limitations hinder clinical translation. First, the majority of existing models rely on within-participant (WP) training protocols, which inflate accuracy and do not generalise to unseen individuals [6], [7]. Second, high-accuracy models rely on laboratory equipment or computationally intensive feature extraction methods incompatible with resource-constrained edge devices [8]–[10]. Third, centralising continuous physiological data raises privacy concerns. The proposed system functions as a personal awareness tool rather than a definitive clinical diagnostic instrument. A cross-participant (XP) accuracy of approximately 53% at inception is sufficient to detect mood trends and support personalisation over time, and the federated architecture preserves individual privacy by ensuring personal data remains local while improving performance progressively through continuous on-device learning. This paper addresses

all three limitations through a unified, systematic experimental framework. The main contributions are as follows:

- A multi-phase windowed HRV classification approach is evaluated across three public affective datasets, DREAMER [7], AMIGOS [11], and EmoWear [12], using Leave-One-Participant-Out cross-validation as the primary generalisation benchmark.
- Class imbalance is identified and characterised as the primary suppressor of low-arousal low-valence detection, a finding with direct clinical relevance as this quadrant corresponds most closely to depressive mood states.
- The approach is validated on an unseen participant using a Polar H10 consumer chest strap, demonstrating real-world deployability on hardware accessible to end users without laboratory infrastructure.
- A privacy-preserving federated learning architecture is implemented and deployed on a six-node Raspberry Pi cluster, showing that distributed training on partitioned physiological data consistently match centralised training while keeping all raw signals local to each device.

The remainder of this paper is organised as follows. Section II reviews related work. Section III presents the methodology. Section IV reports experimental results. Section V describes real-world deployment and live validation. Section VI evaluates performance. Section VII discusses findings and limitations. Section VIII concludes the paper.

II. Related Work

A. ECG-HRV Emotion Classification

Heart rate variability analysis for emotion recognition has been explored since the early work of Valenza et al. [5], who demonstrated that time and frequency domain HRV features extracted from short ECG windows could discriminate between positive and negative emotional states using a point-process modelling framework. The DREAMER dataset, introduced by Katsigiannis and Ramzan [7], established a widely used benchmark for multimodal emotion recognition using low-cost, off-the-shelf ECG and EEG devices. Using a Support Vector Machine on within-participant folds, the original DREAMER paper reported 62.49% valence accuracy, the primary ECG-based reference for this dataset. Subsequent work pursued higher accuracy through increasingly complex methods. Nita et al. [8] applied a seven-layer convolutional neural network with synthetic data augmentation directly to raw ECG waveforms from DREAMER, reporting 95.16% within-participant valence accuracy. Wang et al. [9] proposed an HRV Emotion Recognition method combining Emotion Quantification analysis with Amplitude Level Quantisation to extract frequency-band-specific local amplitude patterns, achieving 84.3% within-participant valence accuracy using Logistic Regression on DREAMER. Ferdinando et al. [6] applied bivariate empirical mode decomposition spectral features to ECG from the Mahnob-HCI database, achieving 42.6% valence accuracy using standard HRV features under 10-fold cross-validation and 59.2% under subject-independent validation on a three-class problem.

Despite significant advances, four critical limitations persist in the current literature. First, within-participant evaluation artificially inflates accuracy and fails to reflect real-world performance on unseen individuals. Second, deep learning on raw ECG waveforms requires GPU computation, making it incompatible with the resource constraints of edge computing devices. Third, novel features such as Amplitude Level Quantisation (ALQ) lack cross-participant validation and likely encode individual cardiac identity rather than shared emotional signal. Finally, no prior study has conducted validation on a live participant using a consumer wearable, which is a prerequisite for determining real-world system efficacy. This work addresses all four gaps, surpassing the 42.6% standard HRV baseline reported by Ferdinando et al. [6] under the stricter Leave-One-Participant-Out (LOPO-CV) protocol and providing live consumer-hardware validation.

B. Federated Learning for Physiological Monitoring

Federated learning is a distributed machine learning paradigm in which a collaborative model is trained across multiple devices. Each participant trains a local model on their own private data and shares only the resulting model weight updates with a central aggregation server, which combines them into an improved global model. Raw physiological signals never leave the local device. Federated learning was formally introduced by McMahan et al. [13] through the Federated Averaging (FedAvg) algorithm. ECG-based physiological signals are inherently non-Independent and Identically Distributed (non-IID) across individuals, with each person presenting a distinct cardiac autonomic baseline [4], meaning centralising data often leads to models that struggle to generalise across diverse individuals. Federated learning resolves this by allowing each node to develop a locally calibrated model before contributing to the global update, thereby improving robustness across heterogeneous participants. The Flower framework [14] provides a hardware-agnostic federated implementation applicable to heterogeneous edge devices. Prior work has demonstrated federated approaches that consistently match or exceed centralised accuracy for physiological signal classification whilst preserving privacy [15], [16]. The present study extends this literature with a physical hardware deployment tested across three independent datasets.

