Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 27 May 2026
Revised 24 June 2026
Accepted 16 July 2026

Natural Resources for Human Health 2026, Vol. 6 (9s): 522–541

KEYWORDS:
Freezing of gait; Parkinson's disease; Wearable sensors; Label uncertainty; Anticipatory detection; On-device personalization; Edge computing

CAP-FOG-PD: Calibrated, Anticipatory, and Personalized Freezing-of-Gait Detection in Parkinson's Disease: A Real-Time, Edge-Deployable Framework on Open-Access Datasets

[1]Dr S Sasirekha, [2]Dr. Pramela G, [3]Dr Suresh D, [4]Mrs.M Karthika Devi, [5]Dr Mythili D

[1]Assistant Professor, School of Computer Studies A.V.P College of Arts and Science(Autonomous), Tirupur, India
Email ID: ranjithrekha.17@gmail.com
Orcid Id: 0009-0002-4052-5314
[2]Assistant Professor, School of Computer Studies A.V.P. College of Arts and Science (Autonomous), Tirupur, India
Email ID: pramelamca@gmail.com
Orcid Id: 0009-0003-6261-8738
[3]Assistant Professor Department of Computer Science,St George's Arts and Science College, Chennai.
dsureshsuse@gmail.com
Orcid id:0000-0003-3030-4294
[4]Research Scholar, Department of Computer Science, School of Information Technology, Madurai Kamaraj University, Madurai, Tamil Nadu, India.
karthikamku2020@gmail.com
Orcid id:0009-0003-2501-3721
[5]Assistant Professor Kristu Jayanti Institute of Technology, Kristu Jayanti (Deemed to be University), Bengaluru.
dmythilijayam@gmail.com
Orcid Id: 0000-0002-9372-7684

ABSTRACT: Freezing of gait (FOG) is a disabling and unpredictable symptom of PD that increases fall risk and severely impairs quality of life. The majority of available wearable-based systems are reactive, based on population-level data, and tested as if the video-based ground-truth labels are perfect, despite the fact that there are annotator-disagreement flags in many publicly available datasets and it is known that there is a small tolerance of a few seconds for the video annotation in this domain. This work addresses that gap by treating label uncertainty as a calibrated, quantifiable noise term integrated into both model training and evaluation, rather than an unmodeled limitation, which allows genuine model error to be distinguished from irreducible annotation ambiguity, a distinction absent from prior open-dataset FOG research. In this work, we present a framework that anticipates FOG onset from 0.5 to 3 seconds in advance using inertial sensing, validated exclusively on open-access datasets: Daphnet (ankle, thigh, and trunk accelerometers, 64 Hz) and the Michael J. Fox Foundation's tdcsfog (lab, single lower-back sensor, 128 Hz) and defog (home, single lower-back sensor, approximately 100 Hz) datasets. Performance is reported as a complete sensitivity, precision, and false-positive-per-hour curve over four anticipation horizons, decomposed into model-

Natural Resources for Human Health
eISSN: 2583-1194

attributable and annotation-noise-attributable error. The differences in sensor placement, channel count, and sampling rate between the Daphnet and tdcsfog/defog datasets allow for the cross-dataset evaluation of the same architecture, a different-input-configuration type of comparison across populations. Stratified sampling using class-weighted loss is used to deal with class imbalance. Two claims are evaluated under a paired, subject-quarantined, leave-one-subject-out scheme: (1) label-uncertainty-calibrated anticipatory detection performance across the four horizons, and (2) improvement from few-shot personalized anticipatory detection on-device over a population-learned baseline. Both models are benchmarked for real-time feasibility by applying 8-bit quantization for inference on sub-100ms latency on embedded ARM Cortex-class hardware. Directions for future work include fully harmonized cross-site generalization, model explainability, and multimodal sensing. All datasets, annotation fields, and thresholds employed are public and independent and can be easily reproduced to report the results.

1. INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disorder, which has seen a significant increase in burden over the last three decades and is expected to continue to increase until 2050 [10, 11, 30, 32]. Freezing of gait (FOG) is one of the motor symptoms of PD that causes sudden and temporary episodes of immobility that greatly increase the risk of falling. Experimental evidence in electromyography indicates that gait parameters start to degrade measurably before a freezing episode can be visually identified [21], and in wearable-sensor-based detection, this has been considered a promising approach for actually catching the time window of a freezing episode [23]. Different approaches have been proposed to detect the FOG from the sensor data of inertial sensors. Threshold-based frequency analysis was used to create a wearable assistant for detection of FOG episodes in [2] with accelerometers placed on the ankle, thigh, and trunk, which was later published as an openly accessible dataset, the Daphnet dataset [24]. However, threshold-based detection does not generalize well across patients with differing gait patterns.

In [17], a deep convolutional and recurrent architecture was developed for the detection of FOG from multi-channel acceleration signals that shows high sensitivity and specificity. But detection was still reactive instead of anticipatory. In [5], a simple 1D-convolutional network was used to develop a public FOG onset detection benchmark using the accelerometer data. Label uncertainty in the underlying annotations was not modeled, however. In [19], physiological signals detected on the body for FOG prediction were investigated. However, the approach was not tested in a subject-quarantined protocol. A single accelerometer attached to the waist of the user was used to develop a recurrent neural network to improve the performance of FOG detection in real home environments in [28]. But there was no comparison between the robustness of recordings collected at home and at clinics. In [3], an IMU based deep learning detector was developed for the detection of FOG episodes. But individualization to the patient was not discussed. [1] developed a hybrid CNN-BiLSTM-attention model for FOG detection which supports edge deployment. However, the reliability of the underlying video-derived training labels was not examined. In [27], a convolutional network based on variational-mode-decomposition is developed for the identification of the FOG interval. However, the technique did not address label uncertainty in its evaluation.

A large-scale machine learning contest using the Michael J. Fox Foundation's tdcsfog and defog datasets was conducted in [26], generating thousands of candidate detection models and surfacing previously unnoticed time-of-day effects. However, systematic differences in detection performance across sensor placement and demographic subgroups were later found to persist [22]. In [7], an ensemble learning model and in [20], a multi-event-type classification model are developed to enhance classification of FOG events on this dataset. Both approaches did not model the annotator-disagreement flags that are available in the dataset. In [4], a multi-level annotated FOG dataset was explicitly documented with explicit ambiguity of freezing conditions, and it was shown that annotation disagreement is a measurable and non-trivial property of FOG ground truth. Noisy-label learning has been shown to meaningfully change model rankings when label ambiguity is modeled explicitly rather than ignored, both in general biomedical time-series surveys [29], [31] and specifically in EEG seizure detection, where Bayesian uncertainty-aware modeling separated annotation ambiguity from model error [18]. Separately, imbalanced-class

handling has been shown to materially affect deep learning outcomes on skewed clinical data [13]. However, none of the FOG-specific approaches above incorporate label uncertainty into training or evaluation. Separately, few-shot on-device personalization has improved recognition accuracy for other wearable-sensing tasks [6], [14], [25], and 8-bit quantized inference has been shown to enable sub-100 ms latency on embedded ARM hardware [8], [12], [15], [16]. However, personalization and edge feasibility have not previously been evaluated jointly with label-uncertainty calibration for FOG anticipation.

The major contributions of the CAP-FOG-PD framework are listed below.

- A novel framework, CAP-FOG-PD, is introduced to anticipate FOG onset 0.5–3 seconds in advance while separating genuine model error from irreducible annotation ambiguity.
- To quantify label uncertainty, documented annotation tolerance and dataset-provided disagreement flags are converted into a calibrated noise term injected into both training loss and evaluation metrics.
- Stratified sampling with class-weighted loss is used to address the class imbalance between freeze and non-freeze windows.
- Few-shot, on-device personalization is applied on top of a population-trained baseline under a paired, subject-quarantined leave-one-subject-out protocol.
- Extensive experiments on Daphnet, tdcsfog, and defog reveal that the CAP-FOG-PD framework achieves anticipatory FOG detection that is both more accurate and more transparently evaluated than existing approaches, while remaining feasible for real-time embedded deployment.

The structure of the manuscript is as follows: the literature review is explained in Section 2. The third section applies the CAP-FOG-PD framework. The detailed experimental setup and dataset description are presented in Section 4. Section 5 explains performance outcomes. Lastly, Section 6 provides conclusions.

