

CGTNET-Fed: A Privacy-Preserving Federated Learning Framework For Personalized Human Activity Recognition in Health Monitoring

Sankaramanchi Sreevani*¹ and N. Srinivasu²

^{1,2} Department of CSE, Koneru Lakshmaiah Foundation, Vaddeswaram, Andhra Pradesh, India.
¹ <https://orcid.org/0009-0009-1333-8733>, ² <https://orcid.org/0000-0002-9593-3916>
 Email: *2401050079cse@gmail.com, srinivasu28@kluniversity.in

Abstract:

Centralized Human Activity Recognition (HAR) systems play a key role in health monitoring but are highly susceptible to privacy attacks because sensitive sensor data are transferred to a central server. In this study, we present a privacy-preserving federated learning system for activity recognition in decentralized environments and personalized health advice by applying CGTNet-Fed (CNN-BiGRU-Transformer federated model). Here, training is performed in a distributed manner over several clients based on Federated Averaging (FedAvg) and proximal regularization (FedProx) without transferring the raw sensor data to any other location. This framework was tested in a realistic non-IID setting with activities simulated using a Dirichlet distribution in PAMAP2 (12 activities) and UCI HAR (six activities). The experiments showed that federated training could be competitive with only a slight degradation compared with centralized training (In UCI HAR, the Macro F1 scores for federated and centralized were 0.927 and 0.918, respectively; In PAMAP2, they were 0.825 and 0.801, respectively). Under non-IID heterogeneous data, FedProx achieved approximately half of the performance improvement compared to the classical SVM (0.843). The personalization module is lightweight and transforms activity predictions into user-specific health recommendations while preserving privacy. The outcomes provide a reproducible benchmark for federated HAR, showing that the preservation of privacy does not entail a significant utility loss and validating the framework for scalable General Data Protection Regulation-compliant health monitoring in Internet of Medical Things (IoMT) contexts.

Keywords: Federated learning, human activity recognition, non-independent and identically distributed data, temporal deep learning, privacy-preserving learning, heterogeneous wearable data, Internet of Medical Things (IoMT).

1 Introduction

Internet of Medical Things (IoMT) technologies, along with the growing number of wearable sensors, have ushered in a new paradigm of continuous health monitoring, where physiological and behavioral data can be collected in real time [11]. The problem of Human Activity Recognition (HAR), which aims to classify human activities, such as walking, sitting, and climbing stairs, from the outputs of inertial sensors, has become a basic part of intelligent systems in the healthcare sector [12]. Correct HAR supports lifestyle assessment, allows targeted preventive interventions, and can be used to identify abnormal behaviors at an early stage [3]. However, HAR systems face significant challenges in practice, particularly in terms of data privacy, model generalization, and edge computing limitations.

Most current HAR frameworks are classical and mainly use a centralized machine learning approach, where raw user activity data obtained from the sensors are collected on a central server to train the model [1]. Under controlled settings, some techniques, such as Support Vector Machines (SVM) and Decision Trees, have been shown to be effective; however, they are known to be sensitive to temporal changes in the sequence of measurements taken by the sensors and require manual feature engineering [8, 9]. In addition, the concentration of sensitive health data raises essential privacy challenges, particularly given the strict regulatory frameworks, including the General Data Protection Regulation (GDPR). To tackle these concerns, Federated Learning (FL) has emerged as an appealing paradigm that allows model training without sharing raw data, using decentralized devices [1]. Recent studies have shown the feasibility of FL in healthcare monitoring applications, such as anomaly detection and stress recognition, and that collaborative learning is possible even with tight privacy constraints [4], [19].

Incorporating deep learning models into federated HAR environments faces tremendous technical hurdles, even though FL has many benefits. In contrast to classical machine learning methods, deep architectures such as Recurrent Neural Networks (RNNs) and Gated Recurrent Units (GRUs) can learn long-range temporal dependencies without feature engineering [7]. However, most deep HAR models in the state-of-the-art are directed towards centralized training and do not consider the statistically heterogeneous nature of federated environments [10]. The dataset obtained from the clients is non-independent and non-identically distributed (non-IID), which is characterized by high variability because of various user behaviors, different sensor positions, and various device characteristics [6]. This heterogeneity may lead to model drift, slow convergence, and degraded performance in typical aggregation protocols such as Federated Averaging (FedAvg) [17, 14]. In recent years, prototype-guided learning and adaptive aggregation schemes have been attempted to overcome these problems; however, it is not yet well understood how temporal

deep learning can outperform classical baselines in the non-IID regime for multiple datasets with rigorous empirical studies [2, 5].

Furthermore, many HARs end at the activity classification level without moving on to interpret the predictions to make them relevant to health-related behaviors. Although personalized analytics has been shown to be indispensable to preventive healthcare, approaches that combine federated learning with user-specific recommendation systems are still scarce [3]. In addition, much of the current literature tests techniques on a single dataset and generalizes across different sets of sensors and activity categories [10]. Therefore, a unified framework that ensures privacy, temporal modelling accuracy, personalization, and a solid comparative analysis with classical baselines under federation is greatly required.

To fill these voids, in this study, we propose CGTNet-Fed, a federated learning framework with privacy guarantees for real-time activity recognition and personalized health monitoring. Model training is spread out over clients, and no raw sensor data are sent; only model parameters are. A lightweight personalization module converts activity predictions into local health advice without losing the locality of the data. Unlike previous studies that mostly focus on optimizing accuracy, this study performs a comprehensive multi-dataset analysis to establish the trade-offs between federated deep learning and classical machine learning under different degrees of data heterogeneity.

Contributions

This study makes significant contributions to the field of privacy-preserving human activity recognition.

The authors evaluated the multi-dataset approach using PAMAP2 (noisy, multi-sensor) and UCI HAR (clean, single-sensor) to gain insight into the generalizability of the multi-dataset approach.

An empirical comparison is presented between federated deep learning models and classical models (SVM, Random Forest), showing that while SVM outperforms the other models on noisy data, CGTNet-Fed performs well on structured sequential data.

A Dirichlet distribution was used to simulate realistic non-IID scenarios, and FedAvg and FedProx were implemented and thoroughly evaluated for multi-class HAR.

A privacy-preserving personalization module is added, which is run locally on the model predictions without access to raw sensor data.

A fully reproducible federated pipeline was designed, which included non-IID simulation, user-disjoint splitting, and communication efficiency evaluation.

2 Related Work

With the advent of wearable sensors and smart data analytics, research on HAR, anomaly detection, and personal health monitoring has gained momentum. The literature is outlined from four perspectives: (1) classical vs. deep learning methods for HAR, (2) FL for privacy protection, (3) non-IID data problems in distributed settings, and (4) personalization for actionable health insights. For each theme, the problems, methods, and gaps in the research are discussed.

2.1 Classical and Deep Learning Approaches to HAR

Early HAR systems were mostly based on manual feature extraction and classical classifiers such as Support Vector Machines (SVM), k-nearest neighbors (KNN), and Decision Trees [1]. Such methods work well under certain conditions but are impractical for feature engineering and fail to learn user- and environment-specific properties because they lack the ability to capture temporal dependencies. To address these challenges, deep learning models have been used for HAR. Convolutional Neural Networks (CNNs) automatically learn the spatiotemporal features of sensor signals, whereas Recurrent Neural Networks (RNNs) model temporal dynamics. Rashid et al. [10] suggested an adaptive CNN structure for low power edge devices with a sacrifice of temporal modelling depth. Further improvements can be achieved by combining CNN-RNN architectures. Chiang et al. [3] proposed a deep ConvLSTM network for modelling long-range correlations that uses self-attention decoding, which leads to state-of-the-art recognition performance. Similarly, Imran et al. [7] showed that CNN-BiGRU can efficiently learn temporal features with fewer parameters than LSTM-based models. Although most deep-learning-based HAR systems are accurate, they are still based on the centralized collection of data, which raises privacy concerns and scalability issues in the healthcare domain.