III. Methodology

Fig. 1 shows the proposed system architecture, covering the full process from raw ECG acquistion through to federated deployment. The models are initially developed using existing datasets, after which ECG data is collected using a Polar H10 chest strap, with signal processing executed locally on the recording device. For both the public datasets and the live recordings, each continuous recording is segmented into 45-second windows from which HRV features are extracted and passed to a classification model. For cross-participant evaluation, participant-level partitioning is performed before model training to ensure that data from the held-out participant are never exposed during training. Finally, in the federated

learning stage, only model parameter updates, rather than raw signals, are transmitted to the aggregation server.

A. Temporal Windowing Strategy

The original DREAMER study extracted a single fixed window from the final 60 seconds of each stimulus clip [7]. While this provides a baseline, it discards the emotional information present in the rest of the clip and produces only one training sample per clip, limiting data available for cross-participant learning. To overcome this, our approach divides each clip into non-overlapping 45-second segments. Although, European Society of Cardiology recommends 2 to 5 minutes for short-term HRV spectral analysis [4], this applies to the cumulative recording duration rather than internal segmentation. Since stimulus clip durations range from 65 to 393 seconds in DREAMER and are 180 seconds in live validation, all recordings satisfy this requirement. Each 45-second window produces an independent prediction, and the final clip-level classification is determined by the majority voting across all windows. The 45-second duration was selected empirically. To identify the optimal window length, five different window durations were evaluated using the same Random Forest classifier under identical preprocessing, feature extraction, and evaluation settings as shown in Table I.The 45-second window achieves the highest performance under both the LOPO-CV and within-participant evaluation protocols and was therefore adopted for all subsequent phases. The temporal window selection process is illustrated in Fig. 2.

TABLE I
EFFECT OF WINDOW DURATION ON ACCURACY (DREAMER DATASET, FOUR-CLASS, RANDOM FOREST, BALANCED DATA)

Duration	LOPO-CV	Within-Participant
15 seconds	33.38%	40.05%
30 seconds	32.41%	40.52%
45 seconds	**38.27%**	**43.93%**
60 seconds	34.26%	41.02%
90 seconds	38.19%	40.54%

B. Feature Extraction

ECG signals were preprocessed using NeuroKit2 [17], including bandpass filtering, R-peak detection using the Pan-Tompkins algorithm, and RR interval extraction. Each 45-second window produces a single HRV feature vector. Each feature vector includes nine standard time-domain and frequency-domain HRV features: MeanNN, SDNN, RMSSD, pNN50, LF, HF, LF/HF, SD1, and SD2. Let RR_i denote the interval between the i^{th} and $(i+1)^{\text{th}}$ consecutive R-peaks in the ECG signal, where N is the total number of RR intervals within each 45-second window. The primary time-domain HRV features are computed as follows:

$$\text{RMSSD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2} \tag{1}$$

$$\text{SDNN} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(RR_i - \overline{RR})^2} \tag{2}$$

$$\text{pNN50} = \frac{|\{i : |RR_{i+1} - RR_i| > 50\text{ms}\}|}{N-1} \times 100 \tag{3}$$

These features are not arbitrary signal statistics. Each captures a distinct aspect of autonomic nervous system function that is physiologically linked to emotional and mood states. Root Mean Square of Successive Differences (RMSSD) and the Percentage of successive Normal-to-Normal intervals exceeding 50 ms (pNN50) reflect parasympathetic activity, specifically vagal tone, which decreases under psychological stress and negative affect and is chronically reduced in clinical depression [4], [18]. SDNN captures overall autonomic regulatory capacity, a global indicator of the body's ability to adapt to environmental demands, which is similarly diminished in prolonged low-mood states. LF power and the LF/HF ratio reflect sympathovagal balance, shifting toward sympathetic dominance during arousal and stress. MeanNN captures average cardiac cycle length, which lengthens during relaxation and shortens during activation. SD1 and SD2 from the Poincaré plot represent short-term and longer-term beat-to-beat variability respectively, providing complementary views of the dynamic interplay between sympathetic and parasympathetic regulation [18].