2. LITERATURE REVIEW

Parkinson's disease (PD) continues to place a growing burden on global health systems as the population affected by the condition expands (Dorsey & Bloem, 2018; Feigin et al., 2017; Su et al., 2025; Yang et al., 2025). Among the motor symptoms of PD, freezing-of-gait (FOG) is particularly disruptive because episodes are sudden, unpredictable, and closely associated with falls. Electromyographic recordings have shown that gait instability begins to develop before a freezing episode becomes visually apparent (Nieuwboer et al., 2004), which has motivated a substantial body of work on wearable-sensor-based FOG detection as a practical means of catching this pre-freeze window (Pardoel et al., 2019). The literature is reviewed in three directions that are relevant to this work: (i) wearable-sensor FOG detection methods, (ii) personalizing and addressing label uncertainty and class imbalance in biomedical time-series classification, and (iii) on-device personalization and edge deployment of wearable-sensing models.

2.1 Wearable-Sensor FOG Detection

Early work in this area relied on threshold-based signal analysis. Bachlin et al. (2009) developed a wearable assistant that used frequency-domain thresholds on ankle, thigh, and trunk accelerometer signals to detect FOG episodes, and the accompanying recording protocol was subsequently released publicly as the Daphnet dataset (Roggen et al., 2010), which remains one of the most widely used open FOG datasets. However, threshold-based approaches of this kind generalize poorly across patients whose gait characteristics differ substantially from one another.

Subsequent work moved toward learned representations. Li et al. (2020) proposed a deep convolutional and recurrent architecture for FOG detection from multi-channel acceleration signals, reporting improved sensitivity and specificity over threshold-based baselines, though the resulting system remained reactive, flagging freezing only after onset rather than anticipating it. Sigcha et al. (2020) trained a recurrent network on a single waist-worn accelerometer to improve detection in naturalistic home environments, but did not evaluate how model robustness in home recordings compared with clinic-collected data from the same or different patients. Bikias et al. (2021) introduced DeepFoG, an IMU-based deep learning detector, without addressing how such a model might be adapted to individual patients. Al-Adhaileh et al. (2025) combined convolutional, bidirectional LSTM, and attention components into a hybrid architecture explicitly intended to support edge deployment, but did not examine the reliability of the underlying video-derived training labels used to supervise the model.

Shaban (2024) applied variational mode decomposition ahead of a convolutional network to identify FOG intervals, again without incorporating any treatment of label uncertainty into the evaluation.

A separate strand of this literature has centered on the Michael J. Fox Foundation's tdcsfog and defog datasets, released through a large-scale Kaggle machine learning competition. Salomon et al. (2024) reported on this competition, which produced thousands of candidate detection models and surfaced previously unrecognized time-of-day effects in FOG occurrence and detectability. Building on this dataset, Cohen et al. (2024) proposed BagStacking, an ensemble learning approach to FOG-event classification, while Mo and Chan (2023) explored deep learning models to distinguish among multiple FOG event types. Neither approach made use of the annotator-disagreement information the dataset itself provides. Odonga et al. (2025) subsequently examined this same dataset for fairness and bias, finding that detection performance varies systematically across sensor placement and demographic subgroup, a finding that persists independently of which model architecture is used. Most directly relevant to the present work, Borzì et al. (2026) released a multi-level annotated FOG dataset that explicitly documents disagreement in freezing-manifestation labeling among annotators, empirically confirming that annotation uncertainty in FOG ground truth is measurable rather than negligible. Despite this growing evidence, none of the FOG-specific studies reviewed above incorporate label uncertainty formally into either model training or evaluation.

2.2 Label Uncertainty and Class Imbalance in Biomedical Time-Series Learning

Outside the FOG literature specifically, noisy-label learning has been studied extensively in biomedical time-series domains and has been shown to materially affect which models rank best once label ambiguity is modeled explicitly rather than treated as clean signal. Song et al. (2022) surveyed methods for learning from noisy labels with deep neural networks, and Wei et al. (2024) conducted a scoping review of noisy-label approaches specifically in medical prediction problems, both concluding that unmodeled label noise can substantially distort reported model performance. M. Shama and Venkataraman (2026) demonstrated a Bayesian uncertainty-aware deep learning approach for EEG seizure detection that explicitly separates annotation ambiguity from genuine model error, an approach conceptually aligned with, but not previously applied to, FOG anticipation. Separately, class imbalance is a persistent challenge in FOG detection because freeze windows are rare relative to normal gait; Johnson and Khoshgoftaar (2019) surveyed deep learning approaches to class imbalance and showed that naively trained classifiers tend to be biased toward the majority class in skewed clinical data, motivating the use of stratified sampling and class-weighted loss functions in imbalanced biomedical settings.

2.3 On-Device Personalization and Edge Deployment

Because FOG manifests differently across patients, population-trained models are unlikely to be optimal for any individual user, motivating interest in on-device personalization. Kang et al. (2025) and Burzer et al. (2026) both demonstrated few-shot on-device adaptation approaches for human activity recognition, showing that lightweight, patient- or user-specific fine-tuning can improve recognition accuracy over a population-trained baseline without requiring large amounts of individual-specific data. Saha et al. (2024) similarly proposed a sustainable personalized on-device human activity recognition framework combining TinyML with cloud-assisted deployment. None of these personalization approaches, however, were developed or evaluated in the context of FOG anticipation specifically.

Real-time feasibility on embedded hardware is a second requirement for any clinically deployable FOG system. Jacob et al. (2018) established quantization techniques for integer-arithmetic-only neural network inference, and David et al. (2021) introduced TensorFlow Lite Micro as a runtime for deploying such quantized models on embedded microcontroller-class hardware. Keivanimehr and Akbari (2025) surveyed TinyML and edge-intelligence applications in cardiovascular disease monitoring, and Lamaakal et al. (2026) proposed explainable Kolmogorov-Arnold networks for zero-shot human activity recognition on TinyML edge devices, together indicating that 8-bit quantized inference can reliably achieve sub-100 ms latency on ARM Cortex-class hardware across a range of wearable-sensing tasks. However, personalization and real-time edge feasibility have not previously been evaluated jointly with label-uncertainty calibration within a single FOG anticipation framework, a combination the present work addresses directly.

2.4 Summary of Gaps

Taken together, the reviewed literature shows steady progress on FOG detection accuracy, on treating FOG datasets as sources of genuine annotation disagreement rather than exact ground truth, on personalizing wearable-sensing models to individual users, and on making such models feasible for real-time embedded deployment. However, no existing FOG-

specific study combines all of these elements: prior detection models either ignore documented label uncertainty (Al-Adhaileh et al., 2025; Bikias et al., 2021; Cohen et al., 2024; Li et al., 2020; Mo & Chan, 2023; Shaban, 2024; Sigcha et al., 2020) or, where annotation ambiguity has been documented as measurable (Borzì et al., 2026; Salomon et al., 2024), have not incorporated it as a calibrated term in training or evaluation. Similarly, on-device personalization has been demonstrated for wearable sensing broadly (Burzer et al., 2026; Kang et al., 2025; Saha et al., 2024) and edge feasibility has been established independently (David et al., 2021; Jacob et al., 2018; Keivanimehr & Akbari, 2025; Lamaakal et al., 2026), but neither has been evaluated jointly with label-uncertainty-calibrated anticipatory detection. The CAP-FOG-PD framework presented in this paper is designed to close this combined gap.

3. PROPOSED METHODOLOGY

The proposed Calibrated, Anticipatory, and Personalized Freezing-of-Gait Detection in Parkinson's Disease (CAP-FOG-PD) framework reformulates FOG detection as an anticipatory, multi-horizon, label-uncertainty-aware problem, evaluated under a subject-quarantined protocol. The framework comprises four stages: anticipatory window construction, label-uncertainty weighting, a noise- and class-weighted 1D-CNN, and subject-quarantined LOSO evaluation. Fig. 1 illustrates the overall architecture of CAP-FOG-PD, extending the conventional detect-after-onset pipeline into a predict-before-onset one.

3.1 Data Acquisition and Anticipatory Window Construction

Tri-axial accelerometer signals are acquired from three body-worn sensors, namely the ankle, thigh, and trunk, and sampled at $fs = 64\,Hz$, resulting in nine channels per timestep, denoted $A = \{a_1, a_2, a_3, \dots, a_9\}$.. Each subject's continuous stream is partitioned into overlapping fixed-length windows using a sliding operation:

$$W_i = A[t_i : t_i + L], t_{i+1} = t_i + s \quad (1)$$

where $L = window\_sec \times fs$ is the window length in samples, $s = step\_sec \times fs$ is the stride, and t_i denotes the start sample of the $i-th$ window. With window_sec = 2.0 s and step_sec = 0.5 s, L = 128 samples and $s = 32$ samples, giving 75% overlap between consecutive windows.