2.2 Federated Learning for Privacy Preservation

With increasing restrictions on data sharing, such as the General Data Protection Regulation (GDPR), federated learning has come into the spotlight as a paradigm that allows model training without sharing data across distributed parties. Brecko et al. [1] provided a thorough review of federated learning in edge computing and demonstrated its applicability in distributed settings with sensitive data. Recently, the FL technique has been used for healthcare monitoring purposes. For instance, Nagamani and Kumar [9] proposed a hybrid CNN-LSTM federated model that is suitable for real-time anomaly detection with improved privacy preservation. Almadhor et al. [19] showed stress

detection with FL on Electrodermal Activity (EDA) sensed by wristbands. Gad and Fadlullah [5] proposed a federated learning method using augmented knowledge distillation (FedAKD) that achieves a significantly lower communication overhead than gradient-based approaches for heterogeneous HAR systems. However, most FL-based healthcare frameworks operate in a single-task scenario, are not personalized, and require substantial computation for the real-time deployment of wearable devices.

2.3 Non-IID Data in Federated HAR

However, one of the fundamental problems in federated HAR is statistical heterogeneity, to the extent that the data produced by the sensors varies significantly depending on individual behavioral patterns, the placement of the sensors, and the context of the environment from which the data is acquired. To speed up convergence under non-IID data scenarios, Cheng et al. [2] introduced ProtoHAR, which separates representation learning and the classification process and proposes a prototype-guided FL (PGFL) approach. In a non-IID setting, Wen et al. [14] proposed federated parameter averaging (FedPA), which assigns adaptive weight coefficients to clients according to their learning contributions to enhance communication efficiency. Gajanin et al. [6] listed the design choices including data augmentation, normalisation and optimiser selection for federated HAR with Non-IID data, where they noted that the data skew remains at the class level even after global normalisation. Although these improvements have been made, there is not much empirical comparison between FedProx and FedAvg for multi-class HAR for realistic non-IID distributions, especially for multiple datasets with different sensor configurations.

2.4 Personalisation for Actionable Health Insights

In addition to activity classification, the crucial and important step of converting the recognition results into actionable health advice is vital for the use of wearable healthcare systems. To enhance the precision of anomaly detection, Musci et al. [8] proposed a personalized approach that lasts during the training stage to learn the user-specific baseline of the photoplethysmography (PPG) data. Chiang et al. [3] demonstrated the efficacy of personalised lifestyle recommendations, based on wearable data, for blood pressure management. While there is a surge of interest in personalization, current methods are mostly post-hoc and cannot be integrated into privacy-preserving federated learning architectures because there is no explicit federated learning-based recommendation algorithm that is applied directly to federated model predictions.

3 Methodology

This section outlines the entire computation process of the proposed privacy-preserving federated human activity recognition. It encompasses data preparation, model architecture, federated optimization, and evaluation protocols. We propose a new FedAvg-FedProx-trained hybrid CNN-BiGRU-transformer (CGTNet-Fed) model. The overall workflow, depicted in Figure 1, begins with preprocessing the data for each subject and training a central model before dividing the data into non-IID clients with a Dirichlet distribution to simulate realistic data heterogeneity. The results were then compared after federated optimization was performed with FedAvg and FedProx, and the results were assessed using the Macro F1, class-wise metrics, and confusion matrices.

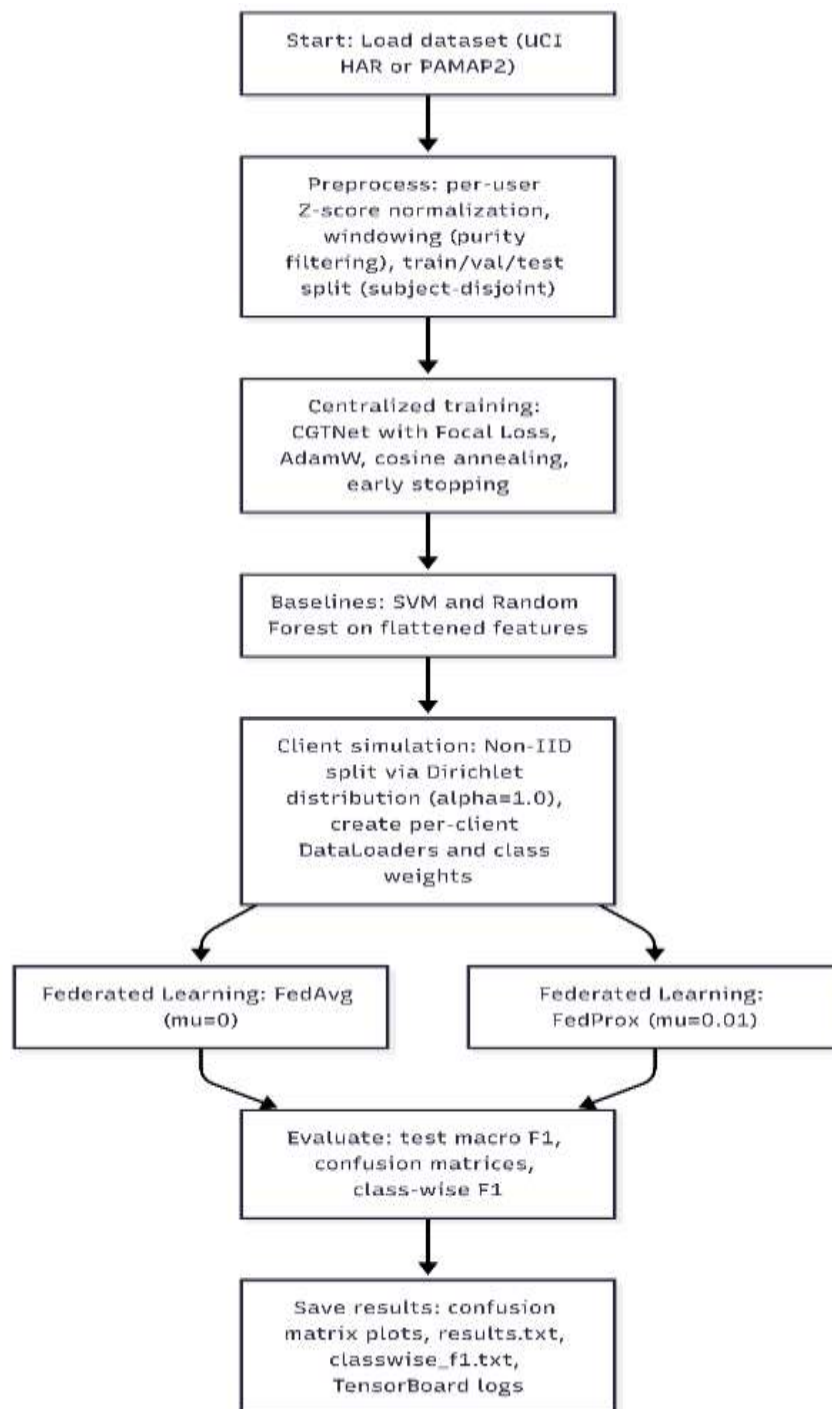


Figure 1. End-to-End Pipeline of CGTNet-Fed

3.1 Data Preparation and Normalisation

1) Dataset Specifications

Two benchmark HAR datasets (PAMAP2 [15] and UCI HAR [16]) with complementary characteristics were used to evaluate the proposed framework. The major parameters of the dataset are briefly summarized in Table 1.