C. Evaluation Protocol

The proposed framework is evaluated using two evaluation protocols that is LOPO-CV to assess subject-independent generalisation and within-participant evaluation to compare with the previous works. Under the LOPO-CV protocol, the evaluation consists of K folds. In each fold k, participant k is held out of testing while the model is trained on the rest of the participants. After completing all folds, the classification accuracy is computed as the average accuracy across all K folds, as defined in Eq. (4).

$$\text{LOPO-CV} = \frac{1}{K}\sum_{k=1}^{N}\text{Accuracy}(M_{-k}, P_k) \tag{4}$$

where M_{-k} denotes the model trained without participant k and P_k the held-out test set.

For the within-participant evaluation, 5-fold cross-validation was performed independently for each participant, using an 80%/20% training/testing split in each fold. The reported results are averaged across all five folds and participants.

D. Classification and Equalisation

To investigate the effect of class imbalance on four class emotion classification a Random Forest (RF) classifier with class_weight='balanced' hyperparameter and 200 estimators was used. Balanced weighting adjusts the contribution of each class to training in an inverse proportion to its frequency, providing a baseline correction for imbalance within the classifier itself. To further investigate the impact of class imbalance

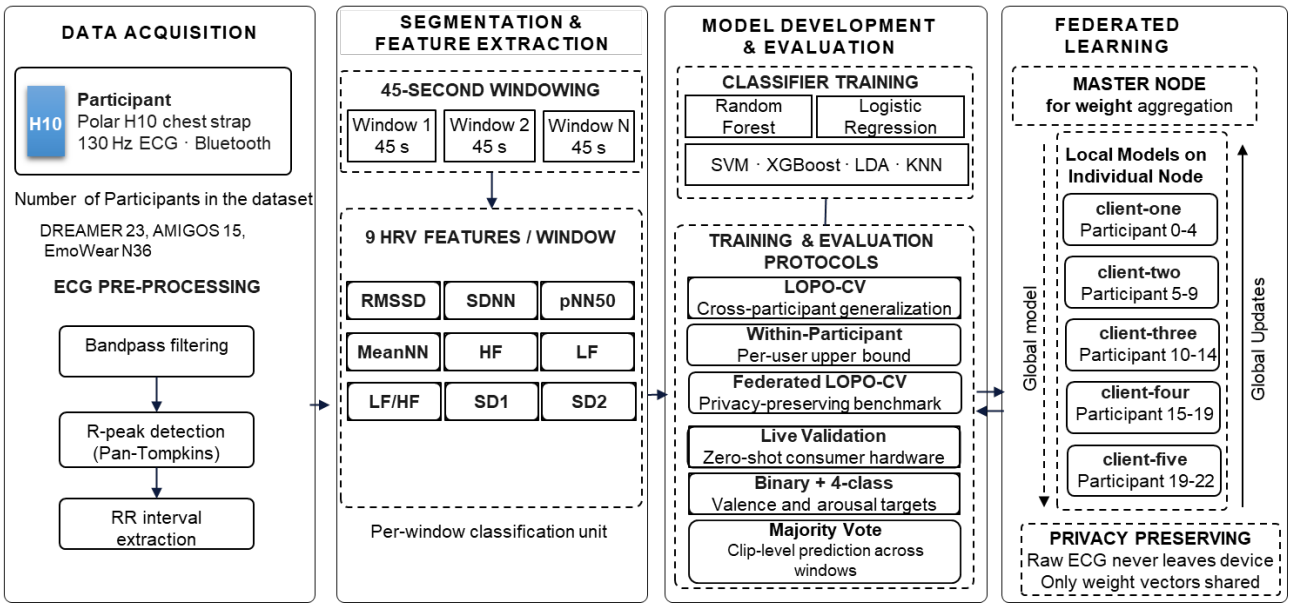


Fig. 1. Proposed federated system architecture for ECG-HRV based mood detection.