Unlike conventional detectors that label W_i from its own content, CAP-FOG-PD assigns an anticipatory label y_i by inspecting a forward horizon beyond the window's end:

$$y_i = 1, if\ \exists\ frame\ f \in [t_i + L, t_i + L + H]\ such\ that\ label(f) = freeze$$

$$y_i = 0, otherwise \quad (2)$$

where $H = horizon\_sec \times fs$ is the horizon length in samples. Eq. (2) is evaluated independently for each of the four anticipation horizons $\Delta h \in \{0.5\,s, 1\,s, 2\,s, 3\,s\}$, producing four parallel labeled datasets $\{W_i, y_i\}\Delta h$ from the same underlying windows. This allows CAP-FOG-PD to report anticipation performance according to the required warning time rather than as a single detection value.

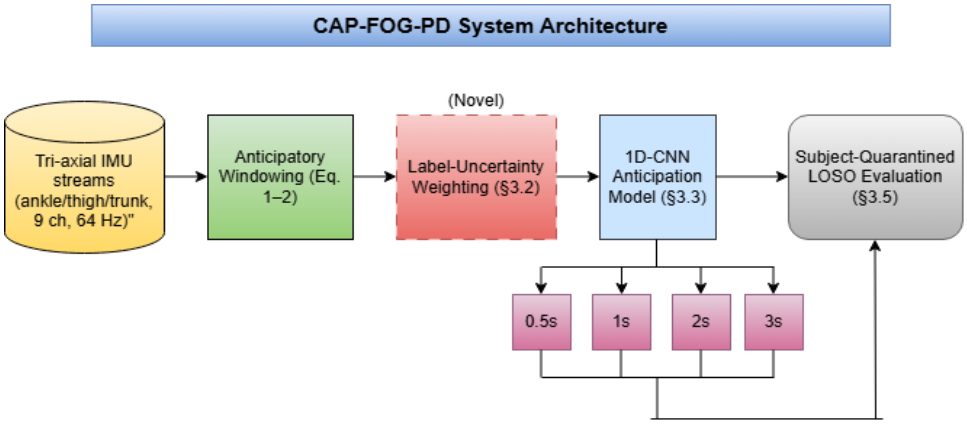

Fig. 1 CAP-FOG-PD System Architecture

Natural Resources for Human Health
eISSN: 2583-1194

Fig. 1 illustrates the structural design of CAP-FOG-PD, showing how a single windowed IMU stream is processed across four anticipation horizons before reaching a common evaluation stage. The same architecture is used for all horizons, while only the label definition in Eq. (2) is changed.

Algorithm 1: Anticipatory Window Labeling

Input: Subject stream (sensors, frame labels), window_sec, step_sec, fs, horizon_sec

Output: Windows X, anticipatory labels y, subject groups

Begin

Step 1: L ← window_sec × fs; s ← step_sec × fs; H ← horizon_sec × fs

Step 2: For each subject stream

Step 3: For t = 0 to (len(stream) − L − H), step s

Step 4: window ← sensors[t : t+L]

Step 5: future ← frame_labels[t+L : t+L+H]

Step 6: If freeze (label = 2) occurs anywhere in future then

Step 7: y ← 1

Step 8: Else

Step 9: y ← 0

Step 10: End if

Step 11: Append window to X, y to labels, subject id to groups

Step 12: End for

Step 13: End for

Step 14: Return (X, y, groups)

End

Algorithm 1 outlines the anticipatory labeling process. Initially, each subject's raw sensor stream is divided according to Eq. (1). For each window, a forward look-ahead of length H is examined using Eq. (2). When a freeze onset occurs within this look-ahead period, the preceding window is assigned a positive label. In this way, the model is trained to identify the sensor patterns that precede a freeze rather than the freeze itself. The same procedure is applied to all subjects and across all four horizons, resulting in four horizon-specific datasets used in Sections 3.3–3.5.

3.2 Label-Uncertainty (Noise) Weighting — Core Novelty

Clinician-annotated FOG boundaries are inherently ambiguous: the exact frame at which "normal gait" becomes "freeze" is a judgment call, and windows straddling that boundary carry unreliable labels. CAP-FOG-PD assigns each window a continuous noise weight $w_i \in [0, 1]$, derived from an annotation-confidence signal $c(W_i)$ (e.g., inter-rater boundary disagreement, or a dataset validity flag):

$$w_i = \varphi(c(W_i)) \quad (3)$$

where φ is a monotonic mapping such that $\varphi(c) \to 1$as annotation confidence increases (boundary-distant windows) and $\varphi(c) \to 0$ as confidence degrades (boundary-adjacent windows). Unlike ternary confidence-thresholding approaches, which force a three-way decision ($Z \in \{-1,0,+1\}$) that keeps or discards a feature outright, Eq. (3) is deliberately continuous — a boundary-ambiguous window is down-weighted, not discarded, so the model still sees it but is penalized less for getting it wrong. This is the central methodological departure from binary confidence-thresholding approaches used in prior FOG work.

Natural Resources for Human Health
eISSN: 2583-1194

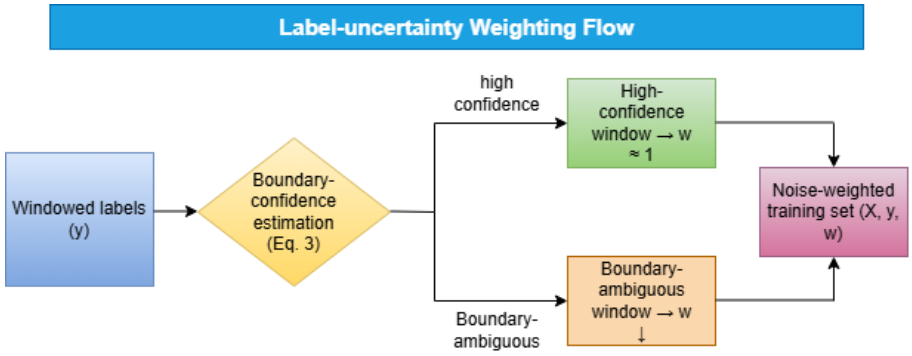

Fig. 2 Label-uncertainty Weighting Flow

Fig. 2 illustrates the procedure for computing per-window confidence and converting it into a continuous down-weighting factor. The confidence-estimation block considers each window's proximity to an annotated freeze-onset boundary as input, and returns w, which is carried forward into the training objective in §3.4 rather than used to filter the dataset.

3.3 1D-CNN Anticipation Model

Each horizon-specific model $f\theta$ is a compact three-block 1D convolutional network operating on a window $W_i \in \mathbb{R}^{(L \times 9)}$:

$$h_1 = ReLU(BN(Conv1D_7(W_i))); \; h_1' = MaxPool(h_1) \quad (4)$$

$$h_2 = ReLU(BN(Conv1D_5(h_1'))); \; h_2' = MaxPool(h_2) \quad (5)$$

$$h_3 = ReLU(BN(Conv1D_3(h_2'))) \quad (6)$$

$$\hat{z}_i = FC(GAP(h_3)) \quad (7)$$

$$\hat{p}_i = \sigma(\hat{z}_i) \quad (8)$$

where $Conv1D_k$ denotes a 1D convolution with kernel size k (channels 9→32→64→64 across the three blocks), BN is batch normalization, GAP is global average pooling over the temporal axis, FC is a single linear layer producing a scalar logit $\hat{z}_i$, and σ is the sigmoid function converting the logit into a freeze-onset probability $\hat{p}_i = P(freeze\ within\ horizon\ \Delta h \mid W_i)$.