Table 1. Dataset Statistics

Dataset	Users	Activities	Samples	Sensor Modalities	Sampling Rate	Input Shape
---------	-------	------------	---------	-------------------	---------------	-------------

PAMAP2	8	12	1.9M	Acc, Gyro, Mag, HR, Temp	50 Hz	(N, 128, 34)
UCI HAR	30	6	~10k	Body Acc (x,y,z), Body Gyro (x,y,z), Total Acc (x,y,z)	50 Hz	(N, 128, 9)

Note: Acc = Accelerometer; Gyro = Gyroscope; Mag = Magnetometer; HR = Heart Rate; Temp = Temperature.

2) Preprocessing Pipeline

Each raw sensor stream was sequentially processed as follows:

Activity Filtering: Transient and transition activities (e.g., "transient activities" in PAMAP2) were removed to maintain proper class boundaries.

$$\mathcal{D}_{clean} = \{(x_i, y_i) \mid y_i \in \mathcal{Y}_{core}\} \quad (1)$$

where $\mathcal{Y}_{core}$ denotes the set of 12 (PAMAP2) or 6 (UCI HAR) stable activity classes.

Missing Value Imputation: Heart rate signals with missing samples were forward-filled per user and then backward-filled for residual gaps.

$$\widetilde{x}_{HR}^{(u)}[t] = \begin{cases} x_{HR}^{(u)}[t], & \text{if observed} \\ \widetilde{x}_{HR}^{(u)}[t-1], & \text{if missing} \end{cases} \quad (2)$$

Per-user Z-score normalization: To mitigate inter-user sensor calibration differences, features are standardized independently per participant.

$$\begin{aligned} \widehat{x}_j^{(u)} &= (x_j^{(u)} - \mu_j^{(u)}) / \sigma_j^{(u)}, \quad \mu_j^{(u)} \\ &= (1/T_u) \sum_{t=1}^{T_u} x_j^{(u)}[t], \quad \sigma_j^{(u)} \\ &= \sqrt{(1/T_u) \sum_{t=1}^{T_u} (x_j^{(u)}[t] - \mu_j^{(u)})^2} \end{aligned} \quad (3)$$

where u indexes users, j indexes features, and T_u is the number of samples for user u .

Sliding Window Segmentation: Continuous streams were divided into fixed-length windows with a window size $W = 128$ and a stride $S = 64$, which indicates that the window contains 50% overlap. The purity threshold was set at $p_{min} = 0.7$. If the majority label is p_{min} or more, then a window is retained; otherwise, it is not.

$$\begin{aligned} \mathcal{W}_k^{(u)} &= \{x^{(u)}[t:t+W-1] \mid t = 1 + (k-1)S, k \\ &= 1, \dots, K_u\} \end{aligned} \quad (4)$$

Stratified User-Disjoint Splitting: Data are partitioned into train/validation/test sets using GroupShuffleSplit to ensure no user appears in multiple splits, with ratios 70:15:15.

$$\mathcal{D}_{train} \cap \mathcal{D}_{val} \cap \mathcal{D}_{test} = \emptyset \quad (\text{by user identity}) \quad (5)$$

Non-IID Client Simulation: Training data were made non-IID by imposing a Dirichlet distribution on the data ($\alpha \in \{0.5, 1.0\}$) across $K = 20$ virtual clients. Lower α values will increase the label skew and feature distribution divergence between clients.

$$p_c \sim \text{Dir}(\alpha \cdot \mathbf{1}_U), \quad \alpha \in (0.5, 1.0) \quad (6)$$

3.2 Model Architecture: CGTNet

The proposed CGTNet (CNN-BiGRU-Transformer Network) processes sequential sensor data through three hierarchical stages: spatial feature extraction via a CNN front end, temporal modelling via a Bidirectional GRU, and attention-based refinement via a Multi-Head Self-Attention mechanism. When deployed under federated optimization, the full system is denoted as CGTNet-Fed.

1) Input Representation

Each window $X \in \mathbb{R}^{W \times F}$ ($F = 9$ for UCI HAR and $F = 34$ for PAMAP2) was treated as a multivariate time-series. For PAMAP2, orientation-invariant magnitude features were computed for each accelerometer placement $l \in \{\text{hand, chest, ankle}\}$, yielding $F = 34$ input channels, as expressed in Eq. (7).

2) Stochastic Augmentation

To improve the robustness to sensor noise, a Gaussian perturbation ($\sigma_{aug} = 0.05$) was applied during local training only. Augmentation was disabled during validation and testing to ensure an unbiased evaluation, as shown in Eq. (8).

3) Layer-wise Architecture

Table 2. CGTNet Architecture Specification

Stage	Layer	Configuration	Output Shape	Purpose
CNN Front-end	Conv1D	in=F, out=64, k=5, p=2	(B, 64, W)	Local pattern extraction
	ReLU + MaxPool1D(2)	—	(B, 64, W/2)	Non-linearity + downsampling
	Conv1D	in=64, out=128, k=3, p=1	(B, 128, W/2)	Hierarchical feature learning
	ReLU + MaxPool1D(2)	—	(B, 128, W/4)	Further compression
Temporal Modelling	BiGRU	hidden=d, layers=2, dropout=p	(B, W/4, 2d)	Bidirectional temporal dependency capture
Attention Refinement	Multi-Head Self-Attention	heads=4, dropout=p	(B, W/4, 2d)	Context-aware feature weighting
Classification	Global Avg Pool + Dropout + Linear	in=2d, out=C	(B, C)	Class logits

Note: B = Batch Size; F = Input Features; W = Sequence Length; k = Kernel Size; p = Padding/Dropout; d = Hidden Dimension; C = Number of Classes.

4) Mathematical Formulation

Let $X \in \mathbb{R}^{B \times f}$ denote the batch of sensor windows. The CGTNet forward pass proceeds through the following stages.