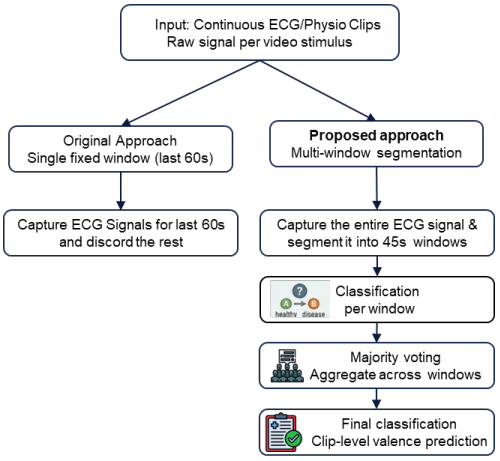


Fig. 2. Comparison of the original single-window (left) and the proposed multi-window segmentation approach (right).

beyond class weighting, explicit class equalisation was additionally applied in selected phases by capping all classes at the window count of the smallest class. For DREAMER four-class classification, the class distribution includes High Arousal–Low Valence (HALV, 304), High Arousal–High Valence (HAHV 261), Low Arousal–High Valence (LAHV, 274), and Low Arousal–Low Valence (LALV, 109) windows. Consequently, the LALV class contains only 36% of the windows in the largest class, indicating substantial class imbalance.

To mitigate this imbalance, two balancing strategies were considered: downsampling and synthetic oversampling. Downsampling to the minority class count was selected instead of synthetic oversampling methods such as SMOTE (Synthetic Minority Oversampling Technique) for two reasons. First, SMOTE generates synthetic HRV feature vectors by interpolating between existing minority class samples. Given the high inter-individual variability of HRV signals, such synthetic samples risk introducing physiologically implausible feature combinations that do not correspond to any real cardiac pattern, as HRV features exhibit high inter-individual variability that makes interpolation between participants unreliable [4], [18]. Second, empirical comparison on the DREAMER confirmed that class-weighted RF with equalisation produced better minority class recall than SMOTE augmentation whilst maintaining interpretable training data. Equalisation was not applied to binary valence experiments on the DREAMER dataset because the HIGH and LOW classes were already approximately balanced, containing 544 and 560 windows respectively. This 1.5% difference was considered insufficient to warrant additional class balancing.

E. Federated Learning Framework

The proposed framework was deployed using a federated learning architecture, with Logistic Regression (LR) adopted as the local classifier on each client node. The federated learning uses FedAvg algorithm to aggregate the locally trained model's parameters received from each client into a global model. This aggregation strategy requires that every client's model be representable as a single continuous numerical weight vector. LR satisfies this directly: its parameters are a compact weight matrix of fixed dimensions ($d \times C$, where d is the number of features and C the number of classes) that can be directly averaged across clients. The LR classifier was configured with regularisation strength C=1.0 (L2 penalty), the LBFGS solver, and a maximum of 1000 iterations, consistent across all centralised and federated experiments. Random Forest (RF), whilst used in the centralised four-class experiments, consists of an ensemble of independent decision trees. Decision tree structures are not representable

as fixed-length numerical vectors and therefore cannot be averaged by FedAvg in a principled way. Whilst federated variants of RF using alternative aggregation strategies exist [19], standard FedAvg as used here requires a continuous weight representation. The FedAvg algorithm [13] computes the global model by averaging locally trained weight vectors

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_k^t \tag{5}$$

where w_k^t is the weight vector of client k after local training at round t, n_k is the number of local training examples, and $n = \sum_k n_k$ is the total across all clients.

Accordingly, logistic regression was used throughout the federated learning experiments. For the federated learning experiments, the entire dataset was partitioned by participants with each client training a local logistic regression model and returning only the model weight updates for the FedAvg aggregation.

F. Datasets

Three publicly available affective computing datasets were selected to evaluate the proposed framework, namely DREAMER [7], AMIGOS [11], and EmoWear [12]. Table II summarises their key characteristics. In all three these datasets, emotional states are labelled through self-report on a two-dimensional valence–arousal space after each stimulus clip. For DREAMER (1–5 scale), the midpoint threshold is 3. For AMIGOS and EmoWear (1–9 scale), the threshold is 5. Clips rated exactly at the midpoint on either dimension were labelled ambiguous and excluded from training and evaluation.

TABLE II
DATASET CHARACTERISTICS. HZ = ECG SAMPLING RATE; LALV = LOW AROUSAL LOW VALENCE WINDOW COUNT.

Dataset	participants	Hz	Scale	LALV
DREAMER [7]	23	256	1–5	109
AMIGOS [11]	15	128	1–9	49
EmoWear [12]	36	250	1–9	219

IV. EXPERIMENTAL RESULTS

This section evaluates our proposed ECG-HRV emotion classification framework. We first establish baseline performance for four-class valence-arousal classification on the DREAMER dataset using a Random Forest (RF) classifier, applying both subject-independent (LOPO-CV) and within-participant protocols. We then examine the impact of class imbalance through equalization. Finally, we validate the framework on the AMIGOS and EmoWear datasets, compare six machine learning classifiers for binary valence classification, and benchmark our results against existing ECG-HRV emotion recognition studies.