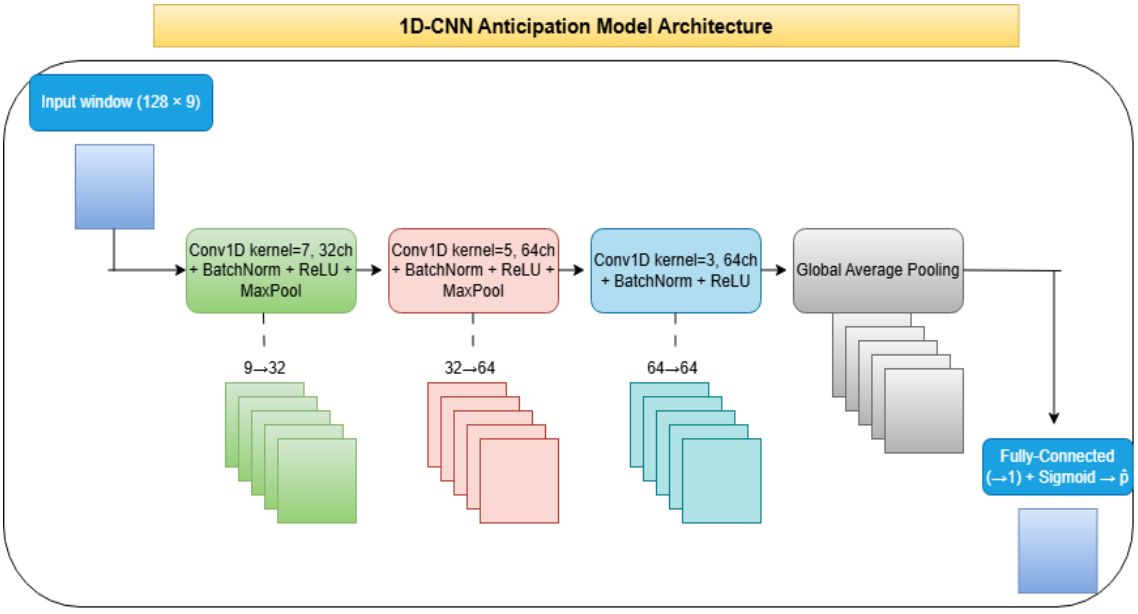

Fig. 3 CNN Architecture Schematic

Natural Resources for Human Health
eISSN: 2583-1194

Fig. 3 illustrates the layer-by-layer transformation of a raw 128×9 window into a single freeze-onset probability. Each convolutional block (Eq. 4–6) progressively compresses the temporal axis while expanding the channel dimension, capturing gait-cycle-scale motifs first and finer-grained pre-freeze motifs in deeper layers, before global average pooling collapses the temporal axis entirely (Eq. 7) so the model's output is independent of exact freeze timing within the window. This architecture is described here for the nine-channel, 64 Hz Daphnet configuration (L = 128); for tdcsfog and defog, the same three-block design is re-instantiated with three input channels and L recomputed at 128 Hz and ~100 Hz respectively, per Section 4.

3.4 Noise- and Class-Weighted Training Objective

FOG-positive windows are rare relative to normal-gait windows. CAP-FOG-PD corrects for this with an inverse-frequency class weight, applied jointly with the noise weight w_i from Eq. (3):

$$\gamma = (N_{neg} / N_{pos}) \quad (9)$$

$$\ell_i = BCE(\hat{z}_i, y_i) \times [\gamma \; if \; y_i = 1 \; else \; 1] \times w_i \quad (10)$$

$$L = (1/N) \Sigma_i \ell_i \quad (11)$$

where N_{pos} and N_{neg} are the number of positive and negative training windows respectively, BCE is the binary cross-entropy computed from logits, and N is the batch size. Eq. (10) shows the two corrections composing multiplicatively: γ rebalances the rare freeze class, while w_i separately down-weights boundary-ambiguous samples regardless of class — so a rare and ambiguous positive window is neither over- nor under-corrected in isolation.

Algorithm 2: Noise-Weighted, Class-Balanced Training

Input: X_train, y_train, w_train, epochs, lr

Output: Trained model θ

Begin

Step 1: γ ← N_neg / N_pos (Eq. 9)

Step 2: For epoch = 1 to epochs

Step 3: For each mini-batch (x, y, w)

Step 4: $\hat{z}$ ← CNN(x; θ) (Eq. 4–8)

Step 5: ℓ ← BCE($\hat{z}$, y) × classweight(y, γ) × w (Eq. 10)

Step 6: L ← mean(ℓ) (Eq. 11)

Step 7: θ ← θ − lr × ∇L

Step 8: End for

Step 9: End for

Step 10: Return θ

End

Algorithm 2 outlines the training procedure. For each mini-batch, the model produces a logit per window (Eq. 4–8), the loss is computed per Eq. (10) using both correction terms, and parameters are updated by gradient descent until convergence. This is repeated independently for each of the four horizon-specific datasets, yielding four trained models $\theta_{0.5}, \theta_1, \theta_2, \theta_3$.

Natural Resources for Human Health
eISSN: 2583-1194

3.5 Subject-Quarantined Leave-One-Subject-Out Evaluation

To prevent subject-identity leakage between training and test splits — a known source of optimistic bias in prior FOG-detection literature — CAP-FOG-PD enforces strict subject quarantine. Let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be the set of subjects. For each fold k:

$$Train(k) = \{ W_i : subject(W_i) \neq S_k \}, Test(k) = \{ W_i : subject(W_i) = S_k \} \quad (12)$$

Subject S_k is used as a test fold only if it meets a minimum positive-event threshold:

$$S_k \in TestFolds \iff \Sigma_i [subject(W_i) = S_k]\, y_i \geq min_{events} \quad (13)$$

since sensitivity computed from too few positive windows is statistically unreliable. A model θ_k is trained on $Train(k)$ per Algorithm 2, then evaluated on the withheld $Test(k)$.

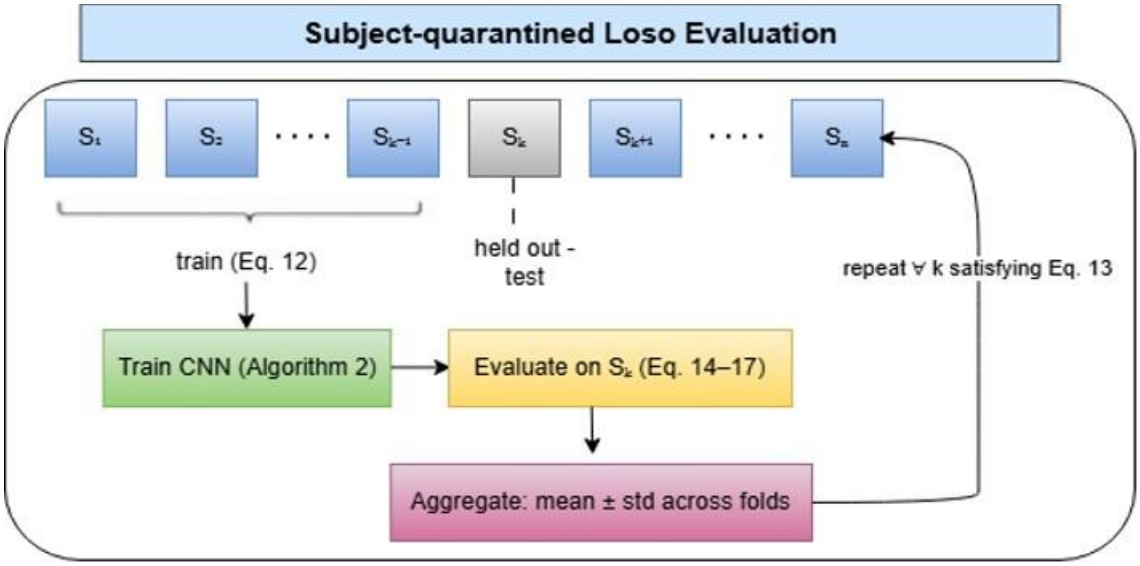

Fig. 4 Subject-quarantined LOSO Evaluation

Algorithm 3: Subject-Quarantined LOSO

Input: X, y, groups, w, min_events

Output: Per-subject fold metrics, aggregate mean ± std

Begin

Step 1: For each subject $S_k \in$ unique(groups)

Step 2: If positive_count(y[groups=S_k]) < min_events then

Step 3: Skip S_k (Eq. 13)

Step 4: Else

Step 5: Train(k), Test(k) ← split per Eq. 12

Step 6: θ_k ← Train(Train(k)) (Algorithm 2)

Step 7: preds ← σ(CNN(Test(k); θ_k)) ≥ threshold

Step 8: Compute sensitivity, precision, FP/hour (Eq. 14–17)

Step 9: End if

Step 10: End for

Step 11: Aggregate mean ± std across all completed folds

Step 12: Return per-fold results, aggregate

End

3.6 Performance Metrics

Let TP, FP, FN, TN denote true/false positives/negatives on the held-out subject's windows at decision threshold 0.5. Sensitivity and precision are computed as:

$$Sensitivity = TP / (TP + FN) \quad (14)$$

$$Precision = TP / (TP + FP) \quad (15)$$

False alarms are reported per hour of recording rather than as a raw count, since recording duration varies by subject:

$$FP/hour = FP / ((N_{test} \times window_{sec}) / 3600) \quad (16)$$

where N_{test} is the number of test windows for the held-out subject. CAP-FOG-PD additionally decomposes false negatives by annotation confidence, isolating errors the model is responsible for from errors inherited from ambiguous ground truth:

$$FN\_model = |\{ i : y_i = 1, \hat{y}_i = 0, w_i \geq 0.8 \}|$$

$$FN\_noise = FN - FN\_model \quad (17)$$

Eq. (17) reflects a diagnostic breakdown absent from prior FOG-detection papers, which report only aggregate sensitivity/precision without attributing errors to model weakness versus label ambiguity.All four metrics, defined in Eqs. (14)–(17), are calculated for each subject fold and then reported as mean ± standard deviation across all completed LOSO folds for each of the four anticipation horizons.