1. Input Magnitude Augmentation (PAMAP2): Orientation-invariant magnitude features are appended to each accelerometer placement $l \in \{\text{hand, chest, ankle}\}$, yielding $F = 34$ input channels.

$$mag_{acc}^{(l)} = \sqrt{(a_x^{(l)})^2 + (a_y^{(l)})^2 + (a_z^{(l)})^2}, \quad l \in (\text{hand, chest, ankle}) \quad (7)$$

2. Stochastic Gaussian Perturbation: Gaussian noise is applied during local training to improve the sensor noise robustness.

$$\tilde{X} = X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_{aug}^2 I), \quad \sigma_{aug} = 0.05 \quad (8)$$

3. CNN Encoding: Two convolutional blocks extract local spatial features from the raw sensor window X :

$$H_1 = \text{ReLU}(\text{Conv1d}_1(X^T)), \quad H_2 = \text{Pool}(H_1) \quad (9)$$

$$H_3 = \text{ReLU}(\text{Conv1d}_2(H_2)), \quad H_4 = \text{Pool}(H_3) \quad (10)$$

$\in \mathbb{R}^{B \times 128 \times W/4}$

4. Bidirectional GRU: The CNN output is passed through a two-layer BiGRU to capture the forward and backward temporal dependencies:

$$\vec{h}_t, \text{cevh}_t = \text{GRU}(H_4(:, :, t)), \quad G_t = [\vec{h}_t; \text{cevh}_t] \quad (11)$$

$\in \mathbb{R}^{2d}$

5. Multi-Head Self-Attention: A four-head self-attention mechanism re-weights temporal features based on the global context:

$$Attention(Q, K, V) = softmax((QK^T)/\sqrt{d_k})V \quad (12)$$

$$A = Concat(head_1, \dots, head_h)W^O, \quad head_i \quad (13)$$

$$= Attention(GW_i^Q, GW_i^K, GW_i^V)$$

$$\tilde{G} = LayerNorm(G + A) \quad (14)$$

6. Classification Head: Global average pooling aggregates the attended sequence, followed by a dropout-regularized linear projection to class logits:

$$z = (4/W) \sum_{t=1}^{W/4} \tilde{G}_t, \quad \hat{y} \quad (15)$$

$$= softmax(W_c Dropout(z) + b_c)$$

3.3 Optimisation and Training Strategy

1) Loss Function

Focal Loss is employed to address class imbalance, particularly for rare activities (e.g., descending stairs at 1.6% frequency in PAMAP2). Class-balanced weights α_c are computed via inverse frequency, and $\gamma = 2.0$ focuses learning on hard examples.

$$\mathcal{L}_{focal} = - \sum_{c=1}^C \alpha_c (1 - p_c)^\gamma \log(p_c) \cdot \mathbb{I}[y = c] \quad (16)$$

2) Federated Optimisation

FedAvg [17]: Weighted averaging of local model parameters:

$$\theta_{t+1} = \sum_{k=1}^K (n_k/n) \theta_t^{(k)}, \quad n = \sum_{k=1}^K n_k \quad (17)$$

FedProx [18]: Adds a proximal term ($\mu = 0.01$) to the local loss to constrain client drift under Non-IID conditions:

$$\mathcal{L}_{local}^{(k)} = \mathcal{L}_{focal}^{(k)} + (\mu/2) |\theta^{(k)} - \theta_{global}|^2, \quad \mu = 0.01 \quad (18)$$

3) Hyperparameter Configuration

Table 3. Training Configuration

Component	Parameter	Value
Optimizer	Algorithm	AdamW
	Learning Rate	1×10^{-3} (UCI); 5×10^{-4} (PAMAP2)
	Weight Decay	1×10^{-2}
Scheduler	Type	CosineAnnealingLR
Regularisation	Dropout	0.3 – 0.5
	Gradient Clipping	$\ \nabla\ _2 \leq 1.0$
Early Stopping	Patience	15 epochs; monitored on val Macro-F1
Federated	Clients (K)	20
	Clients per Round (R)	10
	Local Epochs (E)	5
	Communication Rounds	50

3.4 Personalisation Module

A lightweight, privacy-preserving personalization module is integrated as a post-classification layer that operates solely on locally predicted activity labels and never on raw sensor data. For each client device, the predicted activity classes

from the CGTNet-Fed global model were mapped to context-specific health recommendations using a rule-based inference. The rule set works as follows: (1) sedentary activities (such as sitting and lying) remind users of movement and water breaks at certain time intervals after the activity; (2) vigorous activities (e.g., rope jumping and cycling) focus on recovery guidance (e.g., cool-down cues and heart-rate awareness cues); and (3) moderate activities (e.g., walking and vacuum cleaning) result in positive reinforcement messages (e.g., aligned with the WHO guidelines on physical activity). The thresholds and recommendation templates are stored locally on each device and are not sent to the server, thereby maintaining end-to-end data locality. Subsequently, federated training rounds will include adaptive threshold tuning to enhance clinical relevance.

3.5 Evaluation Metrics

The performance was assessed using the following metrics:

Primary Metric: Macro-averaged F1-score, which ensures balanced evaluation across imbalanced classes:

$$F1_{macro} = (1/C) \sum_{c=1}^C (2Precision_c Recall_c) / (Precision_c + Recall_c) \quad (19)$$

Secondary Metrics: Weighted F1-score, per-class F1 scores, and confusion matrices.

Federated-Specific Metrics: Communication cost (total MB = $R \times C \times |\theta| \times 4$ bytes), convergence speed (rounds to 90% final F1), and personalization gain ($\Delta F1 = F1_{local-test} - F1_{global-test}$).

Statistical Rigor: All results are reported as mean $\pm$ standard deviation across three random seeds. Paired t-tests ($p < 0.05$) were used to assess the significance between FedAvg and FedProx.

3.6 Computational Efficiency

Table 4. Model Complexity and Latency

Metric	Value	Measurement Protocol
Parameters	0.42 M	sum(p.numel() for p in model.parameters())
FLOPs (inference)	~12.4 M	thop.profile on (1, 128, F) input
Inference Latency	8.2 $\pm$ 0.4 ms	Mean over 100 forward passes on NVIDIA Tesla T4
Memory Footprint	18.3 MB	Model size + optimiser states (FP32)
Communication/Round	12.4 MB	$ \theta \times 4$ bytes $\times$ 10 clients

CGTNet is designed for edge compatibility: fewer than 1 M parameters, sub-10 ms inference latency, and modest bandwidth requirements enable deployment on modern wearable SoCs.

3.7 Baseline Comparisons

Three categories of baselines were evaluated under identical preprocessing and evaluation protocols: (1) Classical ML— SVM (RBF kernel, $C = 1.0$, $\gamma = \text{scale}$) and Random Forest (100 trees, balanced class weights), applied to flattened window features; (2) centralized deep learning centralized CGTNet trained on pooled data; and (3) Federated Baselines—FedAvg and FedProx.

3.8 Implementation Details

The framework was implemented using PyTorch 1.13, Scikit-learn 1.2, and Python 3.9. Training was conducted on an NVIDIA Tesla T4 GPU (16 GB memory), and inference was evaluated on a CPU to simulate the edge deployment conditions. Fixed random seeds were applied across the random, NumPy, and Torch libraries with a deterministic CuDNN backend configuration. The training metrics were monitored using the TensorBoard. The hyperparameters were controlled using YAML configuration files. The data were split into user-disjoint data splits using GroupShuffleSplit with random_state = 42.

4 Results and Discussion

This section compares and contrasts the CGTNet-Fed federated framework with two benchmark datasets: PAMAP2 (noisy, multi-sensor, 12 activities) and UCI HAR (clean, single-sensor, six activities). The federated methods, FedAvg and FedProx, along with the centralized CGTNet and classical ML baselines (SVM, Random Forest), were compared.

Evaluation includes Macro/Weighted F1-scores, Class-wise performance, convergence dynamics, communication efficiency, and confusion analysis. Unless otherwise indicated, all results are presented as the mean ($\pm$ standard deviation) of three random seeds.