A. Baseline and the Class Imbalance Finding

Following selection of the optimal window and HRV features, the baseline four-class classification performance was evaluated on the DREAMER dataset. The LOPO-CV accuracy is 38.27% while the within-participant accuracy is 43.93%. LALV recall is 6.5% under LOPO-CV (7.8% within-participant), indicating that the classifier correctly identifies only a small proportion of LALV samples. This happens due to the underrepresentation of the LALV class (only 12% of training examples), the classifier was unable to learn an accurate decision boundary for the minority class.

To isolate the effect of class imbalance, class equalisation was applied by capping all four classes at 109 windows. All other experimental settings remain the same. Table III summarises the quantitative effect of class equalisation on overall accuracy and class-wise recall, while Fig. 3 presents the corresponding row-normalised confusion matrices, illustrating how the class-wise prediction patterns change following equalisation. Together, the results show that class equalisation substantially improves recognition of the minority LALV class while producing more balanced classification performance across all four classes. Overall accuracy increased from 43.93% to 51.40%, while LALV recall improved from 7.8% to 56.3%, demonstrating substantially improved recognition of the minority class.

TABLE III
EFFECT OF CLASS EQUALISATION (DREAMER, FOUR-CLASS, WITHIN-PARTICIPANT)

Condition	Acc.	LALV	HALV	HAHV	LAHV
No equalisation	43.93%	7.8%	53.8%	58.9%	7.0%
With equalisation	51.40%	56.3%	50.0%	46.9%	51.6%
Change	+7.47pp	+48.5pp	−3.8pp	−12.0pp	+44.6pp

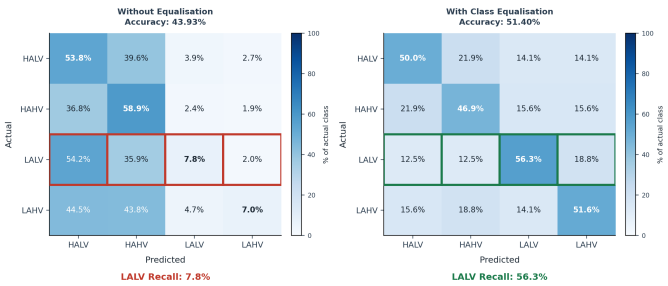


Fig. 3. Confusion matrices for four-class DREAMER classification.

B. Multi-Dataset Evaluation

Having demonstrated the benefit of class equalisation on the DREAMER dataset, the proposed framework was subsequently evaluated on AMIGOS and EmoWear to assess cross-dataset and cross-device generalisation. AMIGOS presented an additional challenge: with only 15 participants in the included subset and approximately 24 windows per participant, each LOPO-CV fold trains on fewer than 336 examples. This lower number of windows relative to DREAMER's approximately 70

windows per participant likely contributed to the lower cross-participant accuracy of 28.82% before equalisation. These results show that participant level data availability is an important factor influencing the performance. EmoWear, with 36 participants and a more balanced LALV distribution of 219 windows, produced the most consistent four-class results.

C. Binary Valence Classification and Algorithm Comparison

Following the evaluation of four-class emotion classification, binary valence classification was investigated to determine whether reducing the classification task to positive and negative valence improves subject-independent performance. Binary valence classification collapses the four quadrants into HIGH (V above midpoint) and LOW (V at or below midpoint). This reduction is motivated by both practical and empirical considerations. From a practical prospective, distinguishing positive from negative emotional state is sufficient for mood monitoring. Six classifiers were evaluated under identical conditions to determine the most suitable classifier for subject-independent binary valence classification. LR achieved the highest LOPO-CV accuracy 53.17% while SVM achieved the lowest 49.91%. The summary of results is presented in Table IV.

TABLE IV
ALGORITHM COMPARISON: DREAMER BINARY VALENCE LOPO-CV

Classifier	Accuracy	HIGH recall	LOW recall
Logistic Regression	**53.17%**	41.4%	64.6%
Random Forest	52.45%	46.5%	58.2%
LDA	52.36%	41.0%	63.4%
XGBoost	51.90%	48.0%	55.7%
KNN (k=5)	51.45%	48.5%	54.3%
SVM (RBF)	49.91%	26.3%	72.9%

To evaluate the generalisability of binary valence classification across different datasets, the two best-performing classifiers (LR and RF) were subsequently evaluated on the DREAMER, AMIGOS, and EmoWear datasets. Table V summarises the corresponding LOPO-CV results. The results demonstrate consistent binary valence classification across the three datasets under the LOPO-CV protocol. LR achieved the highest accuracy on DREAMER (53.17%) and AMIGOS (56.76%), whereas RF performed best on EmoWear (55.51%) and the combined dataset (53.37%).