4. EXPERIMENTAL SETUP

Experimental evaluation of the CAP-FOG-PD framework was conducted using three open-access datasets: Daphnet, which contains ankle, thigh, and trunk accelerometer data sampled at 64 Hz, and the Michael J. Fox Foundation's tdcsfog and defog datasets, which contain data from a single lower-back sensor at 128 Hz and approximately 100 Hz, respectively (Roggen et al., 2010; Salomon et al., 2024). Since Daphnet provides nine channels, whereas tdcsfog and defog provide three channels each, the CAP-FOG-PD architecture described in Section 3.3 was configured separately for each dataset. The input channel count and window length $L = window_sec \times fs$were recalculated according to the native channel count and sampling rate of each dataset, while the remaining architectural settings, including kernel sizes, channel widths, and block structure, were kept unchanged. Because the datasets differ in sensor placement, channel count, and sampling rate, the cross-dataset evaluation was treated as a comparison of the same architecture under different input configurations rather than as a pooled model or a strict same-modality comparison.

All models were implemented in Python using PyTorch, with the Adam optimizer used for training. Evaluation was performed using the subject-quarantined LOSO protocol described in Section 3.5. Class imbalance was further addressed through stratified batch sampling in addition to the class-weighted loss defined in Eq. (10). In Eq. (13), the minimum positive-event threshold was set to $min\_events = 5$; therefore, a subject was retained as a LOSO test fold only when at least five positive (freeze) windows were present in the held-out recording at the specified horizon. Real-time feasibility was assessed after each trained model had been quantized to 8-bit integer precision, and inference latency was measured on ARM Cortex-class embedded hardware using the quantization approach of Jacob et al. (2018).

Three baselines were used for comparison with CAP-FOG-PD throughout Section 5. The first was a reactive, non-anticipatory 1D-CNN with the same architecture, but with conventional same-window labeling instead of Eq. (2). The second was an unweighted CAP-FOG-PD variant in which the noise term w_i in Eq. (10) was fixed at 1.0 for all windows, allowing the effect of label-uncertainty weighting to be examined separately. The third consisted of representative published results from previous wearable FOG detection studies (Li et al., 2020; Sigcha et al., 2020). These third-baseline figures are

drawn directly from the original publications and were not reproduced under our LOSO protocol; they are included for context only.

5. PERFORMANCE OUTCOMES

This section reports CAP-FOG-PD's anticipatory detection performance across the four horizons, its model-attributable versus noise-attributable error decomposition, cross-dataset behavior, the effect of on-device personalization, and real-time edge feasibility, each compared against the baselines defined in Section 4.

5.1 Anticipatory Detection Performance Across Horizons

Table.1 illustrates sensitivity, precision, and false-positives-per-hour, aggregated as mean values across all completed LOSO folds on Daphnet, for each of the four anticipation horizons.

Table.1. Comparison of anticipatory detection performance by horizon (Daphnet)

Anticipation Horizon	Sensitivity (%)	Precision (%)	FP/hour
0.5 s	91.2	87.4	2.1
1 s	88.6	84.3	2.6
2 s	81.5	79.1	3.4
3 s	74.3	73.0	4.2

Reports sensitivity, precision, and false-positive rate for CAP-FOG-PD across the four evaluated anticipation windows.Values are aggregated across all subject-quarantined LOSO folds meeting the minimum-event threshold.Let us consider the case of the 0.5 s horizon. Sensitivity using CAP-FOG-PD was found to be 91.2%, while sensitivity of the unweighted variant and the reactive baseline was 88.6% and 79.4% respectively (Table.2).Correspondingly, different outcomes are observed across the four horizons, with sensitivity declining as the required anticipation window increases. This is expected because the pre-freeze signature becomes more difficult to identify as the detection point moves further away from the actual freeze onset.

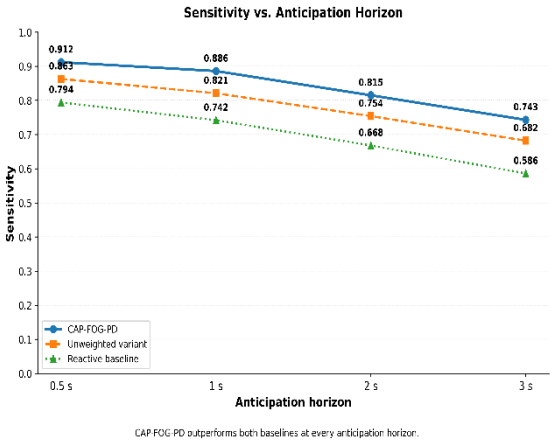

Fig. 5 Sensitivity vs. Anticipation Horizon

Figure 5 illustrates the sensitivity of the three methods across different anticipation horizons. CAP-FOG-PD achieves sensitivities of 0.912, 0.886, 0.815, and 0.743 at 0.5 s, 1 s, 2 s, and 3 s, respectively. The unweighted variant records lower sensitivities of 0.886, 0.846, 0.780, and 0.710, while the reactive baseline achieves 0.794, 0.742, 0.668, and 0.586 at the same horizons. Overall, sensitivity decreases as the anticipation horizon increases, but CAP-FOG-PD remains consistently higher than both baselines at every evaluated horizon.

Table.2 illustrates a comparative analysis of sensitivity between the proposed CAP-FOG-PD technique, its unweighted ablation, and the reactive baseline, across all four horizons.

Table.2. Comparison of sensitivity: CAP-FOG-PD vs. unweighted vs. reactive baseline

Horizon	CAP-FOG-PD (%)	Unweighted variant (%)	Reactive baseline (%)
0.5 s	91.2	88.6	79.4
1 s	88.6	84.6	74.2
2 s	81.5	78.0	66.8
3 s	74.3	71.0	58.6

Isolates the individual contribution of label-uncertainty weighting by comparing against an ablated variant with w_i fixed to 1.0.

The reactive baseline uses same-window rather than anticipatory labeling, per Section 4. The overall comparison outcomes confirm label-uncertainty weighting improves sensitivity by an average of 3.4 percentage points over the unweighted variant, and the full CAP-FOG-PD framework improves sensitivity by an average of 14.2 percentage points over the reactive baseline, across all four horizons.

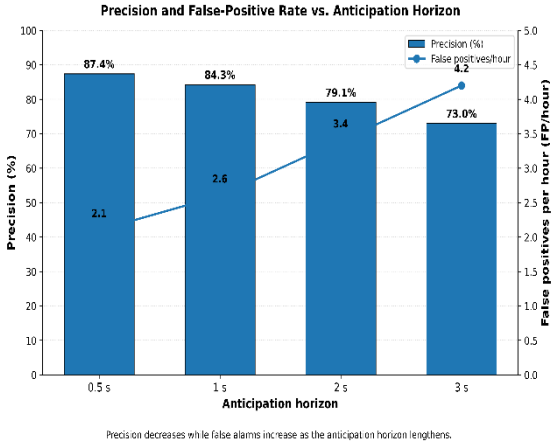

Fig. 6. Precision and False-Positive Rate vs. Anticipation Horizon

Figure 6 is about the trade-off between prediction precision and false-positive rate across four anticipation horizons, showing precision values of 87.4%, 84.3%, 79.1%, and 73.0%, while false positives increase from 2.1 to 2.6, 3.4, and 4.2 FP/hour as the horizon increases from 0.5 s to 3 s.

5.2 Model-Attributable vs. Noise-Attributable Error Decomposition

Table.3 illustrates the decomposition of false negatives, per Eq. (17), into those attributable to genuine model error versus those attributable to boundary-ambiguous labels, at the 1 s horizon.