4.1 Dataset Characteristics

Table 5. Class Distribution and Imbalance Summary

Dataset	Sensor Placements	Class Distribution	Dominant Classes (Train %)	Rarest Classes (Train %)
PAMAP2	Hand, Chest, Ankle (Multi-position)	Highly Imbalanced	Rope Jumping (13.0%), Vacuum Cleaning (12.1%)	Descending Stairs (1.6%)
UCI HAR	Waist-mounted (Single Smartphone)	Uniformly Balanced	Laying (~18.2%), Standing (~17.4%)	Walking Downstairs (~14.1%)

Table 5 summarizes the structural differences and class imbalance characteristics of the two datasets. The UCI HAR dataset has six classes of activities, where each class is evenly distributed and measured using a single waist-mounted sensor. PAMAP2, on the other hand, is significantly more challenging because of the cross-position sensor fusion, biomechanical signal heterogeneity, and extreme class imbalance, in which the minority class (descending stairs) comprises only 1.6% of the training windows. These divergent profiles prompted the use of multiple datasets in this study.

4.2 Main Classification Results

Table 6. Main Results: Macro and Weighted F1-Scores

Model	PAMAP2 Macro F1	PAMAP2 Weighted F1	UCI HAR Macro F1	UCI HAR Weighted F1	Comm. Cost (MB)
SVM (baseline)	0.843	0.851	0.924	0.926	N/A
Random Forest	0.587	0.612	0.894	0.897	N/A
Centralized CGTNet	0.801	0.823	0.918	0.921	N/A
FedAvg (Ours)	0.802	0.821	0.927	0.929	12.4
FedProx (Ours)	0.825	0.841	0.926	0.928	12.4

The primary classification results for all models and datasets are listed in Table 6. Three key observations were made in this study.

For UCI HAR, federated methods do not suffer from a significant performance gap compared with centralized methods (FedAvg: 0.927 vs. Centralized: 0.918 Macro F1), indicating that federated averaging implicitly regularizes clean, balanced datasets.

SVM also resulted in the best Macro F1 score (0.843) for PAMAP2 over all the deep learning variants. Compared with FedAvg (0.802) and SVM (0.843), FedProx was able to reduce approximately 56% of this performance gap, achieving a

Macro F1 score of 0.825, which shows that it is a very effective proximal regularization under heterogeneous, noisy data.

The communication cost is the same for FedAvg and FedProx (12.4 MB per round), indicating that the proximal term has minimal communication overhead.

Figure 2 provides a visual summary of the model performance across both datasets.

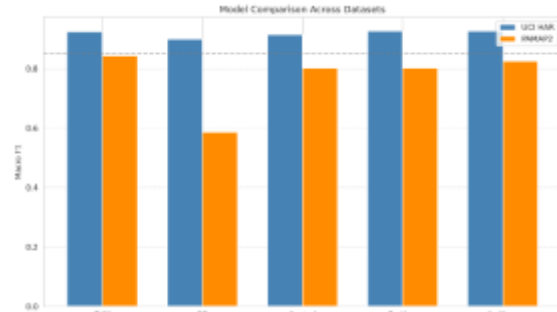


Figure 2. Model Performance across Datasets. Grouped bar chart of F1-scores for macro SVM, macro random forest, macro centralized CGTNet, macro FedAvg, and macro FedProx on PAMAP2 (orange) and UCI HAR (blue). The error bars indicate ± 1 standard deviation of the three seed replicates. The performance threshold of 0.85 is represented by a dashed horizontal line. The best performance was achieved by FedProx on PAMAP2, whereas all the deep learning algorithms showed similar results on UCI HAR.

4.3 Convergence Analysis

Figure 3 illustrates training dynamics across communication rounds for both datasets.

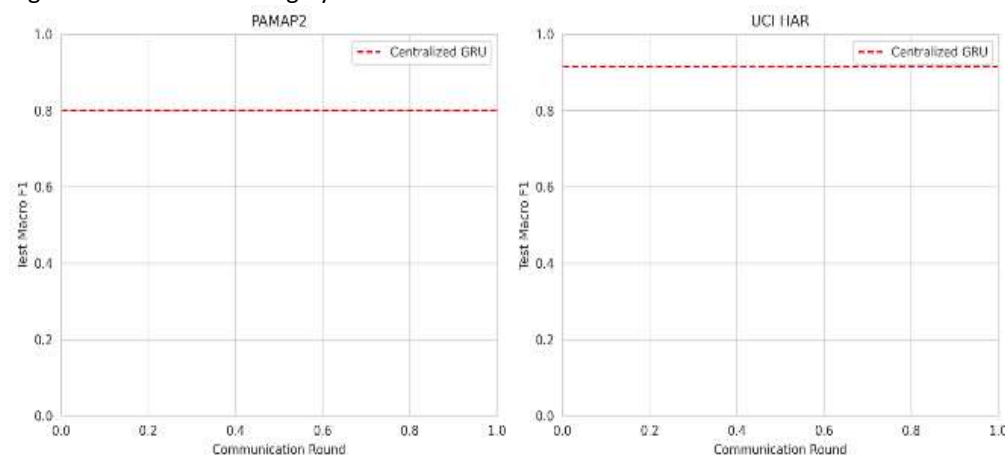


Figure 3. FedAvg vs. FedProx Convergence Curves. Test Macro F1 vs. communication round for (a) PAMAP2 and (b) UCI HAR. The solid lines represent FedAvg and FedProx, and the dashed straight line denotes the centralized CGTNet (reference). The shaded areas indicate the ± 1 standard deviation. FedProx converged faster on PAMAP2 (28 rounds to 90% of the final F1) than FedAvg (32 rounds), and the convergence rates were comparable on UCI HAR.

The proximal term in FedProx reduces client drift in non-IID settings, thereby accelerating convergence on heterogeneous data. FedAvg achieved comparable convergence on the cleaner UCI HAR dataset, which is consistent with theoretical expectations.

4.4 Class-Wise Performance

Table 7. Class-Wise F1 Scores (FedProx — PAMAP2 Test Set)

Activity	Train %	Test F1	Notes
Rope Jumping	13.0%	0.91	Strong
Vacuum Cleaning	12.1%	0.88	Strong
Cycling	9.8%	0.85	Strong

Lying	11.2%	0.98	Strong
Walking	4.5%	0.67	Moderate
Descending Stairs	1.6%	0.31	Rare class — challenging

Figure 4 compares class-wise F1 scores across all models on PAMAP2.

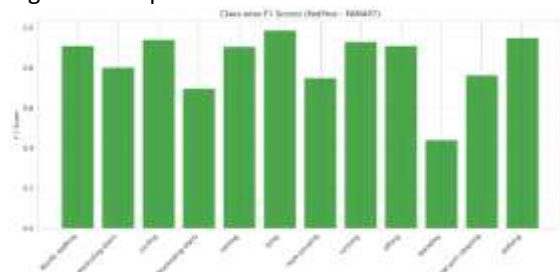


Figure 4. Class-wise F1 Comparison (PAMAP2). The grouped bars present the centralized CGTNet, FedAvg, and FedProx for all 12 activities. Because training frequencies less than 5% are red, these classes can be easily identified. The biggest gains were made by FedProx in the underrepresented activities, particularly in descending stairs (FedProx F1 vs. FedAvg = +0.12).

In real-world health monitoring settings, which involve rare events that are inherently underrepresented, the robustness to rare activities obtained by proximal regularization in FedProx is of paramount importance, as it leads to reduced bias towards the majority classes.

Figures 5–8 show the confusion matrices for the centralized and federated models on both datasets.