TABLE V
BINARY VALENCE LOPO-CV RESULTS ACROSS DATASETS

Dataset	RF	LR	Class Balance
DREAMER	52.45%	53.17%	544H / 560L
AMIGOS	50.12%	56.76%	Near-balanced
EmoWear	55.51%	50.87%	632H / 456L
Combined	53.37%	N/A	All three

Table VI compares the proposed framework with representative ECG-HRV emotion recognition studies. The within-participant accuracy on the DREAMER dataset reaches 65.13%, exceeding the original SVM baseline of 62.49%, using multi-window HRV features framework. within-participant studies train and test on data from the same individuals, which have generally produced higher accuracies than subject independent evaluations. Furthermore, the highest-accuracy prior work relies on raw ECG waveforms processed through deep learning on laboratory-grade equipment, or on person-specific feature extraction methods that are computationally unsuitable for edge deployment.

TABLE VI
COMPARISON OF ECG-HRV VALENCE CLASSIFICATION METHODS.

Study	Dataset	Method	Protocol	Accuracy
Katsigiannis [7]	DREAMER	SVM	WP	62.49%
Nita et al. [8]	DREAMER	CNN	WP	95.16%
Wang et al. [9]	DREAMER	LR+ALQ	WP	84.30%
Ferdinando [6]	Mahnob-HCI	SVM+EMD	XP	42.60%
This study	DREAMER	RF	LOPO-CV	52.45%
This study	DREAMER	LR	LOPO-CV	53.17%
This study	EmoWear	RF	LOPO-CV	55.51%
This study	DREAMER	RF	WP	65.13%

V. DEPLOYMENT AND REAL-WORLD VALIDATION

A. Protocol for Live Validation

To bridge the gap between dataset evaluation and real-world deployment, a live validation session was conducted using a Polar H10 chest strap to record ECG signals at 130 Hz. The participant, one of the study authors, had no personal data included in the training dataset and received no calibration session, constituting a genuine zero-shot deployment scenario.

Four video clips were selected to elicit each of the four intended valence-arousal quadrants: a Bob Ross painting tutorial (LAHV), the opening montage of *Up* (HAHV), a performance scene from *Whiplash* (HALV), and a scene from *Schindler's List* (LALV). After each clip, the participant self-reported valence and arousal on the DREAMER 1–5 scale.

The recorded ECG was processed using the same preprocessing, feature extraction, classification and majority voting based final clip classification procedures used during the training sessions. RF classifier was used across four-class valence classification and binary class classification on DREAMER dataset to evalute using both evalaution protocols including LOPO-CV and within the participant. Self-reported ratings indicated that the participant's emotional responses did not always correspond to the intended valence-arousal quadrant.

Fig. 4 presents the results using participant's self-reported emotional states as the ground truth. The four-class model achieved 25%, which is consistent with the LOPO-CV result of 24.02%. The binary valence model achieved 75%, exceeding the equalised LOPO-CV result (52.81%).

VI. PROPOSED METHOD PERFORMANCE IN FEDERATED LEARNING

The federated system was implemented using the Flower framework [14] and deployed on a physical cluster of six Raspberry Pi 4 devices. One device (master server) runs the Flower server managing round coordination, weight aggregation, and global model storage. Each of the remaining five devices holds a private data partition and returns weight updates to the

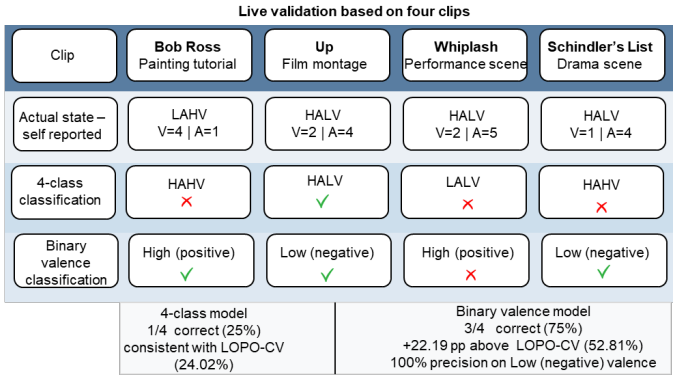


Fig. 4. Live Polar H10 validation results for four video clips.