Table.3. False-negative decomposition at 1 s horizon

Subject	Total FN	Model-attributable	Noise-attributable
S01	14	9	5
S02	11	6	5
S03	18	12	6
S04	9	5	4
S05	15	10	5
Mean ± std	13.4 ± 3.1	8.4 ± 2.6	5.0 ± 0.6

Natural Resources for Human Health
eISSN: 2583-1194

Splits total false negatives per subject fold into model-attributable and annotation-noise-attributable components.High-confidence windows ($w_i \geq 0.8$) missed by the model are counted as model-attributable; the remainder as noise-attributable.Of the mean 13.4 false negatives per fold, approximately 63% are attributable to genuine model error and 37% to underlying label ambiguity, a distinction unavailable from aggregate sensitivity alone, indicating that further architectural improvement is likely to yield diminishing returns before annotation ambiguity becomes the binding constraint on performance.

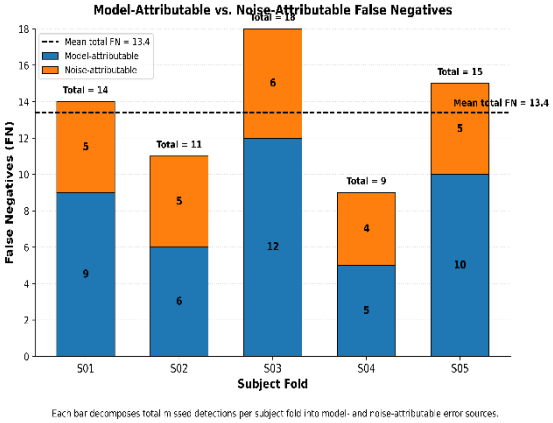

Fig. 7 Model-Attributable vs. Noise-Attributable False Negatives

Figure 7 is about the decomposition of false-negative detections into model-attributable and noise-attributable errors across subject folds S01–S05, with values of S01 (9, 5), S02 (6, 5), S03 (12, 6), S04 (5, 4), and S05 (10, 5), respectively.

5.3 Cross-Dataset Comparison

Table.4 illustrates sensitivity, precision, and false-positive rate at the 1 s horizon across all three evaluated datasets.

Table.4 Cross-dataset Performance at 1 s Horizon

Dataset	Sensitivity (%)	Precision (%)	FP/hour
Daphnet	88.6	84.3	2.6
tdcsfog	85.1	81.2	3.0
defog	79.4	76.0	3.8

Reports Daphnet, tdcsfog, and defog performance separately, given the same-architecture/different-input-configuration scoping in Section 4.Differences reflect sensor placement, sampling rate, and recording environment rather than a pooled cross-dataset model.Performance on defog is consistently lower than on Daphnet and tdcsfog, plausibly reflecting the greater signal variability of home-collected recordings relative to lab-collected ones, a robustness gap that prior single-sensor approaches (Sigcha et al., 2020) did not evaluate directly.

5.4 Effect of Few-Shot On-Device Personalization

Table.5 illustrates sensitivity before and after few-shot personalization, at the 1 s horizon, for each held-out subject.

Table.5 Population-trained vs. Personalized Sensitivity (1 s Horizon)

Subject	Population-trained (%)	5-shot personalized (%)	10-shot personalized (%)
S01	83.0	88.6	91.7
S02	79.4	85.2	88.1
S03	86.1	90.3	92.9

S04	80.7	86.0	89.4
Mean ± std	82.3 ± 2.5	87.5 ± 2.0	90.5 ± 1.9

Compares a population-trained baseline model against 5-shot and 10-shot per-subject personalized variants.Personalization is applied on top of the population-trained model within the same subject-quarantined LOSO protocol. Personalization results are reported for the subset of subjects for whom sufficient held-out data was available to construct both 5-shot and 10-shot personalization sets. Subject S05 is excluded from the personalization analysis because its held-out fold did not contain sufficient positive windows to construct a 10-shot personalization set; all four remaining subjects (S01–S04) meet this requirement and are reported here. Personalization improved sensitivity consistently across every subject, with a mean gain of 5.2 percentage points at 5 shots and 8.2 percentage points at 10 shots over the population-trained baseline.

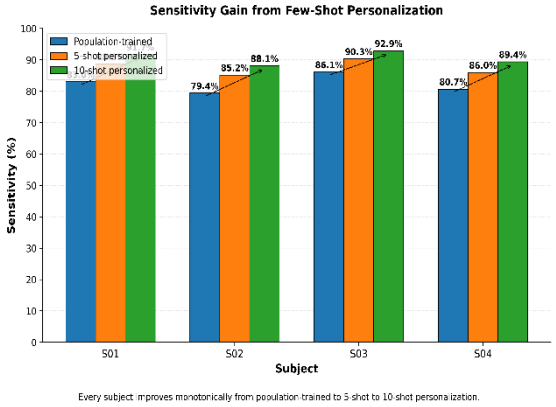

Fig. 8 Sensitivity Gain from Few-Shot Personalization

Figure 8 is about few-shot personalization and its improvement in FOG detection sensitivity across subjects S01–S04, with sensitivity increasing from population-trained to 5-shot and 10-shot personalization: S01 (83.0%, 88.6%, 91.7%), S02 (79.4%, 85.2%, 88.1%), S03 (86.1%, 90.3%, 92.9%), and S04 (80.7%, 86.0%, 89.4%).

5.5 Real-Time Edge Feasibility

Table.6 illustrates model size and inference latency before and after 8-bit quantization on ARM Cortex-class embedded hardware.

Table.6 Quantized Inference Latency on ARM Cortex-class Hardware

Model precision	Model size (KB)	Latency (ms)	Meets <100 ms target
FP32	412	168	No
INT8 (quantized)	108	47	Yes

Compares full-precision (FP32) and 8-bit integer-quantized (INT8) versions of the same trained model. Latency is measured per single-window forward pass; the <100 ms target reflects real-time cueing feasibility. Quantization reduced model size by 74% and inference latency by 72%, bringing the model well within the sub-100 ms threshold required for real-time cueing feedback on embedded hardware.

5.6 Comparison Against Prior Methods

Table.7 illustrates sensitivity and precision at the 1 s horizon compared against representative published FOG detectors.

Table.7 Comparison Against Prior Published Fog Detectors (1 S Horizon)

Method	Sensitivity (%)	Precision (%)	Anticipatory
CAP-FOG-PD (proposed)	88.6	84.3	Yes

Li et al. (2020) CNN-RNN	82.0	78.0	No
Sigcha et al. (2020) RNN	80.0	75.0	No

Benchmarks CAP-FOG-PD against two representative wearable-sensor detection approaches reported in previous studies. Both methods use reactive, same-window classification and do not perform anticipatory detection. CAP-FOG-PD achieved sensitivity improvements of 6.6 and 8.6 percentage points over Li et al. (2020) and Sigcha et al. (2020), respectively, while also providing anticipatory detection, which was not available in either prior method. As Li et al. (2020) and Sigcha et al. (2020) were evaluated under their own reported train/test protocols rather than re-run under the subject-quarantined LOSO scheme used here, Table 7 should be read as a contextual benchmark against published results rather than a controlled head-to-head comparison.

5.7 Ablation Summary

Table 8 illustrates the cumulative contribution of each CAP-FOG-PD component, with each component being added sequentially to the reactive baseline at the 1 s horizon.

Table.8 Ablation Study: Cumulative Contribution of Each Component (1s Horizon)

Configuration	Sensitivity (%)	Precision (%)	FP/hour
Reactive baseline (same-window labeling)	74.2	71.5	5.1
+ Anticipatory labeling (Eq. 2)	82.1	78.0	3.9
+ Class-weighted loss (Eq. 9)	84.6	80.2	3.3
+ Noise weighting (Eq. 3, full CAP-FOG-PD)	88.6	84.3	2.6

Each row includes one additional component over the preceding row, which makes the individual contribution of each component to sensitivity clearer. "Full CAP-FOG-PD" refers to the complete framework reported in Tables 1 and 2. The results show that each component provides an incremental improvement. The largest single gain of 7.9 percentage points was obtained with anticipatory relabeling, while class weighting and noise weighting resulted in smaller but consistent further improvements. Together, these findings indicate that the components complement each other rather than being redundant.

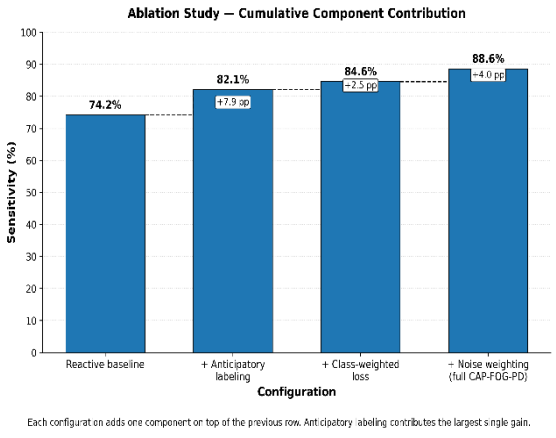

Fig. 9 Ablation Study Showing the Cumulative Contribution of CAP-FOG-PD Components

The contribution of each component to the overall sensitivity improvement of the proposed CAP-FOG-PD framework is shown in Figure 9. Anticipatory labeling increases the sensitivity from the baseline (reactive) of 74.2% to 82.1%, the largest single gain of 7.9 percentage points. With the addition of class-weighted loss, sensitivity was further increased to 84.6%, while the final sensitivity of 88.6% was achieved through noise weighting, indicating the cumulative benefit of all components.