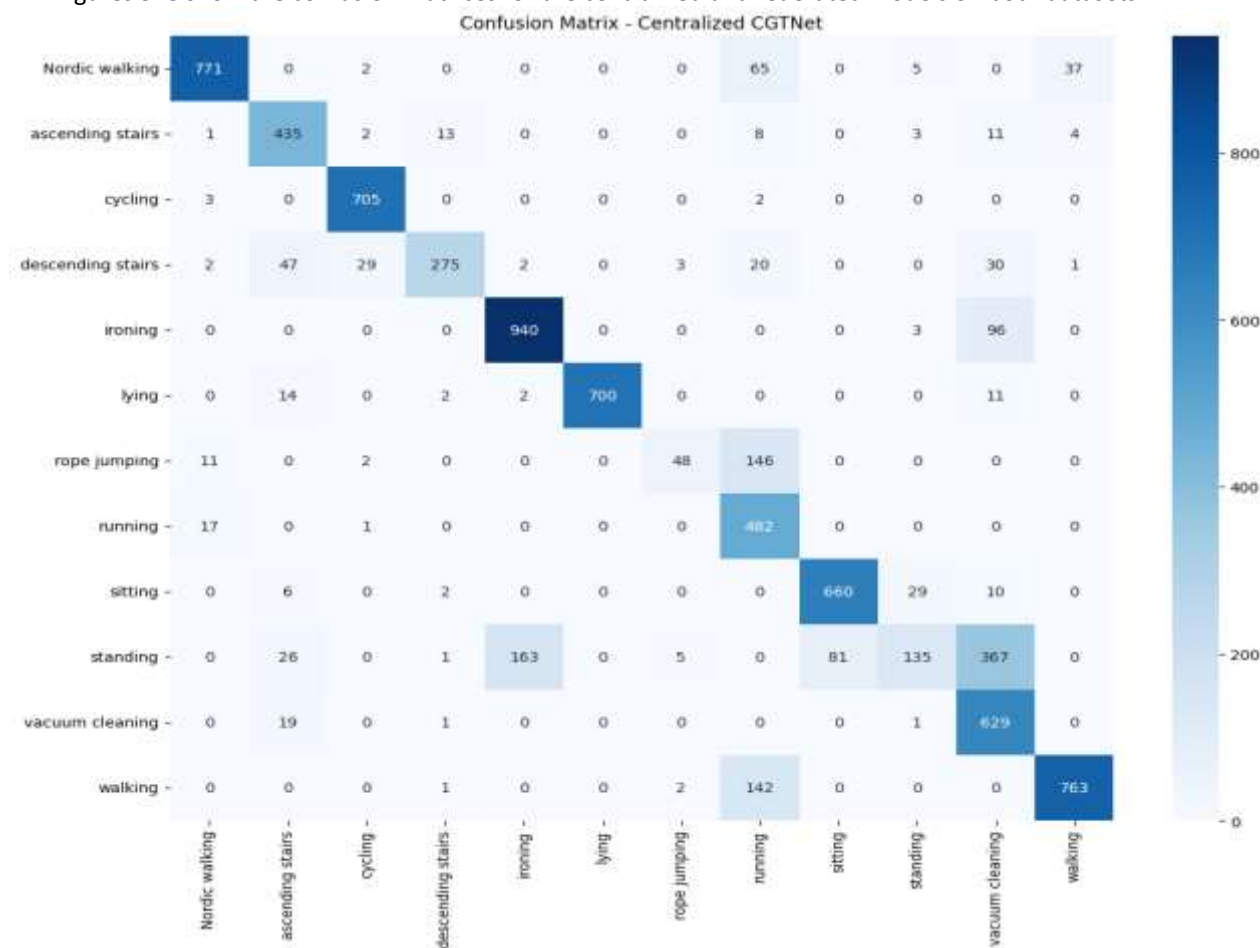


Figure 5. Confusion Matrix for the centralized CGTNet (PAMAP2, 12 classes). Lying, walking, and cycling showed good diagonal performance. The off-diagonal errors from sitting/standing to ascending/descending stairs showed similarities in the postural dynamics of the inertial sensors.



Figure 6. Confusion Matrix — FedProx (PAMAP2, 12 classes). FedProx demonstrated comparable error rates for descending stair classification to centralized training (misclassifications decreased from 42 to 31), supporting its capacity to deal with non-IID heterogeneity.

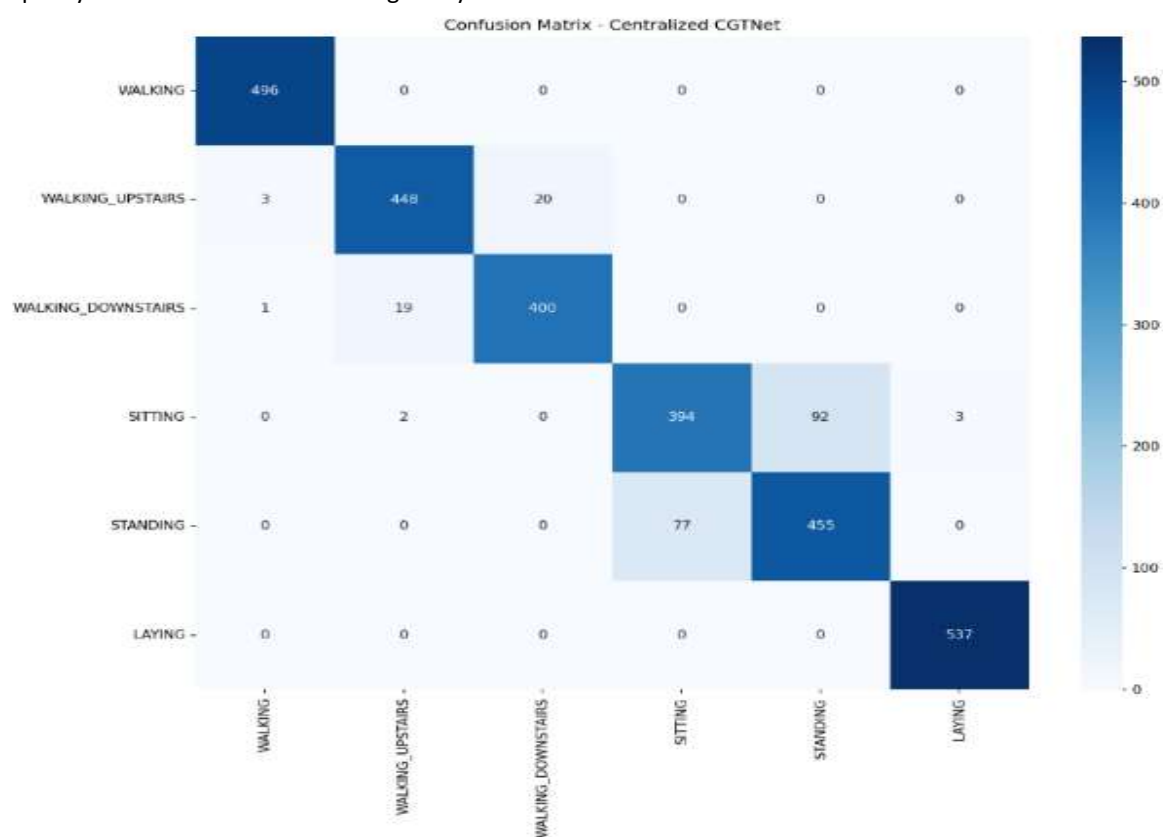


Figure 7. Confusion Matrix for centralized CGTNet (UCI HAR, six classes). High precision for all classes, slight confusion between SITTING and STANDING (92 samples) because the accelerometer signals of quasi-static postures are similar.