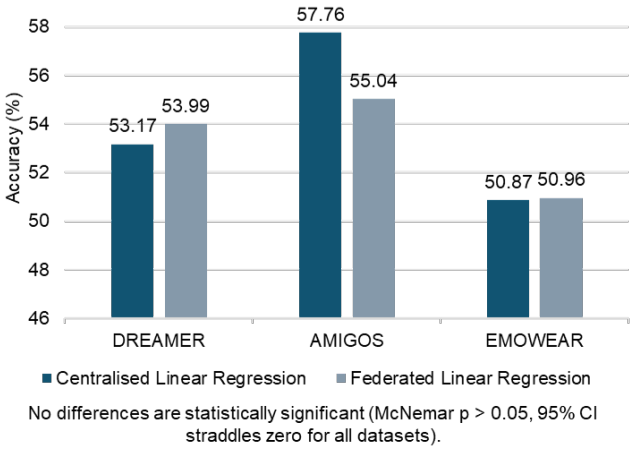


Fig. 5. Centralised vs. federated Logistic Regression performance.

server. The DREAMER dataset was partitioned by participant ID, with each node containing approximately four to five participants. No raw HRV features were shared between nodes at any point. The server received only model weight vectors of dimension $d = 9 \times C$, where C is the number of output classes. To evaluate the effect of federated learning, the federated LR model is compared with its centralised LR counterpart under the same experimental setting.

ECG-HRV features are not independent and identically distributed) across participants [15]. Each individual has a unique cardiac autonomic profile that occupies a different region of the HRV feature space. Therefore, when a centralised LR model is trained on all 23 DREAMER participants simultaneously, it fits a single decision boundary across all overlapping physiological distributions, producing a compromise suboptimal for each individual.

In contrast, each federated node learns a locally calibrated decision boundary from its assigned participant data. FedAvg averages these into a global model informed by five independent optima rather than a single joint fit, with the averaging also acting as an implicit regulariser. As the system matures and each node accumulates data from a single user, local models will increasingly reflect that individual's physiological profile before contributing to the global update. Fig. 5 compares the classification accuracy of centralised and federated LR across the DREAMER, AMIGOS, and EmoWear datasets. Federated learning achieved a modest improvement on the DREAMER dataset (53.99% vs. 53.17%), while centralised learning performed better on AMIGOS (57.76% vs. 55.04%) and both approaches produced comparable performance on EmoWear (50.87% vs. 50.96%).

VII. Discussion

This study demonstrates a proof of concept for continuous, private mood monitoring on consumer hardware, evaluated end-to-end from raw ECG through a physical federated cluster. Starting from the ECG of a completely unseen individual recorded on a chest strap, the approach achieves 75% valence classification accuracy with zero calibration, exceeds the cross-participant benchmark by 22 percentage points, and 100% precision for negative valence classification.The algorithm comparison shows that all six classifiers converge to the same narrow accuracy range between 49.9% and 53.2%, which reframes the research direction for cross-participant HRV emotion classification. The cross-participant performance ceiling appears to be physiological rather than algorithmic. This suggests that personalisation, which the federated architecture enables as personal data accumulates, is a more promising research direction.

The live validation result, while limited to a single participant and four stimuli, provides evidence that dataset-only benchmarks cannot offer: the system works on real physiology, in real time, with real consumer hardware. The 22 % improvement in the live result over the LOPO-CV benchmark suggests that within-session physiological coherence provides a natural form of implicit personalisation even without explicit calibration.The proposed system shows significant clinical and social potential, reaching far beyond laboratory evaluation and academic contexts. A clinician monitoring a patient with depression could receive a continuous longitudinal record of valence trends rather than relying on fortnightly self-report. A parent noticing withdrawal in a teenager could have an objective early signal without intruding on the teenager's daily experience. Individuals managing their own mental health could receive early warnings of mood deterioration before a relapse takes hold. In each of these scenarios, the critical properties are not accuracy in isolation but the combination of deployability on consumer hardware, interpretable features a clinician can reason about, and a privacy model that keeps raw physiological data on the user's own device at all times.

A. Limitations and Ethical Considerations

Despite these promising results, the proposed approach has so far been validated on only a single participant. Future work will involve conducting large-scale live validation. This will require obtaining ethical approval from institutional ethics committee, alongside informed consent from participants, ensuring they are fully aware of the data collected and their right to withdraw at any stage. Therefore, this proposed system is

designed as a personal awareness tool, not a clinical diagnostic instrument. The current study has several limitations. The live validation involved a single participant and four stimuli. The federated implementation used dataset partitions rather than longitudinal personal data from individual users. The current deployment assumes a shared local network. A geographically distributed deployment would require encrypted transmission over HTTPS, but this does not compromise the privacy since only model weights are transmitted, not raw signals [13].