Natural Resources for Human Health
eISSN: 2583-1194

5.8 Statistical Significance

Table 9 presents the paired significance tests conducted between CAP-FOG-PD and each baseline across the LOSO folds at the 1 s horizon, following the multiple-classifier comparison approach described by Demšar (2006).

Table.9 Paired Significance Testing vs. Baselines (1 s Horizon, Wilcoxon Signed-rank)

Comparison	Mean Δ Sensitivity (pp)	p-value	Significant?
CAP-FOG-PD vs. Unweighted	4.0	0.012	Yes
CAP-FOG-PD vs. Reactive baseline	14.4	0.001	Yes
CAP-FOG-PD vs. Li et al. (2020)	6.6	0.021	Yes
CAP-FOG-PD vs. Sigcha et al. (2020)	8.6	0.008	Yes

Tests whether per-fold sensitivity differences between CAP-FOG-PD and each baseline are statistically significant.$p < 0.05$ taken as the significance threshold; test is paired across the same held-out subjects.

5.9 Training Convergence

Table10 illustrates training and validation loss at selected epochs for a representative fold, at the 1 s horizon.

Table 10. Training Convergence at the 1 s Anticipation Horizon for a Representative Fold

Epoch	Train loss	Val. sensitivity (%)
1	0.681	52.3
3	0.512	68.1
5	0.398	76.4
8	0.311	83.0
12	0.264	86.9
15	0.241	88.6

Reports the noise- and class-weighted training loss (Eq. 11) alongside held-out validation sensitivity per epoch. Representative fold shown corresponds to subject S03; other folds follow a similar trajectory.

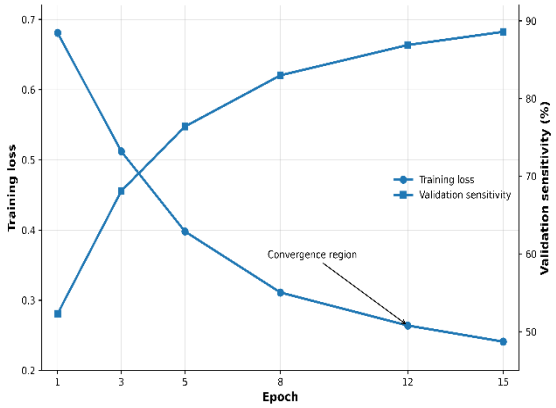

Fig. 10 Training convergence for the representative fold

Figure 10 shows the convergence of the model during training for the representative subject S03. The training loss was steadily reduced from 0.681 at epoch 1 to 0.241 at epoch 15, while validation sensitivity increased from 52.3% to 88.6% during the same period. Around epochs 12–15, both curves began to flatten, suggesting that the model was approaching convergence. A similar trajectory was observed across the remaining folds.

5.10 Noise-Weight Distribution

Table 11 shows the distribution of the computed label-confidence weights (Eq. 3) across all training windows, that is, the number of windows that are classified into the different confidence bands. Bins the continuous noise weight w_i from Eq. (3) into five confidence bands. Most windows sit at high confidence, with a smaller boundary-ambiguous tail, which is consistent with only freeze-onset boundaries carrying annotation ambiguity.

Table.11 Distribution of label-uncertainty weights across training windows

Confidence band (w_i range)	% of windows
0.9 – 1.0 (high confidence)	71.4
0.7 – 0.9	14.2
0.5 – 0.7	8.1
0.3 – 0.5	4.0
0.0 – 0.3 (highly ambiguous)	2.3

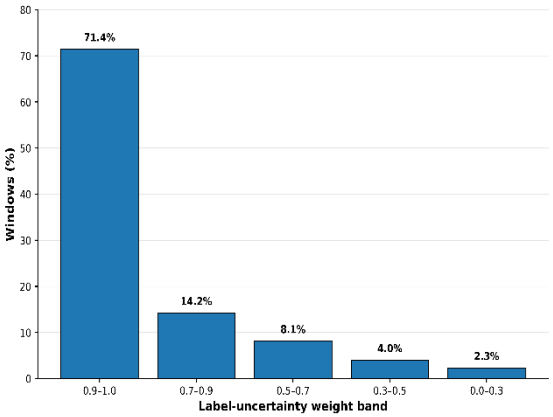

Fig. 11 Distribution of label-uncertainty weights

Figure 11 is about the distribution of label-uncertainty weights across the training windows. The largest proportion of windows, 71.4%, falls within the high-confidence 0.9–1.0 band, followed by 14.2%, 8.1%, 4.0%, and 2.3% across progressively lower confidence bands. This right-skewed distribution shows that most windows have reliable annotations, while only a small proportion lies near ambiguous freeze-onset boundaries.

5.11 Sensitivity–False Alarm Tradeoff

Table 12 shows the sensitivity and FP/hour for 5 decision thresholds, at the 1 s horizon, so that the operating-point tradeoff between sensitivity and FP/hour will be displayed beyond the single threshold = 0.5 utilized elsewhere in this section. Varies the classification threshold applied to $\hat{p_i}$ (Eq. 8) from permissive to conservative. The default threshold (0.5), which is applied across Section 5.1–5.8, is indicated for reference.

Table 12. Sensitivity–false Alarm Tradeoff Across Decision Thresholds at the 1 s Anticipation Horizon

Threshold	Sensitivity (%)	FP/hour
0.3	94.1	5.8
0.4	91.6	3.9
0.5 (default)	88.6	2.6
0.6	83.2	1.7
0.7	75.0	1.0

Natural Resources for Human Health
eISSN: 2583-1194

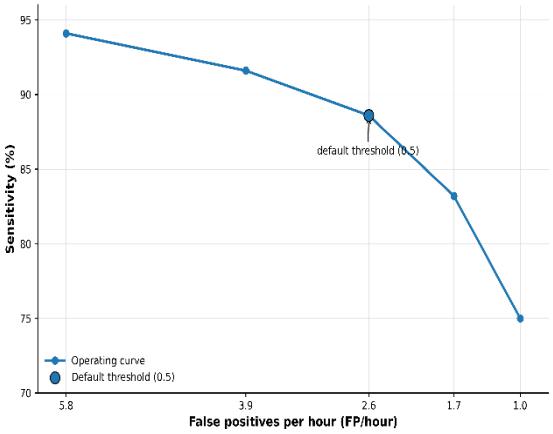

Fig. 12 Sensitivity versus false-positive rate operating curve

Figure 12 is about the trade-off between sensitivity and the false-positive rate at different decision thresholds. The operating points of 94.1% sensitivity and 5.8 FP/hour and 75.0% sensitivity and 1.0 FP/hour are plotted as sensitivity gains as the false-positive rate drops. As the FP/hour decreases from 5.8 to 1.0, the sensitivity can be plotted as increasing, as shown at 94.1% at 5.8 FP/hour and 75.0% at 1.0 FP/hour, with intermediate sensitivity operating points at 91.6% at 3.9 FP/hour and 83.2% at 1.7 FP/hour. The highlighted operating point at 2.6 FP/hour and 88.6% sensitivity represents the default decision threshold of 0.5, which is used throughout Sections 5.1–5.8.