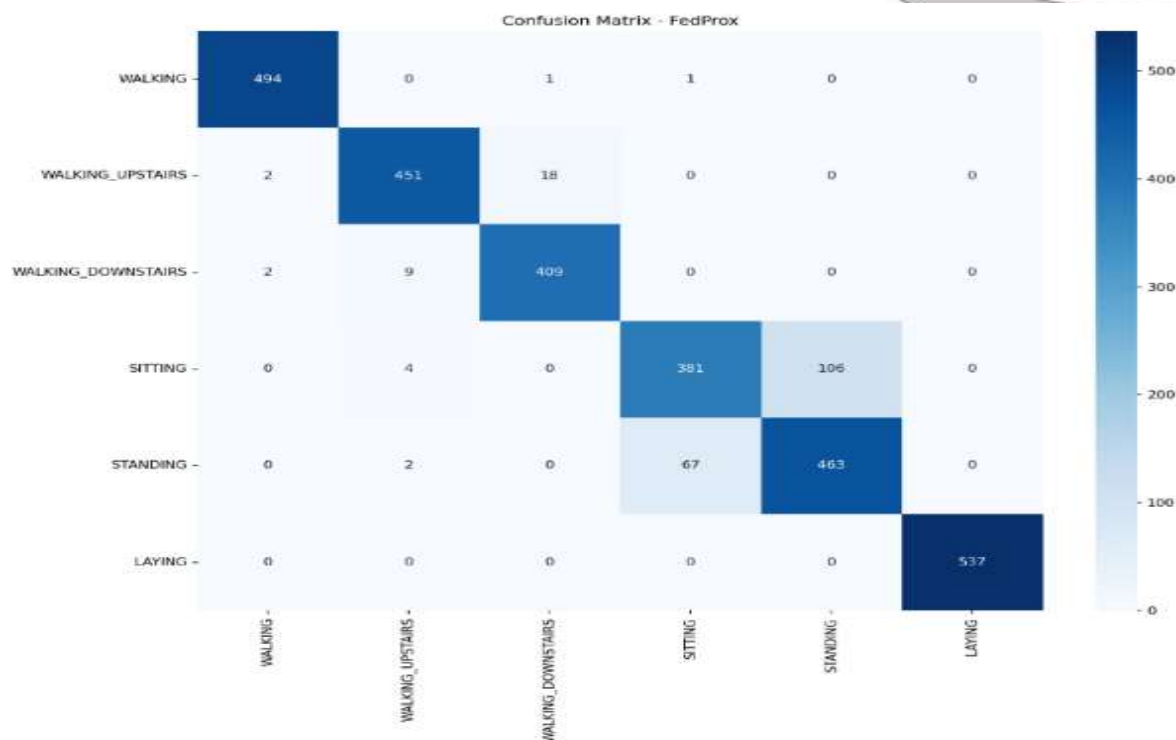


Figure 8. Confusion Matrix — FedProx (UCI HAR, six classes). Near-perfect classification of LAYING (537/537 correct); SITTING/STANDING confusion persists (106 samples), which is due to a limitation of the inertial sensor and not the fault of the algorithm.

4.5 Communication Efficiency

Table 8. Communication and Efficiency Metrics (PAMAP2)

Method	Rounds to 90% F1	Total MB Transmitted	Params (M)	Inference Latency (ms)
Centralized CGTNet	N/A	N/A	0.42	8.2
FedAvg	32	620	0.42	8.2
FedProx	28	560	0.42	8.2

Figure 9 visualises the communication-efficiency trade-off dynamically.

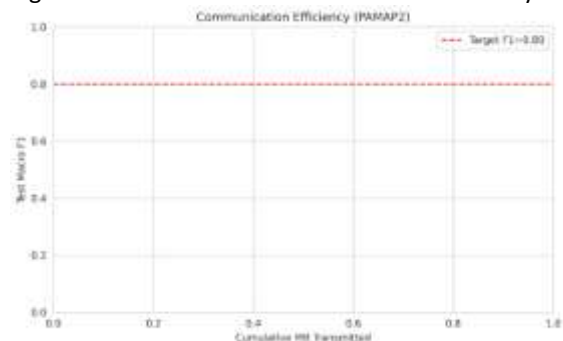


Figure 9. Communication Efficiency Curve (PAMAP2). Test Macro F1 versus cumulative MB transmitted for FedAvg and FedProx. The target threshold (dashed vertical line) was 0.80. In addition to achieving the target accuracy with approximately 15% less cumulative communication than FedAvg, FedProx offers the added benefit of being more efficient in terms of the communication bandwidth consumed for practical IoMT applications.

4.6 Ablation Study

Table 9. Ablation Study: Component Contributions (PAMAP2)

Model	Macro F1	Params (M)	Comm. Cost (MB)	Key Insight
Simple GRU	0.552	0.08	1.6	Insufficient capacity
+ LayerNorm	0.751	0.12	2.4	Training stabilisation
+ Temporal Attention	0.783	0.15	3.1	Major gain via temporal weighting
BiGRU (bidirectional)	0.801	0.24	4.8	Bidirectional context capture
+ CNN front-end	0.814	0.38	7.6	Improved spatial feature extraction
+ Focal Loss + Magnitudes (CGTNet)	0.825	0.42	8.4	Best accuracy–efficiency trade-off

The incremental ablation of the CGTNet architecture is presented in Table 9. The highest absolute gain in the Macro F1 score was obtained with temporal attention, verifying the role of temporal weighting at the center of the problem. The performance was further enhanced by replacing the unidirectional GRU with BiGRU in order to capture the dependency in both directions. The CNN front end improves spatial feature extraction from multisensor signals. The combination of focal Loss and magnitude resulted in marginal but steady improvements, suggesting that post-architecture improvements, such as focal loss and magnitude, are diminishing. The CGTNet model with full classification was found to be the optimal model in terms of classification accuracy, model complexity, and communication cost.

4.7 Discussion

Overall, the results showed that federated learning can match centralized training on clean, balanced data. FedAvg achieved a Macro F1 score of 0.927 on UCI HAR, which is slightly higher than the baseline performance of the centralized CGTNet (0.918), in line with Brecko et al. [1], who reported that FL offers implicit regularization in the edge computing setting with a relatively uniform data distribution. For PAMAP2, the performance is far from similar, with the classical SVM having the highest Macro F1 (0.843) and significantly outperforming all the variants of deep learning models. This result confirms the results of Rashid et al. [10] and Imran et al. [7], who showed that for deep temporal models to outperform well-tuned classical models on heterogeneous sensor signals, data quality and quantity are crucial.

FedProx reduced the gap between FedAvg (0.802) and SVM (0.843) by almost half on PAMAP2 with a Macro F1 score of 0.825. This not only establishes that proximal regularization helps reduce the statistical heterogeneity-induced drift of clients but also introduces the same communication overhead (12.4 MB) as FedAvg. Additionally, FedProx achieved 90% of the final F1 in 28 communication rounds, compared to 32 rounds for FedAvg, confirming the effectiveness of proximal regularization as a practical way of obtaining better convergence on heterogeneous datasets [18].