VIII. Conclusion

This paper presented a systematic, multi-phase investigation of ECG-HRV based emotion classification for mood detection, from dataset evaluation to live wearable deployment to privacy-preserving federated learning. The key contributions are: the identification of class imbalance as the dominant suppressor of depressive mood state detection in four-class valence-arousal problems; the demonstration that standard nine-feature HRV classification matches 78-feature extended sets under cross-participant evaluation while remaining clinically interpretable; the first published standalone ECG-HRV binary valence classification results on the EmoWear dataset; live pilot validation on an unseen participant achieving 75% binary valence accuracy on consumer hardware, a promising result pending broader participant validation; and federated learning achieving accuracy statistically equivalent to centralised training across three independent datasets on a Raspberry Pi cluster, with a privacy preservation support and no sharing of raw physiological data. Future work will extend the federated architecture to incorporate EEG signals alongside the ECG approach, and will include live validation on a broader participant cohort, enabling a multimodal and generalised federated mood detection system.

Acknowledgements

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version of this paper arising from this submission.

References

[1] World Health Organization, "World mental health report: Transforming mental health for all," World Health Organization, Geneva, Switzerland, Tech. Rep., 2022.

[2] ——, "Depression and other common mental disorders: Global health estimates," World Health Organization, Geneva, Switzerland, Tech. Rep. WHO/MSD/MER/2017.2, 2017.

[3] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[4] A. J. Camm, M. Malik, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, R. E. Kleiger, F. Lombardi, A. Malliani, A. J. Moss, J. N. Rottman, G. Schmidt, P. J. Schwartz, and D. H. Singer, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[5] G. Valenza, L. Citi, A. Lanatà, E. P. Scilingo, and R. Barbieri, "Revealing real-time emotional responses: A personalized assessment based on heartbeat dynamics," *Scientific Reports*, vol. 4, p. 4998, 2014.

[6] H. Ferdinando, T. Seppänen, and E. Alasaarela, "Comparing features from ECG pattern and HRV analysis for emotion recognition system," in *Proceedings of the 2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Chiang Mai, Thailand, 2016, pp. 1–6.

[7] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2018.

[8] S. Nita, S. Bitam, M. Heidet, and A. Mellouk, "A new data augmentation convolutional neural network for human emotion recognition based on ECG signals," *Biomedical Signal Processing and Control*, vol. 75, p. 103590, 2022.

[9] S. Wang, K. Li, W. Liu, W. Ding, and J. Wang, "ECG multi-emotion recognition based on heart rate variability signal features mining," *Sensors*, vol. 23, no. 20, p. 8636, 2023.

[10] S. Ali, M. G. Kibria, M. A. Jarwar, S. Kumar, and I. Chong, "Microservices model in woo based iot platform for depressive disorder assistance," in *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, 2017, pp. 864–866.

[11] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2021.

[12] M. H. Rahmani, M. Symons, O. Sobhani, R. Berkvens, and M. Weyn, "EmoWear: Wearable physiological and motion dataset for emotion recognition and context awareness," *Scientific Data*, vol. 11, p. 634, 2024.

[13] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. Proceedings of Machine Learning Research, vol. 54. PMLR, 2017, pp. 1273–1282.

[14] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão, and N. D. Lane, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.

[15] R. Zhao, Y. He, L. Ma, and H. Li, "Federated learning for mental health detection using wearable physiological signals: A survey," *Frontiers in Digital Health*, vol. 6, p. 1495999, 2024.

[16] S. Bagherzadeh, M. R. Norouzi, S. B. Hampa, A. Ghasri, P. T. Kouroshi, S. Hosseininasab, M. A. G. Zadeh, and A. M. Nasrabadi, "Federated learning in emotion recognition systems based on physiological signals for privacy preservation: A review," *Multimedia Tools and Applications*, 2024.

[17] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, 2021.

[18] T. Pham, Z. J. Lau, S. H. A. Chen, and D. Makowski, "Heart rate variability in psychology: A review of HRV indices and an analysis tutorial," *Sensors*, vol. 21, no. 12, p. 3998, 2021.

[19] A.-C. Hauschild and J. Baumbach, "Federated random forests can improve local performance of predictive models for various healthcare applications," *Bioinformatics*, vol. 38, no. 8, pp. 2278–2286, 2022.