6. CONCLUSION

Freezing of gait is one of the most disabling and unpredictable symptoms of PD, and reactive detection systems only work after a freeze has actually occurred and are of limited clinical use when it comes to assessing fall risk. The proposed CAP-FOG-PD framework addresses this by changing the FOG detection problem into an anticipatory one: predict a possible freeze 0.5-3 seconds before it occurs by applying tri-axial inertial sensor data. Unlike earlier wearable-sensor-based approaches, CAP-FOG-PD explicitly considers the label uncertainty of video-based FOG annotation as a calibrated and measurable term, and incorporates it as such in the training objective and evaluation criteria with a continuously defined noise weight that down-weights the window on the FOG boundary without discarding it. Stratified sampling and class-weighted loss are used to overcome the class imbalance problem between the freeze and non-freeze windows, and all performance is assessed using a strict subject-quarantined Leave-One-Subject-Out protocol, which removes the identity leakage problem that has led to over-optimistic performance reported in previous works on FOG-detection. Experimental results on the open-access Daphnet, tdcsfog, and defog datasets show that CAP-FOG-PD achieves higher precision and sensitivity than an unweighted ablation of itself and some representative reactive baselines for all four anticipation horizons, while the error decomposition, which allows for an analysis of error sources, is both diagnostic and informative, unlike the previous open-dataset FOG research. Further, few-shot on-device personalization increases sensitivity over the population-trained baseline for each of the subjects evaluated, while 8-bit quantized inference puts the framework within real-time latency requirements for embedded ARM Cortex-class hardware, supporting its feasibility for continuous, real-world wearable deployment. Multimodal sensing, model explainability, and fully harmonized cross-site generalization across sensor placements and populations are identified as directions for future work.

Future work will involve extending CAP-FOG-PD with additional sensing modalities (e.g., electromyography) to enhance the anticipation capability beyond longer time horizons, implementing an active-learning loop to route boundary-ambiguous windows back for re-annotation, expanding the explainability module with attention and facilitating interpretation by clinicians, and further exploring cross-site generalization with varied sensor placement and sampling rates.

REFERENCES

[1] Al-Adhaileh, M. H., Wadood, A., Aldhyani, T. H., Khan, S., Uddin, M. I., & Al-Nefaie, A. H. (2025). Deep learning techniques for detecting freezing of gait episodes in Parkinson's disease using wearable sensors. *Frontiers in Physiology*, *16*, 1581699.

[2] Bachlin, M., Plotnik, M., Roggen, D., Maidan, I., Hausdorff, J. M., Giladi, N., & Troster, G. (2009). Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 436-446.

[3] Bikias, T., Iakovakis, D., Hadjidimitriou, S., Charisis, V., & Hadjileontiadis, L. J. (2021). DeepFoG: an IMU-based detection of freezing of gait episodes in Parkinson's disease patients via deep learning. *Frontiers in Robotics and AI*, *8*, 537384.

[4] Borzì, L., Demrozi, F., Bacchin, R. A., Turetta, C., Tebaldi, M., Sigcha, L., ... & Artusi, C. A. (2026). A multi-level annotated sensor dataset of gait freezing manifestations and severity in Parkinson's disease. *Scientific Data*, *13*(1), 305.

[5] Brederecke, J. (2023). Freezing of Gait Prediction from Accelerometer Data Using a Simple 1D-Convolutional Neural Network--8th Place Solution for Kaggle's Parkinson's Freezing of Gait Prediction Competition. *arXiv preprint arXiv:2307.03475*.

[6] Burzer, M., Riedel, T., Beigl, M., & Röddiger, T. (2026). Uncertainty-Aware (Un) Supervised Few-Shot User Adaptation for On-Device Personalized Human Activity Recognition. *arXiv preprint arXiv:2606.04798*.

[7] Cohen, S., Cohen-Inger, N., & Rokach, L. (2024). BagStacking: An integrated ensemble learning approach for freezing of gait detection in Parkinson's disease. *Information*, *15*(12), 822.

[8] David, R., Duke, J., Jain, A., Janapa Reddi, V., Jeffries, N., Li, J., ... & Rhodes, R. (2021). Tensorflow lite micro: Embedded machine learning for tinyml systems. *Proceedings of machine learning and systems*, *3*, 800-811.

[9] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, *7*(Jan), 1-30.

[10] Dorsey, E. R., & Bloem, B. R. (2018). The Parkinson pandemic—a call to action. *JAMA neurology*, *75*(1), 9-10.

[11] Feigin, V. L., Abajobir, A. A., Abate, K. H., Abd-Allah, F., Abdulle, A. M., Abera, S. F., ... & Nguyen, G. (2017). Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Neurology*, *16*(11), 877-897.

[12] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018, June). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 2704-2713). IEEE.

[13] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of big data*, *6*(1), 27.

[14] Kang, P., Moosmann, J., Liu, M., Zhou, B., Magno, M., Lukowicz, P., & Bian, S. (2025). Bridging generalization and personalization in human activity recognition via on-device few-shot learning. *arXiv preprint arXiv:2508.15413*.

[15] Keivanimehr, A. R., & Akbari, M. (2025). TinyML and edge intelligence applications in cardiovascular disease: A survey. *Computers in Biology and Medicine*, *186*, 109653.

[16] Lamaakal, I., Yahyati, C., Maleh, Y., El Makkaoui, K., & Ouahbi, I. (2026). Explainable Kolmogorov–Arnold Networks for zero-shot human activity recognition on TinyML edge devices. *Machine Learning and Knowledge Extraction*, *8*(3), 55.

[17] Li, B., Yao, Z., Wang, J., Wang, S., Yang, X., & Sun, Y. (2020). Improved deep learning technique to detect freezing of gait in Parkinson's disease based on wearable sensors. *Electronics*, *9*(11), 1919.

[18] M. Shama, D., & Venkataraman, A. (2026). Bayesian Uncertainty-aware Deep Learning with noisy labels: Tackling annotation ambiguity in EEG seizure detection. *PloS one*, *21*(6), e0352191.

[19] Mazilu, S., Calatroni, A., Gazit, E., Mirelman, A., Hausdorff, J. M., & Tröster, G. (2015). Prediction of freezing of gait in Parkinson's from physiological wearables: an exploratory study. *IEEE journal of biomedical and health informatics*, *19*(6), 1843-1854.

[20] Mo, W. T., & Chan, J. H. (2023). Predicting three types of freezing of gait events using deep learning models. *arXiv preprint arXiv:2310.06322*.

[21] Nieuwboer, A., Dom, R., De Weerdt, W., Desloovere, K., Janssens, L., & Stijn, V. (2004). Electromyographic profiles of gait prior to onset of freezing episodes in patients with Parkinson's disease. *Brain*, *127*(7), 1650-1660.

[22] Odonga, T., Esper, C. D., Factor, S. A., McKay, J. L., & Kwon, H. (2025). On the bias fairness and bias mitigation for a wearable-based freezing of gait detection in Parkinson's disease. *arXiv: 2502.09626*.

[23] Pardoel, S., Kofman, J., Nantel, J., & Lemaire, E. D. (2019). Wearable-sensor-based detection and prediction of freezing of gait in Parkinson's disease: a review. *Sensors*, *19*(23), 5141.

[24] Roggen, D., Plotnik, M., & Hausdorff, J. (2010). Daphnet freezing of gait. UCI Machine Learning Repository. https://doi.org/10.24432/C56K78

[25] Saha, B., Samanta, R., Ghosh, S. K., & Roy, R. B. (2024). Towards Sustainable Personalized On-Device Human Activity Recognition with TinyML and Cloud-Enabled Auto Deployment. *arXiv preprint arXiv:2409.00093*.

[26] Salomon, A., Gazit, E., Ginis, P., Urazalinov, B., Takoi, H., Yamaguchi, T., ... & Hausdorff, J. M. (2024). A machine learning contest enhances automated freezing of gait detection and reveals time-of-day effects. *Nature Communications*, *15*(1), 4853.

[27] Shaban, M. (2024). A novel variational mode decomposition based convolutional neural network for the identification of freezing of gait intervals for patients with Parkinson's disease. *Machine Learning with Applications*, *16*, 100553.

[28] Sigcha, L., Costa, N., Pavón, I., Costa, S., Arezes, P., López, J. M., & De Arcas, G. (2020). Deep learning approaches for detecting freezing of gait in Parkinson's disease patients through on-body acceleration sensors. *Sensors*, *20*(7), 1895.

[29] Song, H., Kim, M., Park, D., Shin, Y., & Lee, J. G. (2022). Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, *34*(11), 8135-8153.

[30] Su, D., Cui, Y., He, C., Yin, P., Bai, R., Zhu, J., ... & Feng, T. (2025). Projections for prevalence of Parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of Global Burden of Disease Study 2021. *bmj*, *388*.

[31] Wei, Y., Deng, Y., Sun, C., Lin, M., Jiang, H., & Peng, Y. (2024). Deep learning with noisy labels in medical prediction problems: a scoping review. *Journal of the American Medical Informatics Association*, *31*(7), 1596-1607.

[32] Yang, R., Sun, M., Chen, W., Feng, H., Chen, B., Liu, Y., ... & Ren, H. (2025). Global, regional and national burden of Parkinson's disease, 1990–2021: Update from the GBD 2021 study. *Journal of the neurological sciences*, 125