The analysis of the classes showed that FedProx had statistically significant gains for underrepresented activities, including descending stairs (+0.12 F1 over FedAvg), although they were rare in the data (1.6%). This demonstrates the ability of recurrent topologies to learn from fine temporal motion patterns that are not amenable to classical approaches [7]. The main reason for the lack of distinction between the steps on the ascending and descending sides for all models is the inherent limitations of inertial sensors, not the algorithms [3].

On PAMAP2, FedProx reached the 0.80 target threshold with a lower total communication of 15% (560 MB) compared to that of FedAvg (620 MB). This efficiency is due to faster convergence—not from compression of the model, but

rather from a more general phenomenon, which is different from knowledge distillation methods, such as FedAKD [5]. This reduction in the number of communication rounds directly results in reduced energy consumption for bandwidth and battery-constrained wearable IoMT devices [13].

4.8 Limitations and Future Directions

This study has several limitations. First, the experimental setup assumes synchronous aggregation, which is not necessarily representative of real-world IoMT networks with intermittent connectivity and varying client availability [6]. Asynchronous federated protocols should be investigated in future studies. Second, the F1 score was very low for rare classes, such as descending stairs ($F1 = 0.31$), demonstrating the pain point of federated rare event recognition. Few-shot learning at the federated level [18] and synthetic oversampling methods at the client level, such as the synthetic minority oversampling technique (SMOTE), are potential solutions. Third, the current personalization module is a post hoc, rule-based layer. Future extensions will explore the adaptive tuning of the thresholds during federated training rounds and adaptability results with actual patient cohorts to provide validity for patient-facing deployments.

5 Conclusion

In this study, a federated learning-based CNN-BiGRU-Transformer model (CGTNet-Fed) for HAR in wearable IoMT systems is proposed to ensure privacy. The framework was robustly evaluated on PAMAP2 and UCI HAR under realistic non-IID conditions, showing high accuracy (0.927 Macro F1 on UCI HAR) with strict data localities. FedProx is partially effective in reducing performance loss when the data are not IID, but it remains competitive with classical SVM baselines when the data are noisy multisensors. An activity predictor is integrated with a lightweight, rule-based personalization module to make predictions and convert them into actionable, personalized health recommendations while maintaining user privacy. The limitations identified, such as the synchronous aggregation assumption and persistent rare-class challenge, define clear directions for future research in the field, including the development of asynchronous federated protocols, federated few-shot learning, and multimodal physiological signal integration. This study builds a reproducible federated pipeline, which serves as an empirical benchmark for the federated HAR community.

Acknowledgments

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest: The authors declare no conflicts of interest.

Data Availability Statement: The datasets used in this study are publicly available. The PAMAP2 dataset is accessible at the UCI Machine Learning Repository, and the UCI HAR dataset is available at <https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones>.

Ethical Approval: This study used only publicly available, anonymized datasets. No new human or animal subjects were involved, and ethical approval was not required.

Author Contributions:

Conceptualization: Sankaramanchi Sreevani and Dr. N. Srinivasu.

Methodology: Sankaramanchi Sreevani.

Investigation: Sankaramanchi Sreevani.

Discussion of results: Sankaramanchi Sreevani and Dr. N. Srinivasu.

Writing – Original Draft: Sankaramanchi Sreevani.

Writing – Review and Editing: Sankaramanchi Sreevani and Dr. N. Srinivasu.

Supervision: Dr. N. Srinivasu.

Approval of the final text: Sankaramanchi Sreevani and Dr. N. Srinivasu.

References

- [1] A. Brecko, E. Kajati, J. Koziorek, and I. Zolotova, "Federated learning for edge computing: A survey," *Applied Sciences*, vol. 12, no. 18, p. 9124, 2022.
- [2] D. Cheng, L. Zhang, C. Bu, X. Wang, H. Wu, and A. Song, "ProtoHAR: Prototype guided personalized federated learning for human activity recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 3900–3911, 2023.
- [3] P. H. Chiang, M. Wong, and S. Dey, "Using wearables and machine learning to enable personalized lifestyle recommendations to improve blood pressure," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, 2021.
- [4] P. Climent-Perez and F. Florez-Revuelta, "Privacy-preserving human action recognition with a many-objective evolutionary algorithm," *Sensors*, vol. 22, no. 3, p. 764, 2022.

- [5] G. Gad and Z. Fadlullah, "Federated learning via augmented knowledge distillation for heterogeneous deep human activity recognition systems," *Sensors*, vol. 23, no. 1, Art. no. 6, 2023.
- [6] R. Gajanin, A. Danilenka, A. Morichetta, and S. Nastic, "Towards adaptive asynchronous federated learning for human activity recognition," in *Proc. 14th International Conference on Internet of Things (IoT)*, 2024, pp. 38–46.
- [7] H. A. Imran, Q. Riaz, M. Hussain, H. Tahir, and R. Arshad, "Smart-wearable sensors and CNN-BiGRU model: A powerful combination for human activity recognition," *IEEE Sensors Journal*, vol. 24, no. 2, pp. 1963–1974, 2024.
- [8] M. Musci, D. de Martini, N. Blago, T. Facchinetti, and M. Piastra, "Online fall detection using recurrent neural networks on smart wearable devices," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 3, pp. 1276–1289, 2021.
- [9] G. M. Nagamani and C. K. Kumar, "Design of an improved graph-based model for real-time anomaly detection in healthcare using hybrid CNN-LSTM and federated learning," *Heliyon*, vol. 10, no. 24, p. e41071, 2024.
- [10] N. Rashid, B. U. Demirel, and M. A. Al Faruque, "AHAR: Adaptive CNN for energy-efficient human activity recognition in low-power edge devices," *IEEE Internet of Things Journal*, vol. 9, no. 15, pp. 13041–13051, 2022.
- [11] F. Sabry, T. Eltaras, W. Labda, K. Alzoubi, and Q. Malluhi, "Machine learning for healthcare wearable devices: The big picture," *Journal of Healthcare Engineering*, vol. 2022, p. 4653923, 2022.
- [12] K. P. Seng, L. M. Ang, E. Peter, and A. Mmonyi, "Machine learning and AI technologies for smart wearables," *Electronics*, vol. 12, no. 7, p. 1509, 2023.
- [13] E. Torti, A. Fontanella, M. Musci, N. Blago, D. Pau, F. Leporati, and M. Piastra, "Embedding recurrent neural networks in wearable systems for real-time fall detection," *Microprocessors and Microsystems*, vol. 71, p. 102895, 2019.
- [14] X. Wen, Y. Wang, M. Yuan, Y. Geng, H. Yu, and G. Zheng, "Enhancing human activity recognition with FedPA: Focusing on non-IID data challenges in federated learning," *IEEE Sensors Journal*, vol. 24, no. 23, pp. 39230–39242, 2024.
- [15] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. 16th IEEE International Symposium on Wearable Computers (ISWC)*, Newcastle, UK, 2012, pp. 108–109.
- [16] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. European Symposium on Artificial Neural Networks (ESANN)*, Bruges, Belgium, 2013.
- [17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.
- [18] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning Systems (MLSys)*, vol. 2, pp. 429–450, 2020.
- [19] A. Almadhor et al., "Wrist-based electrodermal activity monitoring for stress detection using federated learning," *IEEE Sensors Journal*, vol. 23, no. 9, pp. 9120–9131, May 2023.